Case 5: Causality and Observational Data, Application to Pharmaceutical Detailing

Dr. Avery Haviv
University of Rochester
GBA436R

Fall, 2023

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].

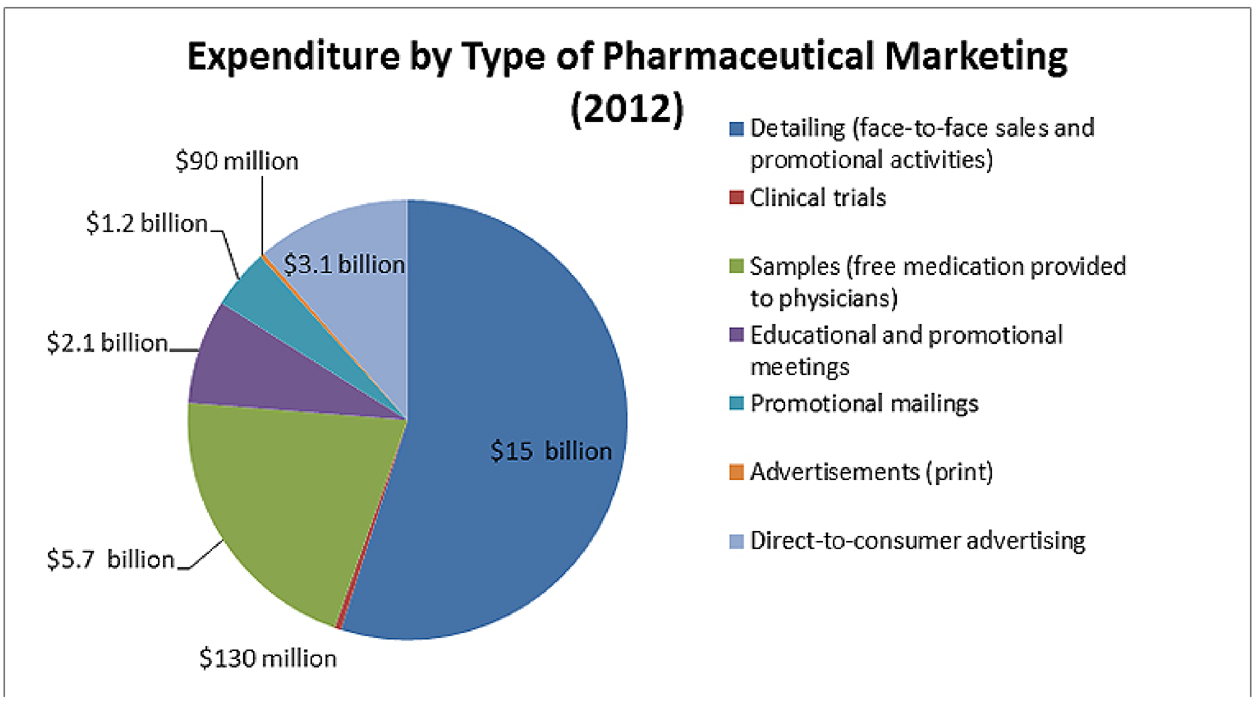

Figure 1: Marketing Spending by the Pharmaceutical Industry.

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

[1] This plot comes from the Pew Institute

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

1.1 How to use this case

- Code will be marked using the monospaced Courier New font. For example, we will run regressions with the `lm` function.
- At the end of each section, I will provide some discussion questions. In a separate document, I will provide the solutions to these discussion questions. To get the most out of the case, I recommend you attempt to solve the questions, in writing, and then check your answer afterward.
- I will elaborate on some points using footnotes. These footnotes are explicitly not testable material. They might help your understanding or provide some interesting facts.

2 Basic Descriptives and Correlations

First, we need to change the directory, and load the dataset into R. I am doing setting the directory with `setwd` function in R, but you can also do this using the 'Session' menu in RStudio:

```
setwd('C:/Dropbox/Teaching Lectures/Detailing Case')
detailData = read.csv('Detailing Case Data.csv')
```

You will need to make adapt that code based on where you stored the file. The loaded dataset should have 23000 observations and 27 variables.

