LPPA 7160: Research Methods and Data Analysis II

Midterm 1

On my honor as a student, I have neither given nor received aid on this exam. _____
Please initial on the line in your answer sheet.

Overall Instructions

- You may use a calculator or you can leave results in fractions.
- If you are turning in something that looks handwritten, make sure it is legible and organized. This is important!
- Do your best for each question, but make sure you are allocating your time across all questions.
- You will get points deducted if as part of your answer there are both right and wrong answers. We encourage you to write only what is needed to respond to the question without any additional fluff.
- Good Luck!

A Conceptual questions: Who is well?

1. HoosWell is a real voluntary program at UVA aimed at improving the wellness of their employees. This is a costly program, so understanding whether it works or not is important since resources could be allocated somewhere else. The HoosWell team will measure success of the program based on two metrics: annual medical expenditures and self-reported health. Annual medical expenditure is the total dollar amount spent on health care[1]. Self-reported health is a scale from 0 to 10, where individuals rank their health from 0 (poor health) to 10 (excellent health). The team has information from before, during and after the program was implemented. The program was implemented in 2017. A goal of the program would be for employees to have lower medical expenditures and better self-reported health.

Table 1: HoosWell First Results

		Before the Program Started		Outcomes After the Program	
	Sample Size (N)	Annual Medical Expenditures	Self-Reported Health	Annual Medical Expenditures	Self-Reported Health
Enrolled in HoosWell	2,208	$433	7.7	$580	5.8
Not Enrolled in HoosWell	1,031	$548	5.6	$569	5.9

[1]The team can observe this since all HoosWell participants are part of UVA's insurance plan

(a) UVA wants to save money, so they ideally want a program that reduces medical expenditures, and the idea is that this means employees have better health. Are medical expenditures a good measure of health? Provide a counterexample. (2pt)

We can spend more on medical care, which would imply better health because we are receiving more care. Less medical expenses could mean less healthcare, which could imply worse health if people are forgoing healthcare.

(b) The first empirical exercise the team makes is a simple comparison of health outcomes before and after the program for the people enrolled in HoosWell. Based on this comparison, what would we conclude about the effectiveness of the program for the people enrolled in Hoos Well? Write down the non-technical conclusion, instead of interpreting estimates. (1pt)

The overall conclusion is that the program is not effective. The program is making people have worse outcomes. People are spending more money on health, and their self-reported health decreases.

(c) Imagine the program does have beneficial effects (lower medical expenditures and better self-reported health), what is a hypothesis that can explain the findings in 1b? (3pts)

Since this is a before-and-after comparison, a good way of thinking about a potential bias issue is how the passage of time could be affecting this comparison. A variable that captures the passage of time is age. In other words, in the counterfactual world, had people not received the program, their health would have been much worse than what we observe. Even though we are seeing a deterioration of their health, you can think that the decline without the program would have been even worse. Notice that an argument such as "people who join the program are older than people who do not" is not valid for this exercise. That reasoning is not relevant because the comparison in 1b is not about people in the program vs. not in the program.

(d) The second empirical exercise the team does is to run the following regressions:

$$AnnualMedicalExpenditures_i = \alpha_0 + \beta_1 HoosWell_i + \varepsilon_i \quad (1)$$

$$SelfReportedHealth = \gamma_0 + \delta_1 HoosWell_i + \nu_i \quad (2)$$

Where HoosWell is a binary variable indicating if an individual is in the program or not. This regression is only using information from after the program's conclusion. What is the value of α_0, β_1, γ_0, *and* δ_1? Show your work for full credit (6pts).

$$E(AnnualMedicaidExpenditures|HoosWell) = \alpha_0 + \beta_1 HoosWell_i$$

$$E(AnnualMedicaidExpenditures|HoosWell = 0) = \alpha_0 = 569$$

$$E(AnnualMedicaidExpenditures|HoosWell=1)=\alpha_0+\beta_1=580,\ since\ \alpha_0=569, then\ \beta_1=11$$

$$E(SelfReportedHealth|HoosWell)=\gamma_0+\delta_1 HoosWell_i$$

$$E(SelfReportedHealth|HoosWell=0)=\gamma_0=5.9$$

$$E(SelfReportedHealth|HoosWell=1)=\gamma_0+\delta_1=5.8,\ since\ \gamma_0=5.9,\ then\ \delta_1=-0.1$$

$$\text{In short: } \alpha_0=569,\ \beta_1=11,\ \gamma_0=5.9,\ and\ \delta_1=-0.1$$

