Hamed Nikookalam

PS5

Table of Contents

Part 1: Simulation

A)

I specified the true data-generating process as

$$Y_i = 1 + 0.5X_i + 1.0C_i + \varepsilon_i,$$

where $X_i$is the treatment variable, $C_i$is a confounder, and $\varepsilon_i \sim N(0,1)$. In the simulation, the confounder $C_i$is drawn from a standard normal distribution, and the treatment is generated as $X_i = 0.7C_i + u_i$, with $u_i \sim N(0,1)$, so that $X_i$and $C_i$are correlated. Using a sample size of $N = 500$, I generated one dataset, fit the linear model Y ~ X + C in R with lm (), and reported the regression summary.

To demonstrate the central limit theorem for the treatment coefficient, I repeated this data-generating process 1,000 times. In each simulation I generated new values of C_i, X_i, and Y_i, re-estimated the same regression model, and stored the estimated coefficient on X. The mean of these 1,000 estimated coefficients is approximately 0.5 (very close to the true value), and their standard deviation is about 0.045. The histogram of these 1,000 estimates is bell-shaped and centered near 0.5, showing that the sampling distribution of the estimator $\hat{\beta}_X$is approximately normal. This provides evidence that the coefficient on the treatment variable follows the central limit theorem under the true model.

B)

For the original simulated dataset from part (a), I computed a bootstrapped standard error for the treatment coefficient. I resampled the rows of the dataset with replacement 1,000 times. For each bootstrap sample I re-estimated the model $Y \sim X + C$and recorded the estimated coefficient on X. I then took the standard deviation of these 1,000 bootstrap estimates as the bootstrapped standard error. This yields a bootstrapped standard error for the treatment coefficient of approximately 0.04583152.

C)

When I fit the correct model Y ~ X + C in part (a), the sampling distribution of the treatment coefficient was approximately normal and centered near the true value 0.5. In part (c), I estimated the misspecified model Y ~ X that omits the confounder. The sampling distribution of the treatment coefficient is still approximately normal, but it is now centered around about 0.97, far from the true effect. This shows that even when the sampling distribution of an estimator is normal, omitting a confounder can produce a biased coefficient. Statistical tests (t-tests, p-values, confidence intervals) based on biased sampling distribution can therefore be very misleading.

Part 2: Data Analysis

A)

I simulated a dataset with 200 observations. The variable treat is a binary indicator (0 = control, 1 = treated) and the outcome Y is continuous. I am interested in whether the mean outcome differs between treated and control units: H0: μ1=μ0   vs. HA: μ1≠ μ0.

Because the population variance is unknown and the outcome is continuous, I use a two-sample t-test with equal variances and a two-sided alternative at $\alpha = 0.05$.

The sample mean of Yis about 9.97 in the control group and 12.24 in the treated group, so the observed difference in means (treated − control) is roughly 2.26. The t-test reports a t-statistic of about −5.06 and an extremely small p-value (effectively zero and far below 0.001), with a 95% confidence interval for (control − treated) of [−3.14, −1.38]. This corresponds to a 95% confidence interval for (treated − control) of about [1.38, 3.14]. Statistically, because the p-value is extremely small and the confidence interval does not include 0, I reject the null hypothesis that the two-group means are equal. Substantively, this implies that the treatment is associated with an increase of roughly 2 to 3 units in the outcome variable Yrelative to the control group.

B)

In the linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, the coefficient on treat is about 2.26, which means that, on average, units in the treated group have outcome values roughly 2.26 points higher than units in the control group. The standard error of this coefficient is about 0.45, indicating that across repeated samples we would expect the estimated treatment effect to vary by around 0.45 units due to sampling noise. The resulting t-value is about 5.06, meaning the estimate is more than five standard errors away from zero, which is very large in absolute terms. The associated p-value is extremely small, the probability of observing such an extreme t-value if the true treatment effect were actually zero. Because this p-value is far smaller than a conventional 0.05 significance level, we reject the null hypothesis that the treatment has no effect and conclude that the treatment has a statistically significant and substantively meaningful positive impact on Y.