Optimal Graduate Schools

Michael McCoy

11/3/2021

Markowitz's Efficient Frontier of Optimal ~~Portfolios and Allocation~~ **Graduate Schools**

My loose goal is to find the optimal graduate school using R. I thought this would be good practice for an R beginner like me

Loose Goals:

1. I don't want a school that is a research mill.
2. Yet, I still want a school that has an impact on research.

There is a delicate trade-off between citations and impact. Here, impact is defined as how many citations a university receives divided by how many total papers each university publishes. The following data on impact and citations is across a five-year span from 2012 through the end of 2016.

The data on total papers reflect journals indexed in the following Web of Science Core Collection editions: Science Citation Index Expanded, Social Sciences Citation Index, and Arts and Humanities Citation Index. Data included herein are derived from Clarivate Analytics InCites. For blank categories, those institutions may have received less than 600 citations over the five-year span from 2012 through the end of 2016.

To begin, let's load our necessary packages and load my starting data set of universities that I created in Excel

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.0.5
```

```
Graduate_Programs_Copy <- read_excel("data/Graduate Programs - Copy.xlsx")
```

Here is a brief summary of the data:

```
summary(Graduate_Programs_Copy)
```

```
##  Institution        Clinical Psychology Ph.D. PCSAS Accreditation1
##  Length:160         Length:160
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
##  Clinical Psychology Ph.D. APCS Member Bolder_Boulder_Model   Citations
##  Length:160                            Length:160           Min.   :  600
##  Class :character                      Class :character     1st Qu.: 1323
##  Mode  :character                      Mode  :character     Median : 2459
##                                                             Mean   : 4588
##                                                             3rd Qu.: 6021
##                                                             Max.   :42538
##                                                             NA's   :16
##     Articles          Impact         Valence
##  Min.   : 129.0   Min.   :2.603   Min.   :2.700
##  1st Qu.: 323.8   1st Qu.:3.996   1st Qu.:3.000
##  Median : 523.5   Median :4.799   Median :3.400
##  Mean   : 826.6   Mean   :4.918   Mean   :3.538
##  3rd Qu.:1121.5   3rd Qu.:5.688   3rd Qu.:4.000
##  Max.   :6288.0   Max.   :8.113   Max.   :4.800
##  NA's   :16       NA's   :16      NA's   :31
```

Now I am going to print the names of all the universities in the data set (printing all the other columns won't fit)

```
Graduate_Programs_Copy %>%
  select(Institution) %>%
  print(n=Inf)
```

```
## # A tibble: 160 x 1
##     Institution
##     <chr>
##   1 Hofstra University
##   2 Washington University in St. Louis
##   3 Illinois Institute of Technology
##   4 Vanderbilt University
##   5 University of Virginia
##   6 University of Arizona
##   7 University of California, Berkeley
##   8 Duke University
##   9 University of Pennsylvania
##  10 University of California, Los Angeles
##  11 University of California, Santa Barbara
##  12 University of Iowa
##  13 Boston University
##  14 Harvard University
##  15 Emory University
##  16 Yale University
##  17 University of Wisconsin, Madison
##  18 University of California, San Diego
##  19 University of Colorado at Boulder
##  20 University of Rochester
##  21 Yeshiva University
##  22 New York University
##  23 University of Denver
##  24 University of Pittsburgh
##  25 University of Michigan at Ann Arbor
##  26 University of Washington, Seattle
##  27 University of Oregon
##  28 University of North Carolina at Chapel Hill
##  29 University of Texas Southwestern Medical Center Dallas
##  30 University of Minnesota, Twin Cities
##  31 State University of New York at Stony Brook
##  32 University of Southern California
##  33 University of Nortre Dame
##  34 Northwestern University
##  35 University of Cincinnati
##  36 University of Maryland, College Park
##  37 Michigan State University
##  38 University of Illinois at Chicago
##  39 Northeastern University
##  40 University of Georgia
##  41 University of Vermont
##  42 Florida State University
##  43 Indiana University at Bloomington
##  44 University of Missouri at Columbia
##  45 Kent State University
##  46 University of New Mexico
##  47 Texas A&M University at College Station
##  48 Temple University
##  49 University of Illinois at Urbana-Champaign
##  50 Columbia University, Teachers College
```

