PUBP 8751 - Big Data

Matteo Zullo

January 27, 2020

Problem Set 1

Problem 1

What Alvarez really means is that big data is not a substitute for social science theory and statistical theory, but offers greater power to carry out statistical analysis as well as new data sources. Thus, "the new tools of data science are not themselves leading to a less theoretical social science; rather they can and should be used in conjunction with social science theory" (p.13).

This is better understood by way of an example. The bleeding edge of the social sciences over the last couple of decades has been the so-called "quasi-experiments", mainly Regression Discontinuity Design, Difference-in-Differences, and, recently, the Synthetic Control Method. The question that is of interest to us is: do quasi-experiments *necessarily* require big data? The answer is no, and, actually, the majority of quasi-experimental studies conducted so far use surveys, questionnaires, polls, and other "small data". The statistical theory behind each of these methods, and the identification conditions that each requires, do not depend on the data but on the experimental design. Therefore, answering the original question − whether experimental, statistical, or computational problems are mostly solved by big data analysis − we should say mostly computational. Big data allows researchers to go beyond the limitations of the small *Ns* that hampered social science research so far.

The data analytics revolution, which encompasses big data AND the machine learning tools applied to big or small data, is revolutionary in at least two respects: *i)* it gives researchers greater computational power to get the most out of small samples; and *ii)* gives them new sources of data that were not on the radar twenty or so years ago. Thinking of surveys, they are often times marred with missing data, poorly recorded entries, and other forms of machine/human errors that lead to underpowered studies. Now, data imputation has become a breeze thanks to the software available to researchers, and the complicated calculations involved by Multiple Imputation by Chained Equations now take very little time to run. This is just a quick-and-dirty example of *i)*. Another example is multi-level modeling, a regression-based technique that allows to estimate models with nested data (e.g. students nested within schools). Now, researchers can estimate three- or four-level models in (almost) the blink of an eye, and this thanks to the efficiency of the Maximum Likelihood Algorithms and the speed of modern-day software.

Most importantly, big data provides new sources of data that enable researchers to test implications derived from social science theory. Within my field of expertise, education policy, big data promises to change the game in student college retention (for reference, see the article "Foundations of Dynamic Learning Analytics: Using University Student Data to Increase Retention", De Freitas et al., 2015). It is not yet clear what the determinants of student retention are, and testing causal theories has proved difficult because of correlated inputs (see Manski, 1983). To see what that means, we can go with the mind to our first year in college and think of the plethora of inputs that came into play (e.g. relocation, choice of major, supervisor, etc.). Now, by harvesting institutional databases, researchers can test the effect of small-scale interventions, such as changes in schedule. If the hardest math class is shifted over from the first to the second semester, a RDD framework can be set up, and the change in retention rates between adjacent cohorts computed as an estimate of the treatment effect. This is just one example, but is a good one. Out of the 4 Vs, volume and velocity are clearly the most impacted; universities can micro-dose interventions and swiftly test their impacts with sufficient volume to produce reliable estimates - no surprise that A/B testing is one of the big things in the data analytics community right now. Also, the richness of the institutional databases allow to block many potential confounders - thus boosting the veracity of the estimates - and offers a lot of variety. Compared to traditional student surveys reporting student GPA and few more demographics, institutional databases provide constantly-updating snapshots of the student career (classes taken, grade in each class, etc.).

Wrapping up, the data analytics revolution, which includes big data but is not limited to it, bestows immense computational power upon researchers and gives them *i)* better and *ii)* newer ways to test their theories. What it cannot do is to advance social science theory and methodology. Therefore, it still holds true that results are only as good as the model is; however, the lack of data that murdered many good models in the cradle is less likely to do so as the "revolution" catches on.

Problem 2

A) Article 4 of the GDPR defines as personal data "any information relating to an identified or identifiable natural person ('data subject')". Thus, the definition encompasses IP addresses, location data (e..g work address), biometric data (e.g. weight, height), and behavioral data (e.g. voting behavior, religion). As a matter of fact, any information that can lead to the identification, "directly or indirectly" of individuals may fall under the shield of the GDPR. Clearly, this definition is highly contextual. In my own research, I was denied access to test scores of Italian students because the information (location of the school, class size, etc.) was deemed to be specific enough to enable the identification of individual students, teachers, and school managers. On the other hand, personal information such as individual name and surname might not necessarily lead to a violation of the GDPR if the information is not sufficient to uniquely identify subjects.

