Eco 375: Applied Econometrics
University of Toronto
Department of Economics
Assignment 1

1. Goals of Assignment

In the course so far, we have focused on understanding the basic theoretical aspects of regression analysis. This assignment is an opportunity to apply this knowledge to a particular context and through this, gain complimentary skills in programming, data manipulation, statistical analysis, and interpretation of results.

The first assignment in Eco 375 is designed to complement the theoretical tools we are currently studying in weeks 1-5 of the course: simple linear regression and its extension to estimation of a multiple linear regression model. While the weekly empirical problems acquaint you with the main building blocks of programming and analyzing data, this assignment will expand on this process substantially. The final analysis will be more formal, and your execution will be structured like a short paper.

While the assignment is structured like a paper, the workload is substantially less. This is because the scope of the analysis is suitably narrow. Specifically, like the weekly assigned empirical problems, you will be provided a context and data. Secondly, the assignment requirements are tightly defined (see below) meaning there is less leg work in refining your own idea (with attendant data) and less range for missing the mark in your execution. The aim is that the tight scope will give you practice in coding, analyzing, and writing about econometric findings in a formal format, which can better prepare you to write subsequent papers based on your own ideas and data collection in other courses (e.g. Eco 475) and give you the building blocks to constructing formal analysis in your future careers.

The data and focus of the assignment this year is an Israeli dataset on COVID-19 infection and vaccination by age and week (collected over the last several weeks). A complete description of the data and topic is given in the Context and Data section below.

2. The Context and Data

Israel is the world leader for COVID-19 vaccination, an accomplishment that is born of a deal between the Israeli government and Pfizer-BioNTech. Here, in return for vaccines from Pfizer-BioNTech, the Israeli Government promised a swift roll-out of their vaccination campaign and medical data on their population of 9 million.[1] These data will be used by Pfizer-BioNTech to study the population effect of its vaccine and to determine if "herd immunity is achieved after reaching a certain percentage of vaccination coverage" (Israeli Ministry of Health 2020).

Pfizer-BioNTech, Israel, you, me, and the rest world would like the answer to this question.

The core focus of your assignment will be to analyze the effect of vaccination on COVID-19 infection in Israel. In theoretical terms, vaccination is an economic good that delivers an externality benefit, which, by nature, yields decreasing returns. Specifically, if early vaccines prevent infection and reduce spread to others, then eventually reduced spread leaves little additional benefits to further vaccination (i.e., the so-called herd immunity effect).

While this may be true in theoretical terms, the size of these returns (and the level where herd immunity might be reached) is an empirical question. Importantly, this is a question that cannot be answered by existing randomized control trials on COVID-19 vaccination (likely what sparks Pfizer-BioNTech's interest in this deal with Israel). Furthermore, we cannot use results from other types of vaccine preventable diseases because the returns (and the herd immunity level) depend jointly on the infectiousness of the disease and the effectiveness of the vaccine (all of which vary across disease).

[1] See Figure 1 for a comparison of vaccination in Israel to the U.S. and Canada. You can look at other interesting comparisons here: https://ourworldindata.org/coronavirus.

Your job in this assignment is to use course tools to study this question and analyze this relationship empirically using the attached data. You will consider what you can and can't answer with these data, and you will do this by using your knowledge of econometric theory along with careful consideration of the underlying assumptions of regression analysis.

Data: the dataset includes the following variables, which are measured by week and age group for Israel from September 2020 to the end of January 2021.

Variable	Description
dateWKbegin	Start Date of Week
week	Year-Week
age	Age Group
cases_rate	Weekly COVID-19 Cases (per million people)
cases_log	Natural Log of Weekly COVID-19 Cases
lagvacc_per	Two Week Lag of 1st Dose Vaccination (in percent)
lagvacc_per2	Two Week Lag of 1st Dose Vaccination (in percent) Squared
lagvacc_0	Dummy: Two Week Lag of 1st Dose Vaccination is 0% of Population
lagvacc_0_10	Dummy: Two Week Lag of 1st Dose Vaccination is (0%-10%) of Population
lagvacc_10_20	Dummy: Two Week Lag of 1st Dose Vaccination is [10%-20%) of Population
lagvacc_20_100	Dummy: Two Week Lag of 1st Dose Vaccination is [20%-100%] of Population

3. Structure of Assignment

A "finished product" will be comprised of a concise abstract; two tables: (1) descriptive statistics, and (2) linear regression results; and no more than 4 pages (1.5-spaced) of corresponding text (in 12-point font) outlining and interpreting the empirical results. The assignment should be organized as follows:

Page 1: **Title page**
Including course and student information, and a concise abstract.

Page 2-5: **Text**
Formal academic composition with a 4-page limit, 12-point font, and 1.5 line spacing. The text should NOT include tables or figures.

Page 6: **References**
Includes a list of any references cited in your assignment.

Page 7- : **Tables and figures**
Tables and figures are appended after the list of references. At minimum, this will include two tables, but you may include additional tables and figures if warranted by your discussion and analysis. All included tables and figures must be discussed in the text. DO NOT include undigested STATA output! Results presented in tabular form should have all variables and numbers clearly labeled. Any figures should also be well labeled and clear. Consider each Table/Figure as a stand-alone product, which can be largely understood on its own (i.e. without referring to the written text).

Overall, the assignment should demonstrate application of tools discussed throughout the course, but especially those from week 4 and 5. Background reading includes Chapter 2-6 and Lecture 4 and 5. You will want to pay special attention to the SLR assumptions in Chapter 2, the MLR assumptions in Chapter 3, the discussion of omitted variable bias in section 3-3, and the discussion of functional form in section 6-2a and 6-2b (which is also appended to this overview). The regression specification in textbook example 6.2 is particularly close to our specification in the assignment. It is worth a look. Lastly, Chapter 19, section 19-5 "Writing an Empirical Paper" in the course textbook has some general ideas about style guidelines in economic writing.