```
##  51 Southern Methodist University
##  52 Purdue University at West Lafayette
##  53 Washington State University
##  54 University of Connecticut
##  55 Boston College
##  56 Indiana University–Purdue University Indianapolis
##  57 University of South Alabama
##  58 University of Texas at Austin
##  59 Arizona State University
##  60 Virginia Polytechnic Institute and State University
##  61 Colorado State University
##  62 THE Ohio State University
##  63 Florida International University
##  64 Virginia Commonwealth University
##  65 Uniformed Services University of the Health Sciences
##  66 Loyola University Chicago
##  67 San Diego State University
##  68 University of Florida
##  69 University of North Dakota at Grand Forks
##  70 The New School
##  71 University of Toledo
##  72 Rutgers State University
##  73 State University of New York at Binghamton
##  74 University of Miami
##  75 University of Kentucky
##  76 University of Massachusetts, Amherst
##  77 University of Utah
##  78 University of Delaware
##  79 University of Mississippi
##  80 University of Nevada at Reno
##  81 Pennsylvania State University at University Park
##  82 Case Western Reserve University
##  83 University of North Carolina at Greensboro
##  84 James Madison University
##  85 City University of New York, Queens College
##  86 State University of New York at Buffalo
##  87 University of Wyoming
##  88 Iowa State University
##  89 University of South Carolina
##  90 University of South Florida
##  91 University of Kansas
##  92 Syracuse University
##  93 University of Central Florida
##  94 Miami University
##  95 Sam Houston State University
##  96 University of Akron
##  97 City University of New York, John Jay College of Criminal Justice
##  98 Brigham Young University
##  99 University of Massachusetts, Boston
## 100 Northern Illinois University
## 101 University of Coloardo at Colorado Springs
## 102 George Mason University
## 103 University of Wisconsin, Milwaukee
## 104 University of Alabama at Birmingham
```

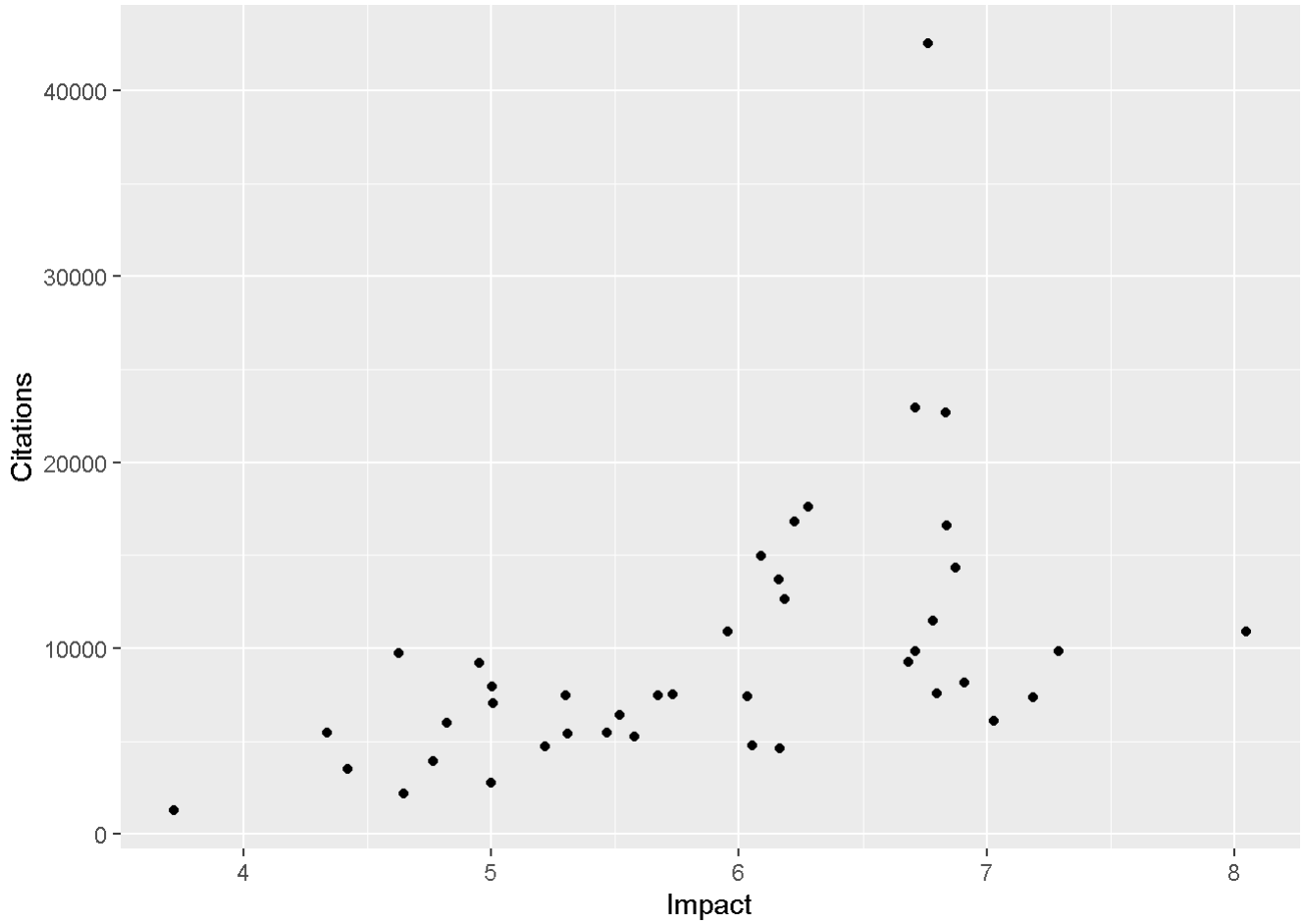
```
## 105 University of Memphis
## 106 University of Alabama at Tuscaloosa
## 107 DePaul University
## 108 American University
## 109 State University of New York at Albany
## 110 Drexel University
## 111 Bowling Green State University
## 112 University of Nevada at Las Vegas
## 113 Fordham University
## 114 Ohio University
## 115 Marquette University
## 116 Georgia State University
## 117 University of Nebraska at Lincoln
## 118 Utah State University
## 119 University of Missouri at Kansas City
## 120 University of Houston
## 121 Oklahoma State University at Stillwater
## 122 Saint Louis University
## 123 University of Southern Mississippi
## 124 University of Maryland at Baltimore County
## 125 University of Colorado at Denver
## 126 University of Missouri at Saint Louis
## 127 George Washington University
## 128 Alliant International University
## 129 University of Tennessee at Knoxville
## 130 Baylor University
## 131 University of North Carolina at Charlotte
## 132 Auburn University
## 133 University of Hawaii at Manoa
## 134 Lehigh University
## 135 University of North Texas at Denton
## 136 University of Louisville
## 137 University of Rhode Island
## 138 West Virginia University
## 139 Wayne State University
## 140 Texas Tech University
## 141 University of Arkansas at Fayetteville
## 142 Adelphi University
## 143 East Carolina University
## 144 Southern Illinois University at Carbondale
## 145 Brandeis University
## 146 City University of New York, City College
## 147 Columbia University
## 148 Cornell University
## 149 Johns Hopkins University
## 150 Loyola University Maryland
## 151 McGill University
## 152 Mount Sinai Hospital
## 153 Pepperdine University
## 154 Ponce Health Sciences University
## 155 Princeton University
## 156 Rutgers State University at New Brunswick
## 157 Stanford University
## 158 University of Chicago
```

