Reading Guide for *Mastering Metrics*, Chapters 1 and 2

Matthew J. Holian*

August 19, 2020

Abstract

This document includes reading guides to Mastering Metrics chapters 1 and 2. The topics covered include analysis of data from randomized experiments which requires a difference-in-means test, usually carried out by estimating a bivariate regression model with a binary independent variable, and regression control, which uses control variables in multivariate regression equations estimated with observational data. Randomized experiments and regression control are the first and second of The Furious Five (Angrist and Pischke, 2015.)

*This document is designed to serve as a reading guide for students or self-directed learners. Please send comments or error corrections to matthew.holian@sjsu.edu.

1 Active Reading Guide, MM Ch 1 and Ch 1 Appendix

To assist students in actively reading Mastering Metrics, Chapter 1, I wrote seventeen questions. Try to think of the answer first before reading my answer. The key statistical concept in this chapter is a difference in means test, and the key conceptual issue relates to how a difference in means estimated with observational data differs from a treatment effect due to selection bias.

1.In Table 1.1, first row, first column, the number 4.01 appears. Describe this number.

This is the average health rating of men who have some health insurance.

2. In Table 1.1, first row, second column, the number 3.7 appears. Describe this number.

This is the average health rating of men who do not have health insurance.

3. In the first row, third column of Table 1.1, what does the number 0.31 represent?

This is the difference in average health between men with and without health insurance.

4. Does this table contain evidence that health insurance causes people to be healthier? Why or why not?

No. We see only that men with insurance are healthier on average, but we don't know from the table why they are healthier.

5. Below the numbers mentioned in questions 1 and 2 above are the numbers 0.93 and 1.01, which appear in square brackets. What do these numbers measure? (Hint: Read page 36 of the Chapter Appendix and note the formula in footnote 16.)

While 4.01 and 3.70 are mean health ratings of men with and without insurance, 0.93 and 1.01 are standard deviations of health ratings of men with and without insurance, respectively.

6. Below the number mentioned in question 3 above, 0.03 appears in parentheses. What does this number measure? (Hint: Take a look at page 45 of the Chapter Appendix and note the formula in footnote 17. The formula that appears here is important and you'll be returning to it again and again in this course; for now we mainly just want to introduce you to the concept and show you where to find the formulas for when you need to calculate these values yourself.)

0.03 is the standard error of the difference in means. Conceptually, this is the standard deviation of the difference in means in repeated samples; it allows us to pin down a range for the statistic (the difference in means) that we should expect to see in repeated samples of men with and without insurance. The difference in means (0.31) is large relative to the standard error.

7. See page 10, first paragraph. What does it mean for "...the difference in means by insurance status [to be] contaminated by selection bias"?

Because individuals select whether or not they have insurance, the difference in means does not tell us what is the causal effect of insurance on health. The difference in means is thus not informative and, as the author's say, is contaminated.

8. See page 10, second paragraph. Why do the individuals with insurance in Table 1.1 have higher levels of health than those individuals without health insurance?

It may be that they have better health because they are better educated.

9. See page 10, third paragraph. What does it mean for "...the principle challenge facing masters of 'metrics is elimination of the selection bias that arises from ... unobserved differences."?

If you only have observational data, then you can't look at a difference in means and conclude much about causal effects. Thus econometricians have devised clever techniques for

estimating causal effects from observational data.

10. See page 11, first paragraph. What does it mean to say, "Experimental random assignment eliminates selection bias."? In your own words, describe how an experiment with random assignment of subjects into control and treatment groups works to overcome the challenges brought about by unobserved factors that plagued the comparison of health outcomes by insurance status in the survey data in Table 1.1.

When people are free to choose whether or not they have insurance, there are many differences between those who do and do not select coverage. When insurance is randomly assignment, then the groups of people who do and do not have insurance are the same, on average. This is because the insurance was randomly assigned. A more educated participant in the experiment would be equally likely to receive coverage as an uneducated, and the same is true for richer participants, older, smarter, prettier, nicer, funnier, or any other trait that we can imagine, whether or not we can measure it. Thus in a health insurance experiment, the only difference between those that do and do no have insurance is their coverage status, and, on average at least, the two groups are identical in every other way. Thus if later on we observe those with insurance are healthier, it must be the insurance that made them so.