The CCPA adopts the GDPR's definition of "personal data" and broadens the pool of recipients to include households. This extension is not trivial: some data that might be easily linked to an household might not be that easily linked to the individuals within the household. However, there are two main fallouts of the CCPA when compared to the GDPR. First of all, the CCPA does not subject companies to invoke a "legal basis" for processing personal data. Second of all, it grants exemptions to categories which include financial services, healthcare services, and clinical trials. These categories are still regulated by industry-specific codes (e.g. HIPAA, CMIA) which are often times underdetermined to protect the privacy of individuals.

B) The CCPA intends to be a transparency act whereas the GDPR aims at minimizing the data that sits idle online without a clear "legal basis". From those lenses follow very different approaches to compliance. The GDPR takes an encompassing, grandiose stance, which subjects noncompliant companies to fines up to €20 million or 4% of the company's yearly turnaround. The CCPA, on the contrary, takes a per consumer-per violation approach, much less grandiose and yet potentially more burdensome. Businesses that fail to comply with any of the three consumer rights sanctioned by the CCPA – the right to access personal data within 12 months of their collection, the right to delete the data, and the right to opt out of sales of the data – are liable for maximum fines of $7,500 per consumer-per violation and $750 per consumer-per breach.

In general, the GDPR follows a rather "bureaucratic" approach whereas the CCPA follows a more "pragmatic" one, which come with major differences. The first of these differences is to be found in the accountability process: under the GDPR, companies are required to appoint a data protection officer who is supposed to produce extensive documentation on the data protection practices. By contrast, the CCPA does not envision preventive measures and grants leeway to companies as long as that they do not violate the three rights of consumers. More importantly, CCPA fines are issued by civil justice courts while GDPR fines are "administered by individual member state supervisory authorities". Everyone familiar with the bureaucratic hodgepodge that the EU is knows that this does not sound too good. More importantly for our discussion, this all means that violations and breaches are fined much more swiftly under the CCPA, and, thanks to the CCPA's narrower scope, the indictment process is much more straightforward. By this token, small and yet frequent penalties stockpile and easily lead to compliance costs which are up to 40% higher than those projected for the GDPR.

C) The definition of an externality in the legal context is that the private costs (benefits) associated with a provision do not match the social costs (benefits) of the same. Thanks to the availability of data, companies have been able to optimize their products and fine-tune their logistics to create a great deal of consumer value. The data analytics revolution is real and is testimonied in the McKinsey's report we discussed in class. Before data scraping and data mining were a thing, companies had to rely on expensive consumer surveys with very little velocity, veracity, volume, and variety. Post big data revolution, companies have been able to use Machine Learning techniques to streamline their production and logistics to better fit the preferences of consumers. Therefore, the size of the pie is likely to reduce if the amount of available data is to dramatically reduce as an effect of the CCPA. The negative externalities for consumers will be extensive: reduced product quality, reduced consumer spending and therefore increased unemployment.

One positive effect of the CCPA is the creation of a market for data, but this is not an externality. By putting a price tag on consumer data, the CCPA establishes what are *de facto* property rights in a market which is now unregulated. However, the benefit that each consumer makes from owning her data is pocketed by her alone.

To find positive externalities, we have to look somewhere else, and specifically at consumer trust and discrimination. The CCPA makes clear that times have changed and companies will be scrutinized for their use/misuse of data, and this expectation can enhance consumer trust in businesses. This qualifies as an externality because when consumer A invokes the CCPA and holds company X accountable for some violation, the positive effects of the sanction spill over on the rest of company's X customers. The second positive externality follows from the inevitable reduction of data that free-floats online and consists in reduced price-discriminatory power of firms. As soon as I got admitted to the MS in Data Analytics at Georgia Tech, a deluge of tech-related ads poured in into my feed, and so did real estate purchase opportunities. The ML models used by the marketers now classified me as tech-oriented, relatively high-SES individual; nothing particularly discriminatory, other than somehow legitimate price-discrimination. Imagine though if I was a member of some group which is deemed to be low-SES, low-purchasing power and so on, and classified as such by the ML algorithms of the marketers. I would have very different feelings about data analytics and how they are used to cluster customers. As a matter of fact, it is unclear where the dividing line between price-discrimination and outright discrimination is to be drawn, and the CCPA has the potential to generate a positive externality by steering away companies from discriminatory targeting.