```
## 159 University of Maine
## 160 University of Toronto
```

Wow, that's 160 universities. Let's have some standards. We'll first filter by 'Bolder' Boulder Model programs which are schools that have high research-related accreditation and membership standards. Then, we'll print the scatter plot of the 'Bolder' Boulder institutions: X = Impact, Y = Citations. As a reminder, Impact = Citations/Web of Science Documents). As another reminder, all citations and web of science documents are across a 5 year period from 2012 to the end of 2016.

```
Bolder_Boulder <- Graduate_Programs_Copy %>%
  filter(!is.na(Bolder_Boulder_Model))
ggplot(Bolder_Boulder, aes(Impact, Citations)) +
  geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

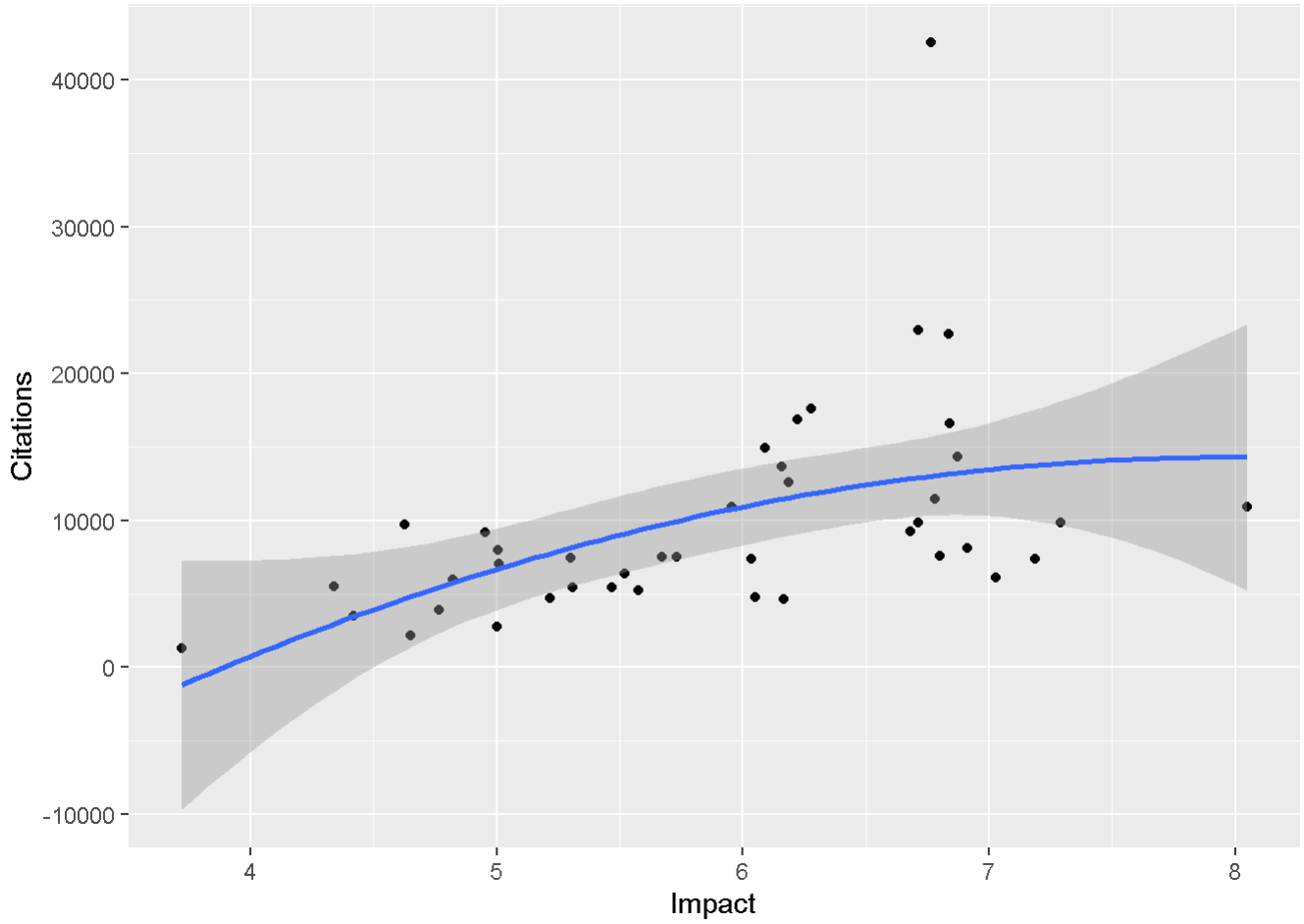


