Columbia Business School
AT THE VERY CENTER OF BUSINESS™

Business Analytics

Homework 1

UP TO AND INCLUDING MODULE 3

This homework will be graded on effort, like every other homework in the class. You must submit your homework as **one single PDF file**, containing a clear description of everything you did, and including as many screenshots from Jupyter Notebook as required to illustrate the work you did. You *may not* submit any other files to support your solutions – anything we should know should be included in your report. This is as much a test of technical communication as technical skill.

Problem 1

Before you start the question, read the case "*Does Detailing Pay?*", provided on Canvas.

In this problem, you will undertake a regression analysis to examine the impact of detailing and sampling on the rate at which physicians prescribe a drug. The data is taken from a study by Mizik and Jacobson. (The data for this exercise is a small sample of the full dataset used in the study. The sample has been selected to be representative of the full dataset, and some values have been adjusted to preserve consistency with the full dataset.) The drug in question, here referred to as Xuris, was considered a rising star at the time of the study; it had been on the market for less than three years but was fast gaining market share and generating sales of half a billion dollars.

You have data on 183 physicians, with 24 months of observations for each physician. For each physician in each month, you have the following information:

- `NewPrescriptions`: the number of new prescriptions for Xuris written by the physician in that month
- `Samples`: the number of free Xuris samples received by the physician in that month
- `Details`: the number of Xuris detailing visits to the physician in that month

Author affiliation
* *Columbia Business School*

- `CompetitorNew`: the number of new prescriptions written by the physician in that month for the main competitor to Xuris

For each month, you also have:

- `MedicalCPI`: the level of the consumer price index for medical expenses in that month
- `DRG`: the level of the DRG index, which tracks pharmaceutical stocks

And for each physician you have a dummy variable:

- `PSYCH` = 1 for psychiatrists, 0 for all other specialties

In addition, you have the changes in the levels of these variables (except the dummy variable) from one month to the next. For example, a value of `NewPrescDiff` for January 1999 is the difference between the number of new Xuris prescriptions written by the same physician in January 1999 and December 1998. `SamplesDiff`, `DetailsDiff`, and the other difference variables work the same way. The 24 months of observations for each physician for the original variables yield 23 monthly differences for each physician for each variable.

Table 1 summarizes regression results with `NewPrescription` as dependent variable and 4,392 observations. Table 2 summarizes regression results with `NewPrescDiff` as dependent variable and 4,209 observations.

TABLE 1.

	Model				
	1	**2**	**3**	**4**	**5**
Intercept	11.20	11.16	-15.15	-15.01	-14.81
	(0.03)	(0.03)	(0.81)	(0.81)	(0.81)
Samples		0.060	0.071	0.070	0.069
		(0.004)	(0.004)	(0.004)	(0.004)
Details	0.448	0.384	0.398	0.394	0.391
	(0.015)	(0.016)	(0.014)	(0.014)	(0.014)
CompetitorNew					0.004
					(0.001)
Psych				-0.862	-0.834
				(0.143)	(0.144)
MedicalCPI			0.103	0.103	0.101
			(0.003)	(0.003)	(0.003)
F test (p-value)	0.000	0.000	0.000	0.000	0.000
R-sq	16.6%	19.9%	35.3%	35.8%	36.0%
se	1.37	1.34	1.21	1.20	1.20

Note: Each column shows results for a separate regression model with *NewPrescription* as dependent variable. For each explanatory variable, the table reports the estimated coefficient for each model; the numbers in parentheses are the standard errors for the coefficient estimates. The last three rows report the p-value for the F test, the R-square, and the standard error of the estimate for each model.

TABLE 2.

	Model			
	6	7	8	9
Intercept	0.080	0.080	0.080	0.081
	(0.004)	(0.004)	(0.004)	(0.010)
SamplesDiff	0.041	0.035	0.035	0.035
	(0.001)	(0.001)	(0.001)	(0.001)
DetailsDiff		0.111	0.111	0.111
		(0.003)	(0.003)	(0.003)
CompetitorNewDiff			0.000	0.000
			(0.000)	(0.000)
MedicalCPIDiff				-0.002
				(0.011)
DRG Diff				0.000
				(0.000)
F test (p-value)	0.000	0.000	0.000	0.000
R-sq	45.6%	61.3%	61.3%	61.3%
se	0.273	0.231	0.231	0.231

Note: Each column shows results for a separate regression model with *NewPrescDiff* as dependent variable. For each explanatory variable, the table reports the estimated coefficient for each model; the numbers in parentheses are the standard errors for the coefficient estimates. The last three rows report the p-value for the F test, the R-square, and the standard error of the estimate for each model.

For the purpose of this exercise, assume that each detailing visit for Xuris has an average unit variable cost of $85 and that the margin from a single new prescription is $115. Each new prescription generates an additional 2 to 3 refills that also yield a $115 margin each. These margin values do not account for the cost of detailing visits.

a) Run regression Model 2. Format the output for legibility. According to Model 2, do free samples and detailing visits have a statistically significant impact on new prescriptions?
b) According to Model 2, what is the expected number of new prescriptions generated by a detailing visit? Report a 95% confidence interval. What is the expected $ margin generated by a detailing visit and what is the 95% confidence interval for the expected margin? Assume each new prescription generates 2.5 additional refills.
c) Answer (b) for Model 7 and compare the answers to those you found for Model 2. Why are the results different? Which model do you consider more informative? The case states that physicians draw on their own knowledge and experience in deciding which drugs to prescribe. How does this bear on the comparison between Models 2 and 7?
d) Overall, of Models 1-9, which one do you consider most reliable? Why? What does your preferred model say about the cost effectiveness of detailing visits?
e) Does `MedicalCPI` appear to have a statistically significant effect? Why might this be? Look at the data before trying to answer. Compare Model 9 with Models 3-5.
f) Interpret the coefficients for `CompetitorNew` and `Psych` in Model 5.

Problem 2

A farmer is trying to optimize the number of eggs his chickens lay every day. He strongly suspects the number of eggs is related to the amount he fed each chicken the day before they laid the eggs, and the daily temperature that day.

The dataset `egg_production.csv` contains information about feed, eggs laid and temperature for a number of days. In this question, you will analyze this dataset and try and figure out how these factors affect egg production.

a) Run a regression of number of eggs on feed and interpret the result. Does it align with your intuition?
b) Now run a regression using both variables. Interpret the result. Does this make sense to you?
c) You suspect that something fishy is going on, and that the amount of feed given to each chicken de- pends on the temperature. Investigate this hypothesis and create a new binary/discrete/categorical variable that captures this phenomenon.
d) Regress number of eggs on feed, temperature, and the new variable you created. Interpret the results.
e) For this model, what is a 90% confidence interval for the prediction of the number of eggs that were produced if the feed was 25 and the temperature was −1. Interpret the results.