Problem 3

A) The marginal probabilities are calculated using R and reported in the table below.

```r
prob.table <- matrix(c(654/1000, 79/1000, 127/1000, 42/1000,
    23/1000, 75/1000), ncol = 3)  # matrix
sum(prob.table)  # making sure it adds up to 1
```

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

```r
marg.prob.x1 <- matrix(apply(prob.table, 1, sum))  # marginal probabilities for x1
prob.table <- cbind(prob.table, marg.prob.x1)
marg.prob.x2 <- t(matrix(apply(prob.table, 2, sum)))  # marginal probabilities for x2
prob.table <- rbind(prob.table, marg.prob.x2)
rownames(prob.table) <- c("P(x1=0)", "P(x1=1)", "MPMF")
colnames(prob.table) <- c("P(x2=0)", "P(x2=1)", "P(x2=3)", "MPMF")
prob.table
```

```
##         P(x2=0) P(x2=1) P(x2=3)  MPMF
## P(x1=0)   0.654   0.127   0.023 0.804
## P(x1=1)   0.079   0.042   0.075 0.196
## MPMF      0.733   0.169   0.098 1.000
```

B) The probability that $P(x_1 + x_2 = 3)$ is the probability that $P(x_1 = 1)$ and $P(x_2 = 2)$ are jointly true. Formally, this is the intersection of two events. We do not have to do any calculation and we can simply look at element $(2,3)$ with $P = 0.075$. This probability represents the joint chance that a consumer is a solar homeowner and primarily searching plug-in electric cars and hybrids in the sample. It does not represent a conditional probability, and this difference will become crucial in the next answer.

C) The argument is flawed because it mixes up causation and correlation, and because it mixes up joint probability and conditional probability. In order for the argument to hold true, it needs to be that *i)* solar homewonership is determined first, and that *ii)* solar homewonership influences car buying searches. Both *i)* and *ii)* are unobserved, thus the claim is unwarranted. It is easy to think of a case where solar homeownership and car buying searches are determined at the same time, in which case neither one influences the other. If that is true, the data scientist should search for the antecedent x_3 that influces both x_1 and x_2. It is also possible - although intuitively less likely - that car buying searches determine solar homeownership. If that is the case, the data scientist would need to find the variable x_3 that really determines car buying searches and should not bother about solar homeownership.
Furthermore, we can imagine a case where both *i)* and *ii)* hold, and yet phasing out the investment is the worse choice. We need to introduce the notion of conditional probability: if solar homeownership determines car buying searches, the probability $P(x_2 = 2|x_1 = 1)$ *conditions* on homeownership status (i.e. the event-space is subsetted to $42 + 23 + 75 = 140$) and is therefore given by $\frac{75}{140} \approx 0.54$. By contrast, if car buying searches determine solar homeownership, the probability $P(x_1 = 1|x_2 = 2)$ *conditions* on car buying searches (i.e. the event-space is subsetted to $127 + 75 = 202$) and reduces to $\frac{75}{202} \approx 0.37$.
Therefore, if solar homeownership is determined first and solar homownership is effectively the link that drives car buying searches, the conditional probability $P(x_2 = 2|x_1 = 1) \approx 0.54$ is the one we should be looking at and would actually suggest to maintain hyper-targeting. Anyways, the claim is neither true or false; it is unwarranted.

Problem 4

I download the dataset and loaded it to R. I required 3 digits of precision for the outputs and avoided scientific notation.

```r
## Reading data
data <- read_csv("leaders.csv", col_types = list(year = col_integer(),
    country = col_factor(), leadername = col_character(), age = col_integer(),
    politybefore = col_double(), polityafter = col_double(),
    interwarbefore = col_factor(), interwarafter = col_factor(),
    civilwarbefore = col_factor(), civilwarafter = col_factor(),
    result = col_factor()))
attach(data)  # attaching data
```

