Formulating Research Designs and Replicating Results

Jerry Hong

September 11, 2024

Abstract

This assignment demonstrates the process of analyzing and developing research designs. Part I focuses on formulating research designs by: identifying the dependent and independent variables, observation and periodical variables, and null and alternative hypotheses. Part II focuses on basic data and regression analysis of health and labor economics data. We will to the best of our ability replicate regression figures from each of these studies.

1 Developing the Research Design

Below are a set of research questions across various domains of interest.Within each question, we will assess the independent and dependent variables, the unit(s) of observation and periodicity, and a hypothesis regarding the relationship of the primary variables of interest.

1.1 Class Size and Test Scores

One research question within the scope of labor and human capital is *How does class size affect test scores?* Intuitively, the primary independent variable of interest is the class size. Does smaller or larger class sizes impact the students' performance in tests? Other supplementary variables of interest include a variety of socioeconomic factors like: race, age, sex, and any disabilities to look out for. However, there can be potential confounding should these variables correlate with the primary independent of interest being class size. Of course, the dependent variable of interest would be the students' test scores. This will indicate whether the treatment of different class sizes has an effect of their test scores. The unit of observation should be based on the class size (how many students in the class) and the accompanied test scores. We'll based this on a quarterly or semester basis, depending on the schedule in which the tests are administered. The null hypothesis would be that there is no statistical difference between class size and test scores. The alternative hypothesis would suggest that there is a statistical difference between class size and test scores.

1.2 Monetary Incentive and HIV Results

A research question within health development and experimental economics is *How do monetary incentives affect the decision to get HIV test results?* The primary independent variable is the use of monetary incentives. However, like the previous question, there can be other socioeconomic factors mainly race, sex, age, location, and previous health conditions. It is important to note as well of possible confounding between these variables and our primary variable being the monetary incentive. We could measure this on a monthly basis and seeing if their willingness to take the tests have changed as a result of receiving the incentive. The null hypothesis is that the monetary does not affect the decision to get HIV test results. The alternative is that the monetary incentive does affect the decision to get HIV test results.

1.3 Job Applicant Names and Callback Success

A research question within discrimination and labor economics is *How does the sound of a job applicant's name (with respect to race/ethnicity) affect the applicant's success in receiving a callback?* The primary independent variable is the name of the job applicant. Albeit, these are simulated resumes and are thus not real. Ultimately, this can be rather subjective given that if we are isolating just

the name, the perception of it can vary depending on the job recruiters. We may include race as another primary variable of interest, but it can introduce confounding like other socioeconomic factors that may influence the decision for callbacks because now it would be based on those factors rather than on the name itself. Nonetheless, given we are applying the sound of the name with respect to race/ethnicity, it may not pose as much of an influence as initially. The dependent variable of interest is the candidate's callback rates. How often are they able to get callbacks? This will be our measurement and assessing any possibilities of discrimination. The unit of observation would primarily be the candidates and can be grouped by race/ethnicity to further explore whether race has to do with the callback rates. This could be measured on monthly basis, so that it gives ample time to receive information regarding callbacks.

2 Replicating Regression Data

In the following two studies, we will be providing some basic summary statistics of variables of interest. In addition, we will to the best of our ability replicate the regression data conducted by the respective authors.

2.1 The Demand for, and Impact of, Learning HIV Status (2008)

The premise of this study conducted by Rebecca Thornton is to explore the effect of implementing a monetary incentive program in rural Malawi to learn their HIV results after being tested. Table 1 shows some summary statistics of the following covariates: Incentive type, Incentive amount, Male, HIV results, Age, and Education. We measure the count, mean, standard deviation, minimum and maximum values.

Table 1: Summary Statistics of Incentive and HIV Test Data

Variable	Count	Mean	Std. Dev	Min	Max
Visits	2894	0.6966	0.4598	0.0	1.0
Any Incentive	2901	0.7659	0.4235	0.0	1.0
Amount of Incentive	2901	104.4881	95.3215	0.0	300.0
Male	4820	0.4689	0.4991	0.0	1.0
HIV	2894	0.0591	0.2555	-1.0	1.0
Age	4379	33.6517	13.1629	11.0	84.0
Education	3154	3.4949	3.6840	0.0	12.0

Note: There are 4,820 observations and 44 variables in total.

Table 2 is a close replication of one of the study's regression data. This is based off of Table 4 from [Tho08]. We will only refer to the first three columns of the original table. In other words, we are replicating to the best of our ability the first three models the author created using various combinations of the covariates. We will primarily use the following regressors: Incentive type, Incentive amount, Incentive amount ^{2}, HIV, Male, Rumphi, and Balaka, which are two districts in Malawi. Overall, the results that we conducted are relatively similar to the ones conducted by Thornton. Generally we are within margin of error as little as 0.001 and up to 0.05. Essentially, we maintained the same regressors Thornton used to model the relationship between the incentives and the visits for HIV results. One thing worth noting is that the author has 17 less observations than the number we used. This may attribute to the slight variation in the test statistics. Nonetheless, the values and magnitude are relatively on pace with the author's findings.