Based on this 'Bolder' Boulder criteria, we go from 160 institutions to 43. Since the data looks like a polynomial function, let's apply a quadratic regression curve of best fit to the scatter plot.

```
Bolder_Boulder %>%
  ggplot(aes(Impact, Citations)) +
  geom_point() +
  #plot line of best fit using quadratic regression
  geom_smooth(method = "lm", formula = y ~ x + I(x^2))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



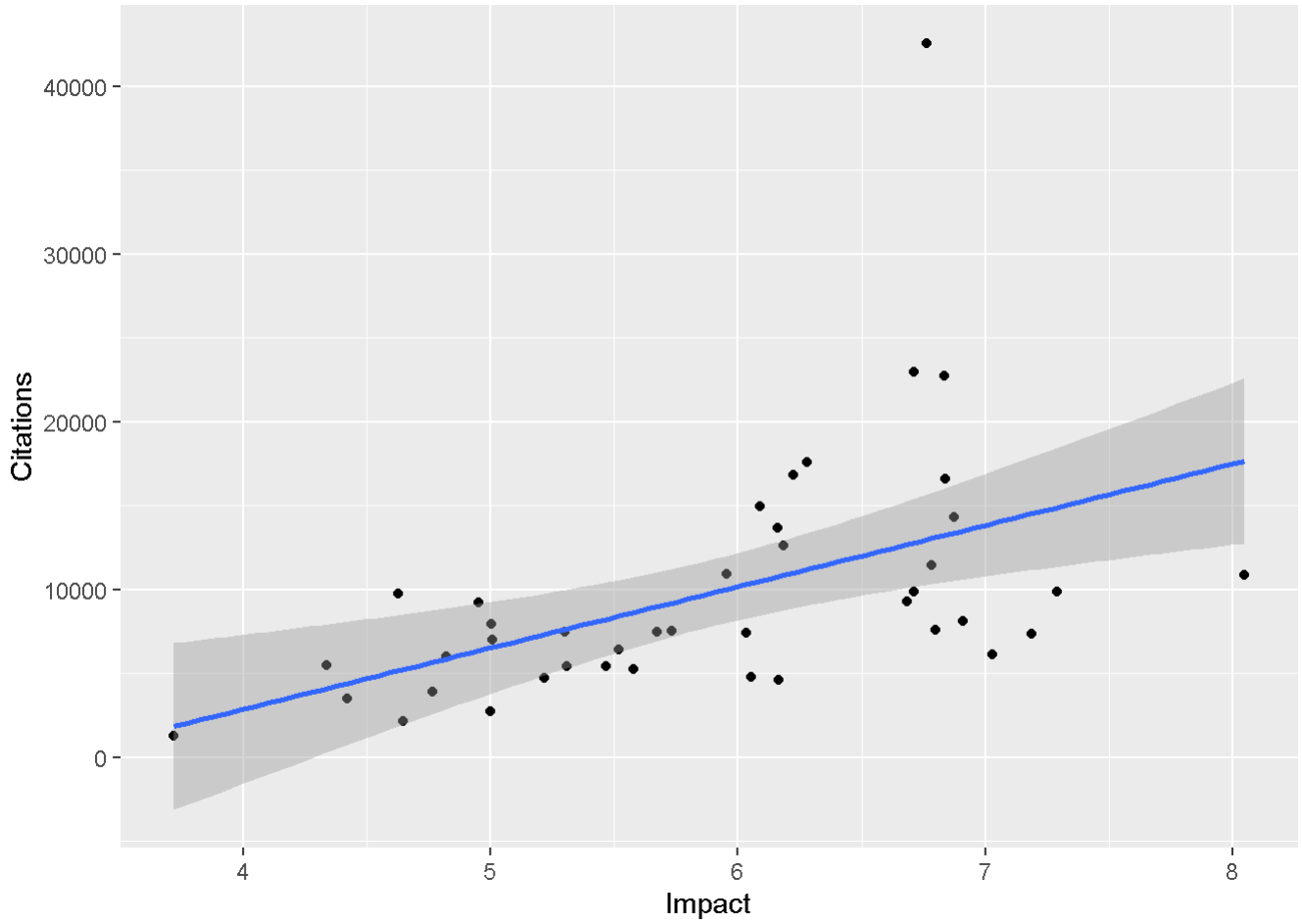
If we were to plot a tangent line, (unfortunately my R knowledge is limited) it would be around x= 6. This tangent line would represent the optimal number of citations and impact. However, now the data doesn't look like a polynomial function. It looks more like a linear function which would completely alter my goal of optimization. Let's plot a regression line of best fit and see how it looks.