4. Content requirements

Following the overall structure outlined in section 3, your assignment should read like a short-term paper. That said, the content requirements and scope in this assignment are tightly defined and should follow the guidelines below.

4.1. Title Page: this will include course and student information, a title, and a concise abstract

4.2. Abstract: a brief summary of your findings (on the title page)

4.3. Introduction: introduce and motivate the analysis.

4.4. Describe the data: briefly describe the data, with reference to, at least, a table of descriptive statistics.

- This table will be named Table 1 and will be included after the references page.
- It will contain summary statistics like the means, s.d.'s, frequencies, etc. for the variables used in the analysis.
- Means and standard deviations allow you to put the size of estimated coefficients in context. A well-written analysis will discuss summary statistics as a prelude to the main point of the analysis and will use summary statistics to put the size of estimated effects in context (e.g. what is a 3% decrease in COVID-19 cases represent from mean COVID-19 cases per million? do COVID-19 rates differ across age group? across time? etc.?).

4.5. Regression analysis section:

Simple linear regression: your assignment will include initial analysis and write up on a simple linear regression: the log of COVID-19 cases regressed on the two-week lag of first dose vaccination (measured in percent): i.e., *case_log* on *lagvacc_per*. Include the following in your write up:

- Report the estimated regression in "equation form" in the body of your assignment.[2]
- Interpret these estimates.
- Test the hypothesis that the effect of vaccination on COVID-19 infections is zero.
- Discuss whether you think the simple linear regression assumptions: SLR.1 to SLR.5 hold in the current context. Discuss each of these in turn.

Multiple linear regression: your assignment will extend this analysis by exploring several different specifications as described below, which you will report in Table 2 (included at the end of your assignment).

- Table 2 will contain the following four specifications:
  (1) The SLR from above: the log of COVID-19 cases regressed on two-week lag of first dose vaccination;
  (2) The log of COVID-19 cases regressed on the two-week lag of first dose vaccination controlling for a set of week dummy variables and a set of age group dummy variables;
  (3) The log of COVID-19 cases regressed on the two-week lag of first dose vaccination, the square of the two-week lag of first dose vaccination and controlling for a set of week dummy variables and a set of age group dummy variables;
  (4) The log of COVID-19 cases regressed the vaccination dummy variables controlling for a set of week dummy variables and a set of age group dummy variables. Use 0% vaccination as the base group for the set of vaccination dummies;

- This table should be formatted and contain at its core the following:

[2] Section 4-6 of Wooldridge includes a discussion about reporting regression results, and there are many examples of "equation form" throughout the textbook. See equation 4.51 for example (these textbook sections are also appended to the end of this overview).

Table 2: Regression Analysis of log COVID-19 infection cases and vaccination

	(1)	(2)	(3)	(4)
Lag Vaccination (in percent)	estimate	estimate	estimate	
	(s.e.)	(s.e.)	(s.e.)	
Lag Vaccination (in percent) squared			estimate	
			(s.e.)	
Lag Vaccination: (0%-10%)				estimate
				(s.e.)
Lag Vaccination: [10%-20%)				estimate
				(s.e.)
Lag Vaccination: [20%-100%]				estimate
				(s.e.)
Age:				
15-19		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
20-29		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
30-39		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
40-49		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
50-59		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
60-69		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
70-79		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
80+		estimate	estimate	estimate
		(s.e.)	(s.e.)	(s.e.)
Year-Week Dummies	No	Yes	Yes	Yes
Adjusted R-Squared	value	value	value	value
N	value	value	value	value

Notes: pertinent details; source data; details on variable definitions; time period; robust standard errors? etc.
***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Discussion of results: be sure to interpret key coefficients, as well as the results of the hypothesis tests. For example, in your discussion of results be sure to mention:

- The size of the estimates in practical terms.
- When adding a set of control variables, discuss whether the estimated coefficient change in terms of both: statistical significance and practical significance.
- Discuss what happens when moving to specification (1) to specification (2). For example, why would including controls for year-week and age group impact results? Is there a confounding time trend in both infection and vaccination? (hint: see Figure 2 attached). Do different age

groups have different chances of COVID-19 infection? Different rates of vaccination? These types of examples can be used to anchor the general discussion.

- Is there a non-linear effect of vaccination? From a theoretical perspective? And is there evidence of a non-linear effect from the estimation? See Section 6-2b in the textbook (also appended to this overview),
- Use these above points to discuss whether the specifications (2) – (4) are able to address any areas where the initial SLR assumptions from specification (1) fail.

4.6. Limitations of results: discuss possible problems with these specifications, especially omitted variables that may still lurk in the residual (i.e. do you interpret your results as causal or are they purely descriptive?). What are the most important remaining threats to the validity of your regression results?

4.7. Conclusion: Based on your analysis what conclusions would you draw about the effects of vaccination on COVID-19 infections in Israel.

5. Evaluation

There is no mechanical grading scheme. But the following should be a helpful guide and explain in more detail what we are looking for in your submission.

5.1. Exposition and Writing (Approximately half the weight)

You will need to explain your empirical results to your reader. This involves describing the data; motivating the regression; clearly explaining what is estimated; posing and interpreting results. You may wish to consult Chapter 19, section 19-5 "Writing an Empirical Paper" in the course textbook for general ideas about style guidelines in economics writing.

Key elements of clear exposition include:

1. Placing results in context (i.e. discuss "economic" significance)
    - Discuss up front (in introduction) and come back to it later (in the results section)

2. Data description (concerning Table 1)
    - What's the range of the variables?
    - Does anything hint at the subsequent empirical results?

