14.32/14.320 Econometric Data Science
Midterm Exam
1pm, Thursday October 26, 2023

1. This exam ends promptly at 2:30 PM.
2. The exam has three parts for a total of 100 points. Please put each part in a separate blue book. Put your name and MIT ID number on the cover of each blue book.
3. You are permitted one two-sided $8\frac{1}{2} \times 11$ sheet of notes, plus a calculator. No computers, wireless, or other electronic devices without prior permission. You may not share resources with anyone else.
4. Some questions ask you to draw a real-world judgment in a problem of practical importance. The quality of that judgment counts. For example, consider the question: "It is 10F outside. In your judgment, why are so many people wearing heavy coats?" The answer, "To stay warm" would receive more points than the answer, "Because they are fashion-conscious."
5. Please return this exam with your completed blue books.

Introduction

Does studying improve your grades? We examine this question using a data set from a liberal arts college in the southeastern United States. The data set consists of survey responses, merged with administrative records, for 210 students selected at random from the freshman class entering in the fall of 2000. All selected students completed the survey. The analysis here looks at grades (semester GPA), average study hours per day during the semester, demographic data, and other related data. The study time data were calculated from daily time diaries kept by the participating students. The data are summarized in Table 1.

Table 1. Variable Definitions and Summary Statistics

Data source: Panel Study of College Students taken in 2000. Unit of observation: individual college student; $n = 210$

Variable	Definition	Mean	Std. Dev.
GPA1	first semester college grade point average (0-4 scale)	3.004	.652
Study	study time, in hours per day (semester average)	3.427	1.631
video	= 1 if roommate brought a video game box = 0 otherwise	0.367	0.033
male	= 1 if male, = 0 if female	0.480	0.034
black	= 1 if black, = 0 otherwise	0.171	0.026
ACT	score on ACT standardized admission test (max = 36)	23.38	3.71
major_ag	= 1 if agriculture major, = 0 otherwise	0.076	0.018
major_bus	= 1 if business major, = 0 otherwise	0.176	0.026
major_ed	= 1 if elementary education major, = 0 otherwise	0.100	0.021
major_hum	= 1 if humanities major, = 0 otherwise	0.223	0.029
major_sci	= 1 if science or math major, = 0 otherwise	0.209	0.028
major_prof	= 1 if pre-professional major (nursing, pre-dental, pre-law) = 0 otherwise	0.119	0.022
major_ss	= 1 if social sciences major, = 0 otherwise	0.071	0.018
major_pe	= 1 if physical education major, = 0 otherwise	0.024	0.011
health_bad	= 1 if health is bad (self-reported) = 0 otherwise	0.067	0.017
health_exc	= 1 if health is excellent (self-reported) = 0 otherwise	0.371	0.033

Majors are most likely majors reported during the first-semester survey, divided into 8 mutually exclusive groups. For the self-reported health students choose one of three options: "bad", "excellent", "neither bad nor excellent"

Table 2. Study Hours and GPA: Regression Results ($n = 210$)

	(1)	(2)	(3)	(4)	(5)
Dependent Variable	GPA1	GPA1	GPA1	GPA1	Study
Regressors coefficients, and standard errors:					
Study	.038 (.025)	-	.038 (.025)	-	-
ln(Study)	-	.114 (.086)	-	-	-
Study ^{2}	-	-	$-.012$ (.054)	-	-
Study ^{3}	-	-	.0076 (.0064)	-	-
video	-	-	-	$-.241^{**}$ (.089)	$-.668^{**}$ (.252)
male	$-.132$ (.084)	$-.133$ (.082)	$-.130$ (.082)	$-.079$ (.086)	$-.155$ (.244)
black	$-.220^{+}$ (.122)	$-.224^{+}$ (.121)	$-.224^{+}$ (.123)	$-.209^{+}$ (.120)	.432 (.341)
ACT	$.062^{**}$ (.013)	$.060^{**}$ (.015)	$.061^{**}$ (.014)	$.062^{**}$ (.012)	$-.019$ (.036)
major_ag	$.834^{**}$ (.298)	$.830^{**}$ (.301)	$.831^{**}$ (.299)	$.906^{**}$ (.293)	1.423^{+} (.828)
major_bus	$.793^{**}$ (.282)	$.794^{**}$ (.281)	$.794^{**}$ (.284)	$.868^{**}$ (.277)	1.421^{+} (.783)
major_ed	$.725^{**}$ (.292)	$.730^{**}$ (.293)	$.729^{**}$ (.290)	$.739^{**}$ (.287)	1.12 (.811)
major_hum	$.796^{**}$ (.283)	$.794^{**}$ (.281)	$.792^{**}$ (.280)	$.889^{**}$ (.277)	1.637^{*} (.784)
major_sci	$.643^{*}$ (.280)	$.641^{*}$ (.285)	$.642^{*}$ (.285)	$.741^{**}$ (.274)	1.575^{*} (.776)
major_prof	$.664^{*}$ (.292)	$.667^{*}$ (.293)	$.667^{*}$ (.291)	$.731^{*}$ (.285)	1.777^{*} (.806)
major_ss	$.901^{**}$ (.304)	$.896^{**}$ (.299)	$.898^{**}$ (.301)	1.002^{**} (.295)	2.128^{*} (.836)
health_bad	.019 (.166)	.020 (.165)	.019 (.167)	.045 (.164)	.209 (.463)
health_exc	.127 (.086)	.123 (.085)	.128 (.087)	.149 (.085)	.095 (.241)
constant	$.719^{+}$ (.408)	.589 (.411)	.678 (.401)	$.793^{*}$ (.398)	2.28^{+} (1.42)
F-statistics testing the hypothesis that all coefficients are zero for:					
Study ^{2}, Study ^{3}	-	-	2.03	-	-
All major variables	3.62	3.64	3.61	5.13	2.30
Regression summary statistics:					
R^2	273	.274	.276	.289	.092
$\bar{R}^2$	.251	.252	.250	.263	.078

