International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Explaining Lord's Paradox in Introductory Statistical Theory Courses

Steven B. Kim[1]

[1] Department of Mathematics and Statistics, California State University, Monterey Bay, USA

Correspondence: Steven B. Kim, Department of Mathematics and Statistics, California State University, Monterey Bay, Seaside, CA, USA.

Received: March 27, 2018 Accepted: April 14, 2018 Online Published: April 27, 2018

doi:10.5539/ijsp.v7n4p1 URL: https://doi.org/10.5539/ijsp.v7n4p1

Abstract

When two groups are compared in a pre-post study, two different conclusions can be drawn between the two-sample t-test and the analysis of covariance (ANCOVA). It is known as Lord's Paradox, and it occurs because the parameter in the two-sample t-test and the parameter of interest in the ANCOVA model are not the same quantity. The difference between the two parameters can be explained by the covariance of linearly combined random variables which is an important topic in introductory statistical theory courses. Lord's paradox is frequently observed in practice, and it is very important for students (future researchers) to have clear understanding of the paradox. The objective of this article is to explain Lord's Paradox using the covariance of linearly combined random variables. The paradox is explained using three scenarios in the context of educational research. The first scenario is when the average baseline (pre-score) is greater in the treatment group than the control group, the second scenario is when the average baseline is lower in the treatment group than the control group, and the third scenario is when the average baseline is same between the two groups by randomization. This article is written at the level of introductory statistical theory courses for undergraduate and graduate statistics students to help understanding the difference between the parameter of interest in the two-sample t-test and the parameter of interest in the ANCOVA model.

Keywords: two-sample t-test, ANCOVA, covariance, linear combination of random variables, pre-post studies

1. Introduction

When two groups are compared in a pre-post study, Lord's Paradox can be observed between two researchers when a researcher compares the average change using the two-sample t-test and the other researcher compares the average post-measurement using the analysis of covariance or simply ANCOVA (Lord 1967; Lord 1969). The paradox has been studied in the context of health sciences, environmental sciences, and psychometrics (Holland & Rubin, 1983; Wainer & Brown, 2006; Glymour et al., 2005; Tu et al., 2008; Pearl, 2016). It is an interesting phenomenon which frequently occurs in practice, but it is not easy to quantify the exact difference between the parameter in the two-sample t-test and the parameter in the ANCOVA model without statistical theory. In this article, we explain Lord's Paradox using the covariance of linearly combined random variables which is discussed in many statistical theory textbooks (Wackerly et al., 2008; Ross, 2012).

2. Motivating Example

The following example is adapted from the example given by Wright (2006). Suppose two groups of students are compared in their mathematics skills. Group 1 is the treatment group of size n_1 (receiving a new teaching method), and Group 0 is the control group of size n_0 (receiving a traditional teaching method). Assume each student took pre-test and post-test.

2.1 Scenario 1 (Wright, 2006)

Suppose each student selects a group by his or her own will. Suppose a student with high motivation (who tends to show high academic performance) is more likely to select Group 1, and suppose a student with relative low motivation is more likely to select Group 0. Wright (2006) illustrated a similar scenario with balanced group sizes $n_1 = 5$ and $n_0 = 5$ for Group 1 and Group 0, respectively. See Table 1 for the hypothetical data with minor modification from the example of Wright (2006).

Table 1. Hypothetical data of a pre-post study (Scenario 1)

ID	Group	Pre	Post	Difference
1	0	20	30	10
2	0	30	35	5
3	0	40	40	0
4	0	50	45	-5
5	0	60	50	-10
6	1	40	50	10
7	1	50	55	5
8	1	60	60	0
9	1	70	65	-5
10	1	80	70	-10