3. Simple Linear Regression: explain, interpret
    - Definition of variables
    - Specify and motivate hypotheses
    - Mapping coefficients into interpretation
    - Discussion of SLR assumptions

4. Multiple Linear Regressions
    - Motivate additional regressions and hypotheses
    - Reflects back on SLR assumptions

5. Quality of the tables
    - We will penalize (heavily!) undigested STATA output (i.e. tables do not look like STATA output).
    - Tables should (ideally) have "English" variable names, not acronyms
    - Ask yourself: "If I had never seen this table before would I still understand what it is

about?"

6. Integration of empirical results from tables to text.

7. Overall writing and coherence
    - Does the assignment draw results together?
    - Is it clear to read?
    - Usually, a confusing assignment reflects underlying confusion.

5.2. Content (Approximately half the weight)

It is not easy to separate exposition from content, but in general characteristics we will be looking for execution of the content guidelines described in section 4.

1. Clear, succinct introduction.

2. Base regression: explanation, interpretation, discussion of assumptions

2. Estimation and discussion of additional regressions/extensions
    - Implementation of hypothesis tests and interpretation
    - When using dummy variables, make sure that the interpretation and specification is correct (i.e., relative to a base group)

3. Potential statistical problems
    - Be sure to discuss why/how a data problem might affect results (as opposed to just stating a problem might exist).

5.3. Due Date and Submission Details

Assignment submission has two main parts:
Part 1: Your answers submitted on Crowdmark (details below). This is for grading.
Part 2: The same work submitted on Quercus (details below). This is to check your Stata support files and to run your work through Turnitin. You must verify (the correct version of) your work was submitted before the deadline. Details in the links below.

- https://crowdmark.com/help/verifying-that-an-assignment-was-submitted/
- https://ctl1.utsc.utoronto.ca/quercus/help/Uploading_an_assignment_to_Quercus.pdf

Use Chrome, clear your cache, restart your browser/computer, close all other tabs and programs before submitting your work.

5.3.1. Part 1: Crowdmark submission

Location: Crowdmark: https://app.crowdmark.com/sign-in.
Due date: Feb 24th 10:00am
Late Penalties: 100% penalty per 1-minute past the due date.
File Format: Complete your work in Microsoft Word. Convert it into a pdf file and submit a pdf file.
General Instructions:

- Follow the required format described in the Assignment Overview (this document).
- You will submit your assignment in Crowdmark in response to two questions: the first question will ask you to upload a scan of your Title page (page 1 as described above) with your student card placed at the top righthand corner. The second question will ask you to upload your pdf file (all pages) to the system.
- Your answers must be TYPED. Any handwritten answers will get zero points.
- Verify your submission: https://crowdmark.com/help/verifying-that-an-assignment-was-submitted/ . This is especially relevant if you are having connection issues.

5.3.2. Part 2: Quercus submission

Location: Quercus Assignments: https://q.utoronto.ca/courses/197735/assignments
Due date: Feb 24th 10:00am
Late Penalties: 100% penalty per 1-minute past the due date.
File Format: One Microsoft Word file. One code file (e.g. Stata do file). One log file (e.g. Stata log file).
General Instructions:

- Your answers must be TYPED. Any handwritten answers will get zero points.
- Your answers must be in English without any other languages embedded in the file. Again, this will be an automatic zero if we cannot process the file through the required checks because of this.
- Verify your submission, including whether you have submitted the correct file: https://ctl1.utsc.utoronto.ca/quercus/help/Uploading_an_assignment_to_Quercus.pdf (scroll down to part 3). This is especially relevant if you are having connection issues.

5.3.3. Plagiarism

Plagiarism is a serious problem (in general) with university writing. Obviously, if we detect this form of academic dishonesty, we deal with it severely. Even "inadvertent" plagiarism is penalized. You should familiarize yourself with the rules regarding the citation of sources, etc.

References:

Roser, Max, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

Israeli Ministry of Health. (2020) REAL-WORLD EPIDEMIOLOGICAL EVIDENCE COLLABORATION AGREEMENT. Accessed February 3, 2021. https://govextra.gov.il/media/30806/11221-moh-pfizer-collaboration-agreement-redacted.pdf.

