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

Malcolm S. Townes

Saint Louis University

Declarations of interest: none.

No sponsors were involved in designing this study.

No third-party financial support was used to conduct this research or prepare this article.

Corresponding author details:

Malcolm S. Townes

7530 Wayne Avenue, University City, MO 63130-1320 USA

malcolm.townes@slu.edu

Abstract

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 in this area. 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 derived from federally-funded R&D that accrue to society. In this study, I have explored a broader conceptualization of technology transfer to include the transfer of knowledge derived from R&D activities. Traditional approaches to studying technology transfer tend to focus only on the transfer of the technology to the private sector as the primary benefit of R&D endeavors. I argue that the transfer of the knowledge is also an important and desirable outcome that is often forgotten or ignored. Using U.S. patent data, I conduct binomial logistic, ordinal logistic, and multiple regression analyses to understand and explain this aspect of technology transfer and demonstrate the feasibility of measuring non-financially-based benefits of R&D activity. The study results indicate that the generality of a patent has very strong positive association with the probability and degree to which the knowledge embodied in the patent is transferred. The originality of a patent has a moderate negative association with the probability and degree to which the knowledge embodied in the patent is transferred. These results have implications for how the benefits of R&D are defined and measured, which could influence public policy regarding federal-funded R&D and technology transfer.

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

**Introduction**

Improving the transfer of discoveries derived from federally-funded research and development (R&D) to the private sector, so called technology transfer, is arguably a very high public policy priority 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 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 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 was 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 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 funding 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. As such, the primary aim of this study was to explore an alternative construct for technology transfer, demonstrate an alternative approach to measuring the payoff from R&D based on such construct, examine the technology transfer process in the context of these alternative constructs and measures, and consider the public policy implications of any new insights about technology transfer that were produced.

**Literature Review**

There is no official or universally accepted definition for technology transfer. While most studies of the topic didn’t explicitly define technology transfer, they generally seemed 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 seemed to conflate the concept of technology transfer with the mechanisms for achieving it. Licensing, new venture formation, research collaboration, and faculty consulting were 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 seemed that most studies didn’t bother to define the term. Generally, academic research related to technology transfer seemed 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 broadened the idea of technology to include academic knowledge (Gonzalez-Pernia, Kuechle & Pena-Legazkue, 2013). In doing so, these studies acknowledged that technology is not the only benefit that is derived from research and development.

Defining success in the context of technology transfer has been problematic for scholarly studies of the subject. Most research studies seemed 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. Theoretically, technology transfer can occur in the absence of a financial transaction. These approaches also carry 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 licenses and university spinout company as 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 categorize all sponsored research as instances of successful technology transfer.

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 produces a societal benefit.

There have also been technology transfer studies that focus on patent data. 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 demonstrated that regression analysis is a useful method for gaining insight into the factors associated with technology transfer success. However, most of studies seemed to focus on exogenous factors. 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**

How we operationalize the construct of technology transfer in studies should reflect the intent of technology transfer policy. While most studies of the topic have seemed to operationalize the construct as a financially-motivated exchange of a “technology” however defined, one can make an argument that this does not completely reflect the intent of technology transfer policy. The presidential memorandum issued by President Obama on October 28, 2011 (Daily Comp. Pres. Doc., 2011-October-28) clearly signals that the intent of policy in this area encompasses broader benefits beyond the production and transfer of technology assets. As such, it’s reasonable to conclude that the intent of technology transfer policy is to maximize all types of benefits derived from R&D efforts. Consequently, the transfer of technologies derived from federally-funded R&D to the private sector should not be the only goal of technology transfer policy. As the literature reveals, there can be other types of benefits derived from federally-funded R&D such as new knowledge, which the government should also seek to transfer to the private sector. However, such transfer need not be based on a financially-motivated exchange. This is context in which this study was conducted.

The purpose of this study was 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 investigated the following questions:

1. How do we measure the broader non-technology benefits produced by federally-funded R&D?
2. Can patent citations be used to evaluate benefits produced by federally-funded R&D?
3. What insights can be derived about the factors that drive the transfer of broader benefits produced by federally-funded R&D to the private sector based on patent citations data?

Most studies of technology transfer have seemed to assume that the generation of technologies is the primary output of R&D and focus only on the transfer of such technologies to the private sector. Moreover, most studies of technology transfer have seemed 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 defined technology transfer more broadly to include other R&D outputs, specifically new knowledge. Patents are often an output of R&D activity. They embody technology and knowledge, both of which can be transferred to other parties. For this study, patents allowed by the United States Patent and Trademark Office (USPTO) were used as a proxy for new knowledge derived from research and development. Citations of patents were used as a proxy for the transfer of that knowledge.

This study posed several hypotheses about the factors that are associated with the transfer of new knowledge derived from research and development. One hypothesis was that the number of claims made by a patent and the generality of the patent 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 was theorized that opportunities for a patent to be cited by other patents increases with the number of claims the patent contains. It was also theorized that the more general a patent (i.e., the greater the breadth of potential influence of a research discovery across fields) the more opportunities there are for that patent to be cited by other patents across multiple fields. The originality of the patent, which one can think of as a measure of the degree to which a research discovery is novel and independent of anything previous, was also expected to be positively associated with the probability that the patent is cited by other patents and the number of citations received by the patent from other patents. It was speculated that patents with higher levels of originality expand new knowledge to a much greater extent than less original patents and therefore create new opportunities for future discoveries and inventions. The year a patent is allowed was expected to be negatively associated with the number of citations a patent receives from other patents. It was proposed that the older a patent is the less relevant it becomes. Machlup (1962) is credited with proposing the concept of the half-life of knowledge, which is the time it takes for half of the knowledge in a field to be rendered irrelevant. Machlup proposed that although patents provide several years of protection and exclusivity (17 years at the time) obsolescence reduces the practical duration of this protection to no more than a few years. After a time, they essentially become non-factors.

**Data and Methods**

**Data Sources**

This study used a subset of 2,000 observations taken from patent data obtained from the National Bureau of Economic Research (NBER) website. The source data contained 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 the average forward citations lag for the patent, which is the average number of years between the year the patent was allowed and the year other patents cited it. BCKGTLAG measures the average backward citations lag for the patent, which is the average number of years between the year the patent was allowed and the year other patents that it cites were allowed. 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 the information contained in them is not available to other researchers and inventors. As such the knowledge contained in them cannot be transferred. The SUBCAT variable was eliminated from the data set because including it would have 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. This was interpreted to mean that the patent was so specific as to have no influence on inventions across all fields. 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. This was interpreted to mean that the patent was so original as to be completely independent of anything previous. 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 was assigned a value of 0 if the value of the CRECEIVE variable was zero and 1 if the value of the CRECEIVE variable was 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 0 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 was assigned 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 and because most of the outcome levels above 15 did not have enough cases to satisfy the requirements for logistic regression analysis. 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 A contains figures and tables while 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. The data for the CRECEIVE variable was highly skewed as was the data for most of the other variables.

Table 1 shows a correlation matrix for the variables. None of the variables chosen for the analysis appeared 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 (Peduzzi, Concato, Holford & Feinstein, 1996; Sileshi, 2015). 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 [] = ln ( =

= 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 positive 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 -2 log likelihood (-2LL) was reduced from 1,774.459 to 9.167 producing a statistic of 1,765.292 for the model. However, none of the coefficients were significant.

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[)] =

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 positive 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 -2LL was decreased from 2,734.8 to 1,927.5 producing a statistic of 807.3 for the model. Taken in whole, these results suggested that the model fit the data well and was efficient 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).

Several variables had strong associations the odds of a patent receiving more than the median number of citations from other patents. 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 holding all other variables constant. The GENERAL variable had a particularly strong association. 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 holding all other variables constant. While the positive association was consistent with the expectations, the magnitude of the association was unexpected. Also, somewhat surprising was the negative association between the originality of the patent as measured by the ORIGINAL variable and the odds of the patent receiving more than the median number of citations. This was counter to initial expectations. It’s possible 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. This is related to the concept of the adjacent possible described by Johnson (2011), which is the notion that extraordinary change is possible but can only be achieved by progressing through a series of first order combinations of potential knew interactions among current possibilities. A highly original patent may represent a leap to a second order combination or higher. For such a patent to be useful, it may be necessary for the current state of knowledge to expanded over some given amount of time so that first order combinations with the highly original patent become possible.

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 at least a given level of citations from other patents as expressed by the following:

=

= Logit-1[] =

where *j* takes on ordinal values from 1 to 15. The alternative hypothesis was HA: β ≠ 0 for at least one of the independent variables. Again, I originally 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 ordinal regression analysis. The coefficients for CAT02 were not significant. The coefficients for BCKGTLAG were significant at the 0.05 level while all remaining independent variables were significant at the 0.001 level. However, the confidence interval at the 0.95 level for the coefficients of CAT02, CAT03, and CAT05 contained the value 1.0 indicating changes in these variables did not change odds of a patent receiving a given level of citations from other patents. The McFadden pseudo-R2 value for the model was 0.189 and the -2LL for the model was decreased from 9,744.127 to 7,904.203 producing a statistic of 1,839.924 for the model. Taken in whole, these results suggested that the model fit the data well and was efficient predicting the odds of a patent receiving at least the specified level of citations from other patents (i.e., the odds and probability associated with achieving a given level of transfer of knowledge). 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 expected. Like the results of the binomial logistic regression, the magnitude of the association for the GENERAL variable was much larger than expected.

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)*

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 positive associations with both the ORIGINAL and GENERAL variables.

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 with the binomial and ordinal logistic regression analyses.

A check of the assumptions of linear regression suggested that the multiple regression model could be significantly improved. As Figure 3 shows, there appeared to be some heteroscedasticity in the data. Moreover, outliers appeared to be influencing the results. The variables had low correlation with the model residuals. The mean of the regression residuals was zero for all intents and purposes. The Durbin-Watson test statistic was 1.9843 with a p value of 0.3622 indicating that there was not sufficient evidence to reject the null hypothesis that true autocorrelation is zero. 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 indicated that applying the transformation to the dependent variable improved the multiple regression model. As Figure 4 shows, the heteroscedasticity in the data appeared to have been eliminated. The mean of the regression residuals was essentially zero. Correlation between the independent variables and the model residuals remained low. Again, the Durbin-Watson test statistic value was 2.0548 with a p value of 0.8753 indicating that there was not sufficient evidence to reject the null hypothesis that true autocorrelation is zero. The variability in the values of the independent variables remained positive. The VIF for all the independent variables were again low indicating a low level of multicollinearity.

**Discussion**

**Policy Implications**

This study provides information and new insights that help both industry professionals and policymakers better understand the drivers of desirable technology transfer outcomes and identify possible factors that policymakers may want to consider when formulating public policy regarding federal R&D funding and technology transfer. First, the study provides support for the notion that the benefits of R&D extend beyond financially-based outcomes and that non-financially-based measures such as knowledge transfer should be considered when evaluating the payoff from federally-funded R&D activities. This may lead policymakers to significantly modify their conceptions of technology transfer and the goals of R&D and technology transfer policy and programs.

The study provides further evidence that measuring knowledge transfer as a benefit of R&D is feasible. This may cause technology transfer professionals and policymakers to modify the metrics used to measure the outcomes and benefits of federally-funded R&D activities.

The results of the analysis suggest that the generality of knowledge derived from R&D initiatives is strongly associated with the odds of the knowledge being transferred and the extent to which the knowledge is transferred. Moreover, the level of originality of knowledge derived from R&D initiatives is negatively associated with the odds of the knowledge being transferred and the extent to which the knowledge is transferred. This could have implications for which specific R&D efforts are funded and pursued by researchers and policymakers.

**Limitations and Future Analysis**

This study presents several opportunities for future research. Since this analysis was focused on U.S. patent data for a five-year period from 1990 to 1995, findings based on the data may not be relevant to time frames before or after this period. Moreover, it included patents from all sources. Repeating the analysis with data covering a more recent period, data that isolates patents with lineages that trace back to federally-funded research, and data from other contexts would be useful in evaluating the generalizability of the results.

There is a truncation effect in the data (Hall, Jaffe & Trajtenberg, 2001). 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. Examining a subset of data buffered by at least 10 years on both sides of the period of study might help to minimize truncation effects. Developing new methods for dealing with truncation effects in the data would also be beneficial for analyzing similar kinds of data.

The measures of originality and generality are highly dependent on the specification of classifications, which is somewhat arbitrary and subjective by nature (Hall, Jaffe & Trajtenberg, 2001). Developing less arbitrary, more objective measures of originality and generality would likely improve the analysis and eliminate potential bias in the data.