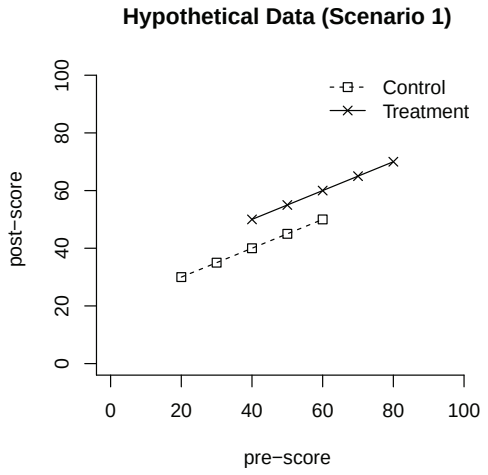

Figure 1. Hypothetical data of a pre-post study (Scenario 1)

The average difference is $(10 + 5 + 0 - 5 - 10) / 5 = 0$ for both groups which can be calculated from Table 1, but the post-score is 10 points greater on average when we condition on the pre-score as shown in Figure 1. (The data in real world may contain random noise around the line.) Using the two-sample t-test, the data are not against the null hypothesis at all (same group average). Using the ANCOVA model, on the other hand, the data are against the null hypothesis and serve as strong evidence for the alternative hypothesis (greater average post-score in Group 1 conditioning on pre-score). This is a traditional example of Lord's Paradox (Lord, 1967; Wright, 2003; Maxwell and Delaney, 2004; Wainer and Brown 2006). In addition to the graphic illustration, an analytic explanation of the paradox can be provided using the covariance of linearly combined random variables.

3. Covariance of Linearly Combined Random Variables

Several textbooks for the first semester of undergraduate statistical theory courses include the following proposition (Wackerly et al., 2008; Ross, 2012).

3.1 Proposition

Let $U_1, \ldots, U_n$ and $W_1, \ldots, W_m$ be random variables. Let $L_1 = \sum_{i=1}^{n} a_i U_i$ and $L_2 = \sum_{j=1}^{m} b_j W_j$ for fixed real numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_m$. Then

$$Cov(L_1, L_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \, Cov(U_i, W_j) \, .$$

Since $V(L_1) = Cov(L_1, L_1)$, a special result for the variance is

$$\begin{aligned} V(L_1) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \, Cov(U_i, U_j) \\ &= \sum_{i=1}^{n} a_i^2 V(U_i) + 2\sum_{i=1}^{n}\sum_{j>i} a_i a_j \, Cov(U_i, U_j)\,. \end{aligned}$$

From these results, we can explain why the two-sample t-test and the ANCOVA model can lead to different conclusions.

3.2 Two-sample t-test

Let Z_i denote the pre-score and Y_i denote the post-score of the i^{th} subject in a sample. Let X_i denote the group indicator for the i^{th} subject, where $X_i = 0$ for Group 0 (control) and $X_i = 1$ for Group 1 (treatment). The two-sample t-test can be formulated as a simple linear model

$$D_i = \beta_0 + \beta_1 X_i + \epsilon_i\,, \tag{1}$$

where $D_i = Y_i - Z_i$ is the change in test score (hence a positive value of D_i is a desirable outcome), and $\epsilon_i \sim N(0, \sigma^2)$ is a random variable which is independent of X_i. In Equation (1), the parameter of interest is the difference in the two group averages

$$\beta_1 = E(D_i \mid X_i = 1) - E(D_i \mid X_i = 0)\,.$$

The null hypothesis is H_0: $\beta_1 = 0$, and the one-sided alternative hypothesis is H_1: $\beta_1 > 0$. An alternative expression of β_1 is

$$\beta_1 = \frac{Cov(X_i, D_i)}{V(X_i)} \tag{2}$$

because

$$\begin{aligned} Cov(X_i, D_i) &= Cov(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) \\ &= Cov(X_i, \beta_0) + Cov(X_i, \beta_1 X_i) + Cov(X_i, \epsilon_i) \\ &= \beta_1 \, V(X_i) \end{aligned}$$

by the proposition in Section 3.1.