11. What is observational data? What is experimental data? Is the data analyzed in Table 1.1 observational or experimental? Is the data analyzed in Table 1.3 observational or experimental?

Observational data is data collected by observing (people's) characteristics. More commonly, we'll ask people questions and record (observe) their answers. On the other hand, experimental data is data gathered from a controlled experiment with random assignment into control and treatment group. The data in Table 1.1 is observational, the data in Table 1.3 is experimental.

12. Consider Table 1.3 from Mastering Metrics. What is the average value of Family

Income for individuals in the Catastrophic plan? (Hint: Answer is shown in the first column.)

31,603

13. Consider again Table 1.3. What is the average value of Family Income for individuals with any level of insurance? (Hint: this value is not reported in the table, but it can be determined with the information that is reported there.)

From the information in the table, we know that:

(Average family income of those with any sort of insurance) - (Average family income of those in the catastrophic plan) = -654.
Plugging in 31,603 on the left hand side we have:

(Average family income of those with any sort of insurance) - (31,603) = -654.

Rearranging, Avg family income of those with any sort of insurance = 31,603 - 654 = 30,949.

14. Consider now Table 1.1, and compare and contrast differences in average income between groups with and without insurance there and in Table 1.3. In other words, in Table 1.1, are average incomes different among respondents with and without insurance? In Table 1.3, are average incomes different among respondents with insurance versus those with the catastrophic-only plan? Are your answers to these two questions the same or different? Why?

In Table 1.1 average incomes are higher among respondents with insurance. This makes sense as richer people find it easier to purchase insurance. In Table 1.3, average income is essentially the same between those with and without insurance; this also makes sense as being richer did not make experimental participants more likely to receive insurance as insurance coverage was randomly assigned. Thus the answer to the two questions is different, and the reason is because Table 1.1 analyzes observational data and Table 1.3 analyzes experimental.

15. Consider a test of the null hypothesis that there is no difference in Average Family Income among individuals with any insurance plan, and individuals with the catastrophic plan. Examine the statistics reported in Table 1.3. What is the absolute value of the test statistic for this test?

In the middle of the page on page 45 of Mastering Metrics we see the test statistic for a difference in means test is

"When the null hypothesis is one of equal means...", the formula for the test statistic simplifies to: t(0) = (diff in means - 0) / (standard error). From Table 1.3, we know that (diff in means)=-654. Plug this in the numerator of the right-hand side, and find the corresponding standard error from the table, and plug this in the denominator of the right-hand side. Then divide the top by the bottom to find the test statistic. So we have: test statistic = -654 / 1,181. This equals -0.55. Thus the absolute value is 0.55.

16. On p. 45, When the t-statistic is large enough to reject a difference of zero, we say the estimated difference is statistically significant. How large is large enough? Hint: see p. 21.

From p. 21, "Differences that are larger than about two standard errors are said to be statistically significant" "In other words, when a difference in sample averages is smaller than about two standard errors, the difference is typically judged to be a chance finding..."

17. Considering your answer to Question 16 above, do you reject the null hypothesis of equal family income? Is the observed difference in family income between subjects in the control and treatment group statistically significant?

We found the test statistic to be 0.55. This is less than 2. (0.55 ¡ 2), so we cannot reject the null of equal means. Given the difference in means between the two groups is small relative to its standard error, it is likely the two groups have the same average income (we don't have enough evidence to say they do not have the same income, on average.)

2 Active Reading Guide, MM Ch 2

This week we will discuss Chapter 2 from Mastering Metrics, on multiple regression. This is a technique we can use with observational data to eliminate or at least reduce bias in our coefficient estimates.

Earlier I have emphasized that simple regression provides an alternate, and often more convenient way of doing difference in means tests. (In fact, all the difference in means tests presented in Ch 1 of MM were carried out in a regression framework, which you can see if you download the Stata files the authors used in creating the tables.)