We can use the `names` function to see the variables in this dataset, and the `summary` function to get more information on the contents of each variable.

```
names(detailData)
```

```
## [1] "X"              "scripts"        "detailing"      "lagged_scripts"
## [5] "mean_samples"   "doctorType"     "doctorID"
```

```
summary(detailData)
```

```
##        X            scripts         detailing      lagged_scripts
##  Min.   :    1   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 5751   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median :11500   Median : 3.000   Median : 2.000   Median : 3.000
##  Mean   :11500   Mean   : 5.061   Mean   : 1.883   Mean   : 5.093
##  3rd Qu.:17250   3rd Qu.: 6.000   3rd Qu.: 3.000   3rd Qu.: 6.000
##  Max.   :23000   Max.   :96.000   Max.   :18.000   Max.   :96.000
##   mean_samples     doctorType           doctorID
##  Min.   :0.0000   Length:23000       Min.   :   1.0
##  1st Qu.:0.1424   Class :character   1st Qu.: 250.8
##  Median :0.3435   Mode  :character   Median : 500.5
##  Mean   :0.5640                      Mean   : 500.5
##  3rd Qu.:0.7783                      3rd Qu.: 750.2
##  Max.   :4.9609                      Max.   :1000.0
```

The descriptions of the variables are as follows:

- **scripts** the number of perscriptions the doctor wrote in this month
- **detailing** the number of detailing visits the doctor receive this month
- **lagged_scripts** the number of perscriptions the doctor wrote last month
- **mean_samples** the average number of free samples the doctor received
- **doctorType** factor. Is this doctor general practioner, a specialist in this therapeutic class, or a specialist in a different area
- **doctorID** factor. An indicator for each doctor.

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

Discussion Questions:

1. Given these variables, speculate on what the relationships between these variables might be. What relationships do you think you will find? What do you think is going on here? How might you test your theory? 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.
2. Given the firm's question, are we performing a descriptive, predictive, or causal analysis? Why?

First, we can 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:

```
cor(detailData$scripts,detailData$detailing)
```

```
## [1] 0.2175696
```

```
cor.test(detailData$scripts,detailData$detailing)
```

```
##
##  Pearson's product-moment correlation
##
## data:  detailData$scripts and detailData$detailing
## t = 33.804, df = 22998, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2052229 0.2298471
## sample estimates:
##       cor
## 0.2175696
```

There is a positive correlation! Furthermore, the correlation is statistically significant since the p-value is less than 0.05. This means that these variables tend to move together, which is good to know. It does not mean

- that detailing increases scripts. Correlation is not causation. We will have a better idea of the potential causal relationship as we incorporate more independent variables.
- that detailing has a large effect on the number of prescriptions. Correlation only tells you how predictable the relationship is. The regression coefficient tells you how much one variable affects the other

We can look at multiple correlations at the same time using the same `cor` function. Below, we we look at the correlations between `scripts`, `detailing`, and `mean_samples` simultaneoulsy.

```
cor(detailData[,c('scripts','detailing','mean_samples')])
```

```
##                scripts detailing mean_samples
## scripts      1.0000000 0.2175696    0.4140847
```

```
## detailing    0.2175696 1.0000000    0.3766691
## mean_samples 0.4140847 0.3766691    1.0000000
```

While `scripts` is correlated with `detailing`, `mean_samples` shows a stronger correlation with both variables. [2]

3 Interpretation of a Univariate Regression

As you learned in your statistics classes, we will first run a regression with `scripts` as the dependent variable and `detailing` as the independent variable. We will use the `summary` function to get the standard errors:

```
summary(lm(scripts~detailing,data=detailData))
```

```
##
## Call:
## lm(formula = scripts ~ detailing, data = detailData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.448  -3.990  -2.231   0.889  90.829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.29142    0.07081   46.48   <2e-16 ***
## detailing    0.93977    0.02780   33.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.232 on 22998 degrees of freedom
## Multiple R-squared:  0.04734,    Adjusted R-squared:  0.0473
## F-statistic:  1143 on 1 and 22998 DF,  p-value: < 2.2e-16
```

We can see that there is a significant relationship here! Each detailing visit increases the number of scripts written by an average of 0.93977. If a doctor does not receive any detailing visits, then they write an average of 3.29 scripts.

