Problem I.

Q1. The dataset I am assigned to is the U.S. patent data managed by National Bureau of Economics (NEBR).

(1) Data Access: We can obtain the dataset from NBER website http://nber.org/patents/. There are two compressed versions are available for us, SAS transport (.tpt) files and ASCII comma-separated variable (.csv) files. There are a couple of files available but I think we should download the data in the line "**Patent data, including constructed variables**".
(2) Data Storage: Even though everyone can access the dataset from the NBER website, the source of the data is stored in the database on the NBER internal network. The NBER is working on some useful expansion of the dataset which will be released soon.
(3) Curators: The dataset is managed by NBER staff members. The PI for this project is Iain Cockburn. Besides that, a group of researchers and software engineers, led by Bronwyn Hall, Adam Jaffe, and Manuel Trajtenberg (2001), is responsible for managing this data.

Q2. There are a couple of key papers that have used this data, some key papers are:

(1) B. Hall, A. Jaffe, and M. Trajtenberg (2005), who researched on the relationship between market value and patent citations. This paper is cited 2764 times.
Citation:
Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38. http://www.jstor.org/stable/1593752
(2) J. Leskovec, J. Kleinberg, and C. Faloutsos used this dataset to build a dynamic graph model on innovation growth (2005). This paper is cited 1949 times.
Citation:
Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005, August). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 177-187). ACM.
(3) Z. Acs , L. Anselin, and A. Varga used this dataset to develop a new measurement of development of new knowledge (2002). This paper is cited 1537 times.
Citation:
Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. Research policy, 31(7), 1069-1085.

Q3. There are a couple of procedures on data collection:

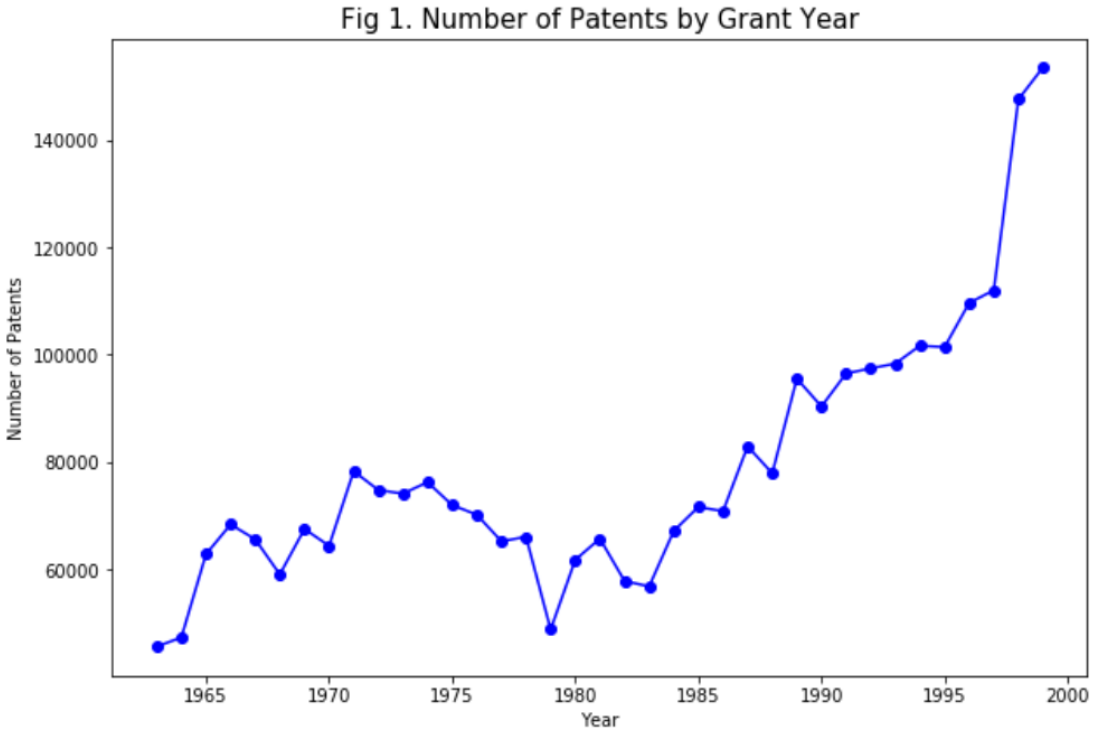
(1) The source of the data comes from the database of United States Patent and Trademark Office (USPTO), who had detailed information on almost three million patents granted between January 1963 and December 1999. It also includes all citations made to these patents between 1975 and 1999. Finally, the dataset is linked to Compustat that includes all firms traded in the US stock market.
(2) The original dataset contains 10 variables such as patent(Patent Number), gyear(Grant year), and appyear(Application Year). However, thinking these variables are not sufficient for reaching meaning results (as researches relied exclusively on simple patent counts as indicators of some sort of innovative output), Bronwyn Hall, Adam Jaffe, and Manuel Trajtenberg used an algorithm to compute some other variables including "originality" and "generality".
For detailed information on the variables included in the dataset, please refer to http://nber.org/patents/pat63_99.txt.