```
Bolder_Boulder %>%
  ggplot(aes(Impact, Citations)) +
  geom_point() +
  #plot line of best fit using regression
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Hmmmm. Let's see the significance of these regression lines.

```
cor.test(Bolder_Boulder$Impact, Bolder_Boulder$Citations)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Bolder_Boulder$Impact and Bolder_Boulder$Citations
## t = 3.5302, df = 40, p-value = 0.001062
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2153567 0.6892222
## sample estimates:
##       cor
## 0.4873868
```

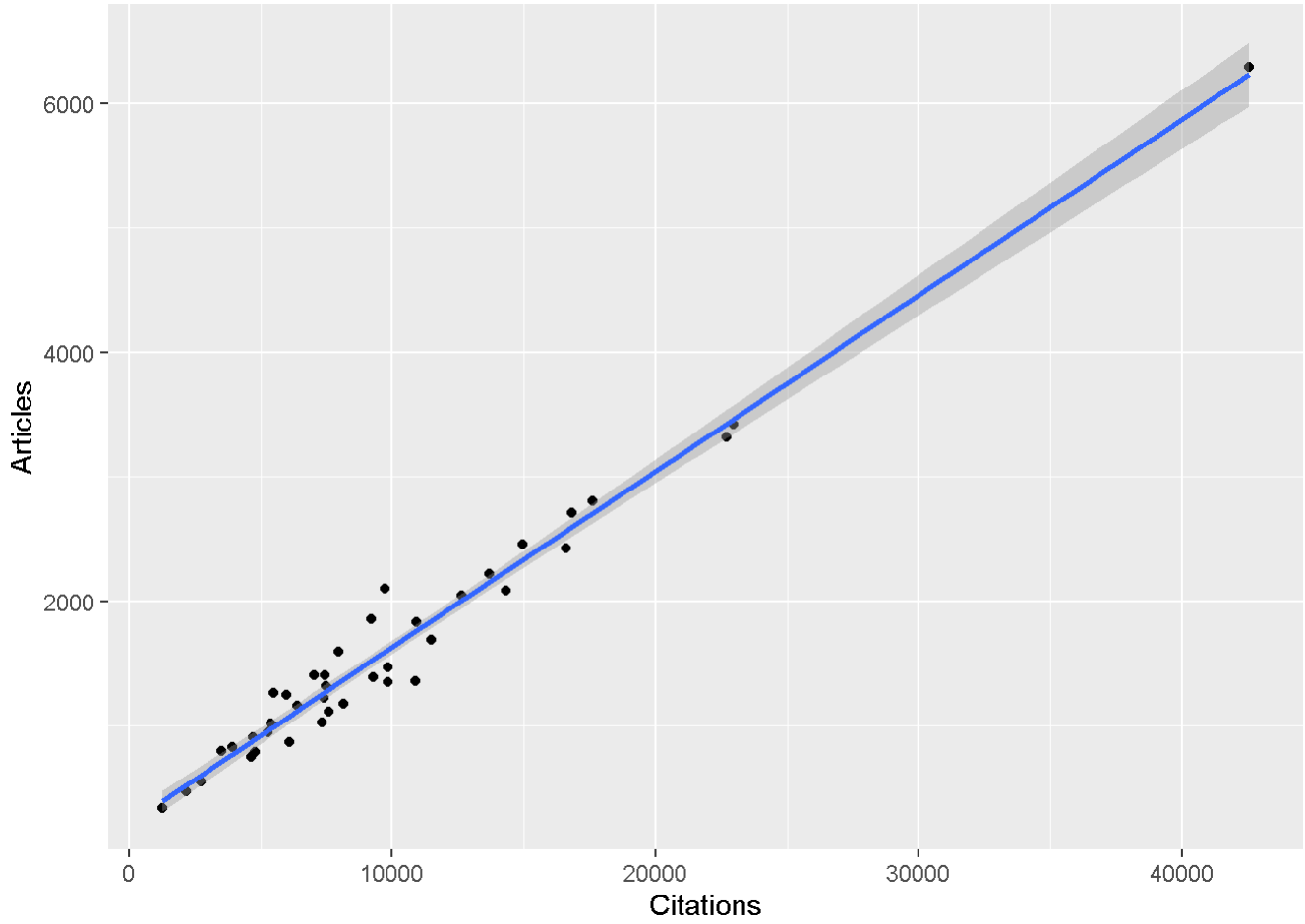
Okay, well, yes it makes sense that they're correlated because impact is directly derived from citations. Although, now I'm interested in seeing Articles by Citations of our 'Bolder' Boulder model universities. Then seeing if that is correlated. In other words, if you have