Statistics, and Observational Data, Application to Pharmaceutical Detailing

Prof X

October, 2019

1 Introduction and Context

This case centers around a real-world, observational dataset from a pharmaceutical firm. The firm markets to doctors through *detailing* visits, where pharmaceutical representatives meet directly with doctors who might prescribe the drug to inform them of the drug's capabilities. Detailing is a massive part of the American pharmaceutical industry, and more is spent on detailing than on clinical trials, free samples, educational meetings, and on other forms advertising *combined*[1].

The dataset tracks how many prescriptions each doctor writes for the drug in a given month, how many detailing visits they received, and a few other characteristics. The firm is interested in more effectively targeting their detailing visits by figuring out which doctors are most likely to increase the number of prescriptions. The problem is particularly relevant as recent masters students have worked on similar problems at their new jobs. In total, the dataset tracks 1,000 physicians over 23 months. During this time these physicians wrote over 100,000 prescriptions for this drug and received over 40,000 detailing visit.

We will explore this dataset and business problem to practice:

1. Thinking through the analysis in the context of the business problem
2. Using R to investigate correlations in our data
3. Interpreting regression coefficients and categorical variables
4. The use and interpretation of interaction effects

2 Basic Descriptives and Correlations

1. Change the working directory to location where the file `Detailing Case Data.csv` is located. Use the command `setwd`.
2. Load the data in a data table named `detailData`. The loaded dataset should have 23000 observations and 27 variables.
3. Use the commands `names` and the `summary` function to get more information on the contents of each variable. Variables descriptions are as follows:
    - **scripts**: Number of perscriptions the doctor wrote in this month
    - **detailing**: Number of detailing visits the doctor receive this month
    - **lagged_scripts**: Number of perscriptions the doctor wrote last month
    - **mean_samples**: Average number of free samples the doctor received
    - **doctorType**: Factor variable indicating is this doctor general practioner, a specialist in this therapeutic class, or a specialist in a different area

This is the complete dataset, other information, such as the month of the visit, is not available.

Please answer the following questions:

[1] This plot comes from the Pew Institute

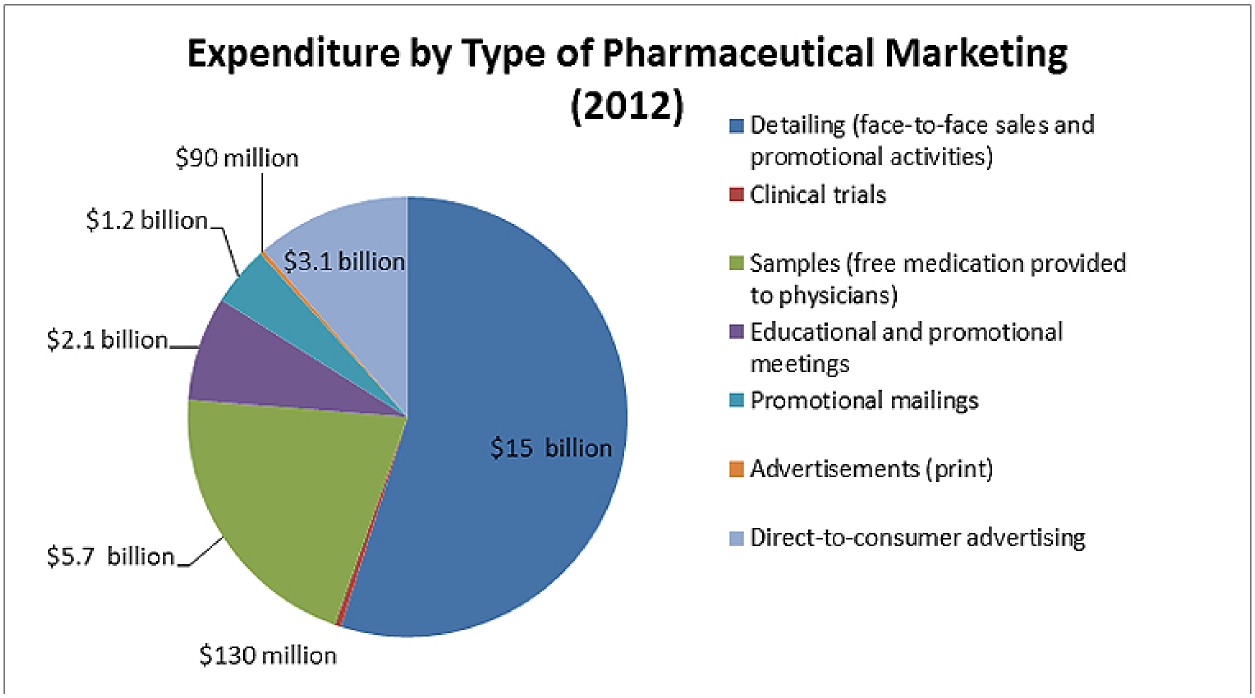

Figure 1: Marketing Spending by the Pharmaceutical Industry.