Finally, an interesting and somewhat surprising result of the study was the lack of association of the number of claims made by a patent (i.e., CLAIMS variable) had with the number of citations received by the patent. Claims define what a patent is asserting to be new and novel. As such, one might expect it to be more strongly associated with the number of citations a patent receives from other patents. However, the analysis results suggested that the number of claims were among the least influential of the factors considered. Patents contain two types of claims – independent and dependent. An independent claim stands alone and do not refer to any other claim. A dependent claim refers to one or more independent or dependent claims. The data made no such distinction between the two types of claims. Introducing variables to distinguish between independent and dependent claims may help to better isolate any association between the number of claims in a patent and the number of citations the patent receives from other patents. Additionally, patent claims can be broad or narrow. Developing a method of capturing and quantify this distinction may also improve the analysis.

**Conclusion**

In this study, I have explored a broader conceptualization of technology transfer to include the transfer of knowledge derived from R&D activities. Patents are a primary output R&D activity. They embody technology and new knowledge. Traditional approaches to studying technology transfer tend to focus only on the transfer of the technology to the private sector as the primary benefit of R&D endeavors. I argue that the transfer of the knowledge is also an important and desirable outcome that is often forgotten or ignored. Using U.S. patent data, I conducted binomial logistic, ordinal logistic, and multiple regression analyses to understand and explain this aspect of technology transfer and demonstrate the feasibility of measuring non-financially-based benefits of R&D activity. The study results indicated that the generality of a patent, which represents the breadth of influence of a patent across fields, had very strong positive association with the probability and degree to which the knowledge embodied in the patent is transferred as measured by the number of citations a patent receives from other patents. The originality of a patent, which one can think of as a measure of the degree to which a research discovery is novel and independent of anything previous, had a moderate negative association with the probability and degree to which the knowledge embodied in the patent is transferred. These results have implications for how the benefits of R&D are defined and measured, which could influence public policy regarding federal R&D funding and technology transfer.

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/

Johnson, S. (2011). *Where good ideas come from: The natural history of innovation*. New York, NY: Penguin.

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

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, *49*(12), 1373-1379.

Sileshi, G. W. (2015). The relative standard error as an easy index for checking the reliability of regression coefficients [Discussion paper]. Retrieved from https://www.researchgate.net/profile/Gudeta\_Sileshi2/publication/281371440\_The\_relative\_standard\_error\_as\_an\_easy\_index\_for\_checking\_the\_reliability\_of\_regression\_coefficients/links/55e415b608ae2fac472143ea/The-relative-standard-error-as-an-easy-index-for-checking-the-reliability-of-regression-coefficients.pdf

Solow, R. (1957). Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics,* *39*(3), 312-320. doi:10.2307/1926047

U.S. Department of the Treasury. (2018a). Monthly Statement of the Public Debt of the United States, October 31, 2018. Retrieved from https://www.treasurydirect.gov/govt/reports/pd/mspd/2018/2018\_oct.htm

U.S. Department of the Treasury. (2018b). Monthly Treasury Statement: Receipts and Outlays of the United States Government, For Fiscal Year 2019 Through October 31, 2018, and Other Periods. Retrieved from https://www.fiscal.treasury.gov/reports-statements/mts/previous.html

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



Figure 2. Residuals Plot for Multiple Regression Model Using CRECEIVE

Figure 3. 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 those 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

Additional Variables Created for the Analysis

| Variable | Variable Type | Extended Name | Description |
| --- | --- | --- | --- |
| CAT02 | Numerical  Nominal | CAT 02 Indicator | Indicator variable  0 indicates patent not assigned to CAT 02  1 indicates patent assigned to CAT02 |
| CAT03 | Numerical  Nominal | CAT 03 Indicator | Indicator variable  0 indicates patent not assigned to CAT 03  1 indicates patent assigned to CAT03 |
| CAT04 | Numerical  Nominal | CAT 04 Indicator | Indicator variable  0 indicates patent not assigned to CAT 04  1 indicates patent assigned to CAT04 |
| CAT05 | Numerical  Nominal | CAT 05 Indicator | Indicator variable  0 indicates patent not assigned to CAT 05  1 indicates patent assigned to CAT05 |
| CAT06 | Numerical  Nominal | CAT 06 Indicator | Indicator variable  0 indicates patent not assigned to CAT 06  1 indicates patent assigned to CAT06 |
| CRECbinary | Numeric  Nominal | Number of Citations Received | 0 indicates 0 citations  1 indicates 1 or more  Takes on an integer value of 0 or 1. |
| CRECmdnSplt | Numeric  Nominal | Median Citations Received | 0 indicates less than or equal to median  1 indicates greater than median  Takes on an integer value of 0 or 1. |
| CRECordinal | Numeric  Ratio | Ordinal Level of Citations Received | The level of citations in other patents that reference the patent.  1 indicates 1 citation  2 indicates 2 citations  3 indicates 3 citations  4 indicates 4 citations  5 indicates 5 citations  6 indicates 6 citations  7 indicates 7 citations  8 indicates 8 citations  9 indicates 9 citations  10 indicates 10 citations  11 indicates 11 citations  12 indicates 12 citations  13 indicates 13 citations  14 indicates 14 citations  15 indicates 15 or more  Takes on integer values between 0 and 15. |
| CRECEIVEsqrt | Numerical  Ratio | Square Root of CRECEIVE | The square root of the value of the CRECEIVE variable |

Table 4

Correlation Matrix

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1. PATENT | 1.0000 |  |  |  |  |  |  |  |  |  |
| 2. GYEAR | 0.9856 | 1.0000 |  |  |  |  |  |  |  |  |
| 3. CRECEIVE | -0.1590 | -0.1494 | 1.0000 |  |  |  |  |  |  |  |
| 4. CAT | -0.0335 | -0.0268 | -0.0956 | 1.0000 |  |  |  |  |  |  |
| 5. CLAIMS | 0.0407 | 0.0402 | 0.1317 | -0.0172 | 1.0000 |  |  |  |  |  |
| 6. CMADE | 0.1023 | 0.0942 | 0.0632 | 0.0402 | 0.1655 | 1.0000 |  |  |  |  |
| 7. GENERAL | -0.1264 | -0.1188 | 0.4175 | -0.1054 | 0.1200 | 0.0881 | 1.0000 |  |  |  |
| 8. ORIGINAL | 0.0803 | 0.0809 | 0.0013 | -0.0606 | 0.0349 | 0.2530 | 0.2147 | 1.0000 |  |  |
| 9. FWDAPLAG | -0.1112 | 0.1051 | -0.1980 | 0.0149 | -0.0833 | -0.0733 | -0.2969 | -0.0092 | 1.0000 |  |
| 10. BCKGTLAG | -0.0112 | -0.0074 | -0.1356 | 0.1760 | -0.0714 | 0.0131 | -0.1063 | 0.2359 | 0.1309 | 1.0000 |

Table 5

Binomial Logistic Regression for CRECbinary

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient Estimate | Std. Error | | Z value | | p value | | Sig. | Exp (Coef) | | | C.I. Lower Exp(Coef) | | C.I. Upper Exp(Coef) |
| Intercept | 18.62 | 4.195(104) | | 0.000 | | 1.000 | |  | 1.222(108) | | | 0 | | Inf |
| GYEAR | 0.31 | 8.802(10-1) | | 0.355 | | 0.723 | |  | 1.366(100) | | | 0.3094 | | 22.7335 |
| CAT02 | 51.87 | 2.259(105) | | 0.000 | | 1.000 | |  | 3.365(1022) | | | 0.0000 | | Inf |
| CAT03 | 57.59 | 3.690(104) | | 0.002 | | 0.999 | |  | 1.023(1025) | | | 0.0000 | | Inf |
| CAT04 | 76.29 | 3.578(104) | | 0.002 | | 0.998 | |  | 1.359(1033) | | | 0.0000 | | Inf |
| CAT05 | 18.54 | 4.344(104) | | 0.000 | | 1.000 | |  | 1.130(108) | | | 0.0000 | | Inf |
| CAT06 | 61.74 | 3.574(104) | | 0.002 | | 0.999 | |  | 6.530(1026) | | | 0.0000 | | Inf |
| CMADE | -197.90 | 1.877(10-1) | | -0.105 | | 0.916 | |  | 9.804(10-1) | | | 0.5223 | | 1.3160 |
| CLAIMS | -236.40 | 1.741(10-1) | | -0.136 | | 0.892 | |  | 9.766(10-1) | | | 0.5599 | | 1.2531 |
| ORIGINAL | 2.19 | 4.565(100) | | 0.480 | | 0.631 | |  | 8.956(100) | | | 0.0017 | | 2.66(106) |
| GENERAL | 70.98 | 3.623(104) | | 0.002 | | 0.998 | |  | 6.686(1030) | | | 0.0000 | | Inf |
| FWDAPAG | -68.76 | 2.461(103) | | -0.028 | | 0.978 | |  | 1.376(10-30) | | | 0.0000 | | 0.0000 |
| BCKGTLAG | 0.0028 | 5.339(10-2) | | 0.052 | | 0.959 | |  | 1.003(100) | | | 0.8447 | | 1.1268 |
| \*\*\* 0.001 \*\* 0.01 \* 0.05 | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | |
| -2 log likelihood null | | | 1,744.459 | |  | |  | | |  |  | |  | |
| -2 log likelihood residual | | | 9.167 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |
| McFadden R2 | | | 0.9948 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |

Table 6

Binomial Logistic Regression for CRECmdnSplt

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient Estimate | Std. Error | | Z value | | p value | | Sig. | Exp (Coef) | | | C.I. Lower Exp(Coef) | | C.I. Upper Exp(Coef) |
| Intercept | 530.049 | 71.396 | | 7.424 | | 0.0000 | | \*\*\* | 1.575(10230) | | | 6.937(10169) | | 2.817(10291) |
| GYEAR | -0.266 | 0.036 | | -7.435 | | 0.0000 | | \*\*\* | 0.766 | | | 0.7138 | | 0.8215 |
| CAT02 | 1.002 | 0.223 | | 4.496 | | 0.0000 | | \*\*\* | 2.722 | | | 1.7645 | | 4.2277 |
| CAT03 | 0.725 | 0.218 | | 3.323 | | 0.0008 | | \*\*\* | 2.065 | | | 1.3482 | | 3.1728 |
| CAT04 | 0.401 | 0.184 | | 2.178 | | 0.0294 | | \* | 1.494 | | | 1.0417 | | 2.1464 |
| CAT05 | 0.115 | 0.182 | | 0.634 | | 0.5262 | |  | 1.122 | | | 0.7860 | | 1.6035 |
| CAT06 | 0.350 | 0.186 | | 1.878 | | 0.0603 | |  | 1.418 | | | 0.9856 | | 2.0449 |
| CMADE | 0.024 | 0.007 | | 3.226 | | 0.0013 | | \*\* | 1.024 | | | 1.0100 | | 1.0402 |
| CLAIMS | 0.014 | 0.006 | | 2.441 | | 0.0146 | | \* | 1.015 | | | 1.0030 | | 1.0275 |
| ORIGINAL | -1.101 | 0.221 | | -4.978 | | 0.0000 | | \*\*\* | -0.223 | | | 0.2148 | | 0.5114 |
| GENERAL | 4.277 | 0.231 | | 18.487 | | 0.0000 | | \*\*\* | 72.030 | | | 46.0593 | | 114.1223 |
| FWDAPAG | -0.199 | 0.027 | | -7.276 | | 0.0000 | | \*\*\* | 0.819 | | | 0.7755 | | 0.8635 |
| BCKGTLAG | -0.014 | 0.005 | | -2.999 | | 0.0027 | | \*\* | 0.986 | | | 0.9761 | | 0.9948 |
| \*\*\* 0.001 \*\* 0.01 \* 0.05 | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | |
| -2 log likelihood null | | | 2,734.8 | |  | |  | | |  |  | |  | |
| -2 log likelihood residual | | | 1,927.5 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |
| McFadden R2 | | | 0.2951 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |

Table 7

Ordinal Logistic Regression

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Coefficient Estimate | Std. Error | | t value | | p value | | Sig. | Exp (Coef) | | | C.I. Lower Exp(Coef) | | C.I. Upper Exp(Coef) |
| Variables: |  |  | |  | |  | |  |  | | |  | |  |
| GYEAR | -0.3659 | 0.00010 | | -375.053 | | 0.0000 | | \*\*\* | 0.694 | | | 0.6935 | | 0.6936 |
| CAT02 | 0.9424 | 0.11150 | | 8.453 | | 0.0000 | | \*\*\* | 2.566 | | | 1.8637 | | 3.5377 |
| CAT03 | 0.9285 | 0.10950 | | 8.480 | | 0.0000 | | \*\*\* | 2.531 | | | 1.8336 | | 3.4959 |
| CAT04 | 0.4930 | 0.09072 | | 5.434 | | 0.0000 | | \*\*\* | 1.637 | | | 1.2525 | | 2.1411 |
| CAT05 | 0.0345 | 0.08923 | | 0.386 | | 0.6995 | |  | 1.035 | | | 0.7950 | | 1.3477 |
| CAT06 | 0.2972 | 0.09029 | | 3.292 | | 0.0010 | | \*\*\* | 1.346 | | | 1.0329 | | 1.7550 |
| CMADE | 0.0242 | 0.00501 | | 4.839 | | 0.0000 | | \*\*\* | 1.025 | | | 1.0143 | | 1.0350 |
| CLAIMS | 0.0176 | 0.00455 | | 3.865 | | 0.0001 | | \*\*\* | 1.018 | | | 1.0087 | | 1.0268 |
| ORIGINAL | -0.8521 | 0.12358 | | -6.895 | | 0.0000 | | \*\*\* | 0.427 | | | 0.3132 | | 0.5804 |
| GENERAL | 4.6444 | 0.06795 | | 68.351 | | 0.0000 | | \*\*\* | 104.002 | | | 72.4010 | | 150.0700 |
| FWDAPAG | -0.4885 | 0.02004 | | -24.372 | | 0.0000 | | \*\*\* | 0.614 | | | 0.5878 | | 0.6397 |
| BCKGTLAG | -0.0067 | 0.00324 | | -2.060 | | 0.0395 | | \*\* | 0.993 | | | 0.9870 | | 0.9997 |
|  |  |  | |  | |  | |  |  | | |  | |  |
| Intercepts: |  |  | |  | |  | |  |  | | |  | |  |
| 0|1 | -732.991 | 0.00198 | | -3,695,549.5 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 1|2 | -730.923 | 0.10504 | | -6,958.3 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 2|3 | -729.860 | 0.11145 | | -6,549.0 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 3|4 | -729.079 | 0.11439 | | -6,373.8 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 4|5 | -728.525 | 0.11660 | | -6,248.5 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 5|6 | -728.049 | 0.11891 | | -6,122.7 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 6|7 | -727.700 | 0.12091 | | -6,018.7 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 7|8 | -727.412 | 0.12283 | | -5,922.2 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 8|9 | -727.149 | 0.12487 | | -5,823.4 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 9|10 | -726.933 | 0.12652 | | -5,745.7 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 10|11 | -726.699 | 0.12874 | | -5,644.7 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 11|12 | -726.499 | 0.13029 | | -5,575.9 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 12|13 | -726.323 | 0.13116 | | -5,537.6 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 13|14 | -726.188 | 0.13145 | | -5,524.3 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| 14|15 | -726.030 | 0.13180 | | -5,508.8 | | 0.0000 | | \*\*\* | 0.000 | | | 0.000 | | 0.000 |
| \*\*\* 0.001 \*\* 0.01 \* 0.05 | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | |
| -2 log likelihood null | | | 9,744.1 | |  | |  | | |  |  | |  | |
| -2 log likelihood residual | | | 7,904.2 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |
| McFadden R2 | | | 0.1888 | |  | |  | | |  |  | |  | |
|  | | |  | |  | |  | | |  |  | |  | |

Table 8

Adjusted R2 for Regression Subsets

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Adjusted R2 | 0.23 | X | X | X | X | X |  |  |  | X | X | X | X |  |
| 0.23 | X | X | X | X | X |  |  |  | X |  | X | X | X |
| 0.23 | X | X | X | X | X |  |  |  | X |  | X | X |  |
| 0.23 | X | X | X | X |  |  |  |  |  | X | X | X |  |
| 0.23 | X | X | X | X | X |  |  |  |  |  | X | X |  |
| 0.22 | X | X | X | X |  |  |  |  |  |  | X | X |  |
| 0.22 | X | X | X | X |  |  |  |  | X |  | X |  |  |
| 0.21 | X | X | X | X |  |  |  |  |  |  | X |  |  |
| 0.21 | X |  | X | X |  |  |  |  |  | X | X |  |  |
| 0.20 | X |  | X | X |  |  |  |  |  |  | X |  |  |
| 0.20 | X | X |  | X |  |  |  |  |  |  | X |  |  |
| 0.19 | X |  |  | X |  |  |  |  |  |  | X |  |  |
| 0.18 | X |  | X |  |  |  |  |  |  |  | X |  |  |
| 0.17 | X |  |  |  |  |  |  |  |  |  |  |  |  |
|  | | Intercept | GYEAR | CAT02 | CAT03 | CAT04 | CAT05 | CAT06 | CMADE | CLAIMS | ORIGINAL | GENERAL | FWDAPLAG | BCKGTLAG |

Table 9

Multiple Regression Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | CRECEIVE |  |  | CRECEIVEsqrt |  |  |
|  | Coefficient Estimate | p-value | Sig. | Coefficient Estimate | p-value | Sig. |
| Intercept | 998.65051 | 0.000 | \*\*\* | 169.336058 | 0.000 | \*\*\* |
| GYEAR | -0.49997 | 0.000 | \*\*\* | -0.084001 | 0.000 | \*\*\* |
| CAT02 | 3.05106 | 0.000 | \*\*\* | 0.135559 | 0.004 | \*\* |
| CAT03 | 3.45773 | 0.000 | \*\*\* | 0.141919 | 0.002 | \*\* |
| CAT04 | 1.37359 | 0.000 | \*\*\* | 0.072884 | 0.034 | \* |
| CLAIMS | 0.05989 | 0.000 | \*\*\* | 0.004406 | 0.003 | \*\* |
| ORIGINAL | -1.55882 | 0.002 | \*\* | -0.249616 | 0.000 | \*\*\* |
| GENERAL | 9.57013 | 0.000 | \*\*\* | 1.473686 | 0.000 | \*\*\* |
| FWDAPAG | -0.21028 | 0.000 | \*\*\* | -0.158777 | 0.000 | \*\*\* |

\*\*\* 0.001 \*\* 0.01 \* 0.05

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Residual SE | DF | Multiple R2 | Adjusted R2 | F-statistic | DF | p-value |
| CRECEIVE | 6.357 | 1,989 | 0.2373 | 0.2342 | 77.4 | 8 & 1,989 | 0.000\*\*\* |
| CRECEIVEsqrt | 0.544 | 1,754 | 0.6417 | 0.6401 | 392.7 | 8 & 1,754 | 0.000\*\*\* |

\*\*\* 0.001 \*\* 0.01 \* 0.05

Appendix B. R Notebook and Output

**R Notebook: Improving Construct Validity in Studies of Technology Transfer**

Malcolm S. Townes

(July 21, 2019)

**Introduction**

This is an R Notebook for an investigation that explores possiblities for improving construct validity in studies of technology transfer.

**Project Set Up**

The following code chunk enables the R Notebook to integrate seemlessly with the project organization format. This is normally included in the R Notebook to simplify file calls and enable file portability but it has been causing an error. To work around this problem, I’ve embedded the here() function where I enter a file path when necessary.

knitr**::**opts\_knit**$set**(root.dir = here**::here**())

**Load Dependencies**

The following code chunk loads package dependencies required to perform the necessary tasks. Basic tasks include importing, reading, wrangling, and cleaning data; selecting a subset of the data; checking for unique observations; analyzing missing data; and performing various types of regression analyses.

**library**(tidyverse) *# loads the basic R packages*  
**library**(here) *# enables file portability*  
**library**(readr) *# functions for reading data*  
**library**(dplyr) *# functions for data wrangling*  
**library**(janitor) *# functions for data cleaning*  
**library**(naniar) *# functions for analyzing missing data*  
**library**(ggplot2) *# functions for data visualizations*  
**library**(boot) *# functions for regression analysis*  
**library**(ordinal) *# functions for regression models for ordinal data*  
**library**(MASS) *# functions for ordered logistic or probit regression*  
**library**(broom) *# functions for tidying ordinal logistic regression models*  
**library**(gvlma) *# functions for global validation of linear model assumptions*  
**library**(lmtest) *# functions for testing linear regression models*  
**library**(leaps) *# functions for regression subset selection*  
**library**(car) *# companion to applied regression*  
**library**(aod) *# functions to analyze overdispersed data counts and proportions*  
**library**(pscl) *# contains function for pseudo R2 measures for logistic regression*  
**library**(ResourceSelection) *# contains function for Hosmer-Lemeshow goodness of fit test*

**Load Raw Data**

The following code chunk imports the raw data from the txt file for the NBER data set for the period 1963 to 1999.

DataRaw <- **read.table**(**here**("DataRaw","NBERpatents1963to1999/apat63\_99.txt"),   
 sep = ",", header = TRUE, fill = TRUE, dec = ".")

**Subset Data**

The following code chunk creates a subset of the data for the period 1990 through 1995.

DataRaw **%>%** *# subset data*  
 **filter**(GYEAR**>=**1990) **%>%**  
 **filter**(GYEAR**<=**1995) -> DataSubset90to95  
DataSubset90to95 <- **as\_tibble**(DataSubset90to95) *# convert data frame to tibble*

**Extract Sample Data**

The following code chunk takes a sample of 2,000 cases from the data subset for the period 1990 through 1995.

**set.seed**(1972)  
Sample90to95 <- **sample**(1**:nrow**(DataSubset90to95), size = 2000,   
 replace = TRUE, prob = NULL)  
Sample90to95 <- DataSubset90to95[Sample90to95,]  
Sample90to95 <- **as\_tibble**(Sample90to95)

**Clean Data 01**

The following code chunk reorganizes the variables and eliminates variables not used in the analysis.

Sample90to95 **%>%**  
 dplyr**::select**(PATENT, GYEAR, CRECEIVE, CAT, CLAIMS, CMADE, GENERAL,   
 ORIGINAL, FWDAPLAG, BCKGTLAG) -> Sample90to95A   
*# Another package also has a `select()` function*

**Inspect Sample Data**

The following code chunk evaluates the data sample to determine if additional data cleaning is necessary. It first checks for missing data for each variable. It then checks for missing data for each variable in each case. Then it checks for duplicate cases with the PATENT variable to determine if that variable can be used as a unique identifier for each case. Finally, it checks for duplicate cases across all variables to ensure that each case is unique.

**miss\_var\_summary**(Sample90to95A, order = TRUE)

## # A tibble: 10 x 3  
## variable n\_miss pct\_miss  
## <chr> <int> <dbl>  
## 1 GENERAL 327 16.4  
## 2 FWDAPLAG 327 16.4  
## 3 ORIGINAL 48 2.4  
## 4 BCKGTLAG 34 1.7  
## 5 PATENT 0 0   
## 6 GYEAR 0 0   
## 7 CRECEIVE 0 0   
## 8 CAT 0 0   
## 9 CLAIMS 0 0   
## 10 CMADE 0 0

**miss\_case\_summary**(Sample90to95A, order = TRUE)

## # A tibble: 2,000 x 3  
## case n\_miss pct\_miss  
## <int> <int> <dbl>  
## 1 346 4 40  
## 2 516 4 40  
## 3 590 4 40  
## 4 1176 4 40  
## 5 1224 4 40  
## 6 1470 4 40  
## 7 1664 4 40  
## 8 1792 4 40  
## 9 1111 3 30  
## 10 1337 3 30  
## # ... with 1,990 more rows

**get\_dupes**(Sample90to95A, PATENT)

## # A tibble: 4 x 11  
## PATENT dupe\_count GYEAR CRECEIVE CAT CLAIMS CMADE GENERAL ORIGINAL  
## <int> <int> <int> <int> <int> <int> <int> <dbl> <dbl>  
## 1 4.99e6 2 1991 0 1 15 9 NA 0.370  
## 2 4.99e6 2 1991 0 1 15 9 NA 0.370  
## 3 5.30e6 2 1994 0 1 2 6 NA 0.278  
## 4 5.30e6 2 1994 0 1 2 6 NA 0.278  
## # ... with 2 more variables: FWDAPLAG <dbl>, BCKGTLAG <dbl>

**get\_dupes**(Sample90to95A)

## # A tibble: 4 x 11  
## PATENT GYEAR CRECEIVE CAT CLAIMS CMADE GENERAL ORIGINAL FWDAPLAG  
## <int> <int> <int> <int> <int> <int> <dbl> <dbl> <dbl>  
## 1 4.99e6 1991 0 1 15 9 NA 0.370 NA  
## 2 4.99e6 1991 0 1 15 9 NA 0.370 NA  
## 3 5.30e6 1994 0 1 2 6 NA 0.278 NA  
## 4 5.30e6 1994 0 1 2 6 NA 0.278 NA  
## # ... with 2 more variables: BCKGTLAG <dbl>, dupe\_count <int>

**Adjust for Missing Data**

The following code chunk modifies cases with missing data, removes duplicate cases, and then evaluates the data sample to determine if additional cleaning is necessary. It first assigns a value of 0 to instances of NA in the data for the GENERAL variable. It then assigns a value of 1 to instances of NA in the data for the ORIGINAL variable. For the FWDAPLAG and BCKGTLAG variables it assigns the maximum value in the data for each variable to instances of missing data. It then removes duplicate cases. The code chunk then checks for missing data for each variable in each case and missing data for each case. Then it checks for duplicate cases with the PATENT variable to determine if that variable can be used as a unique identifier for each case. Finally, it checks for duplicate observations across all variables to ensure that each case is unique.

