HW 5

Kevin Lin

2/13/2026

1

We are given $X_i \in \{0, 1\}$ (0 means control, 1 means treatment) and Y_i be the outcome for observation i. Let $n_0 = \sum_{i=1}^{n} \mathbf{1}\{X_i = 0\}$ and $n_1 = \sum_{i=1}^{n} \mathbf{1}\{X_i = 1\}$ be the number of control and treatment observations, respectively. Let $\bar{Y}_0$ and $\bar{Y}_1$ be the sample means of the control and treatment groups, respectively, and $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ the sample variance, defined as

$$\hat{\sigma}_g^2 = \frac{1}{n_g - 1} \sum_{i:X_i=g} (Y_i - \bar{Y}_g)^2, \quad g \in \{0, 1\}$$

The pooled variance two sample t-statistic is given by:

$$t = \frac{\bar{Y}_1 - \bar{Y}_0}{s_p \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

where the pooled standard deviation is:

$$s_p = \sqrt{\frac{(n_0 - 1)\hat{\sigma}_0^2 + (n_1 - 1)\hat{\sigma}_1^2}{n_0 + n_1 - 2}}$$

Lastly, we are given the linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$.

(a) We can show that the OLS slope estimate equals the difference between the sample mean of the two groups as follows:
OLS satisfies the normal equations:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0, \quad \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

We know the group means should be:

$$\bar{Y}_0 = \frac{1}{n_0} \sum_{i:X_i=0} Y_i, \quad \bar{Y}_1 = \frac{1}{n_1} \sum_{i:X_i=1} Y_i$$

Using the second normal equation and that X_i is binary ($X_i^2 = X_i$), we have:

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$

Again, because X_i is binary, any summation terms with $X_i = 0$ will vanish, so we can simplify the above to:

$$\begin{aligned}
\sum_{i:X_i=1} Y_i - \hat{\beta}_0 \sum_{i:X_i=1} 1 - \hat{\beta}_1 \sum_{i:X_i=1} 1 = 0 \\
\sum_{i:X_i=1} Y_i - \hat{\beta}_0 n_1 - \hat{\beta}_1 n_1 = 0 \\
\frac{1}{n_1} \sum_{i:X_i=1} Y_i = \hat{\beta}_0 + \hat{\beta}_1 \\
\bar{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1
\end{aligned}$$

Now we go back to the first equation and simplify by splitting into their respective groups:

$$\begin{aligned}
\sum_{i:X_i=0} (Y_i - \hat{\beta}_0) + \sum_{i:X_i=1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1) = 0 \\
\sum_{i:X_i=0} Y_i - n_0 \hat{\beta}_0 + \sum_{i:X_i=1} Y_i - n_1 \hat{\beta}_0 - n_1 \hat{\beta}_1 = 0 \\
\sum_{i:X_i=0} Y_i + \sum_{i:X_i=1} Y_i = (n_0 + n_1)\hat{\beta}_0 + n_1 \hat{\beta}_1
\end{aligned}$$

Substitute in the group means and the equation for $\bar{Y}_1$:

$$\begin{aligned}
n_0 \bar{Y}_0 + n_1 \bar{Y}_1 = (n_0 + n_1)\hat{\beta}_0 + n_1 \hat{\beta}_1 \\
n_0 \bar{Y}_0 + n_1 (\hat{\beta}_0 + \hat{\beta}_1) = (n_0 + n_1)\hat{\beta}_0 + n_1 \hat{\beta}_1 \\
n_0 \bar{Y}_0 = n_0 \hat{\beta}_0 \\
\bar{Y}_0 = \hat{\beta}_0
\end{aligned}$$

Finally, we can substitute $\hat{\beta}_0$ back into the equation for $\bar{Y}_1$ to get:

$$\begin{aligned}
\bar{Y}_1 = \bar{Y}_0 + \hat{\beta}_1 \\
\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0
\end{aligned}$$

which proves that the OLS slope estimate is the difference between the sample means of the two groups.

(b) We can show that the OLS intercept estimate equals the sample mean of the control group in the same steps we took above with the first normal equation:

$$\sum_{i:X_i=0}(Y_i - \hat{\beta}_0) + \sum_{i:X_i=1}(Y_i - \hat{\beta}_0 - \hat{\beta}_1) = 0$$
$$\sum_{i:X_i=0} Y_i - n_0\hat{\beta}_0 + \sum_{i:X_i=1} Y_i - n_1\hat{\beta}_0 - n_1\hat{\beta}_1 = 0$$
$$\sum_{i:X_i=0} Y_i + \sum_{i:X_i=1} Y_i = (n_0 + n_1)\hat{\beta}_0 + n_1\hat{\beta}_1$$

Substitute in the group means and the equation for $\bar{Y}_1$:

$$n_0\bar{Y}_0 + n_1\bar{Y}_1 = (n_0 + n_1)\hat{\beta}_0 + n_1\hat{\beta}_1$$
$$n_0\bar{Y}_0 + n_1(\hat{\beta}_0 + \hat{\beta}_1) = (n_0 + n_1)\hat{\beta}_0 + n_1\hat{\beta}_1$$
$$n_0\bar{Y}_0 = n_0\hat{\beta}_0$$
$$\bar{Y}_0 = \hat{\beta}_0$$