3.3 ANCOVA

Preserving the same notation used in Section 3.2, the ANCOVA model assumes

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \delta_i\,, \tag{3}$$

where $\delta_i \sim N(0, \tau^2)$ is a random variable which is independent of X_i and Z_i. Under the ANCOVA model, the parameter of interest is γ_1, the difference in the expected post-score when we compare a randomly selected subject in Group 1 to a randomly selected subject in Group 0 of the same pre-score. The null hypothesis is H_0: $\gamma_1 = 0$, and the one-sided alternative hypothesis is H_1: $\gamma_1 > 0$. An alternative expression of the ANCOVA model is

$$D_i = \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \delta_i$$

by subtracting Z_i on both sides of Equation (3). Using the proposition in Section 3.1,

$$\begin{aligned} Cov(X_i, D_i) &= Cov(X_i, \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \delta_i) \\ &= \gamma_1 V(X_i) + (\gamma_2 - 1)\, Cov(X_i, Z_i)\,, \end{aligned}$$

so the parameter of interest can be written as

$$\begin{aligned} \gamma_1 &= \frac{Cov(X_i, D_i) + (1 - \gamma_2)\, Cov(X_i, Z_i)}{V(X_i)} \\ &= \beta_1 + (1 - \gamma_2)\left(\frac{Cov(X_i, Z_i)}{V(X_i)}\right) \end{aligned}$$

from Equation (2). Using the same argument of the two-sample t-test, we can write

$$\kappa_1 \equiv \frac{Cov(X_i, Z_i)}{V(X_i)} = E(Z_i \mid X_i = 1) - E(Z_i \mid X_i = 0)\,,$$

which is interpreted as the difference in the average pre-score when we compare Group 1 to Group 0.

3.4 Summary

In general, the two-sample t-test and the ANCOVA model have different parameters of interest, and they are related as

$$\begin{aligned} \gamma_1 &= \beta_1 + (1-\gamma_2)\,\kappa_1\,, \\ \beta_1 &= \gamma_1 + (\gamma_2 - 1)\,\kappa_1\,. \end{aligned} \tag{4}$$

They are the same quantity (i.e., $\beta_1 = \gamma_1$) if $\kappa_1 = 0$ or $\gamma_2 = 1$. The first condition $\kappa_1 = 0$ can be satisfied by randomization (i.e., conducting an experimental study instead of an observational study), but the second condition $\gamma_2 = 1$ is out of researcher's control. In most pre-post studies, pre- and post-scores are positively correlated in both groups, so $\gamma_2 > 0$. In addition, we often have $0 < \gamma_2 < 1$ because of regression toward the mean (Stigler, 1997; Barnett et al., 2005).

4. Hypothetical Scenarios

In this section, using the relationship between β_1 and γ_1 in Equation (4), three scenarios are discussed in the context of the educational research. The first scenario is when the average baseline (pre-score) is greater in the treatment group than in the control group (Section 2.1), the second scenario is when the average baseline is lower in the treatment group than in the control group, and the third scenario is when the average baseline is the same between the treatment group and the control group by randomization. The control group is referred to as Group 0, and the treatment group is referred to as Group 1.

4.1 Revisiting Scenario 1

In Scenario 1 (from Section 2.1), the ordinary least square estimation (OLSE) results in $\hat{\gamma}_1 = 10$ and $\hat{\gamma}_2 = 0.5$. Due to self-selection by students, the pre-score is greater in Group 1 by 20 points on average when compared to Group 0, so

$$\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 10 + (0.5 - 1)\,(20) = 0$$

for the two-sample t-test. This is an example of Lord's Paradox when the ANCOVA model can reject the null hypothesis, whereas the two-sample t-test cannot reject the null hypothesis even though the new teaching method seems significantly more effective than the traditional teaching method when we compare two randomly selected students from each group with the same baseline score.