Sample90to95B <- Sample90to95A  
Sample90to95B**$**GENERAL[**is.na**(x=Sample90to95B**$**GENERAL)] <- 0  
Sample90to95B**$**ORIGINAL[**is.na**(x=Sample90to95B**$**ORIGINAL)] <- 1  
Sample90to95B**$**FWDAPLAG[**is.na**(x=Sample90to95B**$**FWDAPLAG)] <- **max**(Sample90to95B**$**FWDAPLAG, na.rm = TRUE)  
Sample90to95B**$**BCKGTLAG[**is.na**(x=Sample90to95B**$**BCKGTLAG)] <- **max**(Sample90to95B**$**BCKGTLAG, na.rm = TRUE)  
  
Sample90to95B **%>%**  
 **distinct**() -> Sample90to95B  
  
**miss\_var\_summary**(Sample90to95B, order = TRUE)

## # A tibble: 10 x 3  
## variable n\_miss pct\_miss  
## <chr> <int> <dbl>  
## 1 PATENT 0 0  
## 2 GYEAR 0 0  
## 3 CRECEIVE 0 0  
## 4 CAT 0 0  
## 5 CLAIMS 0 0  
## 6 CMADE 0 0  
## 7 GENERAL 0 0  
## 8 ORIGINAL 0 0  
## 9 FWDAPLAG 0 0  
## 10 BCKGTLAG 0 0

**miss\_case\_summary**(Sample90to95B, order = TRUE)

## # A tibble: 1,998 x 3  
## case n\_miss pct\_miss  
## <int> <int> <dbl>  
## 1 1 0 0  
## 2 2 0 0  
## 3 3 0 0  
## 4 4 0 0  
## 5 5 0 0  
## 6 6 0 0  
## 7 7 0 0  
## 8 8 0 0  
## 9 9 0 0  
## 10 10 0 0  
## # ... with 1,988 more rows

**get\_dupes**(Sample90to95B, PATENT)

## # A tibble: 0 x 11  
## # ... with 11 variables: PATENT <int>, dupe\_count <int>, GYEAR <int>,  
## # CRECEIVE <int>, CAT <int>, CLAIMS <int>, CMADE <int>, GENERAL <dbl>,  
## # ORIGINAL <dbl>, FWDAPLAG <dbl>, BCKGTLAG <dbl>

**get\_dupes**(Sample90to95B)

## # A tibble: 0 x 11  
## # ... with 11 variables: PATENT <int>, GYEAR <int>, CRECEIVE <int>,  
## # CAT <int>, CLAIMS <int>, CMADE <int>, GENERAL <dbl>, ORIGINAL <dbl>,  
## # FWDAPLAG <dbl>, BCKGTLAG <dbl>, dupe\_count <int>

**Calculate Measures of Central Tendency**

The following code chunk calculates measures of central tendency in the sample data for each of the variables.

**summary**(Sample90to95B)

## PATENT GYEAR CRECEIVE CAT   
## Min. :4890423 Min. :1990 Min. : 0.000 Min. :1.000   
## 1st Qu.:5034806 1st Qu.:1991 1st Qu.: 1.000 1st Qu.:2.000   
## Median :5185746 Median :1993 Median : 3.000 Median :4.000   
## Mean :5184975 Mean :1993 Mean : 4.952 Mean :3.725   
## 3rd Qu.:5336132 3rd Qu.:1994 3rd Qu.: 6.000 3rd Qu.:5.000   
## Max. :5479597 Max. :1995 Max. :99.000 Max. :6.000   
## CLAIMS CMADE GENERAL ORIGINAL   
## Min. : 1.00 Min. : 0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 6.00 1st Qu.: 4.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 10.00 Median : 7.000 Median :0.0907 Median :0.4444   
## Mean : 12.69 Mean : 8.398 Mean :0.2578 Mean :0.3828   
## 3rd Qu.: 17.00 3rd Qu.: 11.000 3rd Qu.:0.5000 3rd Qu.:0.6250   
## Max. :101.00 Max. :158.000 Max. :0.8800 Max. :1.0000   
## FWDAPLAG BCKGTLAG   
## Min. : 0.000 Min. : 0.00   
## 1st Qu.: 3.000 1st Qu.: 6.50   
## Median : 4.000 Median :10.75   
## Mean : 4.959 Mean :15.13   
## 3rd Qu.: 6.000 3rd Qu.:18.50   
## Max. :10.500 Max. :85.14

**Prepare Histograms**

The following code chunk displays histograms for the variables to enable visual inspection of the data to evaluate whether or not they fit normal distributions. The code chunk generates separate png files for each histogram, which are saved in the Results folder.

**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(GYEAR))



**ggsave**(**here**("results", "histogramGYEAR.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(CRECEIVE))



**ggsave**(**here**("Results", "histogramCRECEIVE.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(CAT))



**ggsave**(**here**("Results", "histogramCAT.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(CLAIMS))



**ggsave**(**here**("Results", "histogramCLAIMS.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(CMADE))



**ggsave**(**here**("Results", "histogramCMADE.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(GENERAL))



**ggsave**(**here**("Results", "histogramGENERAL.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(ORIGINAL))



**ggsave**(**here**("Results", "histogramORIGINAL.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(FWDAPLAG))



**ggsave**(**here**("Results", "histogramFWDAPLAG.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_histogram**(Sample90to95B, mapping = **aes**(BCKGTLAG))



**ggsave**(**here**("Results", "histogramBCKGTLAG.png"), dpi = 300)

**Prepare Scatter Plots**

The following code chunk displays scatter plots with CRECEIVE as the dependent variable against each of the the independent variables to visually inspect for linear relationships between the dependent variable and each of the independent variables. The code chunk generates separate png files for each scatter plot, which are saved in the Results folder.

**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = GYEAR, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyGYEAR.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = CAT, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyCAT.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = CLAIMS, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyCLAIMS.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = CMADE, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyCMADE.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = GENERAL, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyGENERAL.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = ORIGINAL, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyORIGINAL.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = FWDAPLAG, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyFWDAPLAG.png"), dpi = 300)  
  
**ggplot**() **+**  
 **geom\_point**(Sample90to95B, mapping = **aes**(x = BCKGTLAG, y = CRECEIVE))



**ggsave**(**here**("results", "scatterCRECEIVEbyBCKGTLAG.png"), dpi = 300)

**Prepare Q-Q Plots**

The following code chunk displays Quantile-Quantile (Q-Q) plots to check for normal distribution in the data sample for each variable. The code chunk generates separate png files for each Q-Q plot, which are saved in the Results folder.

**ggplot**(Sample90to95B)**+**  
 **aes**(sample = GYEAR)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("GYEAR Q-Q Plot")



**ggsave**(**here**("Results", "QQplotGYEAR.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = CRECEIVE)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("CRECEIVE Q-Q Plot")



**ggsave**(**here**("Results", "QQplotCRECEIVE.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = CLAIMS)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("CLAIMS Q-Q Plot")



**ggsave**(**here**("Results", "QQplotCLAIMS.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = CMADE)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("CMADE Q-Q Plot")



**ggsave**(**here**("Results", "QQplotCMADE.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = GENERAL)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("GENERAL Q-Q Plot")



**ggsave**(**here**("Results", "QQplotGENERAL.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = ORIGINAL)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("ORIGINAL Q-Q Plot")



**ggsave**(**here**("Results", "QQplotORIGINAL.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = FWDAPLAG)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("FWDAPLAG Q-Q Plot")



**ggsave**(**here**("Results", "QQplotFWDAPLAG.png"))  
  
**ggplot**(Sample90to95B)**+**  
 **aes**(sample = BCKGTLAG)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("BCKGTLAG Q-Q Plot")



**ggsave**(**here**("Results", "QQplotBCKGTLAG.png"))

**Calculate Pairwise Correlation Coefficients**

The following code chunk calculates the pairwise correlation coefficients for all variables in the sample data using the Pearson product-moment correlation function.

Sample90to95corrmatrix <- **cor**(Sample90to95B)  
**print**(Sample90to95corrmatrix)

## PATENT GYEAR CRECEIVE CAT CLAIMS  
## PATENT 1.00000000 0.985635583 -0.159041534 -0.03356222 0.04070183  
## GYEAR 0.98563558 1.000000000 -0.149458127 -0.02684555 0.04029299  
## CRECEIVE -0.15904153 -0.149458127 1.000000000 -0.09560927 0.13170051  
## CAT -0.03356222 -0.026845547 -0.095609270 1.00000000 -0.01724595  
## CLAIMS 0.04070183 0.040292995 0.131700513 -0.01724595 1.00000000  
## CMADE 0.10233427 0.094237930 0.063242195 0.04021272 0.16558795  
## GENERAL -0.12640786 -0.118898816 0.417511656 -0.10545725 0.12005496  
## ORIGINAL 0.08031646 0.080904546 0.001329185 -0.06068721 0.03499034  
## FWDAPLAG -0.11129042 -0.105161070 -0.198067341 0.01490230 -0.08333314  
## BCKGTLAG -0.01126732 -0.007455822 -0.135654022 0.17606285 -0.07141466  
## CMADE GENERAL ORIGINAL FWDAPLAG BCKGTLAG  
## PATENT 0.10233427 -0.12640786 0.080316461 -0.111290419 -0.011267319  
## GYEAR 0.09423793 -0.11889882 0.080904546 -0.105161070 -0.007455822  
## CRECEIVE 0.06324219 0.41751166 0.001329185 -0.198067341 -0.135654022  
## CAT 0.04021272 -0.10545725 -0.060687208 0.014902300 0.176062850  
## CLAIMS 0.16558795 0.12005496 0.034990339 -0.083333140 -0.071414662  
## CMADE 1.00000000 0.08818194 0.253063640 -0.073355840 0.013103136  
## GENERAL 0.08818194 1.00000000 0.214772596 -0.296932164 -0.106377914  
## ORIGINAL 0.25306364 0.21477260 1.000000000 -0.009235051 0.235975356  
## FWDAPLAG -0.07335584 -0.29693216 -0.009235051 1.000000000 0.130915221  
## BCKGTLAG 0.01310314 -0.10637791 0.235975356 0.130915221 1.000000000

**Modify Data 01**

The following code chunk creates additional variables needed for the binary logistic regression, ordinal logistic regression, and multiple regression analyses. It first creates a new variable called CRECbinary that converts the CRECEIVE variable into a dichotomous variable. It then creates a series of dummy variables for the nominal CAT variable to use in multiple regression analysis.

Sample90to95B **%>%**  
 **mutate**(CRECbinary = **ifelse**(CRECEIVE **==** 0, 0, 1)) **%>%**  
 **mutate**(CAT01 = **ifelse**(CAT **==** 1, 1, 0)) **%>%**  
 **mutate**(CAT02 = **ifelse**(CAT **==** 2, 1, 0)) **%>%**  
 **mutate**(CAT03 = **ifelse**(CAT **==** 3, 1, 0)) **%>%**  
 **mutate**(CAT04 = **ifelse**(CAT **==** 4, 1, 0)) **%>%**  
 **mutate**(CAT05 = **ifelse**(CAT **==** 5, 1, 0)) **%>%**  
 **mutate**(CAT06 = **ifelse**(CAT **==** 6, 1, 0)) -> Sample90to95C

**Count Observations 01**

The following code chunk calculates the number of observations for each outcome of each nominal and ordinal variable to determine if the sample size is large enough for logistic regression analysis, which requires at least 10 observations for the least frequent outcome for each variable.