Q4. Please refer to Table 1 below.

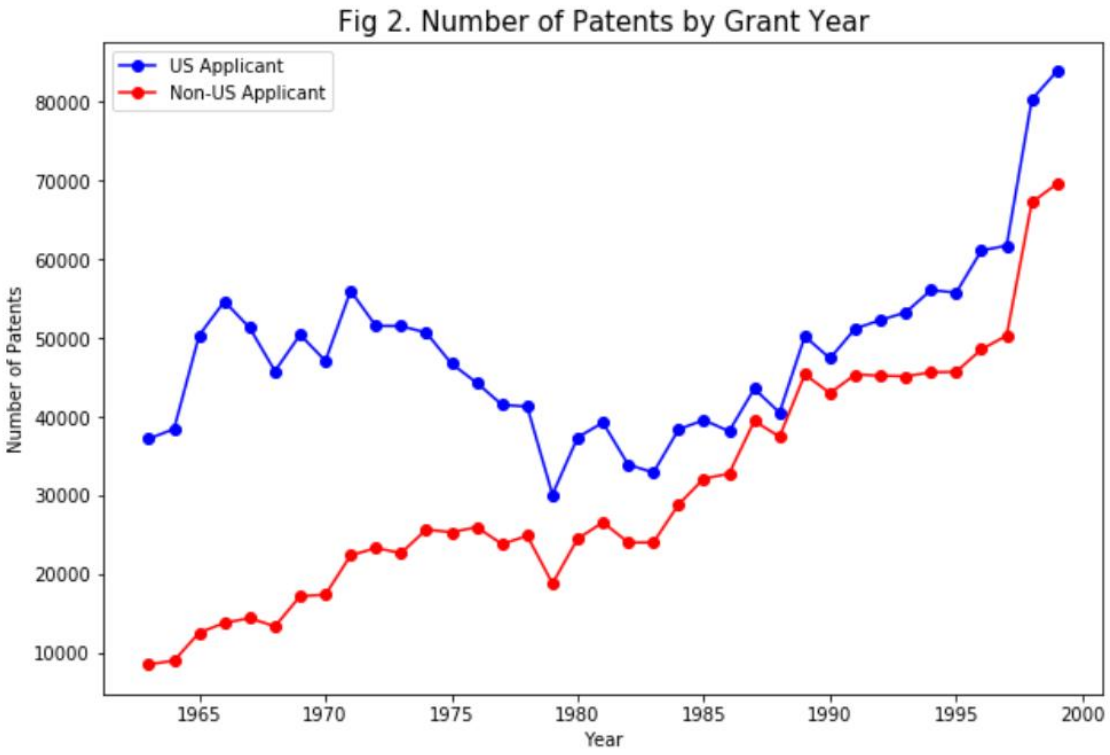
Table 1. Descriptive Statistics

Variables	Label	Counts	Mean	Std.	Min	Max
CLAIMS	Number of Claims	1984055	12.08	10.27	1	868
CMADE	Number of Citations Made	2139314	7.72	9.00	0	770
CRECEIVE	Number of Citations Received	2923922	4.78	7.35	0	779
RATIOCIT	Percent of Citations Made to Patents Granted	2088795	0.84	0.25	0.00	1.00
GENERAL	Measure of Generality	2240348	0.32	0.28	0.00	0.94
ORIGINAL	Measure of Originality	2042151	0.35	0.28	0.00	0.95
FWDAPLAG	Mean Forward Citation Lag	2074641	8.31	5.80	0.00	96.00
BCKGTLAG	Mean Backward Citation Lag	2088785	14.10	11.77	0.00	154.00

Q5. I made a plot to examine the relationship between year and number of patents granted in each year. Fig 1 shows that there are two periods when the number of granted patents increased rapidly: Late 1980s and late 1990s.



Q6. I split the dataset into two parts: US applicants and Non-US applicants. I am interested in the number of patents by grant year among US and non-US applicants separately. Please refer to Fig 2 for more details.



As we can see, the number of US applicants is way more than non-US applicants before the late 1980s, but the latter slowly catches up during the late 1980s. We may interpret the increase of patents granted during the late 1980s is caused by the rapid innovation outside of the US. Or it can be caused by the fact that many international companies tried to establish their business in the US in that era so as a part of protecting their interests, they submitted many applicants during that time.

We see the number of patents granted increases rapidly among both US applicants and Non-US applicants during the late 1990s. We may see the boom of innovation during the 1990s may be caused by the innovative efforts of people in and outside of the US.

References:

Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge. *Research policy*, *31*(7), 1069-1085.

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005, August). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 177-187). ACM.

Problem 2.

Q1. Research Question

General Topic: Why do insurance agents in India continue to give low-quality advice to their consumers?

Research Question 1: Do consumers with more sophistication (generally more knowledge about the insurance) tend to get better advice from their insurance agents?