1. 250 assasination attempts are recorded in the data, as many as there are rows in the "leaders.csv" dataset. 88 unique countries experienced attempts, and their average number per year/per country is 2.45.

```r
nrow(data)  # no. of assassination attempts
```

```
## [1] 250
```

```r
length(unique(country))  # no of. target countries
```

```
## [1] 88
```

```r
mean(tapply(country, year, length))  # no. of avg. attempts/year
```

```
## [1] 2.45
```

2. I created the new variable success, and it turned out that 54 out the 250 recorded attacks were successful (21.6%). Although we might have wanted a more even split to suggest that the "treatment" is effectively random, this results actually offers very little evidence for either case. Clearly, the event of drawing a

red card from a deck of 52 cards (i.e. $\frac{1}{2} = 0.5$) needs to happen with $P = 0.5$ if it is a random event indeed. However, the event of drawing an ace needs to happen with probability $\frac{4}{52} \approx 0.08$ if the draw is be truly random because there are only 4 aces in the deck. This shows that an even split is not necessarily required for an event to be classified as random. Therefore, the event "succesfully murdering a political leader" might not have a probability of 0.5 and yet be random. The true probability might be 0.3 because the weaponry available in the 1930 is rather poor, or because leaders have access to healthcare facilities that enhance their chance of surviving. The point is, this result does not prove nor disproves that the success of an assasination is randomly determined.

```r
levels(result)  # obtaining levels for result
```

```
##  [1] "not wounded"
##  [2] "dies within a day after the attack"
##  [3] "survives, whether wounded unknown"
##  [4] "wounded lightly"
##  [5] "plot stopped"
##  [6] "hospitalization but no permanent disability"
##  [7] "dies between a day and a week"
##  [8] "dies, timing unknown"
##  [9] "survives but wounded severely"
## [10] "dies between a week and a month"
```

```r
data <- data %>% mutate(success = factor(ifelse(result %in%
    c("dies within a day after the attack", "dies between a day and a week",
        "dies, timing unknown", "dies between a week and a month"),
    1, 0)))
table.success <- table(data$success)
table.success  # no. of successes
```

```
##
##   0   1
## 196  54
```

```r
prop.table(table.success)  # prop of successes
```

```
##
##     0     1
## 0.784 0.216
```

3. The test used for both investigations is a t-test for the difference of means. The test is appropriate because two groups are compared (i.e. success vs fail) and the politybefore variable is continuous. The sample mean of the politybefore index for successful attempts is -0.7 and the sample mean of the same index for failed attempts is -1.7, meaning that the average polity score three years prior to an assasination attempts was approximately one point lower in the case of successfull attempts. In other words, attempts appear to have greater chances of success when countries lean towards autochratic. That being said, we really want to pay attention to the t-statistic, which is associated with a p-value of 0.3 and therefore fails to reject the null hypothesis that the difference is due to chance alone. This might be due to the small sample size which is factored into the t-statistic calculation, or it might be that the true difference is not different than zero. Turning to age, it seems that this predictor matters. Murdered leaders are older, at a mean age of 56.5 yrs relative to 52.7 yrs of survivors, a difference which is significant at the 5% confidence level (p-value = 0.03). However, we need to be very careful about infering that this result threatens the experimental validity. The very definition of "bias" suggests caution: bias occurs when the treatment variable z is correlated with a predictor x which is also correlated with the outcome variable y. If the outcome variable y is, say, number of trade deals signed, and this is not correlated with the age of the leader (i.e. x) - please note that I making it up - the correlation of x and y would not be a cause for concern. Therefore, our conclusion is that age

is positively correlated with successfull attempts, but this would only constitute a threat to validity if the outcome variable(s) chosen is/are correlated with the age of the leader.

```r
# difference-of-means tests
t.test(politybefore[success == 1], politybefore[success == 0])  # politybefore
```

```
##
##  Welch Two Sample t-test
##
## data:  politybefore[success == 1] and politybefore[success == 0]
## t = 1, df = 83, p-value = 0.3
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.952  3.031
## sample estimates:
## mean of x mean of y
##    -0.704    -1.743
```

```r
t.test(age[success == 1], age[success == 0])  # age
```

```
##
##  Welch Two Sample t-test
##
## data:  age[success == 1] and age[success == 0]
## t = 2, df = 98, p-value = 0.03
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.433 7.065
## sample estimates:
## mean of x mean of y
##      56.5      52.7
```