4.2 Scenario 2 (Lower Average Baseline Score in the Treatment Group)

In the second scenario, assume the instructor allocates each student to Group 0 (control) or Group 1 (treatment) believing that the new teaching method would benefit students particularly with low academic performance. See Table 2 for hypothetical data, and see Figure 2 for the scatter plot of pre-score and post-score by group. Note that the pre-score is lower in Group 1 by 20 points on average when compared to Group 0 (i.e., $\hat{\kappa}_1 = -20$).

Table 2. Hypothetical data of a pre-post study (Scenario 2)

ID	Group	Pre	Post	Difference
1	0	40	45	5
2	0	50	50	0
3	0	60	55	-5
4	0	70	60	-10
5	0	80	65	-15
6	1	20	35	15
7	1	30	40	10
8	1	40	45	5
9	1	50	50	0
10	1	60	55	-5

From the data, the OLSE provides $\hat{\gamma}_1 = 0$ and $\hat{\gamma}_2 = 0.5$. In this scenario, the ANCOVA model cannot reject the null hypothesis because $\hat{\gamma}_1 = 0$. From Equation (4), for the two-sample t-test, we estimate $\hat{\beta}_1 = 0 + (0.5 - 1)\,(-20) = +10$ which can lead to the rejection of $\beta_1 = 0$ in favor of $\beta_1 > 0$ (i.e., greater benefit from the new teaching method). This is another example of Lord's Paradox when the two-sample t-test can reject the null hypothesis even though the new teaching method seems ineffective conditioning on the pre-score.

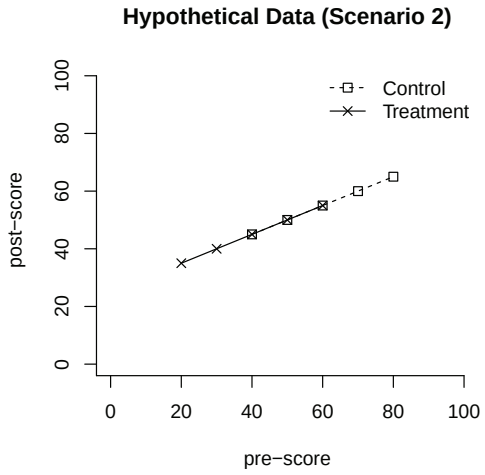

Figure 2. Hypothetical data of a pre-post study (Scenario 2)

4.3 Scenario 3 (Same Average Baseline Score between the Two Groups)

Suppose students are randomized (or controlled to match the average pre-score between the two groups) so that $\kappa_1 = 0$. In this case, the result from Equation (4) leads to $\beta_1 = \gamma_1$. As shown in Table 3 and Figure 3, we have $\hat{\kappa}_1 = 0$, so $\hat{\beta}_1 = \hat{\gamma}_1 = 10$, but the strength of statistical evidence for the alternative hypothesis is stronger in the ANCOVA model than in the two-sample t-test because the standard error is lower in the ANCOVA model. Though the ANCOVA model leads to nearly zero p-value, the two-sample t-test results in a p-value close to 0.05 (for the right-tail H_1: $\beta_1 > 0$). In practice, when students are randomized, the ANCOVA model should have higher statistical power than the two-sample t-test. It is because, while the OLSE is unbiased for both β_1 and γ_1, the variance of $Y_i - \gamma_2 Z_i$ is lower than the variance of $Y_i - Z_i$ conditioning on X_i as discussed in Appendix 1.

Table 3. Hypothetical data of a pre-post study (Scenario 3)

ID	Group	Pre	Post	Difference
1	0	30	40	10
2	0	40	45	5
3	0	50	50	0
4	0	60	55	-5
5	0	70	60	-10
6	1	30	50	20
7	1	40	55	15
8	1	50	60	10
9	1	60	65	5
10	1	70	70	0