Once a student understands that a hypothesis test of a coefficient in a simple regression model with a binary (dummy) independent variable is the same thing as a difference in means test, it is straightforward to extend this understanding to the case of a simple regression model with a continuous independent variable, or to a hypothesis test of a regression coefficient in a model with multiple right-hand side variables, i.e. a multiple regression model, which is the focus of this chapter.

Let's start with a brief review of MM Ch 1. There we learned that with observational data, the difference in means is equal to the treatment effect plus the selection effect. In data from a controlled, randomized experiment, there is no selection effect, so the difference in means can be interpreted as the treatment effect (i.e. the causal effect.)

Equation 1.4 of MM Ch 1 shows:

Difference in Group Means = Average Causal Effect + Selection Bias

Where in the application there, Selection Bias is what the difference in health would be between the two groups in a world where no one has insurance. Of course we don' t observe what the health of someone who has insurance would be if they didn't have insurance, so the value of this equation is in clarifying the concept of selection bias and how a difference in means in observational data is not equal to the causal effect. " Page 11 of Ch 1 contains the following passage that introduces the intuition behind how multiple regression works

to isolate treatment effects. "We wrap up this discussion by pointing out the subtle role played by information like that reported in panel B of Table 1.1. This panel shows that the groups being compared differ in ways that we can observe. As we' ll see *in the next chapter*, if the only source of selection bias is a set of differences in characteristics that we can observe and measure, selection bias is (relatively) easy to fix. Suppose, for example, that the only source of selection bias in the insurance comparison is education. This bias is eliminated by focusing on samples of people with the same schooling, say, college graduates. The subtlety in Table 1.1 arises because when observed difference proliferate, so should our suspicions about unobserved differences...The principle challenge facing masters of metrics is elimination of the selection bias that arises from such unobserved differences." (p. 11, emphasis added).

Usually we don' t have the luxury of running a randomized controlled experiment. Done correctly, estimates obtained from multiple regression models can provide persuasive evidence of causal effects. If the analyst can "control" for enough variables, then the "treatment" is as good as randomly assigned and the coefficient on the "treatment" variable is the causal effect.

Now let' s turn to a passage from the first page of Chapter 2:

"Regression-based causal inference is predicated on the assumption that when key observed variables have been made equal across treatment and control groups, selection bias from things we can't see is also *mostly* eliminated." (p. 47, emphasis added.)

I include this quote because it highlights that evaluating whether estimates from multiple regression models can be interpreted as causal effects is sometimes more of an art than a science. There is no test we can do. Instead, its up to the reader to determine whether they think the analyst has done a good enough job of controlling for confounding variables.

Consider the data in Table 1, adapted from Angrist and Pischke's (2015) Table 2.1:

The R script I wrote that is associated with this chapter uses the following five lines to code in the data:

Table 1: Data from MM Table 2.1

Student	Group	Private	Earnings
1	A	1	110
2	A	1	100
3	A	0	110
4	B	1	60
5	B	0	30

STUDENT = c(1,2,3,4,5,6,7,8,9)

GROUP = c("A", "A", "A", "B", "B", "C", "C", "D", "D")

PRIVATE = c(1,1,0,1,0,1,1,0,0)

EARNINGS = c(110,100,110,60,30,115,75,90,60)

MMCh2Table1 = data.frame(STUDENT, GROUP, PRIVATE, EARNINGS)

If we use these data to estimate the model that appears on p. 57 of MM as equation 2.1, we get the results shown in the Table 2 here, in Column 2 (this uses data from only the first 5 students.) As indicated on page 55, "...the uncontrolled comparisons generates a gap of $20,000" and we see this in column 1 in Table 2 here, which includes the Private school dummy but no ability control.)

Columns 3 and 4 estimate the uncontrolled and controlled models, respectively, this time on the whole sample of nine students. This is to show that the coefficient on Private is the same in (2) and (4). The authors note on page 54, "Groups C and D are uninformative, because...each is composed of either all-treated or all-control individuals." We see this is true in the table.