Sample90to95C **%>%**  
 **group\_by**(GYEAR) **%>%**  
 **summarize**(**n**())

## # A tibble: 6 x 2  
## GYEAR `n()`  
## <int> <int>  
## 1 1990 301  
## 2 1991 339  
## 3 1992 331  
## 4 1993 332  
## 5 1994 347  
## 6 1995 348

Sample90to95C **%>%**  
 **group\_by**(CRECEIVE) **%>%**  
 **summarize**(**n**())

## # A tibble: 48 x 2  
## CRECEIVE `n()`  
## <int> <int>  
## 1 0 325  
## 2 1 307  
## 3 2 265  
## 4 3 234  
## 5 4 172  
## 6 5 142  
## 7 6 98  
## 8 7 73  
## 9 8 60  
## 10 9 44  
## # ... with 38 more rows

Sample90to95C **%>%**  
 **group\_by**(CRECbinary) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CRECbinary `n()`  
## <dbl> <int>  
## 1 0 325  
## 2 1 1673

Sample90to95C **%>%**   
 **group\_by**(CAT) **%>%**  
 **summarize**(**n**())

## # A tibble: 6 x 2  
## CAT `n()`  
## <int> <int>  
## 1 1 380  
## 2 2 207  
## 3 3 211  
## 4 4 376  
## 5 5 432  
## 6 6 392

Sample90to95C **%>%**   
 **group\_by**(CAT01) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT01 `n()`  
## <dbl> <int>  
## 1 0 1618  
## 2 1 380

Sample90to95C **%>%**   
 **group\_by**(CAT02) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT02 `n()`  
## <dbl> <int>  
## 1 0 1791  
## 2 1 207

Sample90to95C **%>%**   
 **group\_by**(CAT03) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT03 `n()`  
## <dbl> <int>  
## 1 0 1787  
## 2 1 211

Sample90to95C **%>%**   
 **group\_by**(CAT04) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT04 `n()`  
## <dbl> <int>  
## 1 0 1622  
## 2 1 376

Sample90to95C **%>%**   
 **group\_by**(CAT05) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT05 `n()`  
## <dbl> <int>  
## 1 0 1566  
## 2 1 432

Sample90to95C **%>%**   
 **group\_by**(CAT06) **%>%**  
 **summarize**(**n**())

## # A tibble: 2 x 2  
## CAT06 `n()`  
## <dbl> <int>  
## 1 0 1606  
## 2 1 392

Sample90to95C **%>%**   
 **group\_by**(CLAIMS) **%>%**  
 **summarize**(**n**())

## # A tibble: 57 x 2  
## CLAIMS `n()`  
## <int> <int>  
## 1 1 46  
## 2 2 55  
## 3 3 101  
## 4 4 107  
## 5 5 117  
## 6 6 123  
## 7 7 112  
## 8 8 133  
## 9 9 113  
## 10 10 107  
## # ... with 47 more rows

Sample90to95C **%>%**   
 **group\_by**(CMADE) **%>%**  
 **summarize**(**n**())

## # A tibble: 49 x 2  
## CMADE `n()`  
## <int> <int>  
## 1 0 34  
## 2 1 108  
## 3 2 159  
## 4 3 182  
## 5 4 179  
## 6 5 175  
## 7 6 152  
## 8 7 151  
## 9 8 131  
## 10 9 118  
## # ... with 39 more rows

**Modify Data 02**

The following code chunk groups cases where the outcome level for CRECEIVE is greater than or equal to 15 citations because most outcome levels above 15 citations do not have enough cases individually for logistic regression analysis, which requires at least 10 cases for the least frequent outcome level of each variable. This was also done to simplify the ordinal logistic regression analysis.

Sample90to95C **%>%**   
 **mutate**(CRECordinal = **ifelse** (CRECEIVE**>=**15,15,CRECEIVE)) -> Sample90to95C  
Sample90to95C <- **as\_tibble**(Sample90to95C) *# convert data frame to tibble*

**Count Observations 02**

The following code chunk calculates the number of observations for each outcome level of the new CRECordinal variable.

Sample90to95C **%>%**  
 **group\_by**(CRECordinal) **%>%**  
 **summarize**(**n**())

## # A tibble: 16 x 2  
## CRECordinal `n()`  
## <dbl> <int>  
## 1 0 325  
## 2 1 307  
## 3 2 265  
## 4 3 234  
## 5 4 172  
## 6 5 142  
## 7 6 98  
## 8 7 73  
## 9 8 60  
## 10 9 44  
## 11 10 43  
## 12 11 33  
## 13 12 26  
## 14 13 18  
## 15 14 19  
## 16 15 139

**Binary Logistic Regression Analysis**

The following code chunk uses the new dichotomous variable CRECbinary as the dependent variable in a binary logistic regression analysis. It then displays the results. It also calculates the odds ratio, various pseudo R-squared measures, confidence intervals for the coefficients, and Hosmer-Lemeshow goodness of fit test.

logitCRECEIVE <- **glm**(CRECbinary **~** GYEAR **+** **as.factor**(CAT) **+** CMADE **+** CLAIMS **+**   
 ORIGINAL **+** GENERAL **+** FWDAPLAG **+** BCKGTLAG,   
 data = Sample90to95C, family = binomial,   
 na.action = na.omit)  
**summary**(logitCRECEIVE)

##   
## Call:  
## glm(formula = CRECbinary ~ GYEAR + as.factor(CAT) + CMADE + CLAIMS +   
## ORIGINAL + GENERAL + FWDAPLAG + BCKGTLAG, family = binomial,   
## data = Sample90to95C, na.action = na.omit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.3879 0.0000 0.0000 0.0000 2.6810   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.862e+01 4.195e+04 0.000 1.000  
## GYEAR 3.122e-01 8.802e-01 0.355 0.723  
## as.factor(CAT)2 5.187e+01 2.259e+05 0.000 1.000  
## as.factor(CAT)3 5.759e+01 3.690e+04 0.002 0.999  
## as.factor(CAT)4 7.629e+01 3.578e+04 0.002 0.998  
## as.factor(CAT)5 1.854e+01 4.344e+04 0.000 1.000  
## as.factor(CAT)6 6.174e+01 3.574e+04 0.002 0.999  
## CMADE -1.979e-02 1.877e-01 -0.105 0.916  
## CLAIMS -2.364e-02 1.741e-01 -0.136 0.892  
## ORIGINAL 2.192e+00 4.565e+00 0.480 0.631  
## GENERAL 7.098e+01 3.623e+04 0.002 0.998  
## FWDAPLAG -6.876e+01 2.461e+03 -0.028 0.978  
## BCKGTLAG 2.785e-03 5.399e-02 0.052 0.959  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1774.459 on 1997 degrees of freedom  
## Residual deviance: 9.167 on 1985 degrees of freedom  
## AIC: 35.167  
##   
## Number of Fisher Scoring iterations: 25

*# Raise e to the coefficients*  
**print**(**exp**(**coef**(logitCRECEIVE)))

## (Intercept) GYEAR as.factor(CAT)2 as.factor(CAT)3   
## 1.222174e+08 1.366461e+00 3.365555e+22 1.023373e+25   
## as.factor(CAT)4 as.factor(CAT)5 as.factor(CAT)6 CMADE   
## 1.358901e+33 1.129561e+08 6.530469e+26 9.804027e-01   
## CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 9.766327e-01 8.956214e+00 6.686125e+30 1.375684e-30   
## BCKGTLAG   
## 1.002789e+00

*# Obtain various pseudo R-squared measures*  
**print**(**pR2**(logitCRECEIVE))

## llh llhNull G2 McFadden r2ML   
## -4.5834781 -887.2297285 1765.2925008 0.9948339 0.5866786   
## r2CU   
## 0.9967854

*# Confidence intervals for the coefficients*  
**print**(**exp**(**confint**(logitCRECEIVE, level = 0.95)))

## 2.5 % 97.5 %  
## (Intercept) 0.000000e+00 Inf  
## GYEAR 3.094650e-01 2.273347e+01  
## as.factor(CAT)2 0.000000e+00 Inf  
## as.factor(CAT)3 0.000000e+00 Inf  
## as.factor(CAT)4 0.000000e+00 Inf  
## as.factor(CAT)5 0.000000e+00 Inf  
## as.factor(CAT)6 0.000000e+00 Inf  
## CMADE 5.223381e-01 1.316047e+00  
## CLAIMS 5.598928e-01 1.253122e+00  
## ORIGINAL 1.719095e-03 2.662451e+06  
## GENERAL 0.000000e+00 Inf  
## FWDAPLAG 3.463248e-228 0.000000e+00  
## BCKGTLAG 8.446546e-01 1.126897e+00

*# Hosmer-Lemeshow Goodness of Fit Test*  
HosLemBinomial <- **hoslem.test**(Sample90to95C**$**CRECbinary,   
 **fitted**(logitCRECEIVE), g=10)  
**print**(HosLemBinomial)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: Sample90to95C$CRECbinary, fitted(logitCRECEIVE)  
## X-squared = 1.149e-09, df = 8, p-value = 1

**print**(**cbind**(HosLemBinomial**$**expected, HosLemBinomial**$**observed))

## yhat0 yhat1 y0 y1  
## [2.22e-16,1.25e-10] 200 1.148984e-09 200 0  
## (1.25e-10,1] 125 1.673000e+03 125 1673

**Modify Data 03**

The following code chunk creates a new variable called CRECmdnSplt using a median split of the CRECEIVE values. It then calculates the number of observations for each outcome level of the new variable.

Sample90to95C **%>%**  
 **mutate**(CRECmdnSplt = **ifelse**(CRECEIVE **<=** **median**(CRECEIVE),0,1)) -> Sample90to95C  
  
Sample90to95C **%>%**  
 **group\_by**(CRECordinal) **%>%**  
 **summarize**(**n**())

## # A tibble: 16 x 2  
## CRECordinal `n()`  
## <dbl> <int>  
## 1 0 325  
## 2 1 307  
## 3 2 265  
## 4 3 234  
## 5 4 172  
## 6 5 142  
## 7 6 98  
## 8 7 73  
## 9 8 60  
## 10 9 44  
## 11 10 43  
## 12 11 33  
## 13 12 26  
## 14 13 18  
## 15 14 19  
## 16 15 139

**Binomial Logistic Regression 02**

The following code chunk uses the new dichotomous variable CRECmdnSplt as the dependent variable in a binary logistic regression analysis. It then displays the results. It also calculates the odds ratio, various pseudo R-squared measures, confidence intervals for the coefficients, and Hosmer-Lemeshow goodness of fit test.

logitCRECEIVE02 <- **glm**(CRECmdnSplt **~** GYEAR **+** **as.factor**(CAT) **+** CMADE **+** CLAIMS **+**   
 ORIGINAL **+** GENERAL **+** FWDAPLAG **+** BCKGTLAG,   
 data = Sample90to95C, family = binomial,   
 na.action = na.omit)  
**summary**(logitCRECEIVE02)

##   
## Call:  
## glm(formula = CRECmdnSplt ~ GYEAR + as.factor(CAT) + CMADE +   
## CLAIMS + ORIGINAL + GENERAL + FWDAPLAG + BCKGTLAG, family = binomial,   
## data = Sample90to95C, na.action = na.omit)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4878 -0.7198 -0.2807 0.7900 2.2788   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 530.049130 71.396218 7.424 1.14e-13 \*\*\*  
## GYEAR -0.266350 0.035824 -7.435 1.05e-13 \*\*\*  
## as.factor(CAT)2 1.001504 0.222754 4.496 6.92e-06 \*\*\*  
## as.factor(CAT)3 0.724961 0.218178 3.323 0.000891 \*\*\*  
## as.factor(CAT)4 0.401390 0.184315 2.178 0.029425 \*   
## as.factor(CAT)5 0.115214 0.181788 0.634 0.526222   
## as.factor(CAT)6 0.349511 0.186079 1.878 0.060342 .   
## CMADE 0.024088 0.007466 3.226 0.001254 \*\*   
## CLAIMS 0.014959 0.006127 2.441 0.014634 \*   
## ORIGINAL -1.101033 0.221183 -4.978 6.43e-07 \*\*\*  
## GENERAL 4.277086 0.231356 18.487 < 2e-16 \*\*\*  
## FWDAPLAG -0.199317 0.027395 -7.276 3.45e-13 \*\*\*  
## BCKGTLAG -0.014514 0.004839 -2.999 0.002707 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2734.8 on 1997 degrees of freedom  
## Residual deviance: 1927.5 on 1985 degrees of freedom  
## AIC: 1953.5  
##   
## Number of Fisher Scoring iterations: 5

*# Raise e to the coefficients*  
**print**(**exp**(**coef**(logitCRECEIVE02)))

## (Intercept) GYEAR as.factor(CAT)2 as.factor(CAT)3   
## 1.575478e+230 7.661711e-01 2.722374e+00 2.064650e+00   
## as.factor(CAT)4 as.factor(CAT)5 as.factor(CAT)6 CMADE   
## 1.493900e+00 1.122114e+00 1.418373e+00 1.024380e+00   
## CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 1.015072e+00 3.325274e-01 7.203026e+01 8.192905e-01   
## BCKGTLAG   
## 9.855905e-01

*# Obtain various pseudo R-squared measures*  
**print**(**pR2**(logitCRECEIVE02))

## llh llhNull G2 McFadden r2ML   
## -963.7701190 -1367.4155161 807.2907942 0.2951885 0.3323889   
## r2CU   
## 0.4458102

