Course Preview: Designing and Running Randomized Evaluations

Answers

This course preview is meant to give prospective learners the opportunity to get a taste of the content and exercises that will be covered in the course. While there are no prerequisites for this online course, it is recommended that learners have some familiarity with economics or statistics. Each question below is tied to concepts that will appear in this course, all of which it would be beneficial to feel comfortable with, as course work will ask you to apply these concepts. If you are new to these subjects, or eager to refresh your memory, please do consult the available resources below, and be prepared to refer to these resources over the course of the class. Try to first answer these questions without consulting the resources, but fear not if you do consult them - being an agile user of outside resources will help you succeed in this course.

A score of 60% or above in this course preview indicates that you are ready to take this course, while a score below 60% indicates that you should further review some concepts in the attached materials before commencing the course.

Useful Resources:

- *Review of key statistical concepts*: Khan Academy: Statistics and probability
- *Introduction to econometrics*: Mastering 'Metrics: The path from cause to effect
- *Overview of RCTs*: Running Randomized Evaluations: The book and the blog

1. **Probability**: A fair coin is tossed three times. What is the probability of getting at least two heads? (1 point)

   Solution: P($\geq$ 2 heads) = P(= 2 heads) + P(= 3 heads) =

   $$\frac{\binom{3}{2}+\binom{3}{3}}{2^3}$$

   where:

   $$\binom{3}{2} = 3 \quad \text{and} \quad \binom{3}{3} = 1$$

   thus:

   $$\frac{3+1}{2^3} \quad = \quad \frac{4}{8} \quad = \quad \frac{1}{2}$$

2. **Exponential Functions**: As an epidemiologist, you realize that you can use an exponential function to map the spread of malaria, as follows:

   $$Malaria_{\text{today}} = 10 + 1.5 * (Malaria_{\text{yesterday}})^2$$

   where Malaria$_{\text{today}}$ is the number of individuals in the population with malaria today, while Malaria$_{\text{yesterday}}$ is the number of individuals in the population with malaria yesterday.

   If you know there are 16 cases of malaria today, how many were there yesterday? (1 point)

   Solution:

   $$16 = 10 + 1.5x^2$$

   Use algebra to find that $x = 2$, indicating that there were 2 cases of malaria yesterday.

3. **Measurement**: In your role as a clinic manager, you conduct monthly tests on your nurses' sanitation practices. In July, only 50% of nurses passed the sanitation test. (1 point)

   Thinking ahead to the test in August, would you rather see a:

   (A) a 50% increase in the number of nurses who passed

   (B) a 50 percentage point increase in the number of nurses who passed

   (C) A and B are equivalent, so both responses are correct

   Solution: You would rather see (B) a 50 percentage point increase, because:

   (A) a 50% increase in the 50% of nurses passing is a 25 percentage point increase = 75% of nurses would pass

   (B) a 50 percentage point increase in the 50% of nurses passing = 100% of nurses would pass

   (C) A and B are equivalent: This is a false statement, because (B) > (A)

As you would rather see 100% of nurses pass than only 75%, (B) is the right response.

4. **Interpreting Regressions**: The administration of a school district in Rajasthan, India wishes to understand the effect of class size on learning outcomes, measured as:

$$score = \alpha + \beta_1 * ratio + \beta_2 * income + \beta_3 * attendance$$

where the variables are defined as:

Dependent Variable (Y) :

- *score* = student's score on last year's final exam (out of 100 points)

Independent Variables (X) :

- *ratio* = number of students per teacher in the student's classroom
- *income* = annual household income of student (in thousands of Rupees)
- *attendance* = number of class days missed by the student each month

Data was collected and analyzed, producing the following suite of regression models. Please read the regression table below and answer the four associated questions.

Table 1: Regressions of Student-Teacher Ratio on Student Test Scores

Variable Label	(1)	(2)	(3)
Student-teacher ratio	-2.070**	-0.649	0.047
	(0.719)	(0.552)	(0.495)
Parental income		1.769**	1.436**
		(0.821)	(0.075)
Days absent			-1.41*
			(0.583)
Constant	48.36**	42.75**	41.12**
	(1.036)	(0.830)	(0.5956)
n	525	525	525

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

4.1. (a) In the first regression, what is the effect of adding one more student to the classroom (e.g. increasing the student-teacher ratio by one degree) on student test scores? Is this result statistically significant? (2 points)

Solution: On average, adding one more student to the classroom would lead to a decrease in test scores of 2.07 points. This result is statistically significant at the 0.05 level.

(b) How does the effect of the student-teacher ratio on test scores differ across the regression models? What does this imply about the true impact of the student-teacher ratio on test scores? (2 points)

Solution: While the *ratio* is both practically and statistically significant in the first regression model, adding more variables, particularly the *income* variable, causes the coefficient for *ratio* to lose its significance in subsequent regression models. This indicates that (1) variation in student test scores may primarily be attributed to parental income, rather than the student-teacher ratio, and (2) parental income may influence student test scores through the mechanism of the student-teacher ratio. For instance, wealthier parents would be more able to place their students in schools with a better student-teacher ratio.

Of note, the first regression is a good example of omitted variable bias, as excluding the *income* variable led researchers to find a significant correlation between *ratio* and *score* in the first regression, and could lead these researchers to falsely believe that student-teacher ratios may be a main driver of student test scores. However, building second and third regression models, which do include the *income* variable, allow researchers to recognize that this initial correlation between *ratio* and *score* in the first regression may, in fact, be attributable to the missing variable of parental income, which influences both the student-teacher ratio and student test scores. In fact, the latter two regressions indicate that parental income may itself be the main driver of student test scores, with student-teacher ratio only serving as a mechanism through which parental income influences student test scores. A regression that directly explores the correlation between parent income and the student-teacher ratio would help further investigate this theory.

4.2. What is the standard error for parental income in the second regression model? (1 point)

(a) 525
(b) 0.821
(c) 0.01
(d) 1.769

Solution: The standard error for each variable can be found in the parentheses underneath the variable's coefficient. In the case of parental income in the second regression model, there is a standard error of 0.821, such that the 95% confidence interval for the parental income coefficient of 1.769 is [0.948, 2.590].

4.3. Consider the third regression model. If a student misses three days of class, what is the predicted change in test scores? (1 point)

(a) A decrease of 4.23 points
(b) A decrease of 1.41 points
(c) An increase of 0.583 points
(d) An increase of 4.23 points

Solution: From the regression table, each absent day is correlated with an average decrease in test scores of 1.41 points, so missing three days would, on average, cause a student's test scores to decrease by 4.23 points; thus the answer is (A).

4.4. What does n represent in the above regression table? (1 point)

Solution: In econometrics, n stands for sample size. In the table above, each regression model uses a sample size of 525 student observations.

(a) alpha level
(b) sample size
(c) standard error
(d) error term