Notes: Heteroskedasticity-robust standard errors are given in parentheses under estimated coefficients, and p-values are given in parentheses under F - statistics. The F-statistics are heteroskedasticity-robust. Coefficients are individually statistically significant at the $^{+}10\%$, $*5\%, **1\%$ significance level.

Part 1 (36 points) - USE BLUE BOOK #1

1. A student is considering increasing studying from $2\frac{1}{2}$ hours per day to 3 hours per day. Compute a 95% confidence interval for the predicted effect of this increase using:
   (a) (5 points) Table 2, regression (1). In real-world terms, is this effect large or small?
   (b) (5 points) Table 2, regression (2). In real-world terms, is this effect large or small?
2. (5 points) Suppose that the first hour of studying is more productive (in terms of grades) than the second, etc; that is, suppose that there are declining marginal returns to studying. Which specification is better suited to model these declining marginal returns, (1) or (2)? Briefly explain.
3. (5 points) Test the null hypothesis that GPA1 is linear in Study, holding constant sex, race, ACT score, major, and health, against the alternative hypothesis that it is a polynomial of degree up to 3, at the 10% significance level. Explain your reasoning (be explicit).
4. (5 points) Test the hypothesis (at the 5% significance level) that Social Studies majors and Physical Education majors have the same GPA on average, holding constant hours studying, sex, race, ACT score, and health using regression (2).
5. (5 points) How would you modify regression (1) to examine whether the effect on grades of studying is the same for men and women, holding constant race, ACT score, major, and health? Be specific.
6. (6 points) Provide an example of a variable omitted from regression (1) that plausibly would result in omitted variable bias in the estimated effect on GPA1 of Study, controlling for the other variables in regression (1).

Part 2 (34 points) - USE BLUE BOOK #2

1. (6 points) Which regression among (1)- (3) would you prefer and why?
2. Consider the following threats to internal validity:
   (a) (7 points) sample selection bias
   (b) (7 points) simultaneous causality bias

   For each of these threats, in your judgment is the coefficient on Study in regression (1) likely to be subject to bias as a result of this threat? If so, is the coefficient in regression (1) arguably biased up (too large) or down (too small)? Explain, using an example as appropriate.

3. (6 points) For this and all following questions assume that roommates are assigned at random the summer before freshman year. Consider the effect on GPA1 of a roommate bringing a video game box to school, holding constant sex, race, ACT, major, and health. In your judgment, is coefficient on video in Table 2, regression (4), a biased or unbiased estimate of this effect? Explain.
4. Victor and Nathan are both new male Freshmen with the same race, ACT score, major, and health status; but Victor has a roommate with a video box whereas Nathan does not.
   (a) (4 points) What is the predicted difference between Nathan's study time and Victor's? Is this difference large or small in a real-world sense?
   (b) (4 points) What is the predicted difference between Nathan's first-semester GPA and Victor's? Is this difference large or small in a real-world sense?

Part 3 (30 points) - USE BLUE BOOK #3 True, False, Uncertain. Explain your answer. Formal derivations are not necessary if you are able to fully explain your answer. Just guessing will earn you zero points, even if your guess is correct:

1. (7 points) Consider the regression:

$$\ln(Wage)_i = \beta_0 + \beta_1 Female_i + \beta_2 College_i + \beta_3 (Female_i \times College_i) + e_i \quad (1)$$

   where $\ln(Wage)_i$ is log of wages, $Female_i$ is a dummy variable equal to one if female, and $College_i$ is a dummy variable equal to one if a person has a college diploma. You also decide to divide your observation in 4 groups: group 1- college educated women, group 2- college educated men, group 3- women without college diploma, and group 4- men without college diploma. You decide to run another regression:

$$\ln(Wage)_i = \gamma_0 + \gamma_1 Group1_i + \gamma_2 Group2_i + \gamma_3 Group3_i + e_i, \quad (2)$$

   where $GroupK$ is a dummy for group $K = 1, 2, 3$. Then the R^2 for regressions (1) and (2) will be the same.
2. (7 points)Suppose you collect data from a survey on wages. In addition you ask for information about marijuana use. The original question is "On how many separate occasions last month did you smoke marijuana?" Write an equation that would allow you to estimate the effect of marijuana use on wages. You should be able to make statements such as "Smoking marijuana 5 more times per month is estimated to change wage by x%"

3. (16 points) Assume that you have a regression $Y_i = \alpha + \beta X_i + e_i$ with heteroskedastic errors. Assume also that $E(e_i \mid X_i) = 0$, (X_i, Y_i) is an i.i.d. sample and $E[X_i^4] < \infty$, $E[e_i^4] < \infty$. Which of the following statements are true?

   i. Running regression in STATA with robust option gives an unbiased estimate of β .

   ii. Running regression in STATA without robust option gives an unbiased estimate of β .

   iii. The OLS estimator of β is inconsistent.

   iv. The OLS estimator of β is not asymptotically normal.