Figure 1: Cumulative COVID-19 vaccination doses administered per 100 people

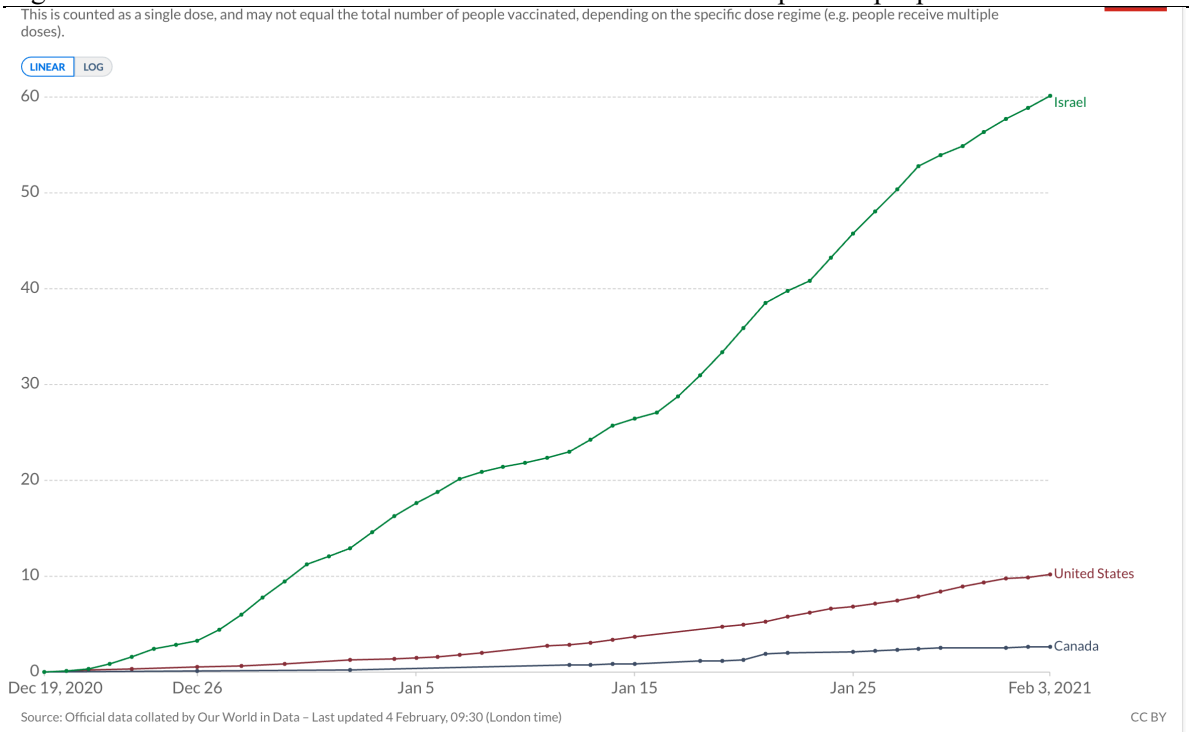
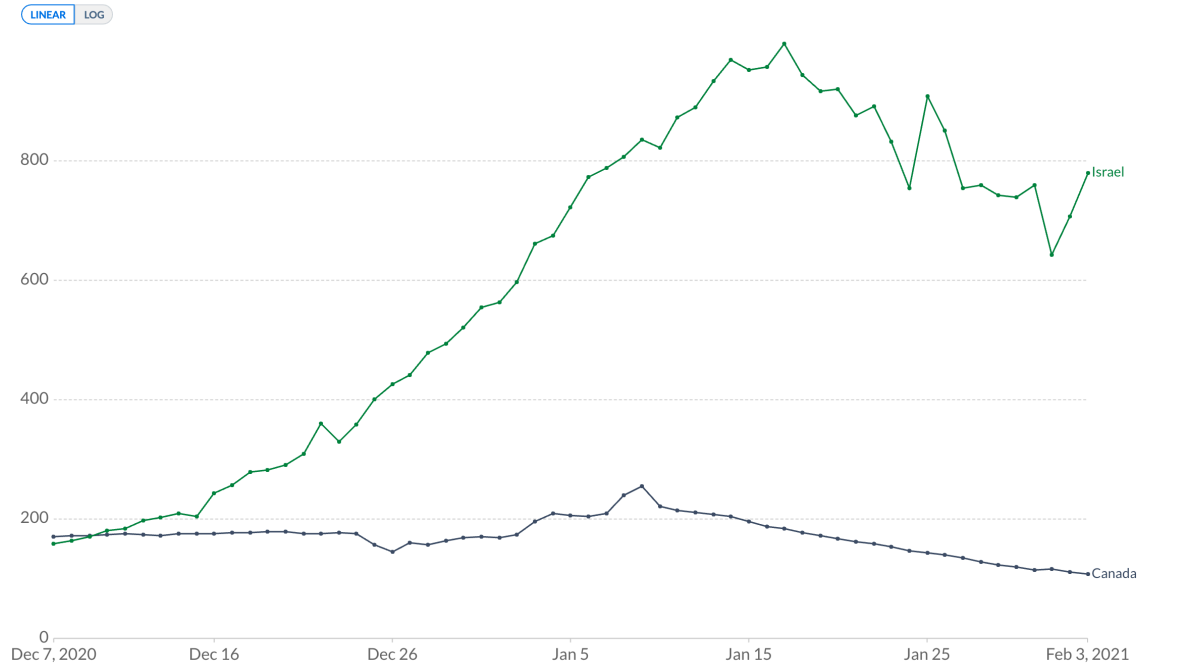


Figure 2: Daily new confirmed COVID-19 cases per million people



4-6 Reporting Regression Results

We end this chapter by providing a few guidelines on how to report multiple regression results for relatively complicated empirical projects. This should help you to read published works in the applied social sciences, while also preparing you to write your own empirical papers. We will expand on this topic in the remainder of the text by reporting results from various examples, but many of the key points can be made now.

Naturally, the estimated OLS coefficients should always be reported. For the key variables in an analysis, you should *interpret* the estimated coefficients (which often requires knowing the units of measurement of the variables). For example, is an estimate an elasticity, or does it have some other interpretation that needs explanation? The economic or practical importance of the estimates of the key variables should be discussed.

The standard errors should always be included along with the estimated coefficients. Some authors prefer to report the t statistics rather than the standard errors (and sometimes just the absolute value of the t statistics). Although nothing is really wrong with this, there is some preference for reporting standard errors. First, it forces us to think carefully about the null hypothesis being tested; the null is not always that the population parameter is zero. Second, having standard errors makes it easier to compute confidence intervals.

The R-squared from the regression should always be included. We have seen that, in addition to providing a goodness-of-fit measure, it makes calculation of F statistics for exclusion restrictions simple. Reporting the sum of squared residuals and the standard error of the regression is sometimes a good idea, but it is not crucial. The number of observations used in estimating any equation should appear near the estimated equation.

If only a couple of models are being estimated, the results can be summarized in equation form, as we have done up to this point. However, in many papers, several equations are estimated with many different sets of independent variables. We may estimate the same equation for different groups of people, or even have equations explaining different dependent variables. In such cases, it is better to summarize the results in one or more tables. The dependent variable should be indicated clearly in the table, and the independent variables should be listed in the first column. Standard errors (or t statistics) can be put in parentheses below the estimates.