*# Confidence intervals for the coefficients*  
**print**(**exp**(**confint**(logitCRECEIVE02, level = 0.95)))

## 2.5 % 97.5 %  
## (Intercept) 6.937303e+169 2.816829e+291  
## GYEAR 7.138246e-01 8.215045e-01  
## as.factor(CAT)2 1.764477e+00 4.227735e+00  
## as.factor(CAT)3 1.348257e+00 3.172838e+00  
## as.factor(CAT)4 1.041740e+00 2.146404e+00  
## as.factor(CAT)5 7.859863e-01 1.603493e+00  
## as.factor(CAT)6 9.856086e-01 2.044875e+00  
## CMADE 1.010048e+00 1.040194e+00  
## CLAIMS 1.003048e+00 1.027454e+00  
## ORIGINAL 2.147779e-01 5.113870e-01  
## GENERAL 4.605930e+01 1.141228e+02  
## FWDAPLAG 7.754603e-01 8.634669e-01  
## BCKGTLAG 9.760505e-01 9.947798e-01

*# Hosmer-Lemeshow Goodness of Fit Test*  
*# Null hypothesis: the model is a good fit for the data*  
*# Alternative hypothesis: the model is NOT a good fit for the data*  
HosLemBinomial02 <- **hoslem.test**(Sample90to95C**$**CRECmdnSplt,   
 **fitted**(logitCRECEIVE02), g=10)  
**print**(HosLemBinomial02)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: Sample90to95C$CRECmdnSplt, fitted(logitCRECEIVE02)  
## X-squared = 30.913, df = 8, p-value = 0.0001456

**print**(**cbind**(HosLemBinomial02**$**expected, HosLemBinomial02**$**observed))

## yhat0 yhat1 y0 y1  
## [0.00453,0.0533] 193.47787 6.522134 200 0  
## (0.0533,0.119] 182.51013 17.489868 191 9  
## (0.119,0.189] 169.43856 30.561440 172 28  
## (0.189,0.279] 152.73414 46.265862 141 58  
## (0.279,0.414] 132.08743 67.912568 120 80  
## (0.414,0.55] 103.11823 96.881768 86 114  
## (0.55,0.659] 78.16828 120.831722 81 118  
## (0.659,0.75] 58.66664 141.333359 74 126  
## (0.75,0.849] 40.15402 159.845977 44 156  
## (0.849,0.978] 20.64470 179.355303 22 178

**Ordinal Logistic Regression Analysis**

The following code chunk performs an ordinal logistic regression analysis on the data sample using CRECordinal as the dependent variable. It then displays the results. It performs the analysis two different ways for comparison.

*# Ordinal Logistic Regression Results - Method 01*  
CRECEIVEordinal01 <- **clm**(**as.factor**(CRECordinal) **~** GYEAR **+** **as.factor**(CAT) **+**   
 CMADE **+** CLAIMS **+** ORIGINAL **+** GENERAL **+** FWDAPLAG **+**   
 BCKGTLAG, data = Sample90to95C)  
**summary**(CRECEIVEordinal01)

## formula:   
## as.factor(CRECordinal) ~ GYEAR + as.factor(CAT) + CMADE + CLAIMS + ORIGINAL + GENERAL + FWDAPLAG + BCKGTLAG  
## data: Sample90to95C  
##   
## link threshold nobs logLik AIC niter max.grad cond.H   
## logit flexible 1998 -3952.10 7958.20 7(2) 4.51e-07 9.9e+13  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## GYEAR -0.365924 0.026953 -13.577 < 2e-16 \*\*\*  
## as.factor(CAT)2 0.942450 0.163457 5.766 8.13e-09 \*\*\*  
## as.factor(CAT)3 0.928549 0.164576 5.642 1.68e-08 \*\*\*  
## as.factor(CAT)4 0.493043 0.136760 3.605 0.000312 \*\*\*  
## as.factor(CAT)5 0.034448 0.134614 0.256 0.798027   
## as.factor(CAT)6 0.297271 0.135203 2.199 0.027899 \*   
## CMADE 0.024240 0.005091 4.761 1.92e-06 \*\*\*  
## CLAIMS 0.017572 0.004548 3.864 0.000112 \*\*\*  
## ORIGINAL -0.852045 0.157334 -5.416 6.11e-08 \*\*\*  
## GENERAL 4.644408 0.185912 24.982 < 2e-16 \*\*\*  
## FWDAPLAG -0.488523 0.021583 -22.634 < 2e-16 \*\*\*  
## BCKGTLAG -0.006676 0.003265 -2.044 0.040926 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Threshold coefficients:  
## Estimate Std. Error z value  
## 0|1 -733.01 53.75 -13.64  
## 1|2 -730.94 53.73 -13.60  
## 2|3 -729.88 53.73 -13.59  
## 3|4 -729.10 53.73 -13.57  
## 4|5 -728.54 53.72 -13.56  
## 5|6 -728.07 53.72 -13.55  
## 6|7 -727.72 53.72 -13.55  
## 7|8 -727.43 53.72 -13.54  
## 8|9 -727.17 53.72 -13.54  
## 9|10 -726.95 53.72 -13.53  
## 10|11 -726.72 53.72 -13.53  
## 11|12 -726.52 53.72 -13.53  
## 12|13 -726.34 53.72 -13.52  
## 13|14 -726.21 53.72 -13.52  
## 14|15 -726.05 53.71 -13.52

*# Ordinal Logistic Regression Results - Method 02*  
CRECEIVEordinal02 <- **polr**(**as.factor**(CRECordinal) **~** GYEAR **+** **as.factor**(CAT) **+**  
 CMADE **+** CLAIMS **+** ORIGINAL **+** GENERAL **+** FWDAPLAG **+**  
 BCKGTLAG, data = Sample90to95C, Hess = TRUE,   
 model = TRUE, method = "logistic")  
**summary**(CRECEIVEordinal02)

## Call:  
## polr(formula = as.factor(CRECordinal) ~ GYEAR + as.factor(CAT) +   
## CMADE + CLAIMS + ORIGINAL + GENERAL + FWDAPLAG + BCKGTLAG,   
## data = Sample90to95C, Hess = TRUE, model = TRUE, method = "logistic")  
##   
## Coefficients:  
## Value Std. Error t value  
## GYEAR -0.365915 9.756e-05 -3750.5274  
## as.factor(CAT)2 0.942448 1.115e-01 8.4527  
## as.factor(CAT)3 0.928546 1.095e-01 8.4802  
## as.factor(CAT)4 0.493042 9.072e-02 5.4347  
## as.factor(CAT)5 0.034456 8.924e-02 0.3861  
## as.factor(CAT)6 0.297260 9.029e-02 3.2922  
## CMADE 0.024240 5.009e-03 4.8389  
## CLAIMS 0.017572 4.546e-03 3.8649  
## ORIGINAL -0.852066 1.236e-01 -6.8949  
## GENERAL 4.644389 6.795e-02 68.3507  
## FWDAPLAG -0.488513 2.004e-02 -24.3722  
## BCKGTLAG -0.006675 3.241e-03 -2.0599  
##   
## Intercepts:  
## Value Std. Error t value   
## 0|1 -732.9911 0.0020 -369549.4531  
## 1|2 -730.9229 0.1050 -6958.2905  
## 2|3 -729.8599 0.1114 -6549.0132  
## 3|4 -729.0789 0.1144 -6373.8457  
## 4|5 -728.5245 0.1166 -6248.4878  
## 5|6 -728.0492 0.1189 -6122.7410  
## 6|7 -727.6995 0.1209 -6018.6841  
## 7|8 -727.4124 0.1228 -5922.1601  
## 8|9 -727.1485 0.1249 -5823.4243  
## 9|10 -726.9330 0.1265 -5745.7368  
## 10|11 -726.6987 0.1287 -5644.7307  
## 11|12 -726.4988 0.1303 -5575.8923  
## 12|13 -726.3228 0.1312 -5537.6347  
## 13|14 -726.1877 0.1315 -5524.3396  
## 14|15 -726.0301 0.1318 -5508.7824  
##   
## Residual Deviance: 7904.203   
## AIC: 7958.203

*# Calculate P-Values for Coefficients*  
coefsOrdinal <- **coefficients**(**summary**(CRECEIVEordinal02))  
pvalues <- **pt**(**abs**(coefsOrdinal)[,"t value"], df=CRECEIVEordinal02**$**df,lower.tail = FALSE)**\***2  
pval <- **pnorm**(**abs**(coefsOrdinal)[,"t value"],lower.tail = FALSE)**\***2  
coefsOrdinal01 <- **cbind**(coefsOrdinal, "p values (t dist)" = **round**(pvalues, 5))  
coefsOrdinal01 <- **cbind**(coefsOrdinal01, "p values (Normal)" = **round**(pval, 5))  
**print**(coefsOrdinal01)

## Value Std. Error t value p values (t dist)  
## GYEAR -3.659147e-01 9.756353e-05 -3.750527e+03 0.00000  
## as.factor(CAT)2 9.424477e-01 1.114961e-01 8.452738e+00 0.00000  
## as.factor(CAT)3 9.285464e-01 1.094952e-01 8.480247e+00 0.00000  
## as.factor(CAT)4 4.930423e-01 9.072128e-02 5.434693e+00 0.00000  
## as.factor(CAT)5 3.445585e-02 8.923836e-02 3.861103e-01 0.69946  
## as.factor(CAT)6 2.972598e-01 9.029238e-02 3.292191e+00 0.00101  
## CMADE 2.424001e-02 5.009381e-03 4.838924e+00 0.00000  
## CLAIMS 1.757160e-02 4.546481e-03 3.864879e+00 0.00011  
## ORIGINAL -8.520660e-01 1.235786e-01 -6.894933e+00 0.00000  
## GENERAL 4.644389e+00 6.794942e-02 6.835069e+01 0.00000  
## FWDAPLAG -4.885132e-01 2.004388e-02 -2.437219e+01 0.00000  
## BCKGTLAG -6.675213e-03 3.240554e-03 -2.059899e+00 0.03954  
## 0|1 -7.329911e+02 1.983472e-03 -3.695495e+05 0.00000  
## 1|2 -7.309229e+02 1.050435e-01 -6.958291e+03 0.00000  
## 2|3 -7.298599e+02 1.114458e-01 -6.549013e+03 0.00000  
## 3|4 -7.290789e+02 1.143860e-01 -6.373846e+03 0.00000  
## 4|5 -7.285245e+02 1.165921e-01 -6.248488e+03 0.00000  
## 5|6 -7.280492e+02 1.189090e-01 -6.122741e+03 0.00000  
## 6|7 -7.276995e+02 1.209067e-01 -6.018684e+03 0.00000  
## 7|8 -7.274124e+02 1.228289e-01 -5.922160e+03 0.00000  
## 8|9 -7.271485e+02 1.248661e-01 -5.823424e+03 0.00000  
## 9|10 -7.269330e+02 1.265169e-01 -5.745737e+03 0.00000  
## 10|11 -7.266987e+02 1.287393e-01 -5.644731e+03 0.00000  
## 11|12 -7.264988e+02 1.302928e-01 -5.575892e+03 0.00000  
## 12|13 -7.263228e+02 1.311612e-01 -5.537635e+03 0.00000  
## 13|14 -7.261877e+02 1.314524e-01 -5.524340e+03 0.00000  
## 14|15 -7.260301e+02 1.317950e-01 -5.508782e+03 0.00000  
## p values (Normal)  
## GYEAR 0.00000  
## as.factor(CAT)2 0.00000  
## as.factor(CAT)3 0.00000  
## as.factor(CAT)4 0.00000  
## as.factor(CAT)5 0.69942  
## as.factor(CAT)6 0.00099  
## CMADE 0.00000  
## CLAIMS 0.00011  
## ORIGINAL 0.00000  
## GENERAL 0.00000  
## FWDAPLAG 0.00000  
## BCKGTLAG 0.03941  
## 0|1 0.00000  
## 1|2 0.00000  
## 2|3 0.00000  
## 3|4 0.00000  
## 4|5 0.00000  
## 5|6 0.00000  
## 6|7 0.00000  
## 7|8 0.00000  
## 8|9 0.00000  
## 9|10 0.00000  
## 10|11 0.00000  
## 11|12 0.00000  
## 12|13 0.00000  
## 13|14 0.00000  
## 14|15 0.00000

*# Raise e to the coefficients*  
**print**(**exp**(**coef**(CRECEIVEordinal01)))

