# Problem set 1

PPHA 31102 Statistics for Data Analysis II: Regressions

Jay Ballesteros

January 16, 2026

## The Oregon Health Insurance Experiment (OHIE) (15 points)

The Oregon Health Insurance Experiment was a randomized experiment run in the 2008 that expanded Medicaid to low-income, uninsured, able-bodied adults aged 19-64 in Oregon through a random lottery drawing. Eligible individuals interested in receiving Medicaid signed up for a lottery; winning the lottery ("treated") provided an opportunity for the individual and up to two additional eligible household members to sign up for Medicaid. 12,229 individuals participated in a long-term study examining outcomes 2 years later.

In this question, you will work with data from the OHIE Experiment.[1] You should use R to answer this question and append your code and output (or Rmarkdown output) as PDF to your submission. Failure to submit code and output will result in a 0 on this question.

Refer to the datafile **OHIE.csv** available on Canvas→Modules→Problem Set 1. This file contains 12,229 rows, where each row is a survey response from a person in the experiment. The dataset on Canvas has the following variables:

| Variable | Description |
|---|---|
| `person_id` | Unique anonymous person identifier |
| `treated` | 1=won lottery to apply for Medicaid |
| `numhh_list` | Number of eligible household members |
| `female` | 1=female sex, 0=male sex |
| `age` | Age in years |
| `race_white` | 1=self-reported race non-Hispanic White, 0=all others |
| `hs_degree` | 1=HS diploma or GED, 0=all others |
| `college_degree` | 1=college degree, 0=all others |
| `health_baseline` | 1=Diagnosis of any major health condition, pre-lottery |

[1] These data come from a study by former Harris Dean and current UChicago provost Katherine Baicker.

| Variable | Description |
|---|---|
| `ever_medicaid` | Ever Enrolled in Medicaid coverage since lottery |
| `visit_dr` | Number of doctor office visits since lottery |
| `visit_er` | Number of emergency room visits since lottery |
| `out_of_pocket_spend` | Amount of out-of-pocket spending ($) since lottery |
| `health_score` | "Framingham risk score"* – summary measure of current health status at endline |
| `happy` | 1=reported happy or pretty happy at endline |

* Note: The Framingham risk score is a function of age, total cholesterol and HDL cholesterol levels, measured blood pressure and use or nonuse of medication for high blood pressure, current smoking status, and blood sugar levels.

Please note that some outcomes have "NA" values if respondent data were unavailable. When analyzing an outcome with an "NA" value, simply exclude those rows from your analysis for that particular outcome. This also means that the sample sizes will be smaller for outcomes with "NA" values.

## 1.

**Fill in the following balance table. In Column (4), calculate the p-value using a two-sample t-test assuming equal variance. Recall, the test-statistic for this test is given by:**[2]

$$t \;=\; \frac{\bar{Y_T} - \bar{Y_C}}{\sqrt{\left(\frac{(N_T-1)s_T^2+(N_C-1)s_C^2}{N_T+N_C-2}\right)\left(\frac{1}{N_C}+\frac{1}{N_T}\right)}}$$

**where $t$ is distributed as Student's $t$ with $N_T + N_C - 2$ degrees of freedom. In terms of notation, $\bar{Y_T}$ is the sample mean of the treated group, $\bar{Y_C}$ is the sample mean of the control group, $N_T$ is the sample size of the treated group, $N_C$ is the sample size of the control group, $s_T^2$ is the sample variance of the treated group, $s_C^2$ is the sample variance of the control group.**

**You can code the t-test up manually yourself or use R's `t.test()` command with the option `var.equal=TRUE`. (3 points)**

[2] (Welch's two-sample t-test could also be used here, which does not assume equal variance. But there's a reason we want you to do it this way.)

| Baseline characteristic | (1) Control Mean | (2) Treated Mean | (3) Difference (2)-(1) | (4) p-value |
|---|---:|---:|---:|---:|
| `numhh_list` | 1.195995 | 1.288555 | 0.093 | 0.00 |
| `female` | 0.5686409 | 0.5686409 | -0.006 | 0.51 |
| `age` | 40.60606 | 40.98638 | 0.380 | 0.07 |
| `race_white` | 0.6898387 | 0.6868322 | -0.003 | 0.72 |
| `hs_degree` | 0.6823006 | 0.6795052 | -0.003 | 0.74 |
| `college_degree` | 0.1116056 | 0.1153906 | 0.004 | 0.51 |
| `health_baseline` | 0.2716535 | 0.2703930 | -0.001 | 0.88 |

## 2.

**Discuss your findings. Do these baseline characteristics appear balanced? (2 points)**

In 6 out of the 7 variables, we can interpret that the baseline characteristics appear balanced (more precisely, we don't have evidence of differences). This is because of $p-values > 0.05$, failing to reject the null hypothesis. In the case of `numhh_list`, we reject the null hypothesis of equal means ($p-value$ close to 0).

## 3.

**Calculate the treatment effect of winning the Medicaid lottery and the statistical significance for each of the outcomes, filling out the table below. Discuss your findings—- What conclusions can we draw about the effectiveness of the Medicaid lottery program? (3 points)**

| Endline characteristic | (1) Control Mean | (2) Treated Mean | (3) Difference (2)-(1) | (4) p-value |
|---|---:|---:|---:|---:|
| `visit_dr` | 5.745862 | 6.141554 | 0.396 | 0.067 |
| `visit_er` | 1.0132302 | 0.9773407 | -0.036 | 0.312 |
| `out_of_pocket_spend` | 563.4565 | 490.4123 | -73.0442 | 0.0008 |
| `health_score` | 0.08159156 | 0.08003334 | -0.0016 | 0.310 |
| `happy` | 0.7456261 | 0.7579987 | 0.0124 | 0.114 |

Given the results, we only reject the null in `out_of_pocket_spend`, determining that there's statistical significant evidence of a difference that individuals in the treatment group spent, on average, 73.04 less than people in the control group. `visit_dr` was a little above 0.05. If we gave some flexibility to our p-value (maybe 0.07), we could also reject the null hypothesis and assess that people in the treatment group visited the doctor 0.396 times more than people in the control group.

## 4.

**Let's try redoing Part (3) using a simple linear regression.**

**For each of the outcomes in Part (3) above, run a simple linear regression of the outcome on a treatment indicator:**

$$Y_i \ = \ \beta_0 + \beta_1 Treated_i + u_i$$

**Compare to your results in Part (3) above.**

**For this question, we recommend using the `lm()` regression function in R and referring to the output in your answer. (3 points)**

Linear regressions for all indicators returned as coefficients the same results that we got in the difference in means analysis from the question above. Therefore, conclusions are the same: we reject the null in the `out_of_pocket_spend`.

## 5.

**In what ways might features of this experiment affect the external validity of the results, say, to thinking about expanding Medicaid to the entire U.S. population of low-income, able-bodied adults? (2 pt)**

It may be affected (external validity) because the experiment was in Oregon, so assignment might differ in characteristics of the population.

## 6.

**Suppose instead of running an experiment, you could get health data on *everyone* in the low-income, able-bodied adult population in 2008. You compare the health outcomes of those with health insurance to those without health insurance. Do you expect you will find similar results to Part (3)? Why or not? (2 pts)**

Not necessarily. Because they might not be comparable, starting to with the idea that people with insurance did not enrolled out of a lottery, but because of willingness to be insured. So this people might have different risk aversion, even different income, health and access to information regarding prevention.