Ignore column 5 for now; though you might just note for now that the dependent variable is the GroupA dummy, not Earnings as in models in columns 1-4.

It is important to become comfortable interpreting tables like the one above, as this is the standard way regression results are reported in articles and term papers like you' re writing.

I next draw your attention to several important points and nuances in the chapter.

Table 2: Regression Analysis of Data from MM Ch 2 Table 1

	Dependent variable:				
	EARNINGS				GROUPA
	(1)	(2)	(3)	(4)	(5)
PRIVATE	20.000	10.000	19.500	10.000	0.167
	(39.907)	(13.924)	(20.248)	(13.209)	(0.576)
GROUPA		60.000***			
		(13.924)			
factor(GROUP)B				−60.000***	
				(13.209)	
factor(GROUP)C				−15.000	
				(22.474)	
factor(GROUP)D				−25.000	
				(20.230)	
Constant	70.000*	40.000***	72.500***	100.000***	0.500
	(36.515)	(13.171)	(17.185)	(12.495)	(0.456)
Observations	5	5	9	9	5
R^2	0.094	0.921	0.125	0.756	0.028
Adjusted R^2	−0.207	0.843	0.0002	0.511	−0.296
Residual Std. Error	39.158	14.142	29.044	20.310	0.624
F Statistic	0.313	11.700*	1.002	3.091	0.086
Note:					*p<0.1; **p<0.05; ***p<0.01

p. 58, "The last component of equation (2.1) is the residual (also called an error term)." Although this is correct, it' s best to think about it as a prediction error here, as there is a conceptually very different concept of error term that is a population concept. This is a nuance, but put it in the back of your mind for now and we' ll return to this point later in the term.

p. 60 The CB data contains 5,583 students that can be placed into 151 selectivity groups. Although this is a much larger number than the four groups we considered in the hypothetical example above, the principle is the same (and note in the R script the "factor" command can be used to easily include large numbers of dummy variables in a model.)

p. 61. Although there are 151 groups, only 150 group dummies are included to avoid perfect multicollinearity (also called the dummy variable trap in this context; footnote 4 on p. 57.) Note also on p. 61 the dependent variable is logged. More on this later.

p. 62, bottom, compares standard errors for regression to diff in means. On p. 63 (top) they note that the private school coefficient is statistically significant. Nonetheless, "some of this gap is almost certainly due to selection bias." (pp. 63-64)

Table 2.3 utilizes a different empirical strategy that enables the authors to use the entire sample. It may have been included as a robustness check or to investigate a related but slightly different hypothesis. Here we see that controlling for the average SAT score among students at the schools applied to results in qualitatively similar findings as the results in Table 2.2 with a more complex control strategy (but smaller sample size.) If all the tables in this chapter overwhelms you, just focus on Table 2.2 and don't worry so much about Tables 2.3-2.5.

Table 2.4 seems to have been included to investigate a related but slightly different hypothesis. Maybe it isn' t the fact that the schools are private that matters but the fact that they have smarter students (stronger peer effects.) The only difference between Tables 2.3 and 2.4 is that 2.3 includes a private school dummy as the main independent variable of interest, while in 2.4 it is replaced with the average SAT score of students at the school

attended.

In all cases, it seems that students who went to private schools, or more selective schools, would have done about the same whether they went to public schools, or less selective schools. In other words, the results are consistent with a world where it is primarily the innate abilities of students that are driving their future earnings, not the decisions they make about whether to attend private or public, selective or less selective. Of course these decisions are made on the margin and if a smart student went to the worst college rather than Harvard their earnings probably would suffer, but generally students who have a decision to make regarding attending an elite school are considering other very selective schools. On the margin, attending a little more selective school doesn' t seem to be causing higher future earnings.

Section 2.3 aims to formalize the idea of selection bias in regression context.

p. 69. "The regression version of the selection bias generated by inadequate controls is called omitted variables bias (OVB), and it' s one of the most important ideas in the metrics canon."