1. Given these variables, speculate on what the relationships between these variables might be.
2. What relationships do you think you will find?
3. One of the most exciting things about data analysis is that you can actually prove yourself wrong. If we already knew all the answers, then we wouldn't have to analyze the data. What do you think is going on here? How might you test your theory?
4. Given the firm's question, are we performing a descriptive, predictive, or causal analysis? Why?
5. Check the correlation between the two variables of interest here, `scripts` and `detailing` with the `cor` and `cor.test` function. These are columns in the `detailData` dataframe. To get a column, we use the `$` symbol:

3 Univariate Regression

Run a regression with `scripts` as the dependent variable and `detailing` as the independent variable. Construct a table with the `estimated` values and the `standard errors`

$$\texttt{scripts} = \texttt{doctorType} + \beta_1\texttt{detailing} + \varepsilon \tag{1}$$

Please answer the following questions:

1. According to this regression, how many prescriptions would a physician write if they received 3 detailing visits?
2. If a detailing visit costs the pharmaceutical company 200 dollars, and a new prescription generated 1000 dollars of revenue, would detailing visits be worthwhile?
3. Calculate an approximate 95% confidence interval for the coefficient of detailing.
4. Would your conclusions change if the coefficient was at the top/bottom of this confidence interval?

4 Multivariate Regression

Estimate using OLS the following model.

$$\texttt{scripts} = \beta_0 + \beta_1\texttt{detailing} + \beta_2\texttt{lagged\_scripts} + \beta_3\texttt{mean\_samples} + \varepsilon \tag{2}$$

Please answer the following questions:

1. Do you think it was a good idea to add those two regressors? Compare the change in magnitude and sign (if any) with the univariate case
2. Are the coefficients of `lagged_scripts` positive or negative? Given the context, Can you find an explanation of why might this be the case?
3. If a detailing visit costs the pharmaceutical company 200 dollars, and a new prescription generated 1000 dollars of revenue, would detailing visits be worthwhile?
4. Calculate an approximate 95% confidence interval for the coefficient of detailing. Would your conclusions change if the coefficient was at the top/bottom of this confidence interval?

5 Categorical Variables

Estimate using OLS the following model.

$$\texttt{scripts} = \texttt{doctorType} + \beta_1\texttt{detailing} + \beta_2\texttt{lagged\_scripts} + \beta_3\texttt{mean\_samples} + \varepsilon \tag{3}$$

Different types of doctor may be more or less likely to prescribe this particular drug. To account for this in the analysis, we need to control for `doctorType`.

However, in this dataset there are three different kinds of doctors (General Physicians, Area Specialists, and Other Specialists), so we will need to treat this as a categorical variable. We do this by using the `factor` function in the regression formula:

Please answer the following questions:

1. What is the interpretation of the coefficient of `factor(doctorType)General Physician`?
2. What is the interpretation of the coefficient of `factor(doctorType)Other Specialist`?
3. Recall that the firm is interested in targeting their detailing more efficiently. Based on this analysis, who should the pharmaceutical firm be targeting?

6 Interaction Effects

$$\texttt{scripts} = \texttt{doctorType} \times \beta_1\texttt{detailing} + \beta_2\texttt{lagged\_scripts} + \beta_3\texttt{mean\_samples} + \varepsilon \tag{4}$$

Please answer the following questions:

1. Are the interactions significant? Comment.
2. How much does a single detailing visit increase prescriptions for a General Physician? How about for an Area Specialist?
3. Why is the coefficient on `detailing` so much higher now that we've controlled for interactions?
4. Are we ready to answer the firm's fundamental question: who should they target?

7 Conclusions and Final Discussion Questions

With this final analysis, we can now plausibly answer the firm's question.

Please answer the following questions:

1. What are the key take aways of this Case?
2. When analyzing observational data, if you don't control for the right variables, you can get a terribly wrong answer. Thinking through the context is crucial to figuring out if you have the right set of controls
3. Categorical variables and interaction effects have real, important effects on the results of an analysis. In some cases, they are strictly required to even be able to answer the question at hand
4. You should directly care about your coefficient estimates since they are what map into the business decision
5. Based on this analysis, who should the firm target? If a detailing visit costs the pharmaceutical company 200 dollars, and a new prescription generated 1000 dollars of revenue, would this targeting strategy be worthwhile?
6. What are some variables *not* in this dataset that we might want to control for in this context?
7. Beyond changes in the detailing strategy, is there anything else you might recommend to the firm do based on this analysis?