more article publications, are you going to have more people citing those publications? You would certainly hope so as a university, otherwise you're operating on diminishing returns. Enough chatter, let's see the data.

```
Bolder_Boulder %>%
  ggplot(aes(Citations, Articles)) +
  geom_point() +
  #plot line of best fit using regression
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
cor.test(Bolder_Boulder$Articles, Bolder_Boulder$Citations)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Bolder_Boulder$Articles and Bolder_Boulder$Citations
## t = 37.542, df = 40, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9741266 0.9925581
## sample estimates:
##       cor
## 0.9861044
```

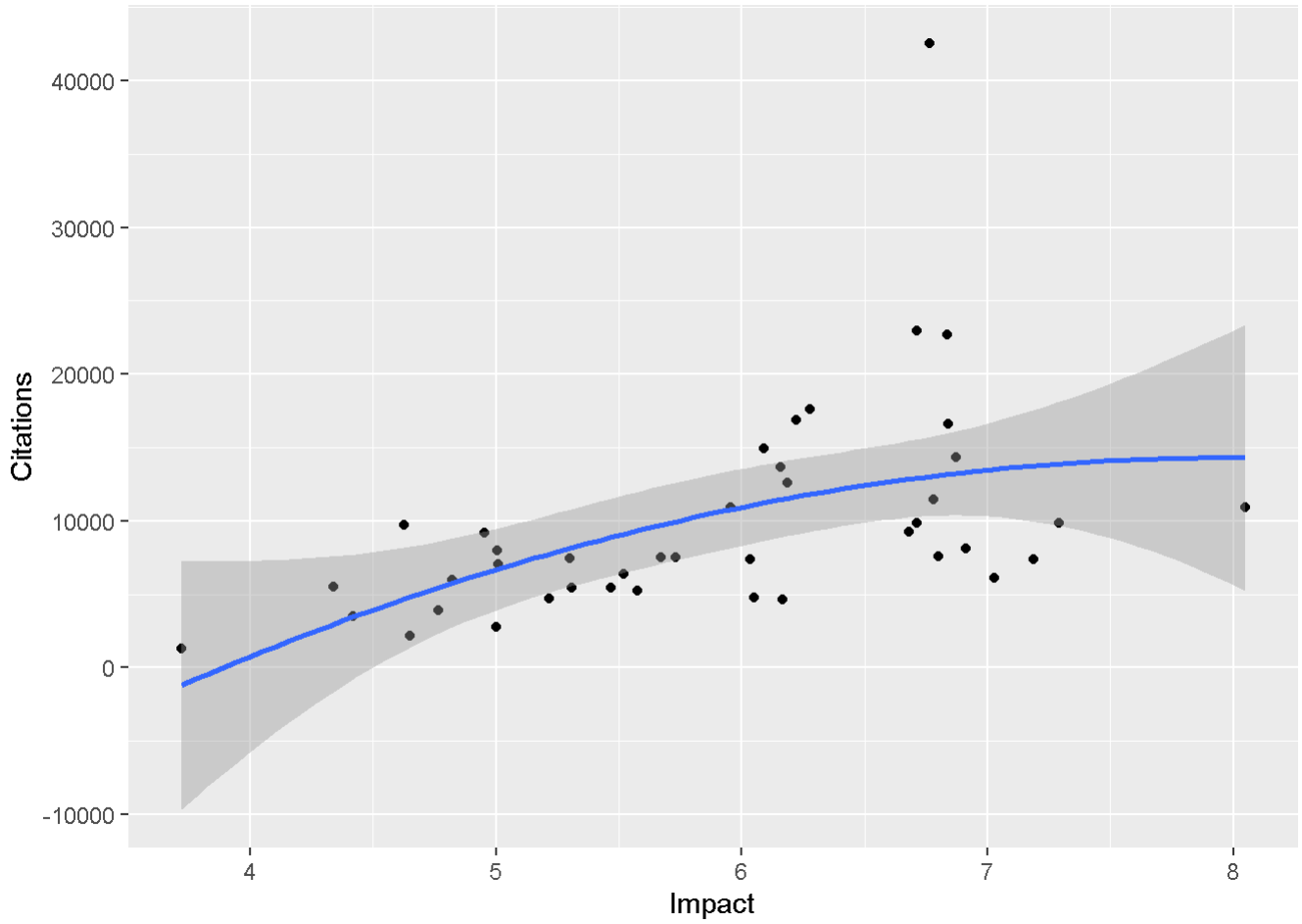
Hmmm. Interesting. Those are some pretty compelling results. Articles and citations are strongly correlated at p-value < 2.2e-16.