4. I created the variable warbefore and run the logistic regression model $success \sim warbefore$. A bivariate logistics model entails the same statistical test of a difference-of-proportions test. The Wald-chi2 statistic associated with the warbefore parameter tests the null-hypothesis that the success rate for countries that experienced wars in the three preceding years is about the same rate of those which did not. The p-value associated with the t-statistic (0.78) lands well into the acceptance area, therefore we cannot reject the null hypothesis. Experiencing a war does not seem to be significantly related to the success of an attempt.

```r
## creating warbefore
data <- data %>% mutate(warbefore = ifelse(interwarbefore ==
    1 | civilwarbefore == 1, 1, 0))
data$warbefore <- factor(data$warbefore)

# difference of prop test - warbefore
diff.war.before <- glm(success ~ warbefore, data = data, family = "binomial")
summ(diff.war.before, model.info = FALSE, model.fit = FALSE)
```

```
## Standard errors: MLE
## --------------------------------------------------
##                        Est.   S.E.    z val.      p
## ----------------- -------- ------ --------- ------
## (Intercept)          -1.26   0.19     -6.56   0.00
## warbefore1           -0.09   0.32     -0.28   0.78
## --------------------------------------------------
```

5. To tackle part a), I required a t-test for polityafter and then run the regression $polityafter \sim success, politybefore$. The t-test requires a difference-of-means for the two groups of countries and does account for differences in the polity score predating assassination attempts. The t-test excludes that two variables are significantly related (p-value = 0.78), and the result is robust to the inclusion of politybefore as a control variable in the subsequent regression. There, the t-statistic on success is not significant with a p-value of 0.64, thus ruling out that succesfull leader assassinations cause democratization in countries with the same democratic score in the pre-treatment period (i.e. once holding politybefore constant).

   Turning to part b) of the question, I created the variable warafter which takes a 1 when countries experienced wars, civil or international, three years after the assassination attempts. I run a bivariate logistic model $warafter \sim success$ (not shown) followed by a multiple regression model $warafter \sim success, warbefore$ which accounts for whether countries went to war before the attempts. Results mirror each other: the relation between assassinations and the chance of experiencing wars is not significant (p-value = 0.18), whether or not warbefore is specified.

   I wanted to dig a little deeper and investigate international and civil wars separately. Results are quite interesting (only multiple regression outputs are shown) and evidence an almost insignificant relation of assassinations and civil wars (!); conversely, the link between assassinations and international wars is just a tad shy of statistical significance (p-value = 0.06). What the last output tells us is that the log-odds of countries waging war internationally decrease by -1.9 (!) upon leader elimination, a result which is significant at the 10% confidence level and accounts for international wars preceding assassinations. The interpretation can only be provisional and would require a whole lot of background reading on the causes and antecedents of civil and international wars which cannot be done here. One possible line of thought is that countries that kill their leaders are in a weaker position to start/join international wars because of the internal turmoil, political and financial. On the flipside, killings do not seem to be a cause for civil wars; this can be because civil wars, differently than international wars, have causes which are rooted in the political and social climate of the country and have less to do with political leaders.

```
# IV model - polityafter
t.test(polityafter[success == 1], politybefore[success == 0])  # t-test
```

```
##
##  Welch Two Sample t-test
##
## data:  polityafter[success == 1] and politybefore[success == 0]
## t = 1, df = 86, p-value = 0.3
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.951  2.912
## sample estimates:
## mean of x mean of y
##    -0.762    -1.743
```

```
diff.polity <- lm(polityafter ~ success + politybefore, data = data)  # IV model
summ(diff.polity, model.info = FALSE, model.fit = FALSE)
```

```
## Standard errors: OLS
## -------------------------------------------------
##                       Est.   S.E.   t val.      p
## ------------------ -------- ------ -------- ------
## (Intercept)          -0.43   0.27    -1.61   0.11
## success1              0.26   0.57     0.46   0.64
## politybefore          0.84   0.04    23.16   0.00
## -------------------------------------------------
```

```
# IV model - war
data <- data %>% mutate(warafter = ifelse(interwarafter == 1 |
    civilwarafter == 1, 1, 0))  # creating warafter
data$warafter <- factor(data$warafter)
diff.war1 <- glm(warafter ~ success, data = data, family = "binomial")
diff.war2 <- glm(warafter ~ success + warbefore, data = data,
    family = "binomial")
summ(diff.war2, model.info = FALSE, model.fit = FALSE)
```

```
## Standard errors: MLE
## ------------------------------------------------
##                       Est.   S.E.   z val.      p
## ----------------- ------- ------ -------- ------
## (Intercept)         -1.72   0.24    -7.19   0.00
## success1            -0.54   0.40    -1.34   0.18
## warbefore1           1.87   0.31     6.00   0.00
## ------------------------------------------------
```

```
# IV model - civilwar
diff.civilwar <- glm(civilwarafter ~ success + civilwarbefore,
    data = data, family = "binomial")
summ(diff.civilwar, model.info = FALSE, model.fit = FALSE)
```

```
## Standard errors: MLE
## ----------------------------------------------------
##                           Est.   S.E.   z val.      p
## --------------------- ------- ------ -------- ------
## (Intercept)               0.01   0.28     0.05   0.96
## success1                 -0.10   0.46    -0.21   0.83
## civilwarbefore0          -2.22   0.37    -6.09   0.00
## ----------------------------------------------------
```

```
# IV model - interwar
diff.interwar <- glm(interwarafter ~ success + interwarbefore,
    data = data, family = "binomial")
summ(diff.interwar, model.info = FALSE, model.fit = FALSE)
```

```
## Standard errors: MLE
## ----------------------------------------------------
##                           Est.   S.E.   z val.      p
## --------------------- ------- ------ -------- ------
## (Intercept)              -2.00   0.24    -8.42   0.00
## success1                 -1.09   0.57    -1.90   0.06
## interwarbefore1           1.58   0.39     4.01   0.00
## ----------------------------------------------------
```