## 0|1 1|2 2|3 3|4   
## 4.544663e-319 3.595172e-318 1.040850e-317 2.272788e-317   
## 4|5 5|6 6|7 7|8   
## 3.956714e-317 6.364810e-317 9.029007e-317 1.203245e-316   
## 8|9 9|10 10|11 11|12   
## 1.566508e-316 1.943268e-316 2.456352e-316 3.000083e-316   
## 12|13 13|14 14|15 GYEAR   
## 3.577103e-316 4.094518e-316 4.793246e-316 6.935552e-01   
## as.factor(CAT)2 as.factor(CAT)3 as.factor(CAT)4 as.factor(CAT)5   
## 2.566261e+00 2.530835e+00 1.637291e+00 1.035048e+00   
## as.factor(CAT)6 CMADE CLAIMS ORIGINAL   
## 1.346181e+00 1.024536e+00 1.017727e+00 4.265416e-01   
## GENERAL FWDAPLAG BCKGTLAG   
## 1.040018e+02 6.135317e-01 9.933467e-01

*# Obtain various pseudo R-squared measures*  
**print**(**pR2**(CRECEIVEordinal02))

## llh llhNull G2 McFadden r2ML   
## -3952.1016446 -4872.0634922 1839.9236952 0.1888239 0.6018326   
## r2CU   
## 0.6064539

*# Confidence intervals for the coefficients*  
**print**(**exp**(**confint**(CRECEIVEordinal01, level = 0.95)))

## 2.5 % 97.5 %  
## GYEAR 0.6935048 0.6936049  
## as.factor(CAT)2 1.8636582 3.5377111  
## as.factor(CAT)3 1.8335726 3.4958786  
## as.factor(CAT)4 1.2524627 2.1410877  
## as.factor(CAT)5 0.7950133 1.3476907  
## as.factor(CAT)6 1.0329291 1.7550461  
## CMADE 1.0143317 1.0349899  
## CLAIMS 1.0087081 1.0268643  
## ORIGINAL 0.3131889 0.5803639  
## GENERAL 72.4010072 150.0700715  
## FWDAPLAG 0.5877900 0.6396990  
## BCKGTLAG 0.9869618 0.9996843

*# Hosemer-Lemeshow Goodness of Fit Test*  
*# Null hypothesis: the model is a good fit for the data*  
*# Alternative hypothesis: the model is NOT a good fit for the data*  
HosLemOrdinal <- **hoslem.test**(Sample90to95C**$**CRECordinal,  
 **fitted**(CRECEIVEordinal01), g=10)  
**print**(HosLemOrdinal)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: Sample90to95C$CRECordinal, fitted(CRECEIVEordinal01)  
## X-squared = 1626200, df = 8, p-value < 2.2e-16

**print**(**cbind**(HosLemOrdinal**$**expected, HosLemOrdinal**$**observed))

## yhat0 yhat1 y0 y1  
## [0.00154,0.032] 196.62033 3.379674 -1711 1911  
## (0.032,0.0542] 191.21231 8.787688 -1419 1619  
## (0.0542,0.0807] 186.71360 13.286396 -998 1198  
## (0.0807,0.112] 179.96381 19.036191 -703 902  
## (0.112,0.137] 175.35476 24.645237 -634 834  
## (0.137,0.187] 167.96449 32.035507 -516 716  
## (0.187,0.258] 155.15712 43.842879 -356 555  
## (0.258,0.444] 131.16791 68.832087 -439 639  
## (0.444,0.732] 87.33144 112.668564 -28 228  
## (0.732,0.956] 31.74051 168.259495 200 0

**Multiple Regression Model Selection**

The following code chunk creates regression subsets using the exhaustive algorithm with CRECEIVE as the dependent variable. It then displays the summary statistics to facilitate selection of the best regression model on which to focus.

CRECregsubsets <- **regsubsets**(CRECEIVE **~** GYEAR **+** **as.factor**(CAT) **+** CMADE **+**   
 CLAIMS **+** ORIGINAL **+** GENERAL **+** FWDAPLAG **+**   
 BCKGTLAG, data = Sample90to95C,   
 nbest = 2, method = "exhaustive")  
**summary**(CRECregsubsets,all.best=FALSE, matrix=TRUE)

## Subset selection object  
## Call: regsubsets.formula(CRECEIVE ~ GYEAR + as.factor(CAT) + CMADE +   
## CLAIMS + ORIGINAL + GENERAL + FWDAPLAG + BCKGTLAG, data = Sample90to95C,   
## nbest = 2, method = "exhaustive")  
## 12 Variables (and intercept)  
## Forced in Forced out  
## GYEAR FALSE FALSE  
## as.factor(CAT)2 FALSE FALSE  
## as.factor(CAT)3 FALSE FALSE  
## as.factor(CAT)4 FALSE FALSE  
## as.factor(CAT)5 FALSE FALSE  
## as.factor(CAT)6 FALSE FALSE  
## CMADE FALSE FALSE  
## CLAIMS FALSE FALSE  
## ORIGINAL FALSE FALSE  
## GENERAL FALSE FALSE  
## FWDAPLAG FALSE FALSE  
## BCKGTLAG FALSE FALSE  
## 2 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## GYEAR as.factor(CAT)2 as.factor(CAT)3 as.factor(CAT)4  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " "\*" " "   
## 3 ( 1 ) " " "\*" "\*" " "   
## 4 ( 1 ) "\*" "\*" "\*" " "   
## 5 ( 1 ) "\*" "\*" "\*" " "   
## 6 ( 1 ) "\*" "\*" "\*" " "   
## 7 ( 1 ) "\*" "\*" "\*" "\*"   
## 8 ( 1 ) "\*" "\*" "\*" "\*"   
## as.factor(CAT)5 as.factor(CAT)6 CMADE CLAIMS ORIGINAL GENERAL  
## 1 ( 1 ) " " " " " " " " " " "\*"   
## 2 ( 1 ) " " " " " " " " " " "\*"   
## 3 ( 1 ) " " " " " " " " " " "\*"   
## 4 ( 1 ) " " " " " " " " " " "\*"   
## 5 ( 1 ) " " " " " " " " " " "\*"   
## 6 ( 1 ) " " " " " " "\*" " " "\*"   
## 7 ( 1 ) " " " " " " "\*" " " "\*"   
## 8 ( 1 ) " " " " " " "\*" "\*" "\*"   
## FWDAPLAG BCKGTLAG  
## 1 ( 1 ) " " " "   
## 2 ( 1 ) " " " "   
## 3 ( 1 ) " " " "   
## 4 ( 1 ) " " " "   
## 5 ( 1 ) "\*" " "   
## 6 ( 1 ) "\*" " "   
## 7 ( 1 ) "\*" " "   
## 8 ( 1 ) "\*" " "

**plot**(CRECregsubsets, scale = "adjr2")



**Multiple Regression Analysis**

The following code chunk performs a multiple regression analysis on the data sample using the selected variables. It then displays the results.

*# Multiple Regression*  
CRECEIVEregression <- **lm**(CRECEIVE **~** GYEAR **+** CAT02 **+** CAT03 **+** CAT04 **+** CLAIMS **+**   
 ORIGINAL **+** GENERAL **+** FWDAPLAG,   
 data = Sample90to95C, na.action = na.omit)  
**summary**(CRECEIVEregression)

##   
## Call:  
## lm(formula = CRECEIVE ~ GYEAR + CAT02 + CAT03 + CAT04 + CLAIMS +   
## ORIGINAL + GENERAL + FWDAPLAG, data = Sample90to95C, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.593 -2.956 -0.850 1.035 87.029   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 998.65051 171.81537 5.812 7.16e-09 \*\*\*  
## GYEAR -0.49997 0.08621 -5.799 7.73e-09 \*\*\*  
## CAT02 3.05106 0.48425 6.301 3.64e-10 \*\*\*  
## CAT03 3.45773 0.47537 7.274 5.01e-13 \*\*\*  
## CAT04 1.37359 0.37734 3.640 0.000279 \*\*\*  
## CLAIMS 0.05989 0.01527 3.921 9.13e-05 \*\*\*  
## ORIGINAL -1.55882 0.50028 -3.116 0.001860 \*\*   
## GENERAL 9.57013 0.56144 17.046 < 2e-16 \*\*\*  
## FWDAPLAG -0.21028 0.05332 -3.944 8.31e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.357 on 1989 degrees of freedom  
## Multiple R-squared: 0.2373, Adjusted R-squared: 0.2342   
## F-statistic: 77.36 on 8 and 1989 DF, p-value: < 2.2e-16

**Check Linear Regression Assumptions**

The following code chunk performs various checks to determine if the model satisfies the assumptions of linear regression.

*# Global check of linear regression assumptions*  
**par**(mfrow=**c**(2,2))  
**gvlma**(CRECEIVEregression)

##   
## Call:  
## lm(formula = CRECEIVE ~ GYEAR + CAT02 + CAT03 + CAT04 + CLAIMS +   
## ORIGINAL + GENERAL + FWDAPLAG, data = Sample90to95C, na.action = na.omit)  
##   
## Coefficients:  
## (Intercept) GYEAR CAT02 CAT03 CAT04   
## 998.65051 -0.49997 3.05106 3.45773 1.37359   
## CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 0.05989 -1.55882 9.57013 -0.21028   
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma(x = CRECEIVEregression)   
##   
## Value p-value Decision  
## Global Stat 210438.23 0.000e+00 Assumptions NOT satisfied!  
## Skewness 9785.54 0.000e+00 Assumptions NOT satisfied!  
## Kurtosis 200585.30 0.000e+00 Assumptions NOT satisfied!  
## Link Function 29.73 4.959e-08 Assumptions NOT satisfied!  
## Heteroscedasticity 37.66 8.424e-10 Assumptions NOT satisfied!

*# View residuals*  
**png**(filename = **here**("Results","MultRegres01ModelResidualsPlotA.png"))  
CRECEIVEresid <- **residuals**(CRECEIVEregression)  
**plot**(CRECEIVEresid)  
**dev.off**()

## png   
## 2

**ggplot**(CRECEIVEregression)**+**  
 **aes**(x=.fitted, y=.resid)**+**  
 **geom\_point**()



**ggsave**(**here**("Results","MultRegres01ModelResidualsPlotB.png"))  
  
*# Check that mean of residuals equals zero*  
**mean**(CRECEIVEregression**$**residuals)

## [1] -3.595121e-16

*# Check for normality of residuals*  
*# Check for homoscedasticity of residuals or equal variance*  
**png**(filename = **here**("Results", "MultRegres01ModelResidualsDistribution.png"))  
**par**(mfrow=**c**(2,2)) *# set 2 rows and 2 column layout for plot*  
**plot**(CRECEIVEregression)  
**dev.off**()

## png   
## 2

*# Check for autocorrelation of residuals using Durbin-Watson test*  
*# Null hypothesis: true autocorrelation is zero*  
*# Alternative hypothesis: true autocorrelation is greater than zero*  
AutoCorr <- **dwtest**(CRECEIVEregression)  
**print**(AutoCorr)

##   
## Durbin-Watson test  
##   
## data: CRECEIVEregression  
## DW = 1.9843, p-value = 0.3622  
## alternative hypothesis: true autocorrelation is greater than 0

*# Check that the independent variables and the residuals are uncorrelated*  
CorrGYEAR <- **cor.test**(Sample90to95C**$**GYEAR, CRECEIVEregression**$**residuals)  
**print**(CorrGYEAR)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$GYEAR and CRECEIVEregression$residuals  
## t = -1.598e-11, df = 1996, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04385287 0.04385287  
## sample estimates:  
## cor   
## -3.576716e-13

CorrCAT <- **cor.test**(Sample90to95C**$**CAT, CRECEIVEregression**$**residuals)  
**print**(CorrCAT)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$CAT and CRECEIVEregression$residuals  
## t = -0.19801, df = 1996, p-value = 0.8431  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04827556 0.03942846  
## sample estimates:  
## cor   
## -0.004432068

CorrCLAIMS <- **cor.test**(Sample90to95C**$**CLAIMS, CRECEIVEregression**$**residuals)  
**print**(CorrCLAIMS)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$CLAIMS and CRECEIVEregression$residuals  
## t = 1.4501e-17, df = 1996, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04385287 0.04385287  
## sample estimates:  
## cor   
## 3.245828e-19

CorrORIGINAL <- **cor.test**(Sample90to95C**$**ORIGINAL, CRECEIVEregression**$**residuals)  
**print**(CorrORIGINAL)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$ORIGINAL and CRECEIVEregression$residuals  
## t = 9.1473e-16, df = 1996, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04385287 0.04385287  
## sample estimates:  
## cor   
## 2.047445e-17

CorrGENERAL <- **cor.test**(Sample90to95C**$**GENERAL, CRECEIVEregression**$**residuals)  
**print**(CorrGENERAL)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$GENERAL and CRECEIVEregression$residuals  
## t = -1.1524e-15, df = 1996, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04385287 0.04385287  
## sample estimates:  
## cor   
## -2.579462e-17