The estimates of equation 2.3 (the long equation) appear in column 2 of the table above, while estimates of the short equation (on the top of the next page) appear in column 1. This section describes an interesting relationship between those estimates:

Effect of P in short = Effect of P in long + [(the relationship between omitted and included) times (the effect of omitted in long).

This sounds a bit confusing but it is the clearest way to describing these relationships. Let' s take these one at a time, while referencing Table 2 in this document:

Effect of P in short = 20 (see column 1)

Effect of P in long = 10 (see column 2)

The effect of omitted in long = 60 (see column 2)

The relationship between omitted and included = 0.167 (see column 5).

The one statistic we haven' t yet discussed is this last one the relationship between omitted and included. To get this, we regress the ability control (A) on the private school dummy (P). These results are in the last column of the table in column 5.

Thus we have:

Effect of P in short = Effect of P in long + [(the relationship between omitted and included) times (the effect of omitted in long)

$$20 = 10 + [(0.167) * (60)].$$

The right-hand side is 10 + 0.167*60 = 10 + 10.02. The right-hand side of this does no t precisely equal 20 due to rounding error only.

This equation is the foundation for the *Omitted Variables Bias* formula presented in MM Ch 2.

Rearranging, the OVB (which is the difference between the effect of P in short versus long) is equal to the relationship between ability on earnings (effect of A in long) times the relationship between earnings ability.

Later in this semester we' ll return to OVB. We' ll see that OVB exists when two conditions are met: 1.) the omitted variable is a determinant of Y (here this is the effect of A in long) and 2.) the omitted variable is correlated with the included independent variable (this is the relationship between A and P). Clearly if either of these two conditions don' t hold, then the OVB will be zero. The formula on the bottom of p. 71 tries to make this clear.

The last part of this section describes the use of the OVB formula. There are some variables we can' t include, such as family size, because we simply don't have data on them. What is the consequence of omitting such factors? The short answer is that our estimate of the effect of P on earnings is overstated. But they found it to be very small. So if it is overstated, then it is even closer to zero. Thus we can use the OVB formula to argue that the estimates we found are an upper-bound. This only strengthens the conclusion that the effect of private school attendance on earnings is zero or very low.

Sensitivity Analysis

The section on sensitivity analysis makes the point that we typically include more than one set of estimates in our papers. If the main results don't change when we modify the list of controls, we say the results are *robust* to alternative specifications. Note the jargon here, robust does not mean *robust standard errors*. Also, a *specification* refers to which variables are included in the model. (Specification also refers to *how* the variables are included in the model, for example, are they included in logs or in levels?)

The final part of this section puts the OVB formula to work in the more complicated models discussed in this chapter. But the idea is the same as I discuss above in the context of the models estimated with the five-person data file described above.

By now all students know that "correlation doesn' t prove causation". Hopefully in addition students can put together some reasonable answers to the question, Well, what does prove causation? My take on this is that nothing really proves causation, but the most suggestive evidence comes from a randomized controlled experiment. Short of that, a regression analysis with a clever control strategy can also be highly suggestive, though perhaps not as compelling as evidence from an experiment.

3 Active Reading Guide, MM Ch 2, Appendix

The appendix is a concise introduction to regression details that will occupy us for most of the semester. There are a few additional concepts I included in this section, along with clarifications and discussion of the key points from their appendix. Don' t worry if you have trouble fully grasping everything, it's complicated and you' ll be seeing it again. Along with the concepts, including hypothesis testing, discussed in the Chapter 1 Appendix, understanding the eight concepts below will take you far in making sense of econometrics research:

1. RSS (Residual Sum of Squares)

2. Measures of Fit (R-squared and Residual Standard Error)

3. Interpreting models with logged variables

4. Polynomial models (using squared and cubed variables)

5. Standard errors for regression coefficients

6. Confidence intervals

7. Test statistics

8. P-values

The first topic is Residual Sum of Squares.