Example 4.11 Evaluating a Job Training Program

We reproduce the simple and multiple regression estimates and now put the standard errors below the coefficients. Recall that the outcome variable, *earn*98, is measured in thousands of dollars:

$$\widehat{earn98} = 10.61 - 2.05\, train \quad [4.51]$$
$$(0.28)\ \ (0.48)$$
$$n = 1{,}130,\ R^2 = 0.016$$

$$\widehat{earn98} = 4.67 + 2.41\, train + .373\, earn96 + .363\, educ - .181\, age + 2.48\, married \quad [4.52]$$
$$(1.15)\ \ (0.44)\ \ \ \ (.019)\ \ \ \ \ (.064)\ \ \ \ (.019)\ \ \ \ (0.43)$$
$$n = 1{,}130,\ R^2 = 0.405$$

As discussed in Example 3.7, the change in the sign of coefficient on *train* is striking when moving from simple to multiple regression. Moreover, the *t* statistic in (4.51) is $-2.05/0.48 \approx -4.27$, which gives a very statistically significant and practically large *negative* effect of the program. By contrast, the *t* statistic in (4.52) is about 5.47, which shows a strongly statistically significant and *positive* effect. It is pretty clear that we prefer the multiple regression results for evaluating the job training program. Of course, it could be that we have omitted some important controls in (4.52), but at a minimum we know that we can account for some important differences across workers.

6-2 More on Functional Form

In several previous examples, we have encountered the most popular device in econometrics for allowing nonlinear relationships between the explained and explanatory variables: using logarithms for the dependent or independent variables. We have also seen models containing quadratics in some explanatory variables, but we have yet to provide a systematic treatment of them. In this section, we cover some variations and extensions on functional forms that often arise in applied work.

6-2a More on Using Logarithmic Functional Forms

We begin by reviewing how to interpret the parameters in the model

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u, \qquad \textbf{[6.6]}$$

where these variables are taken from Example 4.5. Recall that throughout the text log(x) is the *natural* log of x. The coefficient β_1 is the elasticity of *price* with respect to *nox* (pollution). The coefficient β_2 is the change in log(*price*), when $\Delta rooms = 1$; as we have seen many times, when multiplied by 100, this is the approximate percentage change in *price*. Recall that $100 \cdot \beta_2$ is sometimes called the semi-elasticity of *price* with respect to *rooms*.

When estimated using the data in HPRICE2, we obtain

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306\, rooms \qquad \textbf{[6.7]}$$
$$(0.19)\ \ (.066)\qquad\qquad (.019)$$
$$n = 506,\ R^2 = .514.$$

Thus, when *nox* increases by 1%, *price* falls by .718%, holding only *rooms* fixed. When *rooms* increases by one, *price* increases by approximately $100(.306) = 30.6\%$.

The estimate that one more room increases price by about 30.6% turns out to be somewhat inaccurate for this application. The approximation error occurs because, as the change in log(y) becomes larger and larger, the approximation $\%\Delta y \approx 100 \cdot \Delta \log(y)$ becomes more and more inaccurate. Fortunately, a simple calculation is available to compute the exact percentage change.

To describe the procedure, we consider the general estimated model

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2.$$

(Adding additional independent variables does not change the procedure.) Now, fixing x_1, we have $\Delta\widehat{\log(y)} = \widehat{\beta}_2 \Delta x_2$. Using simple algebraic properties of the exponential and logarithmic functions gives the *exact* percentage change in the predicted *y* as

$$\%\Delta\hat{y} = 100\cdot[\exp(\widehat{\beta}_2 \Delta x_2) - 1], \tag{6.8}$$

where the multiplication by 100 turns the proportionate change into a percentage change. When $\Delta x_2 = 1$,

$$\%\Delta\hat{y} = 100\cdot[\exp(\widehat{\beta}_2) - 1]. \tag{6.9}$$

Applied to the housing price example with $x_2 = \text{rooms}$ and $\widehat{\beta}_2 = .306$, $\%\Delta\widehat{price} = 100[\exp(.306) - 1] = 35.8\%$, which is notably larger than the approximate percentage change, 30.6%, obtained directly from (6.7). {Incidentally, this is not an unbiased estimator because $\exp(\cdot)$ is a nonlinear function; it is, however, a consistent estimator of $100[\exp(\beta_2) - 1]$. This is because the probability limit passes through continuous functions, while the expected value operator does not. See Appendix C.}

The adjustment in equation (6.8) is not as crucial for small percentage changes. For example, when we include the student-teacher ratio in equation (6.7), its estimated coefficient is –.052, which means that if *stratio* increases by one, *price* decreases by approximately 5.2%. The exact proportionate change is $\exp(-.052) - 1 \approx -.051$, or –5.1%. On the other hand, if we increase *stratio* by five, then the approximate percentage change in price is –26%, while the exact change obtained from equation (6.8) is $100[\exp(-.26) - 1] \approx -22.9\%$.

The logarithmic approximation to percentage changes has an advantage that justifies its reporting even when the percentage change is large. To describe this advantage, consider again the effect on price of changing the number of rooms by one. The logarithmic approximation is just the coefficient on rooms in equation (6.7) multiplied by 100, namely, 30.6%. We also computed an estimate of the exact percentage change for *increasing* the number of rooms by one as 35.8%. But what if we want to estimate the percentage change for *decreasing* the number of rooms by one? In equation (6.8) we take $\Delta x_2 = -1$ and $\widehat{\beta}_2 = .306$, and so $\%\Delta\widehat{price} = 100[\exp(-.306) - 1] = -26.4$, or a drop of 26.4%. Notice that the approximation based on using the coefficient on *rooms* is between 26.4 and 35.8—an outcome that always occurs. In other words, simply using the coefficient (multiplied by 100) gives us an estimate that is always between the absolute value of the estimates for an increase and a decrease. If we are specifically interested in an increase or a decrease, we can use the calculation based on equation (6.8).