CorrFWDAPLAG <- **cor.test**(Sample90to95C**$**FWDAPLAG, CRECEIVEregression**$**residuals)  
**print**(CorrFWDAPLAG)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95C$FWDAPLAG and CRECEIVEregression$residuals  
## t = -3.946e-15, df = 1996, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04385287 0.04385287  
## sample estimates:  
## cor   
## -8.83242e-17

*# Check that the variability in independent variable values is positive*  
varGYEAR <- **var**(Sample90to95C**$**GYEAR)  
**print**(varGYEAR)

## [1] 2.882845

varCAT02 <- **var**(Sample90to95C**$**CAT02)  
**print**(varCAT02)

## [1] 0.0929164

varCAT03 <- **var**(Sample90to95C**$**CAT03)  
**print**(varCAT03)

## [1] 0.09450036

varCAT04 <- **var**(Sample90to95C**$**CAT04)  
**print**(varCAT04)

## [1] 0.1528499

varCAT05 <- **var**(Sample90to95C**$**CAT05)  
**print**(varCAT05)

## [1] 0.1695516

varCAT06 <- **var**(Sample90to95C**$**CAT06)  
**print**(varCAT06)

## [1] 0.1577822

varCLAIMS <- **var**(Sample90to95C**$**CLAIMS)  
**print**(varCLAIMS)

## [1] 88.60728

varCMADE <- **var**(Sample90to95C**$**CMADE)  
**print**(varCMADE)

## [1] 65.00294

varGENERAL <- **var**(Sample90to95C**$**GENERAL)  
**print**(varGENERAL)

## [1] 0.07826721

varORIGINAL <- **var**(Sample90to95C**$**ORIGINAL)  
**print**(varORIGINAL)

## [1] 0.08673389

varFWDAPLAG <- **var**(Sample90to95C**$**FWDAPLAG)  
**print**(varFWDAPLAG)

## [1] 8.095435

varBCKGTLAG <- **var**(Sample90to95C**$**BCKGTLAG)  
**print**(varBCKGTLAG)

## [1] 209.9605

*# Calculate Variance Inflation Factors to check for perfect multicollinearity among the variables*  
VIFregression <- **vif**(CRECEIVEregression)  
**print**(VIFregression)

## GYEAR CAT02 CAT03 CAT04 CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 1.058934 1.076779 1.055360 1.075535 1.021614 1.072760 1.219227 1.137526

**Modify Data 04**

The following code chunk removes cases in which CRECEIVE is greater than or equal to 10 as outliers and applies a transformation to the CRECEIVE variable in an effort to better satisfy the assumptions of linear regression and improve the model.

Sample90to95C **%>%**   
 **filter**(CRECEIVE **<=** 10) **%>%**  
 **mutate**(CRECEIVEsqrt = **sqrt**(CRECEIVE)) -> Sample90to95D

**Q-Q Plots for Transformed Dependent Variable**

The following code chunk creates a Quantile-Quantile (Q-Q) plot for the transformed dependent variable to check for suitability to use in multiple regression analysis.

**ggplot**(Sample90to95D)**+**  
 **aes**(sample = CRECEIVEsqrt)**+**  
 **stat\_qq**()**+**  
 **stat\_qq\_line**()**+**  
 **ggtitle**("CRECEIVEsqrt Q-Q Plot")



**ggsave**(**here**("Results", "QQplotCRECEIVEsqrt.png"))

**Multiple Regression Using Transformed Dependent Variable**

The following code chunk performs a multiple regression analysis using the transformed dependent variable and displays the results.

*# Multiple Regression with Transformed Dependent Variable*  
CRECEIVEregressionTrfm <- **lm**(CRECEIVEsqrt **~** GYEAR **+** CAT02 **+** CAT03 **+** CAT04 **+**   
 CLAIMS **+** ORIGINAL **+** GENERAL **+** FWDAPLAG,   
 data = Sample90to95D, na.action = na.omit)  
**summary**(CRECEIVEregressionTrfm)

##   
## Call:  
## lm(formula = CRECEIVEsqrt ~ GYEAR + CAT02 + CAT03 + CAT04 + CLAIMS +   
## ORIGINAL + GENERAL + FWDAPLAG, data = Sample90to95D, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50502 -0.37531 -0.09773 0.30909 2.13512   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 169.336058 15.585978 10.865 < 2e-16 \*\*\*  
## GYEAR -0.084001 0.007821 -10.740 < 2e-16 \*\*\*  
## CAT02 0.135559 0.047162 2.874 0.00410 \*\*   
## CAT03 0.141919 0.045456 3.122 0.00182 \*\*   
## CAT04 0.072884 0.034381 2.120 0.03415 \*   
## CLAIMS 0.004406 0.001465 3.008 0.00266 \*\*   
## ORIGINAL -0.249616 0.045047 -5.541 3.46e-08 \*\*\*  
## GENERAL 1.473686 0.052161 28.253 < 2e-16 \*\*\*  
## FWDAPLAG -0.158777 0.004641 -34.214 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5442 on 1754 degrees of freedom  
## Multiple R-squared: 0.6417, Adjusted R-squared: 0.6401   
## F-statistic: 392.7 on 8 and 1754 DF, p-value: < 2.2e-16

**Check Linear Regression Assumptions for Transformed Variables**

The following code chunk performs various checks to determine if the model satisfies the assumptions of linear regression.

*# Global check of linear regression assumptions*  
**par**(mfrow=**c**(2,2))  
**gvlma**(CRECEIVEregressionTrfm)

##   
## Call:  
## lm(formula = CRECEIVEsqrt ~ GYEAR + CAT02 + CAT03 + CAT04 + CLAIMS +   
## ORIGINAL + GENERAL + FWDAPLAG, data = Sample90to95D, na.action = na.omit)  
##   
## Coefficients:  
## (Intercept) GYEAR CAT02 CAT03 CAT04   
## 169.336058 -0.084001 0.135559 0.141919 0.072884   
## CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 0.004406 -0.249616 1.473686 -0.158777   
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma(x = CRECEIVEregressionTrfm)   
##   
## Value p-value Decision  
## Global Stat 362.77778 0.0000000 Assumptions NOT satisfied!  
## Skewness 134.44666 0.0000000 Assumptions NOT satisfied!  
## Kurtosis 13.00083 0.0003114 Assumptions NOT satisfied!  
## Link Function 215.27328 0.0000000 Assumptions NOT satisfied!  
## Heteroscedasticity 0.05701 0.8112811 Assumptions acceptable.

*# View residuals*  
CRECEIVEresidTrfm <- **residuals**(CRECEIVEregressionTrfm)  
**png**(filename = **here**("Results","MultRegresTrfmModelResidualsPlotA.png"))  
**plot**(CRECEIVEresidTrfm)  
**dev.off**()

## png   
## 2

**ggplot**(CRECEIVEregressionTrfm)**+**  
 **aes**(x=.fitted, y=.resid)**+**  
 **geom\_point**()



**ggsave**(**here**("Results","MultRegresTrfmModelResidualsPlotB.png"))  
  
*# Check that mean of residuals equals zero*  
**mean**(CRECEIVEregressionTrfm**$**residuals)

## [1] -4.023402e-18

*# Check for normality of residuals*  
*# Check for homoscedasticity of residuals or equal variance*  
**png**(filename = **here**("Results", "MultRegresTrfmModelResidualsDistribution.png"))  
**par**(mfrow=**c**(2,2)) *# set 2 rows and 2 column layout for plot*  
**plot**(CRECEIVEregressionTrfm)  
**dev.off**()

## png   
## 2

*# Check for autocorrelation of residuals using Durbin-Watson test*  
*# Null hypothesis: true autocorrelation is zero*  
*# Alternative hypothesis: true autocorrelation is greater than zero*  
AutoCorrTrfm <- **dwtest**(CRECEIVEregressionTrfm)  
**print**(AutoCorrTrfm)

##   
## Durbin-Watson test  
##   
## data: CRECEIVEregressionTrfm  
## DW = 2.0548, p-value = 0.8753  
## alternative hypothesis: true autocorrelation is greater than 0

*# Check that the independent variables and the residuals are uncorrelated*  
CorrGYEARtrfm <- **cor.test**(Sample90to95D**$**GYEAR, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrGYEARtrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$GYEAR and CRECEIVEregressionTrfm$residuals  
## t = -1.396e-11, df = 1761, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04668485 0.04668485  
## sample estimates:  
## cor   
## -3.326677e-13

CorrCATtrfm <- **cor.test**(Sample90to95D**$**CAT, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrCATtrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$CAT and CRECEIVEregressionTrfm$residuals  
## t = 2.0335, df = 1761, p-value = 0.04216  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.001719017 0.094870465  
## sample estimates:  
## cor   
## 0.04839998

CorrCLAIMStrfm <- **cor.test**(Sample90to95D**$**CLAIMS, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrCLAIMStrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$CLAIMS and CRECEIVEregressionTrfm$residuals  
## t = -6.185e-16, df = 1761, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04668485 0.04668485  
## sample estimates:  
## cor   
## -1.473869e-17

CorrORIGINALtrfm <- **cor.test**(Sample90to95D**$**ORIGINAL, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrORIGINALtrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$ORIGINAL and CRECEIVEregressionTrfm$residuals  
## t = -6.27e-16, df = 1761, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04668485 0.04668485  
## sample estimates:  
## cor   
## -1.494128e-17

CorrGENERALtrfm <- **cor.test**(Sample90to95D**$**GENERAL, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrGENERALtrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$GENERAL and CRECEIVEregressionTrfm$residuals  
## t = 6.7906e-16, df = 1761, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04668485 0.04668485  
## sample estimates:  
## cor   
## 1.61818e-17

CorrFWDAPLAGtrfm <- **cor.test**(Sample90to95D**$**FWDAPLAG, CRECEIVEregressionTrfm**$**residuals)  
**print**(CorrFWDAPLAGtrfm)

##   
## Pearson's product-moment correlation  
##   
## data: Sample90to95D$FWDAPLAG and CRECEIVEregressionTrfm$residuals  
## t = -1.393e-14, df = 1761, p-value = 1  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04668485 0.04668485  
## sample estimates:  
## cor   
## -3.319594e-16

*# Check that the variability in independent variable values is positive*  
varGYEARtrfm <- **var**(Sample90to95D**$**GYEAR)  
**print**(varGYEARtrfm)

## [1] 2.873931

varCAT02trfm <- **var**(Sample90to95D**$**CAT02)  
**print**(varCAT02trfm)

## [1] 0.0802342

varCAT03trfm <- **var**(Sample90to95D**$**CAT03)  
**print**(varCAT03trfm)

## [1] 0.08487944

varCAT04trfm <- **var**(Sample90to95D**$**CAT04)  
**print**(varCAT04trfm)

## [1] 0.1511625

varCAT05trfm <- **var**(Sample90to95D**$**CAT05)  
**print**(varCAT05trfm)

## [1] 0.175818

varCAT06trfm <- **var**(Sample90to95D**$**CAT06)  
**print**(varCAT06trfm)

## [1] 0.1655888

varCLAIMStrfm <- **var**(Sample90to95D**$**CLAIMS)  
**print**(varCLAIMStrfm)

## [1] 79.42126

varCMADEtrfm <- **var**(Sample90to95D**$**CMADE)  
**print**(varCMADEtrfm)

## [1] 61.15318

varGENERALtrfm <- **var**(Sample90to95D**$**GENERAL)  
**print**(varGENERALtrfm)

## [1] 0.07338131

varORIGINALtrfm <- **var**(Sample90to95D**$**ORIGINAL)  
**print**(varORIGINALtrfm)

## [1] 0.08781611

varFWDAPLAGtrfm <- **var**(Sample90to95D**$**FWDAPLAG)  
**print**(varFWDAPLAGtrfm)

## [1] 8.919473

varBCKGTLAGtrfm <- **var**(Sample90to95D**$**BCKGTLAG)  
**print**(varBCKGTLAGtrfm)

## [1] 217.1713

*# Calculate Variance Inflaction Factors to check for perfect multicollinearity among the variables*  
VIFregressionTrfm <- **vif**(CRECEIVEregressionTrfm)  
**print**(VIFregressionTrfm)

## GYEAR CAT02 CAT03 CAT04 CLAIMS ORIGINAL GENERAL FWDAPLAG   
## 1.045817 1.061693 1.043375 1.062993 1.013499 1.060122 1.187775 1.142786

**Save Data**

The following code chunk saves the final cleaned and modified data that was used in the analysis.

**write.csv**(Sample90to95C, **here**("DataClean","NBERpatents1963to1999","NBERPatCitSample90to95C.csv"), append = FALSE)  
**write.csv**(Sample90to95D, **here**("DataClean","NBERpatents1963to1999","NBERPatCitSample90to95D.csv"), append = FALSE)