Research Question 2: Does the disclosure regulation increase the quality of advice consumers receive from their insurance agents?

Research Question 3: Do the insurance agents know they are continuing to give low-quality advice to their consumers?

Q2. Data

The researchers use field experiments and surveys to answer their research questions, so the data are collected by the researchers themselves.

For the field experiments part, the researchers trained a decent number of middle-aged local men, who had at least high-school education and were in their thirties, as auditors, and sent them to life insurance agents to seek for advice. The auditors went to life agents were with varying levels of sophistication (both needs and biases). The researchers finally collected the advice auditors received and coded them based on the advice's quality. Note, to test the second research problem, the researchers used a similar experiment but conducted them in a subtle timeframe: Before and after the disclosure reform came out. For the survey part, the researchers conducted a paid 30-minute interview survey that involved 32 agents. Agents were not informed of the true research topic.

Q3. Theory

This paper is mainly based on two theories, both of them mainly come from the model proposed by Inderst and Ottaviani (2012).

The first theory is that agents are always commission motivated so that they are always providing consumers with the highest commission possible products which may not benefit the consumer themselves. We may treat the process of signing an insurance contract as bargaining. This means the more information consumers have, the less unbalanced this bargaining process will be, and the higher quality the advice consumers will get.

The second theory is that disclosure requirement can decrease the chance of poor advice from the agents since it helps consumers to realize that agents have a bias to maximize the commissions instead of maximize consumer interests.

Q4. Research Type

This research is mainly an identification exercise, with some levels of descriptive studies.

This research incorporates some levels of descriptive studies since it is trying to measure the quality of advice from insurance agents (See table 1 and table 2).