The point just made about computing percentage changes is essentially the one made in introductory economics when it comes to computing, say, price elasticities of demand based on large price changes: the result depends on whether we use the beginning or ending price and quantity in computing the percentage changes. Using the logarithmic approximation is similar in spirit to calculating an arc elasticity of demand, where the averages of prices and quantities are used in the denominators in computing the percentage changes.

We have seen that using natural logs leads to coefficients with appealing interpretations, and we can be ignorant about the units of measurement of variables appearing in logarithmic form because the slope coefficients are invariant to rescalings. There are several other reasons logs are used so much in applied work. First, when $y > 0$, models using log(y) as the dependent variable often satisfy the CLM assumptions more closely than models using the level of y. Strictly positive variables often have conditional distributions that are heteroskedastic or skewed; taking the log can mitigate, if not eliminate, both problems.

Another potential benefit of using logs is that taking the log of a variable often narrows its range. This is particularly true of variables that can be large monetary values, such as firms' annual sales or baseball players' salaries. Population variables also tend to vary widely. Narrowing the range of the dependent and independent variables can make OLS estimates less sensitive to outlying (or extreme) values; we take up the issue of outlying observations in Chapter 9.

However, one must not indiscriminately use the logarithmic transformation because in some cases it can actually create extreme values. An example is when a variable y is between zero and one (such as a proportion) and takes on values close to zero. In this case, log(y) (which is necessarily negative) can be very large in magnitude whereas the original variable, y, is bounded between zero and one.

There are some standard rules of thumb for taking logs, although none is written in stone. When a variable is a positive dollar amount, the log is often taken. We have seen this for variables such as wages, salaries, firm sales, and firm market value. Variables such as population, total number of employees, and school enrollment often appear in logarithmic form; these have the common feature of being large integer values.

Variables that are measured in years—such as education, experience, tenure, age, and so on—usually appear in their original form. A variable that is a proportion or a percent—such as the unemployment rate, the participation rate in a pension plan, the percentage of students passing a standardized exam, and the arrest rate on reported crimes—can appear in either original or logarithmic form, although there is a tendency to use them in level forms. This is because any regression coefficients involving the *original* variable—whether it is the dependent or independent variable—will have a *percentage point* change interpretation. (See Appendix A for a review of the distinction between a percentage change and a percentage point change.) If we use, say, log(*unem*) in a regression, where *unem* is the percentage of unemployed individuals, we must be very careful to distinguish between a percentage point change and a percentage change. Remember, if *unem* goes from 8 to 9, this is an increase of one percentage point, but a 12.5% increase from the initial unemployment level. Using the log means that we are looking at the percentage change in the unemployment rate: $\log(9) - \log(8) \approx .118$ or 11.8%, which is the logarithmic approximation to the actual 12.5% increase.

One limitation of the log is that it cannot be used if a variable takes on zero or negative values. In cases where a variable *y* is nonnegative but can take on the value 0, log(1+*y*) is sometimes used. The percentage change interpretations are often closely preserved, except for changes beginning at $y = 0$ (where the percentage change is not even defined). Generally, using log(1+*y*) and then interpreting the estimates as if the variable were log(*y*) is acceptable when the data on *y* contain relatively few zeros. An example might be where *y* is hours of training per employee for the population of manufacturing firms, if a large fraction of firms provides training to at least one worker. Technically, however, log(1+*y*) cannot be normally distributed (although it might be less heteroskedastic than *y*). Useful, albeit more advanced, alternatives are the Tobit and Poisson models in Chapter 17.

One drawback to using a dependent variable in logarithmic form is that it is more difficult to predict the original variable. The original model allows us to predict log(*y*), not *y*. Nevertheless, it is fairly easy to turn a prediction for log(*y*) into a prediction for *y* (see Section 6-4). A related point is that it is *not* legitimate to compare *R*-squareds from models where *y* is the dependent variable in one case and log(*y*) is the dependent variable in the other. These measures explain variations in different variables. We discuss how to compute comparable goodness-of-fit measures in Section 6-4.

Going Further 6.2

Suppose that the annual number of drunk driving arrests is determined by

$$\log(arrests) = \beta_0 + \beta_1 \log(pop) + \beta_2 age16\_25 + other\ factors,$$

where *age 16_25* is the proportion of the population between 16 and 25 years of age. Show that β_2 has the following (ceteris paribus) interpretation: it is the percentage change in *arrests* when the percentage of the people aged 16 to 25 increases by one *percentage point*.

Answer

6-2b Models with Quadratics

Quadratic functions are also used quite often in applied economics to capture decreasing or increasing marginal effects. You may want to review properties of quadratic functions in Appendix A.

In the simplest case, *y* depends on a single observed factor *x*, but it does so in a quadratic fashion:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

For example, take $y = wage$ and $x = exper$. As we discussed in Chapter 3, this model falls outside of simple regression analysis but is easily handled with multiple regression.

It is important to remember that β_1 does not measure the change in *y* with respect to *x*; it makes no sense to hold x^2 fixed while changing *x*. If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \qquad \textbf{[6.10]}$$

then we have the approximation

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ so } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad \textbf{[6.11]}$$

This says that the slope of the relationship between *x* and *y* depends on the value of *x*; the estimated slope is $\hat{\beta}_1 + 2\hat{\beta}_2 x$. If we plug in $x = 0$, we see that $\hat{\beta}_1$ can be interpreted as the approximate slope in going from $x = 0$ to $x = 1$. After that, the second term, $2\hat{\beta}_2 x$ must be accounted for.

If we are only interested in computing the predicted change in *y* given a starting value for *x* and a change in *x*, we could use (6.10) directly: there is no reason to use the calculus approximation at all. However, we are usually more interested in quickly summarizing the effect of *x* on *y*, and the interpretation of $\hat{\beta}_1$ and $\hat{\beta}_2$ in equation (6.11) provides that summary. Typically, we might plug in the average value of *x* in the sample, or some other interesting values, such as the median or the lower and upper quartile values.