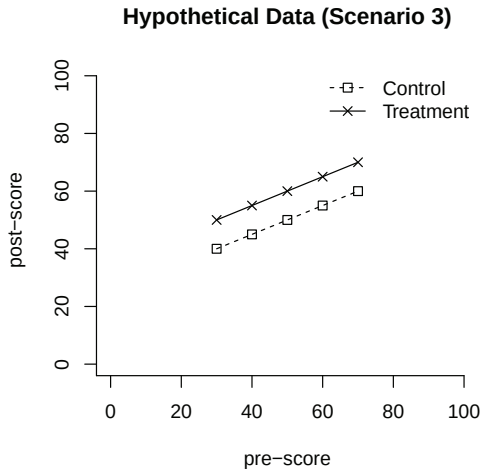

Figure 3. Hypothetical data of a pre-post study (Scenario 3)

5. Examples

In this section, we provide two practical examples. The example in Section 5.1 is to compare the effect of two programs on self-esteem score, and the example in Section 5.2 is to compare the effect of two teaching methods on test score.

5.1 Effect of Exercise on Self-Esteem

This example is from the data in R with `car` package (R Core Team, 2016; Fox & Weisberg, 2011). The data can be seen using the code below.

```
> library(car)
> WeightLoss
```

It has three groups, but we focus on two of the three groups. Twelve subjects ($n_0 = 12$) were treated by a diet program for three months, and this group is referred to as Group 0. Ten subjects ($n_1 = 10$) were treated by an exercise program in addition to the diet program for three months, and this group is referred to as Group 1. From the data presented in Table 4, we can estimate the average self-esteem score 14.8333 for Group 0 and 15.2 for Group 1 at Month 1, so $\hat{\kappa}_1 = 0.3667$.

To formulate hypothesis testing in terms of the expected change in self-esteem (comparing Month 3 to Month 1), the two-sample t-test can be used with H_0: $\beta_1 = 0$ versus H_1: $\beta_1 > 0$, assuming diet and exercise would be more beneficial than diet only, at significance level $\alpha = 0.05$. Using the two-sample t-test, we have a lack of evidence to reject H_0: $\beta_1 = 0$ with observed statistics $\hat{\beta}_1 = 1.0667$, $\widehat{\text{se}} = 0.6568$, $T = 1.624$, and p-value = 0.060.

To formulate hypothesis testing in terms of the expected self-esteem score at Month 3 given the score at Month 1, the ANCOVA model can be used with H_0: $\gamma_1 = 0$ versus H_1: $\gamma_1 > 0$ at $\alpha = 0.05$. Using the ANCOVA model, we have a statistically significance result to conclude H_1: $\gamma_1 > 0$ with observed statistics $\hat{\gamma}_1 = 1.1764$, $\widehat{\text{se}} = 0.6253$, $T = 1.881$, and p-value = 0.038.

In the left panel of Figure 4, the vertical distance between the two parallel lines is $\hat{\gamma}_1 = 1.1764$. In the right panel, the vertical distance between the two horizontal lines is $\hat{\beta}_1 = 1.0667$. Note that $\hat{\gamma}_2 = 0.7006$ in the ANCOVA model, and the estimated parameter in the two-sample t-test is slightly attenuated toward the null value $\beta_1 = 0$ because

$$\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 1.1764 - (0.2994)(0.3667) = 1.0667$$

from Equation (4).

5.2 Comparing Two Teaching Methods

In a mathematics course, two teaching methods were compared for students' learning on set theory, and the learning was quantified by test scores. The first teaching method was based on a traditional lecture (Group 0), and the second teaching method was based on an active-based learning (Group 1). Each of twenty students was randomized into Group 0 or Group 1 by researchers ($n_0 = n_1 = 10$), and each student took a pre-test and a post-test on conceptual thinking.