On page 86 they discuss the Residual Sum of Squares (RSS). They provide a definition:

$$RSS = E[Y_i - a - bX_i]^2$$

In fact, I find this to be incorrect, as the E should be the summation operator (capital Greek sigma, $\sum$). They mention in a sample of data we "...replace expectation with sample average or sum..." but I would point out, average and sum are not the same thing. Thus the strictly speaking correct formula is:

$$RSS = \sum_i [Y_i - a - bX_i]^2$$

In some sense it isn' t important for what they wanted to say, as "the values of a and b that minimize the RSS" are the same values regardless of which version is used. But we' ll see below under Measures of Fit that knowing the correct definnition of RSS matters.

Your regression software knows how to find the values of α and β because it knows the formulas for regression coefficients (equation 2.7). These result from minimizing the RSS. The β coefficient (the slope) is the ratio of covariance between Y and X to the variance of X.

Meanwhile, the α coefficient (the intercept) is the average of Y minus β times the average of X. Thus you have to find the β estimate before you can find the α estimate. (Hint: When you see E[Y], it may be helpful for you to think of this in terms of an actual sample of data, in which case it is $\bar{Y}$)

You could in fact calculate the regression coefficients by hand, if you were inclined. You have the definition for covariance given at top of p. 86. This defines covariance in the population. If we have a sample of data, it is calculated much like variance. (Formulas for sample covariance and correlation are provided in SW equations 3.24 and 3.25.) If you want to derive the OLS estimators to see where the the α and β formulas come from; see the lecture notes at: `https://are.berkeley.edu/courses/EEP118/current/derive_ols.pdf`.

Measures of Fit

MM does not elaborate on how RSS can be used to assess how well the model fits the data, i.e. how well it predicts. There are a few ways. Well discuss the Residual Standard Error and the R-squared .

Residual Standard Error

This is a somewhat unfortunately named term. I prefer to think about it as the "standard deviation of the residuals".

First, and intuitively, if RSS is small, that is a good thing, right? It means your model predictions are close to the actual values. So we calculate Residual Standard Error as the square root of $\frac{RSS}{n}$. Actually it is $\frac{RSS}{n-k-1}$) where n is the number of observations and k is the number of independent variables (k=1 in simple regression.) Using n versus (n-k-1) in the denominator is a nuance called a "degrees of freedom" adjustment and doesn't make a noticeable difference in large samples, but if you're trying to see where the statistics reported in regression output come from, especially with small samples, it's important to have the exact formula. Essentially, the Residual Standard Error can be though of as the standard deviation of the residual.

The table below helps shows how RSS and thus the Residual Standard Error is calculated.

Table 3: Data from MM Table 2.1 and model predictions

Student	Group	Private	Earnings	predicted Earnings	residual	squared residual
1	A	1	110	90	20	400
2	A	1	100	90	10	100
3	A	0	110	70	40	1600
4	B	1	60	90	-30	900
5	B	0	30	70	-40	1600

The final column shows the squared prediction error for each observation. If we sum them together, we find that 400+100+1600+900+1600=4600. This is the RSS. To find the residual standard error, we divide 4600 and take the square root:

$$RSE = \sqrt{\frac{4600}{5-2}}$$

Note also in Table 2 of this document, the RSE is 39.158. This is the same thing you'll find if you solve out the right-hand side of the equation above.

R squared

The most common way to assess models is by the concept of R-squared:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where TSS is the Total Sum of Squares. TSS is the numerator in the formula for the variance of the dependent variable, i.e. the sum of squared differences between the dependent variable and its average.

$$TSS = \sum_i [Y_i - \bar{Y}]^2$$

Using our example data, and noting that the average value of earnings is 82,

$$TSS = [(110-82)^2 + (100-82)^2 + (110-82)^2 + (60-82)^2 + (30-82)^2] = 5080.$$

Above I mentioned TSS is the numerator in the formula for the variance of the dependent variable. Thus variance of Y is 5080/(5-1), or 1,270. The standard deviation of Y is the square root of the variance, and is 35.6.

Now that we have both RSS and TSS in hand, we can calculate R-squared.

$$R^2 = 1 - \frac{4600}{5080} = 0.094$$

.

You can verify this is the same value reported in column 1 of Table 2.