In many applications, $\hat{\beta}_1$ is positive and $\hat{\beta}_2$ is negative. For example, using the wage data in WAGE1, we obtain

$$\begin{aligned} \widehat{\text{wage}} &= 3.73 + .298\, exper - .0061\, exper^2 \qquad \textbf{[6.12]} \\ &\quad (.35)\ \ (.041)\qquad\ \ (.0009) \\ n &= 526,\ R^2 = .093. \end{aligned}$$

This estimated equation implies that *exper* has a diminishing effect on *wage*. The first year of experience is worth roughly 30¢ per hour ($.298). The second year of experience is worth less [about $.298 - 2(.0061)(1) \approx .286$, or 28.6¢, according to the approximation in (6.11) with $x = 1$]. In going from 10 to 11 years of experience, *wage* is predicted to increase by about $.298 - 2(.0061)(10) = .176$, or 17.6¢. And so on.

When the coefficient on *x* is positive and the coefficient on x^2 is negative, the quadratic has a parabolic shape. There is always a positive value of *x* where the effect of *x* on *y* is zero; before this point, *x* has a positive effect on *y*; after this point, *x* has a negative effect on *y*. In practice, it can be important to know where this turning point is.

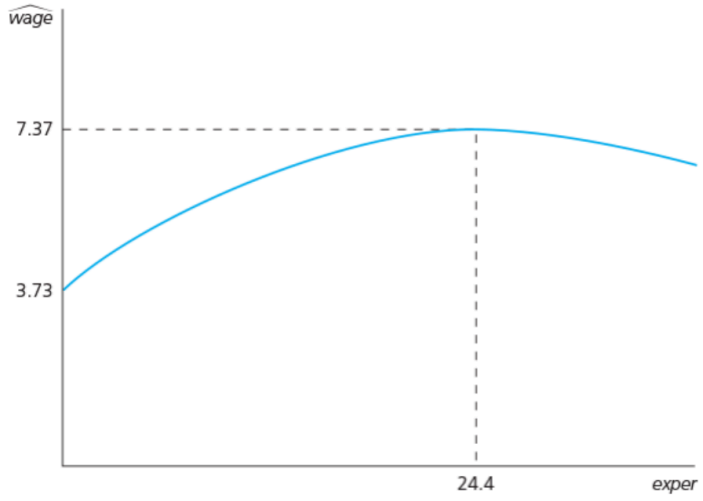
In the estimated equation (6.10) with $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$ the turning point (or maximum of the function) is always achieved at the coefficient on *x* over *twice* the absolute value of the coefficient on x^2:

$$x^* = \left|\hat{\beta}_1/(2\hat{\beta}_2)\right|. \qquad \textbf{[6.13]}$$

In the wage example, $x^* = exper^*$ is $.298/[2(.0061)] \approx 24.4$. (Note how we just drop the minus sign on −.0061 in doing this calculation.) This quadratic relationship is illustrated in Figure 6.1.

Figure 6.1

Quadratic relationship between $\widehat{wage}$ and *exper*.



In the wage equation (6.12), the return to experience becomes zero at about 24.4 years. What should we make of this? There are at least three possible explanations. First, it may be that few people in the sample have more than 24 years of experience, and so the part of the curve to the right of 24 can

be ignored. The cost of using a quadratic to capture diminishing effects is that the quadratic must eventually turn around. If this point is beyond all but a small percentage of the people in the sample, then this is not of much concern. But in the data set WAGE1, about 28% of the people in the sample have more than 24 years of experience; this is too high a percentage to ignore.

It is possible that the return to *exper* really becomes negative at some point, but it is hard to believe that this happens at 24 years of experience. A more likely possibility is that the estimated effect of *exper* on *wage* is biased because we have controlled for no other factors, or because the functional relationship between *wage* and *exper* in equation (6.12) is not entirely correct. Computer Exercise C2 asks you to explore this possibility by controlling for education, in addition to using log(*wage*) as the dependent variable.

When a model has a dependent variable in logarithmic form and an explanatory variable entering as a quadratic, some care is needed in reporting the partial effects. The following example also shows that the quadratic can have a U-shape, rather than a parabolic shape. A U-shape arises in equation (6.10) when $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive; this captures an increasing effect of *x* on *y*.

Example 6.2 Effects of Pollution on Housing Prices

We modify the housing price model from Example 4.5 to include a quadratic term in *rooms*:

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \qquad \textbf{[6.14]}$$

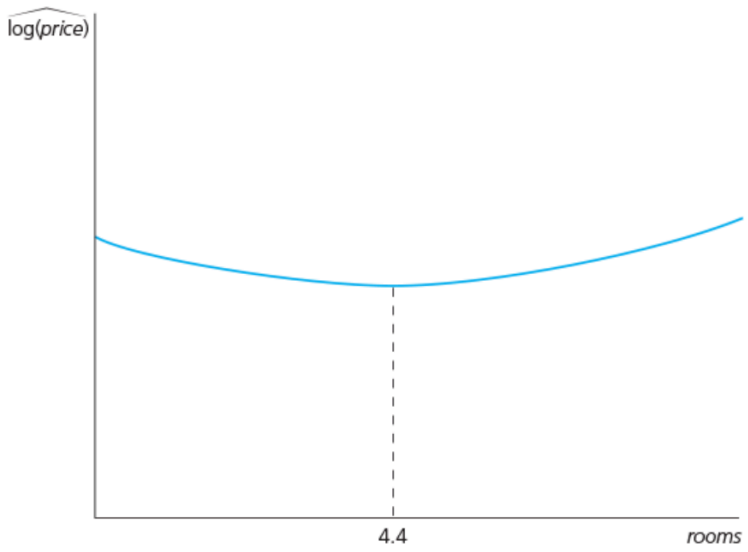
The model estimated using the data in HPRICE2 is

$$\begin{aligned} \widehat{\log(price)} &= 13.39 - .902 \log(nox) - .087 \log(dist) \\ &\qquad (.57) \quad (.115) \qquad\quad (.043) \\ &\quad - .545\, rooms + .062\, rooms^2 - .048\, stratio \\ &\qquad (.165) \qquad (.013) \qquad\quad (.006) \\ &\quad n = 506,\ R^2 = .603. \end{aligned}$$