(e) Write down that comparison done in the previous exercise in Conditional Expectation Language:
$E[Self-reportedHealth|HoosWell=1]-E[Self-reportedHealth|HoosWell=0]$

(f) Use your results in the previous question 1d and the information from table 1 (only using columns from after the program) to assess the magnitudes of the effect of HoosWell on both primary outcomes. (2.5pts)

The effect of HoosWell on annual medical expenditures is $11, and the effect had people not been in HoosWell is 569, this represents a 2% increase in medical expenditures. Meanwhile, the change in self-reported health is -0.1 over the baseline of not being in HoosWell is 1.7% decrease. Both seem relatively small.

2. The team decides to further investigate the effectiveness of the program and runs three different models using annual medical expenditures, and one model to understand HoosWell participation. The results are in Table 2. The standard errors are in parentheses.

Table 2: Choose your model

	2018 Annual Medical Expenditures			HoosWell
	(1)	(2)	(3)	(4)
HoosWell	11	-61.4***	0.4	
		(21.4)	(1.4)	
Female		12.2**	9.2***	0.06
		(5.3)	(3.3)	(0.005)
Black		9.7***	6.7***	
		(1.3)	(2.3)	
Hispanic		-4.4***	-4.9**	
		(2.2)	(5.5)	
Other race		1.5	3.5	
		(0.7)	(0.7)	
Age (in years)		3.2***	3.3***	0.03
		(0.8)	(0.9)	(0.008)
Age^2		-0.04	-0.04	
		(0.003)	(0.003)	
Married		15.1***	15.1***	0.015
		(0.01)	(0.01)	(0.001)
Married × Female		5.40**	5.40**	0.005
		(0.03)	(0.03)	(0.003)
Number of Dr. Visits (2018)			53.2**	
			(15.3)	
Annual Salary (in thousands) (all years)			-20.9*	-0.0029
			(9.8)	(0.0008)
Constant	569	403***	350***	
		(152.2)	(86.3)	
Observations	3,239	3,239	3,239	1,049
R^2	0.04	0.34	0.07	0.16

Notes: Reference group for race is white, and reference group for females is males.

(a) Using the results from model (2), What are the average annual expenditures for a 30-year-old married white woman who is not in the program? Show your work (2 pts)

In order to get this average, we would need to add: 403+3.2×30+-0.04(30*30)+12.2+15.1+5.4=$495.7
Her average annual medical expenditure is $495.7 or approximately $496

(b) Using the results from model (2), What is the marginal effect of being married on annual medical expenditures? Show your work, and provide a complete interpretation with numbers (Hint: think about all potential cases). (3 pts)

The marginal effect of being married can be represented by a derivative: $\frac{\delta Exp}{Married} = \beta_{married} + \beta_{marriedXFemale} Female$, which in our case is $15.1 + 5.40 Female$. This means that for non-female is 15.1 and for females is 20.5. Marriage is associated with higher medical expenditures for women than men.

(c) Employees' ages at UVA range from 18 to 76. Using the results from model (3), provide a full interpretation of the effect of age on annual medical expenditures. Show your work (3 pts)

We take the derivative with respect to age: $\frac{\delta Exp}{\delta Age} = \beta_{age} + 2\beta_{age^2} Age = 3.3 + 2(-0.04) Age$=3.3-0.08Age. Now from here we'll find the local max or min, and then understand if its a max or min. First step, finding the local max or min:

$$\begin{gathered} 3.3 - 0.08Age = 0 \\ -0.08Age = -3.3 \\ Age = \frac{-3.3}{-0.08} \\ Age = 41.25 \\ \frac{\delta Exp}{\delta^2 Age} = -0.08 < 0, \textit{ so it is a local max} \end{gathered}$$

With all of this information we can formulate an answer: As age increases form ranges 18 to 41, then annual medical expenditures increase, however as age increases from 41 to 76 then medical expenditures decrease.