Nevertheless I digress. Let's go back to our optimal universities plot from before.

```
Bolder_Boulder %>%
  ggplot(aes(Impact, Citations)) +
  geom_point() +
  #plot line of best fit using quadratic regression
  geom_smooth(method = "lm", formula = y ~ x + I(x^2))
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Since there are 9 schools that cluster near x = 6. Let's see the names of those schools. I will print out the schools between 5.7 and 6.5. There should be 9.

```
filter(Bolder_Boulder, between(Impact, 5.8, 6.5)) %>%
  select(Institution)
```

```
## # A tibble: 9 x 1
##   Institution
##   <chr>
## 1 University of Pittsburgh
## 2 University of Michigan at Ann Arbor
## 3 University of Washington, Seattle
## 4 University of Oregon
## 5 University of North Carolina at Chapel Hill
## 6 University of Minnesota, Twin Cities
## 7 State University of New York at Stony Brook
## 8 University of Southern California
## 9 Northwestern University
```

I was surprised that local universities, Temple and UPenn, were not one of the 9 universities. After checking the impact number for Temple and Upenn, it was 5.31 and 6.84, respectively. Since the general rule is to apply to 10-12 schools, I decided to expand my impact range to 5.3 to 7. Although, during my grad school search process, I will make reasonable exceptions for other local schools, since I would like to stay local if possible.

```
filter(Bolder_Boulder, between(Impact, 5.3, 7)) %>%
  select(Institution)
```

```
## # A tibble: 26 x 1
##    Institution
##    <chr>
##  1 University of California, Berkeley
##  2 Duke University
##  3 University of Pennsylvania
##  4 University of California, Los Angeles
##  5 University of Iowa
##  6 Boston University
##  7 Harvard University
##  8 Emory University
##  9 Yale University
## 10 University of Wisconsin, Madison
## # ... with 16 more rows
```

Now we have 26 options. Okay cool, but Harvard is one of them. Not that it's a bad thing, but Harvard is an outlier with over 40,000 citations and only an impact of about 6.7. Let's exclude Harvard.

```
filter(Bolder_Boulder, between(Impact, 5.3, 7), Citations < 30000) %>%
  select(Institution)
```

```
## # A tibble: 25 x 1
##    Institution
##    <chr>
##  1 University of California, Berkeley
##  2 Duke University
##  3 University of Pennsylvania
##  4 University of California, Los Angeles
##  5 University of Iowa
##  6 Boston University
##  7 Emory University
##  8 Yale University
##  9 University of Wisconsin, Madison
## 10 University of Pittsburgh
## # ... with 15 more rows
```

Nice, now we are at 25 universities.

Let's store these universities in a new data frame with the relevant info.

```
optimal_uni <- filter(Bolder_Boulder, between(Impact, 5.3, 7),
                      Citations < 30000) %>%
  select(Institution, Citations, Articles, Impact, Valence)
```

Let's print out the data frame that has our optimal universities.