The quadratic term $rooms^2$ has a *t* statistic of about 4.77, and so it is very statistically significant. But what about interpreting the effect of *rooms* on log(*price*)? Initially, the effect appears to be strange. Because the coefficient on *rooms* is negative and the coefficient on $rooms^2$ is positive, this equation literally implies that, at low values of *rooms*, an additional room has a *negative* effect on log(*price*). At some point, the effect becomes positive, and the quadratic shape means that the semi-elasticity of *price* with respect to *rooms* is increasing as *rooms* increases. This situation is shown in Figure 6.2.

Figure 6.2

$\widehat{\log(price)}$ as a quadratic function of *rooms*.



We obtain the turnaround value of *rooms* using equation (6.13) (even though $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive). The absolute value of the coefficient on *rooms*, .545, divided by twice the coefficient on $rooms^2$, .062, gives $rooms^* = .545/[2(.062)] \approx 4.4$; this point is labeled in Figure 6.2.

Do we really believe that starting at three rooms and increasing to four rooms actually reduces a house's expected value? Probably not. It turns out that only five of the 506 communities in the sample have houses averaging 4.4 rooms or less, about 1% of the sample. This is so small that the quadratic to the left of 4.4 can, for practical purposes, be ignored. To the right of 4.4, we see that adding another room has an increasing effect on the percentage change in price:

$$\Delta\widehat{\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

and so

$$\begin{aligned} \%\Delta\widehat{price} &\approx 100\,\{[-.545 + 2(.062)]\,rooms\}\,\Delta rooms \\ &= (-54.5 + 12.4\,rooms)\Delta rooms. \end{aligned}$$

Thus, an increase in *rooms* from, say, five to six increases price by about $-54.5 + 12.4(5) = 7.5\%$; the increase from six to seven increases price by roughly $-54.5 + 12.4(6) = 19.9\%$. This is a very strong increasing effect.

The strong increasing effect of rooms on log(*price*) in this example illustrates an important lesson: one cannot simply look at the coefficient on the quadratic term—in this case, .062—and declare that it is too

small to bother with, based only on its magnitude. In many applications with quadratics, the coefficient on the squared variable has one or more zeros after the decimal point: after all, this coefficient measures how the slope is changing as *x* (*rooms*) changes. A seemingly small coefficient can have practically important consequences, as we just saw. As a general rule, one must compute the partial effect and see how it varies with *x* to determine if the quadratic term is practically important. In doing so, it is useful to compare the changing slope implied by the quadratic model with the constant slope obtained from the model with only a linear term. If we drop $rooms^2$ from the equation, the coefficient on *rooms* becomes about .255, which implies that each additional room—starting from any number of rooms—increases median price by about 25.5%. This is very different from the quadratic model, where the effect becomes 25.5% at $rooms = 6.45$ but changes rapidly as *rooms* gets smaller or larger. For example, at $rooms = 7$, the return to the next room is about 32.3%.

What happens generally if the coefficients on the level and squared terms have the *same* sign (either both positive or both negative) and the explanatory variable is necessarily nonnegative (as in the case of *rooms* or *exper*)? In either case, there is no turning point for values *x* > 0. For example, if β_1 and β_2 are both positive, the smallest expected value of *y* is at $x = 0$, and increases in *x* always have a positive and increasing effect on *y*. (This is also true if $\beta_1 = 0$ and $\beta_2 > 0$, which means that the partial effect is zero at $x = 0$ and increasing as *x* increases.) Similarly, if β_1 and β_2 are both negative, the largest expected value of *y* is at $x = 0$, and increases in *x* have a negative effect on *y*, with the magnitude of the effect increasing as *x* gets larger.

The general formula for the turning point of any quadratic is $x^* = -\widehat{\beta}_1/(2\widehat{\beta}_2)$, which leads to a positive value if $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have opposite signs and a negative value when $\widehat{\beta}_1$ and $\widehat{\beta}_2$ have the same sign. Knowing this simple formula is useful in cases where *x* may take on both positive and negative values; one can compute the turning point and see if it makes sense, taking into account the range of *x* in the sample.

There are many other possibilities for using quadratics along with logarithms. For example, an extension of (6.14) that allows a nonconstant elasticity between *price* and *nox* is

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 [\log(nox)]^2 + \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u. \quad \textbf{[6.15]}$$

If $\beta_2 = 0$, then β_1 is the elasticity of *price* with respect to *nox*. Otherwise, this elasticity depends on the level of *nox*. To see this, we can combine the arguments for the partial effects in the quadratic and logarithmic models to show that

$$\%\Delta price \approx [\beta_1 + 2\beta_2 \log(nox)]\%\Delta nox; \quad \textbf{[6.16]}$$

therefore, the elasticity of *price* with respect to *nox* is $\beta_1 + 2\beta_2 \log(nox)$, so that it depends on log(*nox*).

Finally, other polynomial terms can be included in regression models. Certainly, the quadratic is seen most often, but a cubic and even a quartic term appear now and then. An often reasonable functional form for a total cost function is

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u.$$

Estimating such a model causes no complications. Interpreting the parameters is more involved (though straightforward using calculus); we do not study these models further.