(d) Comparing the model from column 1 vs column 2 vs column 3: Which estimate on the effect of the HoosWell program would you trust more? Provide a justification of why your preferred model is better than the other two, separately? (i.e. Say you prefer model 1 overall, then your answer should be structured as: Model 1 is better than 2 because ...., and Model 1 is better than 3 because...) (4 pts)

We would trust the model from column (2) more than model (1) because it accounts for essential confounders like age, race, and gender. These characteristics can affect the likelihood of joining the program and affect health outcomes. We would trust the model from column (2) more because the model in column 3 includes covariates that could be potential outcomes. Therefore we could be mitigating or biasing the program's effect by including these covariates. Once we include salary and number of doctor visits, we see no effect of the program.

(e) Among smokers at baseline, about 31 percent enrolled in HoosWell, and among non-smokers 35 percent enrolled in HoosWell. Since the models above do not include a covariate for smoking practices before entering the program, we would want to know if by including a covariate related to smoking the effect of the program would be more positive or more negative. Assuming that being a smoker increases annual medical expenditures, would the estimated effect of the HoosWell program be more positive or more negative once controlling for smoking status? Show your work (3pts)

To answer this question, we need to know the sign of two components: $Corr(smoker\ status, Y)$ and $Corr(smoker\ status, HoosWell)$.

The first component we are told is positive since smokers have higher annual expenditures than non-smokers. The second component is negative. We can tell from the statistic in the first sentence that smokers are less likely to enroll in the program, which means a negative correlation.

The sign of these two correlations will give us the sign of the bias. Since $Corr(smoker\ status, Y) > 0 \times$ Corr$(smoker\ status, HoosWell) < 0$, this means the bias is negative $(+ \times - = -)$. Now we have the following set up $\beta_{estimate} = \beta_{true} + (negative\ number)$. After controlling for smoking status, we would be interpreting the β on HoosWell as the β_{true}, since that β comes from the model in which we did control for smoking status. This beta is

more positive. Why is it more positive? Well because $\beta_{estimate} + Bias = \beta_{true}$. In other words, $\beta_{estimate}$ is more negative (since the bias is negative) than β_{true}, which means that β_{true} would be more positive once controlling for smoking status.

3. The team has already collected a number of important variables but they have money to collect *one more variable*. There are many contenders that your peers are considering. Out of the variables in consideration, which one would you pick? Explain why you pick that variable over each of the other ones. (3 pts)

   - Variables already included in database: Variables in Table 2, and the following 2016 binary variables: academic staff or faculty (binary), heavy drinker, whether the person has a chronic condition, whether the person has high blood pressure.
   - **Variables in consideration**: (1) school in which the individual works (i.e. Batten, Arts and Science, etc), (2) years going to the gym before 2017, (3) political affiliation.

     The key of this question is to understand the main criteria for selecting variables that we should control for. In short, we should include variables that we think are correlated with treatment or in other words variables that best explain selection into treatment. In this case, we want to select variables that explain who would select themselves into a program like HoosWell.

     The question can be rephrased as: Among the variables in consideration, which of them is more likely to explain selection into treatment (conditional on the ones we already have)?

     Let's go one by one: whether a person is in a particular school may or may not affect the likelihood of someone enrolling. One tentative argument is "people in certain schools are more likely to enroll because of income differences," but those differences are already captured in the salary variable. Therefore that is not a good argument. Another argument could be "people in schools of public health may be more likely to enroll in HoosWell." That's not terrible but also seems somewhat weak. What about the other schools?

     The other variable is political affiliation. It is hard to think how this would be correlated, and even though one could come up with stories, it also seems potentially weak.

     Finally, whether a person goes to the gym or not, going to the gym carries more information about the person caring about their health than any of the other two variables. Someone who cares about their health or does physical activity would be more likely to enroll in the program. Therefore out of all of the variables in consideration, "going to the gym" is the variable that would mainly explain selection into treatment. Notice that since this variable is measure before the intervention we are not worried about it being a potential outcome.

     Notice that the other variables could have some information about selection into treatment, but among the three, one carries more information than the other.