```
optimal_uni %>% print(n=25)
```

```
## # A tibble: 25 x 5
##    Institution                                Citations Articles Impact Valence
##    <chr>                                          <dbl>    <dbl>  <dbl>   <dbl>
##  1 University of California, Berkeley              8135     1177   6.91     4.6
##  2 Duke University                                14336     2086   6.87     4.2
##  3 University of Pennsylvania                     16606     2428   6.84     4.4
##  4 University of California, Los Angeles          22700     3321   6.84     4.8
##  5 University of Iowa                              7574     1114   6.80     4.2
##  6 Boston University                              11463     1690   6.78     4.1
##  7 Emory University                                9853     1468   6.71     4.2
##  8 Yale University                                22953     3420   6.71     4.2
##  9 University of Wisconsin, Madison                9275     1388   6.68     4.5
## 10 University of Pittsburgh                       17605     2804   6.28     4.4
## 11 University of Michigan at Ann Arbor            16839     2706   6.22     4.3
## 12 University of Washington, Seattle              12623     2041   6.18     4.5
## 13 University of Oregon                            4625      750   6.17     4
## 14 University of North Carolina at Chapel Hill    13682     2221   6.16     4.7
## 15 University of Minnesota, Twin Cities           14957     2456   6.09     4.5
## 16 State University of New York at Stony Brook     4787      791   6.05     4.6
## 17 University of Southern California               7399     1226   6.04     4.1
## 18 Northwestern University                        10926     1835   5.95     4.1
## 19 University of Maryland, College Park            7511     1310   5.73     4
## 20 Michigan State University                       7476     1318   5.67     3.8
## 21 University of Georgia                           5265      944   5.58     3.8
## 22 Indiana University at Bloomington               6403     1160   5.52     4.3
## 23 University of Missouri at Columbia              5452      997   5.47     3.8
## 24 Temple University                               5405     1018   5.31     4.3
## 25 University of Illinois at Urbana-Champaign      7455     1406   5.30     4.2
```

Now, let's pause for a moment. You might notice that I added a column called valence. Valence refers to U.S. News and World Report's Ranking survey of academics at peer institutions. Each variable reflects average rating from 1 (marginal) to 5 (outstanding) in clinical psychology graduate programs. Again, I'm interested in psych research, not so much clinical psych, but it could still be a good measurement of overall. All of the universities except Michigan State and Georgia are less than 4. So valence isn't going to be a useful measure after all.

As it stands, this data frame is sorted in descending order by impact which is what I want. However, you'll notice that there is a large degree of variance by citations (up to 10,000). That's fine, at this point, I'll individually browse each university's program and start excluding based on personal criteria. The average number of citations for our 25 optimum graduate schools across a 5 year span is:

```
mean(optimal_uni$Citations)
```

```
## [1] 10852.2
```

The average number of articles published for our 25 optimum graduate schools across a 5 year span is:

```
mean(optimal_uni$Articles)
```

```
## [1] 1723
```

In sum, I went from 160 universities to 43 to 25. I did this by first choosing universities that have a Bolder Boulder model for graduate programs which is essentially a model that trains scientific rigor. Next, I plotted the 'Bolder' Boulder model universities and applied a linear regression to find the optimal research impact by total number of citations.

Limitations:

The data is not the most up to date (off by about 5 years). This data is not limited to just psychology departments, it is looking across whole university academic departments. This isn't the most thorough way of looking at schools, but it is a good start for getting a start at which schools to look at. In the end, I think this was good practice with programming and using R while I looked at grad schools. As you'll see below, I also had some missing data from some notable R1 universities which could have impacted my total optimal schools.

Appendum:

After rereading through the code, there were 16 schools that I excluded because they didn't have any data on citations and consequently impact. Let's see what schools didn't have that data:

```
Graduate_Programs_Copy %>%
  select(Institution, Valence) %>%
  filter(is.na(Graduate_Programs_Copy$Citations))
```

```
## # A tibble: 16 x 2
##    Institution                              Valence
##    <chr>                                      <dbl>
##  1 Brandeis University                         NA
##  2 City University of New York, City College    3
##  3 Columbia University                         NA
##  4 Cornell University                          NA
##  5 Johns Hopkins University                    NA
##  6 Loyola University Maryland                   2.8
##  7 McGill University                           NA
##  8 Mount Sinai Hospital                        NA
##  9 Pepperdine University                       NA
## 10 Ponce Health Sciences University            NA
## 11 Princeton University                        NA
## 12 Rutgers State University at New Brunswick    3.3
## 13 Stanford University                         NA
## 14 University of Chicago                       NA
## 15 University of Maine                          2.9
## 16 University of Toronto                       NA
```

I was curious if I had any valence data on these universities. Unfortunately for what valence data exists, they do not look so good. I suppose I will briefly look at some of these schools during the grad school search process in addition to the optimal schools to be inclusive.