Problem 5

The log file mylog.txt is a separate upload. There are four blobs in the container: High_data_public.csv, meter-readings-small.csv, plugshare-small.csv, true-sentiments.csv. The session started at 17:23:28.129397 and completed at 17:23:29.595701, therefore it took the file 1.466304 milliseconds to download.

Bonus Problem

A) The predictive question and the causal question(s) are distinct. The predictive question is: what are the predictors of the increase in the Streetscore index over the years 2007-2014? In other words, the question is addressed as long as the most important correlates of Streetchange 2007–2014 are identified. The causal questions test three theories that link an *explanans* to an *explanandum*: the tipping theory of urban change, the economic theory of human capital agglomeration, the invasion theory of sociology. They state that a given *explanans* (x) has a causal effect on physical urban change (y), and that this effect is not merely correlational (i.e. the two variables do not simply move together, x actually *causes* y to move).

The friction between correlation and causation comes up many times in the article, for example when the authors comment on the positive association of share of college-educated population and Streetchange 2007–2014 by acknowledging the possibility that "the relationship reflects the tendency of educated people to be willing to pay for neighborhoods that appear safer, rather than the ability of educated residents to make a neighborhood feel safe" (p.7574).

B) The reason is that human subjects are involved as primary source of data in the surveys administered through Amazon Mechanical Turk and to MIT graduate students. Subjects have to be treated according to the principles of respect, beneficience, justice, and their involvement has to entail informed consent, a sensible assessment of the risks and benefits of the research, and a fair selection of the study subjects.

C) The correlation between density and perceived safety exists and is pinpointed in both the Streetchange 2007 and the Streetchange 2007-2014 model. However, the correlation is far from being a test for the theory that predicts a causal relationship between density and perceived safety, the tipping theory. According to the theory, neighborhoods which are patrolled by the "eyes" of more neighbors should be safer. However, the causal link fails to be established because of two reasons. First of all, the transmission mechanism - i.e. the "eyes on the street" - is unobserved; secondly, the authors cannot generate counterfactuals and therefore the "finding does not imply that dense urban spaces are seen as safer than low-density suburban or rural areas, because we do not have such low-density spaces in our sample" (p.7574).

Interestingly, the researchers fail to make a cogent causal argument for all three theories. As for the human capital theory, the authors cannot rule out that increases in the density of college-educated population are simply correlated with physical urban change. As for the invasion theory, this "is just one of the reasons why areas may improve more when they have attractive neighbors" (p.7575) and it might be that some sort of regression to the mean is happening.