Discussion 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. Would your conclusions change if the coefficient was at the top/bottom of this confidence interval?

4 Multivariate Analysis

The advantage of regressions is they allow you to control for different variables. Therefore, we will now control for `lagged_scripts` and `mean_samples`.

```
summary(lm(scripts~detailing+lagged_scripts+mean_samples,data=detailData))
```

[2] We can look at the full correlation matrix using the `corr` function. However, we have to careful because `doctorType` is a factor (or categorical) variable. We want to go from the single column category to the multicolumn binary variables, discussed in the notes. This can be done with the `model.matrix` function, which I will demonstate in a later case.

```
##
## Call:
## lm(formula = scripts ~ detailing + lagged_scripts + mean_samples,
##     data = detailData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.565  -1.850  -0.413   1.511  32.228
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.330081   0.041177   8.016 1.14e-15 ***
## detailing      0.071813   0.016308   4.404 1.07e-05 ***
## lagged_scripts 0.809921   0.003831 211.427  < 2e-16 ***
## mean_samples   0.834614   0.046108  18.101  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.921 on 22996 degrees of freedom
## Multiple R-squared:  0.7201, Adjusted R-squared:   0.72
## F-statistic: 1.972e+04 on 3 and 22996 DF,  p-value: < 2.2e-16
```

It's a good thing we controlled for these additional variables! Both have a large, significant effect. Furthermore, the coefficient on `detailing` is less than a tenth of its previous estimate[3]. This should demonstrate to you why compiling a thorough list of control variables is essential: If you do not, you can get a completely incorrect answer.

Discussion Questions:

1. The coefficient of `lagged_scripts` is positive. Thinking about the context, why might this be the case?
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. Would your conclusions change if the coefficient was at the top/bottom of this confidence interval?
4. Let's think through why including including `lagged_scripts` and `mean_samples` reduced the coefficient of `detailing` so much.

5 Categorical Variables

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:

```
summary(lm(scripts~detailing+lagged_scripts+mean_samples+factor(doctorType),data=detailData))
```

```
##
## Call:
## lm(formula = scripts ~ detailing + lagged_scripts + mean_samples +
##     factor(doctorType), data = detailData)
##
## Residuals:
```

[3] In general, controlling for additional variables will *not* change the coefficient estimate in an experiment, which is why experiments are so valuable

```
##     Min      1Q  Median      3Q     Max
## -45.384  -1.840  -0.300   1.538  32.859
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          2.09960    0.08008  26.218  < 2e-16 ***
## detailing                            0.09579    0.01615   5.931 3.05e-09 ***
## lagged_scripts                       0.75809    0.00428 177.109  < 2e-16 ***
## mean_samples                         0.88280    0.04586  19.249  < 2e-16 ***
## factor(doctorType)General Physician -1.92859    0.07674 -25.132  < 2e-16 ***
## factor(doctorType)Other Specialist  -1.95691    0.09037 -21.653  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.865 on 22994 degrees of freedom
## Multiple R-squared:  0.7279, Adjusted R-squared:  0.7278
## F-statistic: 1.23e+04 on 5 and 22994 DF,  p-value: < 2.2e-16
```

Discussion 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

Even though we've run a very reasonable analysis, and we've controlled for every available variable in the dataset, we still cannot answer the firm's fundamental question: who should they target? Figuring out who to target *requires* an interaction effect. We need to know how the effect of `detailing` changes for different groups, which is exactly what interactions allow us to do. Interactions are important, and you should care about them!