This research is mainly an identification exercise because it is trying to identify some important variables that have an impact on the quality of advice given by insurance agents. For example, the researchers varied the auditors with different beliefs and sophistication and used regression to show that auditor sophistication has an impact on the quality of advice from insurance agents (See table 3 and table 4). Similarly, researchers also tried to identify the impact of disclosures (See table 6).

Q5. Methods and Results

Overall, the researchers used randomized experiments as their main method.

In terms of computational methods, the researchers applied a series of logit regressions in their experimental design. Note that some models have fixed effect for auditors.

Method 1: A logit regression to test if agents cater to customer beliefs or respond to customer need. There are four dependent binary variables, includes Any Term (advise any term insurance), Only Term (advise term insurance only), Coverage (advise insurance with more risk coverage), and Premium (advise insurance with high premium). The main independent variables are mainly consumers' bias towards term insurance, consumers' need towards term insurance, and the interaction of the two.

Result 1: Results show both bias and need will increase 10% likelihood that agents advise some term insurance ($p<0.01$), increase 2% that agents advise some term insurance only ($p<0.05$), and increase 12% that agents advise insurance with more risk coverage ($p<0.05$).

Conclusion 1: Agents do cater to customer beliefs or customer needs.

Method 2: A logit regression to test if sophistication gets better advice. There are two dependent binary variables, recommended any term (advise any term insurance), and recommended only term (advise term insurance only). The main independent variable is Shopped Around (Sophistication).

Result 2: Some interaction terms are significant at the $p<0.05$ level, but some at not.

Conclusion 2: Shopping around may not be a signal of sophistication.

Method 3: A logit regression to test if disclosure regulations will increase the quality of advice. The main dependent variable is any ULIP recommendation, which is a binary variable. The main independent variable is Post Disclosure.

Result 3: The recommendation of ULIP drops from 65% to 40% after the disclosure reform.

Conclusion 3: Disclosure regulation will increase the quality given by insurance agents.

Method 4, an interview survey is not a computational method, so I am not going to refer it here.

Q6. Suggestions

Suggestion 1.

I am more from the social work side than the economics side, so I am going to discuss my feeling.

The main problem of this research is that it only uses a group of local men with adequate education who are in their thirties as their auditors (That is, the most advantageous populations). That makes the research suffer from external validity since it does not incorporate the disadvantage populations in the society, for example, children, females, minorities, and the old. I guess the researchers are only focusing on testing their hypothesis, but my hunch it that, with similar researches, we shouldn't go too theoretic. We

should have our researches to have the potential to help the disadvantaged populations in the societies, and neglecting them totally runs contrary to this principle.

My suggestion is that we should incorporate more people disadvantageous into this research so that we can find out a way to develop a better policy to alleviate inequality.

Suggestion 2.

This research seems do not have strong internal validity, especially construct validity. I am quite confused when I read the line “Thus, shopping around does not seem to be a signal of self-control or sophistication” (p. 11).

Let me interpret what the researchers are doing here. The researchers conduct a research with an untested construct, having no certainty whether it will fully represent the constructs in the theories they propose, putting the construct into a costly experiment and hope it would work. Such researches without fully tested construct validity are not tenable and are subject to criticisms.

My suggestion would be, conduct a pilot research and involve some insurance agents into the design of the experiment. If agents are hard to find, we may resort to people who were insurance agents but changed to other jobs later. We should have the constructs fully tested before conducting the experiment.

The trick here is, the effects of treatments in field experiments are always unexpected to researchers.

References

Anagol, S., Cole, S., & Sarkar, S. (2017). Understanding the advice of commissions-motivated agents: Evidence from the Indian life insurance market. *Review of Economics and Statistics*, *99*(1), 1-15.

Inderst, R., & Ottaviani, M. (2012). Competition through commissions and kickbacks. *American Economic Review*, *102*(2), 780-809.