(c) We can show that the homoscedastic OLS estimator of the error standard deviation equals the pooled standard deviation as follows: We know $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, so:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

We can split the above summation into the control and treatment groups:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{i:X_i=0}(Y_i - \hat{\beta}_0)^2 + \sum_{i:X_i=1}(Y_i - \hat{\beta}_0 - \hat{\beta}_1)^2\right)$$

Substitute in the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{i:X_i=0}(Y_i - \bar{Y}_0)^2 + \sum_{i:X_i=1}(Y_i - \bar{Y}_1)^2\right)$$

Recall that the corresponding within-group sample variances are:

$$\hat{\sigma}_0^2 = \frac{1}{n_0 - 1}\sum_{i:X_i=0}(Y_i - \bar{Y}_0)^2, \quad \hat{\sigma}_1^2 = \frac{1}{n_1 - 1}\sum_{i:X_i=1}(Y_i - \bar{Y}_1)^2$$

We can rearrange the above to get:

$$\sum_{i:X_i=0}(Y_i - \bar{Y}_0)^2 = (n_0 - 1)\hat{\sigma}_0^2, \quad \sum_{i:X_i=1}(Y_i - \bar{Y}_1)^2 = (n_1 - 1)\hat{\sigma}_1^2$$

Substitute the above into the equation for $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n-2}\left((n_0-1)\hat{\sigma}_0^2 + (n_1-1)\hat{\sigma}_1^2\right)$$
$$\hat{\sigma} = \sqrt{\frac{(n_0-1)\hat{\sigma}_0^2 + (n_1-1)\hat{\sigma}_1^2}{n-2}} = s_p$$

which proves that the homoscedastic OLS estimator of the error standard deviation equals the pooled standard deviation.

(d) We can show that the OLS t-statistic for the slope estimate equals the two-sample t-statistic as follows:
We know our null hypothesis $H_0 : \beta_1 = 0$ and the OLS t-statistic is given by:

$$t_0 = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}$$

We need to prove that:

$$\frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}}$$

From the previous parts, we've already proved that $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$ and $\hat{\sigma} = s_p$. We just need to show that $\hat{se}(\hat{\beta}_1) = s_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}$. We know that the OLS standard error of the slope estimate is given by:

$$\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

Here, we know that $\bar{X} = \frac{n_1}{n_0+n_1}$ since X_i is binary. Then we have:

$$X_i = 0: \quad X_i - \bar{X} = 0 - \frac{n_1}{n_0+n_1} = -\frac{n_1}{n_0+n_1}$$
$$X_i = 1: \quad X_i - \bar{X} = 1 - \frac{n_1}{n_0+n_1} = \frac{n_0}{n_0+n_1}$$

Thus,

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i:X_i=0}\left(-\frac{n_1}{n_0+n_1}\right)^2 + \sum_{i:X_i=1}\left(\frac{n_0}{n_0+n_1}\right)^2 \\
&= n_0\left(\frac{n_1}{n_0+n_1}\right)^2 + n_1\left(\frac{n_0}{n_0+n_1}\right)^2 \\
&= \frac{n_0 n_1^2 + n_1 n_0^2}{(n_0+n_1)^2} \\
&= \frac{n_0 n_1 (n_0+n_1)}{(n_0+n_1)^2} \\
&= \frac{n_0 n_1}{n_0+n_1}
\end{aligned}$$

Plugging this back into our equation for $\hat{se}(\hat{\beta}_1)$, and applying our previous proof that $\hat{\sigma} = s_p$ we have:

$$\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\frac{n_0 n_1}{n_0+n_1}}} = \hat{\sigma}\sqrt{\frac{n_0+n_1}{n_0 n_1}} = s_p\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}$$

(e) In any finite sample, the OLS slope and intercept will always equal the difference in group means and the control group mean, respectively, regardless of the distribution of the error term ϵ. This comes algebraicly from the linear model, the fact that X is binary, and the OLS normal equations. However, the OLS estimator of the error standard deviation will only equal the pooled standard deviation if the error term ϵ is homoscedastic. This is because the pooled variance formula assumes a common population variance across the two groups, which the homoscedastic OLS estimator of the error standard deviation also assumes. If the error term (grouped variances) differ, then the pooled t-test is invalid and the homoscedastic OLS SE is wrong, thus breaking the equality. Likewise, the t-statistic for the slope estimate will only equal the two-sample t-statistic if the error term is homoscedastic, since the OLS t-statistic relies on the pooled SE. It also requires that the correct degrees of freedom (in this case, $n-2$) is used to calculate the OLS SE, which is the same as the degrees of freedom used to calculate the pooled SE (two degrees of freedom for β_0, β_1, and two degrees of freedom considering X_0, X_1 where $(n_0-1)+(n_1-1) = n-2$). If these assumptions are violated, then the OLS t-statistic will not equal the two-sample t-statistic.

2

See attached Jupyter notebook for code and plots.