Building models with logs

This section describes one way to transform data to make the models predict better or easier to interpret. In particular, we can take the natural logarithm. Remember log base 10 from algebra? Well the natural log is log base 2.718 (where 2.718 is the value of the natural exponent e to three decimal places.) It's easy to take a variable, say income in dollars, and take the natural log of it, and then to use the logged value of the variable as either an independent or dependent variable.

This section contains a typo. At the top of p. 94 they write that the β coefficient, in a log-linear model, can be interpreted as the percentage change in Y due to a one-unit change in X. Actually, it is 100 times the β coefficient!

For example, on the bottom of page 82 they write, the "...CEF is strongly upward-sloping, with an averages slope of about 0.1. In other words, each year of schooling is associated with wages that are about 10% higher on average."

But, if you used the formula on page 94, you would say each year of schooling is associated

with wages that are about .1% higher on average, which is off by a factor of 100.

From the figure, it looks like the equation is ln(earnings)=5.8 + 0.1*YrsSchool.

Interpreting models with logs can be confusing. Therefore I suggest you consult the following table of rules for interpreting different log models:

Table 4: How to interpret logged models

model	equation	interpretation
Log-linear	$\ln Y_i = \beta_0 + \beta_1 X_i + e_i$	A one-unit increase in X is associated with a β_1 percent change in Y (on a 0-1 scale).
Linear-log	$Y_i = \beta_0 + \beta_1 \ln X_i + e_i$	A one percent increase in X is associated with a $\beta_1/100$ change in Y.
Log-log	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + e_i$	A one-percent increase in X is associated with a β_1 percent change in Y (on a 0-100 scale)

We have been discussing log models, but there are other types of models that are nonlinear in the variables, such as polynomial modes (e.g. $Y = B_0 + B_1 * X + B_2 * X^2$) and interaction models (e.g. $Y = B_0 + B_1 X + B_2 Z + B_3 (X * Z)$). These models are straightforward to estimate, but like with log models, interpreting them can be a challenge. The best way to interpret such models is to evaluate them at two sets of X values. For example with the polynomial, find fitted values of the model when X equals 1 and then when X equals 2. It is also possible to take a derivative.

Standard Errors for Regression Coefficients

The last topic discussed in MM Ch 2 Appendix addresses the sampling distribution of the β estimator. The key take away is that, to do a hypothesis test of a regression coefficient you need to know the standard error of the β estimate in repeated sampling.

In this chapter we see two statistics with "standard error" in the name. Above under measures of fit we saw the Residual Standard Error, which is Goodness of Fit measure. Now, we're talking about standard errors of coefficient estimates, and they are used in hypothesis testing, which is conceptually different from evaluating how well a model predicts (the point

of assessing goodness of fit.)

We saw in Chapter 1 the formula for the standard error (SE) for a test of a single mean (I gave the example of cans of sardines[1]) We also saw there the SE for a test of a difference in means, and in fact we saw there are two SE formulas we can use in a difference in means test. I told you then you should use the separate variance SE formula (that appears in footnote 17), not the pooled SE formula that appears as Equation 1.7 on p. 45).

Just as it is with difference in means tests, it is with tests of regression coefficients. There are two formulas. The simpler formula is given in equation 2.15 but the one you should actually use is given in 2.16. Both of these look something like the SE formula for a single mean, in that it has N in the denominator. So, in both types of hypothesis tests, larger samples give smaller estimates SEs. (And as we see below, smaller SEs yield larger test statistics.) Your regression software knows these SE formulas, so you don' t have to remember them, but try to understand them conceptually and intuitively.

In the section titled, "Regression Standard errors and Confidence Intervals" there is no discussion of confidence intervals. So here you go.

Confidence Intervals

[B-1.96*SE; B+1.96*SE]

The 95% confidence interval tells us an interval that the coefficient estimate will fall in in 95 out of 100 repeated samples. You might say we are 95% sure the true value of B is in this interval, although the strict definition is the one given previously.

The idea is the same as that discussed in the context of a single mean in Chapter 1 on page 43. We only have to replace the average with the regression coefficient. Note also, 1.96 is the exact value to be used here, while on page 43 they use 2 which is 1.96 rounded up to the nearest whole number.