The following regression includes an interaction between `detailing` and `doctorType` using the `*` symbol, which multiplies the two terms them together. This also includes the normal, linear effects for each variable:

```
summary(lm(scripts~detailing*factor(doctorType)+lagged_scripts+mean_samples,data=detailData))
```

```
##
## Call:
## lm(formula = scripts ~ detailing * factor(doctorType) + lagged_scripts +
##     mean_samples, data = detailData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.623  -1.836  -0.358   1.541  32.337
##
## Coefficients:
##                                               Estimate Std. Error t value
## (Intercept)                                   1.616035   0.098075  16.478
## detailing                                     0.364417   0.035337  10.313
## factor(doctorType)General Physician          -1.327022   0.106831 -12.422
## factor(doctorType)Other Specialist           -1.333558   0.121866 -10.943
## lagged_scripts                                0.753227   0.004314 174.618
## mean_samples                                  0.887155   0.045947  19.308
## detailing:factor(doctorType)General Physician -0.319701   0.039395  -8.115
## detailing:factor(doctorType)Other Specialist  -0.359380   0.050366  -7.135
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## detailing                                     < 2e-16 ***
## factor(doctorType)General Physician           < 2e-16 ***
## factor(doctorType)Other Specialist            < 2e-16 ***
## lagged_scripts                                < 2e-16 ***
## mean_samples                                  < 2e-16 ***
## detailing:factor(doctorType)General Physician 5.09e-16 ***
## detailing:factor(doctorType)Other Specialist  9.94e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.859 on 22992 degrees of freedom
## Multiple R-squared:  0.7288, Adjusted R-squared:  0.7287
## F-statistic:  8825 on 7 and 22992 DF,  p-value: < 2.2e-16
```

The interaction effects are clearly significant, and it turns out they are quite important! However, the analysis is a bit tricky to interpret because it combines categorical variables with an interaction effect.

Discussion Questions:

1. How much does a single detailing visit increase prescriptions for a General Physician? How about for an Area Specialist?
2. Why is the coefficient on `detailing` so much higher now that we've controlled for interactions?

7 Fixed Effects

Do we necessarily believe that all doctors within a type are equally likely to perscribe this drug? Is it possible that even within a doctor type, some doctors are more likely to perscribe the drug, and are also more likely to be detailed? If so, we will want to the identity of each doctor. This is a lot of coefficients to add, but that's okay! If it helps us reduce bias, we should include the variable.

Including a large categorical variable like this is formally called a *fixed effect*. Putting such a large variable into the `lm` function may not be easy computationally. Instead, we will use the `felm` function in the `lfe` package. Note the syntax below - put the large fixed effects after the | symbol:

```
install.packages('fixest', repos='http://cran.us.r-project.org')
```

```
## Installing package into 'C:/Users/Avery/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'fixest' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Avery\AppData\Local\Temp\RtmpyQxyjw\downloaded_packages
```

```
library('fixest')
  summary(feols(scripts~detailing*factor(doctorType)+lagged_scripts+mean_samples|factor(doctorID),data
```

```
## The variables 'factor(doctorType)General Physician', 'factor(doctorType)Other Specialist' and 'mean_s
```

```
## OLS estimation, Dep. Var.: scripts
## Observations: 23,000
## Fixed-effects: factor(doctorID): 1,000
## Standard-errors: Clustered (factor(doctorID))
##                                               Estimate Std. Error  t value
## detailing                                     0.303723   0.094038  3.22978
## lagged_scripts                                0.277893   0.029050  9.56610
## detailing:factor(doctorType)General Physician -0.249304   0.096366 -2.58705
## detailing:factor(doctorType)Other Specialist  -0.244269   0.103792 -2.35344
##                                                Pr(>|t|)
## detailing                                     0.0012793 **
## lagged_scripts                                < 2.2e-16 ***
## detailing:factor(doctorType)General Physician 0.0098205 **
## detailing:factor(doctorType)Other Specialist  0.0187935 *
## ... 3 variables were removed because of collinearity (factor(doctorType)General Physician, factor(doc
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.29645     Adj. R2: 0.793031
##                 Within R2: 0.081323
```

Discussion Questions:

1. Why are so many of the coefficients NA now? Is this a problem?
2. Compare the standard errors in this analysis with the previous one. What happened to them? Why?

8 Conclusion and Final Discussion Questions

With this final analysis, we can now plausibly answer the firm's question. I think the key takeaways from this case is that in an important business problem:

1. 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

2. 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

3. You should directly care about your coefficient estimates since they are what map into the business decision

Discussion Questions:

1. 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?

2. What are some variables *not* in this dataset that we might want to control for in this context?

3. Beyond changes in the detailing strategy, is there anything else you might recommend to the firm do based on this analysis?