B Multiple Choice

- If true, you don't have to explain. If false, you do have to provide reasoning. Each question is worth 2 points. Reasoning on false should be accurate in order to get full two points.

1. Using Table 2, What's the average change in "HoosWell" participation corresponding to an increase in employee annual salary of $10,000?

   (a) 0.29 percent decrease in HoosWell participation

   (b) 2.9 percentage points increase in HoosWell participation

   (c) 0.0029 percent decrease in HoosWell participation

   (d) 2.9 percent decrease in HoosWell participation

   (e) none of the above

2. An econometrician performs the following regression, where:

   lsalary = log salary of major league baseball player

   games = career games played

   runs = career runs scored

```
      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  2,   350) =  142.54
       Model |  220.933197     2  110.466598            Prob > F      =  0.0000
    Residual |  271.242338   350   .77497811           R-squared     =  0.4489
-------------+------------------------------           Adj R-squared =  0.4457
       Total |  492.175535   352  1.39822595           Root MSE      =  .88033

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       games |   .0005747   .0002504     2.29   0.022     .0000822    .0010672
        runs |   .0016433   .0004482     3.67   0.000     .0007619    .0025247
       _cons |   12.64234   .0766381   164.96   0.000     12.49161    12.79307
------------------------------------------------------------------------------
```

   All else equal, an increase of 1000 runs scored in a career has what effect on salary? (2pts)

   (a) Increases salary by 1.64%.

   (b) Increases salary by 0.016%.

   (c) Increases salary by $16,000.

   (d) Has no statistical significant effect on salary.

   (e) Increases salary by 164%.

3. If the exogeneity assumption is true ($Corr(HoosWell, \varepsilon) = 0$) in model (2) from Table 2, then we know that the OLS estimate of the effect of HoosWell on annual medical expenditures will be statistically significant:

   (a) True

(b) False, Explain: The exogeneity assumption only affects the bias (whether the estimate is centered on the true value), but has no effect on the standard errors of the coefficients. An unbiased estimate can still be statistically insignificant if the standard error is large.

4. In model (2) from Table 2, including more variables that are correlated with annual medical expenditures will always reduce the bias on the HoosWell coefficient no matter what:

   (a) True

   (b) False, Explain: We could be adding more bias if the variables are potential outcomes (bad controls). For example, as we saw in model (3), adding Number of Dr. Visits — which could itself be affected by the program — can bias the HoosWell estimate.

5. Suppose every employee at UVA receives the same annual salary of $55,000 (i.e., salary is a constant c). If $HoosWell$ is a random variable, then:

   (a) $Cov(HoosWell, c) \neq 0$

   (b) $Corr(HoosWell, c) \neq 0$

   (c) (a) and (b) are both true

   (d) $Var(HoosWell) \times Var(c) = 0$

   (e) None of the above

6. The sample sizes in Table 1 are different between the treatment and control groups. This is a concern in terms of the causal interpretation of the effect of the program.

   (a) True

   (b) False, Explain: Differences in sample size across T and C groups are not an issue for causal interpretation, we only care about the core assumptions not sample size for unbiasedness.

7. Interested in the relationship between salary and medical expenditures among UVA employees, the HoosWell team estimates the following regression using data from before the program:

$$\ln(AnnualMedicalExpenditures_i) = \beta_0 + \beta_1 \ln(AnnualSalary_i) + \varepsilon_i$$

   The estimated coefficient is $\hat{\beta}_1 = 0.72$. The average annual medical expenditure in the sample is $540. If an employee's salary increases by 10%, what is the *approximate* change in their annual medical expenditures? (2pts)

   (a) $3.89 increase (a 0.72% increase in medical expenditures)

   (b) $38.88 increase (a 7.2% increase in medical expenditures)
   Correct. In a log-log model, a 10% increase in X leads to approximately $0.72 \times 10 = 7.2\%$ increase in Y, and $0.072 \times \$540 = \$38.88$. The exact change is $(1.10^{0.72} - 1) \times \$540 = (1.0693 - 1) \times \$540 \approx \$37.42$.

   (c) $388.80 increase (a 72% increase in medical expenditures)

   (d) A 10% increase in salary leads to a 7.2% decrease in medical expenditures