Test Statistics

[1] http://mattholian.blogspot.com/2015/11/how-not-to-get-ripped-off-by-safeway.html.

In chapter one we encountered two test statistics. One on page 39 was the test statistic for the hypothesis test of a *single mean*. The second on page 45 was the test statistic for the hypothesis test of a *difference in means*. Compare these two formulas, they are different but have the same general form:

Test statistic = (estimated statistic - null hypothesis) / (appropriate standard error)

In the case of a single mean, the estimated statistic is just the mean (say 9.5 sardines) while the null hypothesis is whatever we want it to be (say the package says 10 sardines per can and we want to test this, then the null is 10.) The appropriate standard error for a test of a single mean is the one given on the top of page 39.

In the case of a difference in means, the estimated statistic is the difference in two means, while the null hypothesis is that there is no difference (i.e. $\mu = 0$). The null is almost always zero in the case of a difference in means, although in principle it could be something else. In this class it will probably always be zero. There are two standard error formulas that could go in the denominator, the separate variance SE formula (that appears in footnote 17), and the pooled SE formula, that appears as Equation 1.7 on p. 45. Best practice is to use the separate variance SE formula there.

In both cases, we calculate the test statistic and then compare the absolute value of the test statistic to the critical value, which is 1.96 for a test at the 5% significance level.

A third type of hypothesis test is a test of a regression coefficient. The test statistic has the same general form as shown above. In the specific case of a regression coefficient it is:

$$TestStatistic = \frac{\hat{\beta} - 0}{SE}$$

On the right hand side, we have the estimated β coefficient, minus zero, divided by the SE. As discussed above, there are two SE formulas, the *homoskedastic-only* formula given on page 96, and the *heteroskedastic-robust* formula given on page 97. Have your regression software calculate these, and have it use the robust formula.

Again, just as in the case of a difference in means test, a high test statistic tell you the coefficient is statistically significant, that is, unlikely to have a population value of zero. And again, compare to the critical value (usually 1.96).

p-value

After calculating the test statistic, you can calculate a p-value. A high test stat results in a low p-value

The p-value tells us the probability of incorrectly rejecting the null. (The null, which is often there's no difference between the groups, may be true, but we find a t-statistic that is large and so reject it, thus making an error.) There is an inverse relationship between the t-statistic and the p-value. We'll cover this topic again after the midterm so for now my main goal is for students to develop an intuitive understanding of p-values.

Here are some questions about p-values associated with two-tailed hypothesis tests based off of the standard normal distribution (you probably encountered one-tail tests, and the Student t distribution in your intro stats class; in this class, we'll only do two-tailed tests and use the standard normal distribution.)

1. If the test statistic is 1.96, what is the p-value?
2. If the test statistic is -1.96, what is the p-value?
3. If the test statistic is 2, is the p-value $>$ or $<$ 0.05?
4. If the test statistic is 1.9, is the p-vale $>$ or $<$ 0.05?

Answers:

1. 05. This means, 95% of the area of the std. normal distribution is between -1.96 and +1.96. in other words, 5% of the area is in the tails.

2. also 0.05. Because it is a two-tailed test, we take the area in both tails of the distribution. In particular, the p-value is all the area that is to the left of the negative value of the test statistic, plus all the area that is to the right of the positive test statistic.

3. There's an inverse relationship between the t-stat and the p-value. If the t-stat is greater than 1.96, the p-value will be less than 0.05.

4. Similarly, if the t-stat is less than 1.96, the p-value will be ¿ 0.05. In the case of a t-stat equal to 1.9, we cannot reject the null at the 5

You can calculate the exact p-value by using a spreadsheet or in R. Say your test statistic is 2. You can find the p-value by typing in a spreadsheet like MS Excel: =2*normdist(-2,0,1,true), which is a little less than 0.05. Here, the normdist function is the cumulative density function for the standard normal distribution. It provides the area in the tails of the standard normal (the bell curve, with mean zero and variance of one), to the left of an x-value, where X is specified here as -2.