Table 2: Effect of Monetary Incentives and Learning HIV Test Results

	Dependent variable: Visits for HIV Results		
	(1)	(2)	(3)
Any incentive	0.431***	0.308***	0.214***
	(0.020)	(0.024)	(0.031)
Amount of incentive		0.001***	0.003***
		(0.000)	(0.000)
Amount of incentive^2			-0.000***
			(0.000)
HIV	-0.041	-0.041	-0.040
	(0.031)	(0.031)	(0.031)
Male	-0.014	-0.017	-0.018
	(0.016)	(0.016)	(0.016)
Age	0.002***	0.002***	0.002***
	(0.001)	(0.001)	(0.001)
Rumphi	-0.135***	-0.153***	-0.159***
	(0.020)	(0.020)	(0.020)
Balaka	-0.117***	-0.122***	-0.120***
	(0.020)	(0.020)	(0.019)
Intercept	0.387***	0.396***	0.400***
	(0.029)	(0.029)	(0.029)
R^2	0.180	0.201	0.207
Adjusted R^2	0.178	0.199	0.205
Residual Std. Error	0.419 (df = 2822)	0.414 (df = 2821)	0.412 (df = 2820)
F Statistic	103.339*** (df = 6; 2822)	101.313*** (df = 7; 2821)	92.110*** (df = 8; 2820)

Note: Number of observations shared are 2,829 *p<0.1; **p<0.05; ***p<0.01
Standard errors in parentheses.

2.2 A Field Experiment on Labor Market Discrimination (2004)

The study conducted by Marianne Bertrand and Sendhil Mullainathan [MB04] looks into whether the job applicant's name attributed to race has an effect on the rate of callbacks. Table 3 shows some basic summary statistics of the job demographics and the callback rates.

Table 3: Summary Statistics of Job Demographics and Callback Rates

Variable	Count	Mean	Std. Dev	Min	Max
Callback? (Y=1)	4870	0.080	0.272	0.0	1.0
High-quality resume? (Y=1)	4870	0.502	0.500	0.0	1.0
Race	4870	0.500	0.500	0.0	1.0
Sex	4870	0.230	0.421	0.0	1.0
Number of Jobs	4870	3.661	1.219	1.0	7.0
Years of Exp	4870	7.843	5.045	1.0	44.0

Note: There are 4,870 observations and 67 variables in total.

The following table 4 shows three regression models divided into the following three categories: All resumes, White name resumes, and African-American name resumes. The dependent variables of interest are years of experience and its squared value, Volunteer and Military experience, Email, Employment holes, Work in school, Honors, and any Computer or Special skills. For the sake of simplicity, we will only focus on the regressors presented on the original table and will not include hidden dummy variables as the authors noted. The rows before the Observations is the null hypothesis that the resume characteristics are zero alongside its p-value. This is done through a chi-squared test.

Table 4: Effect of Resume Characteristics on Callbacks

	Dependent variable: callback rates		
	All resumes	White name resumes	African-American name resumes
Years experience (*10)	-0.000***	-0.000**	-0.000**
	(0.000)	(0.000)	(0.000)
Years experience ^{2} (*100)	0.000**	0.000**	0.000
	(0.000)	(0.000)	(0.000)
Volunteer	0.005	-0.004	0.015
	(0.011)	(0.017)	(0.014)
Military	-0.008	0.011	-0.023
	(0.014)	(0.023)	(0.018)
Email	0.014	0.028	-0.001
	(0.011)	(0.017)	(0.014)
Employment holes	0.027***	0.038***	0.016
	(0.009)	(0.014)	(0.012)
Work in school	0.012	0.020	0.001
	(0.010)	(0.015)	(0.012)
Honors	0.068***	0.074***	0.056**
	(0.018)	(0.027)	(0.023)
Computer Skills	-0.027**	-0.039**	-0.011
	(0.011)	(0.016)	(0.014)
Special Skills	0.052***	0.062***	0.041***
	(0.009)	(0.013)	(0.011)
Intercept	0.065***	0.073***	0.054***
	(0.012)	(0.018)	(0.016)
H0: Resume characteristic effects are zero	[[10.940]]	[[7.221]]	[[4.193]]
p-value	1.187e-18	2.514e-11	8.697e-06
Observations	4870	2435	2435
R^2	0.022	0.029	0.017
Adjusted R^2	0.020	0.025	0.013
Residual Std. Error	0.269 (df = 4859)	0.292 (df = 2424)	0.244 (df = 2424)
F Statistic	10.940*** (df = 10; 4859)	7.221*** (df = 10; 2424)	4.193*** (df = 10; 2424)

Note: *p<0.1; **p<0.05; ***p<0.01

References

[MB04] Sendhil Mullainathan Marianne Bertrand. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[Tho08] Rebecca Thornton. The demand for, and impact of, learning hiv status. *American Economic Review*, 98(5):1829–1863, 2008.

Disclaimer

The figures and models presented in this document have been generated with the assistance of artificial intelligence tools. The findings are reviewed and verified by the author(s) to ensure accuracy and clarity.