The left panel of Figure 5 shows the pre-score on x-axis and the post-score on y-axis by group. Random numbers were

Table 4. Self-esteem data for comparing diet group (Group 0) and diet + exercise group (Group 1)

ID	Group (X_i)	Month 1 (Z_i)	Month 3 (Y_i)	Change (D_i)
1	0	12	14	+2
2	0	13	15	+2
3	0	17	18	+1
4	0	16	18	+2
5	0	16	15	−1
6	0	13	18	+5
7	0	12	14	+2
8	0	12	11	−1
9	0	17	19	+2
10	0	19	19	+0
11	0	15	15	+0
12	0	16	18	+2
13	1	15	19	+4
14	1	16	18	+2
15	1	13	17	+4
16	1	16	17	+1
17	1	13	16	+3
18	1	15	18	+3
19	1	15	18	+3
20	1	16	17	+1
21	1	16	19	+3
22	1	17	17	+0

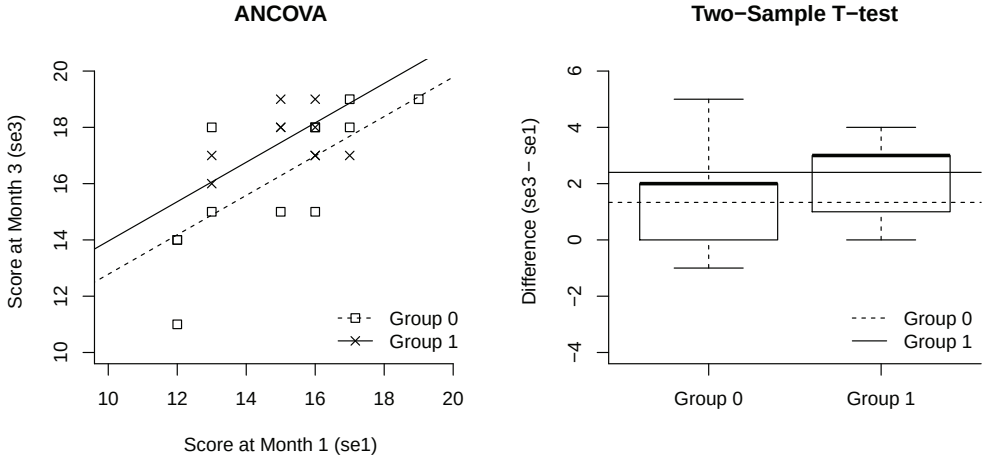

Figure 4. Data comparing diet group (Group 0) and diet + exercise group (Group 1)

generated by $N(0, \eta^2)$ with $\eta = 0.1$, and they were added to original data points for illustration purpose because it was difficult to show all twenty data points without the random noise. Under the ANCOVA model, we estimated $\hat{\gamma}_1 = 1.0283$ (with standard error $\widehat{\text{se}} = 0.3422$) and $\hat{\gamma}_2 = 0.2052$. For the hypothesis testing H_0: $\gamma_1 = 0$ and H_1: $\gamma_1 > 0$ at significance level $\alpha = 0.05$, we could reject H_0 in favor of H_1 with $T = 1.0283/0.3422 = 3.00$ and p-value 0.004.

The right panel of Figure 5 shows the difference in scores (post-score minus pre-score) by group, and the horizontal lines indicate the estimated average difference for each group. Despite the significant result from ANCOVA, the two boxplots look very similar except for one data point in Group 1. Even though the students were randomized, the difference in estimated average pre-score was $\hat{\kappa}_1 = 4.5209 - 3.8075 = 0.7134$ (comparing Group 1 to Group 0). From Equation (4), we can estimate $\hat{\beta}_1 = \hat{\gamma}_1 + (\hat{\gamma}_2 - 1)\,\hat{\kappa}_1 = 1.0283 - (0.7948)(0.7134) = 0.4613$. For the two-sample t-test, the estimated parameter $\hat{\beta}_1 = 0.4613$ was attenuated toward the null value $\beta_1 = 0$, the estimated standard error was $\widehat{\text{se}} = 0.5948$, and the resulting test statistic was $T = 0.4613/0.5948 = 0.776$ with p-value 0.224. Therefore, we could not reject H_0 in the

two-sample t-test at $\alpha = 0.05$.

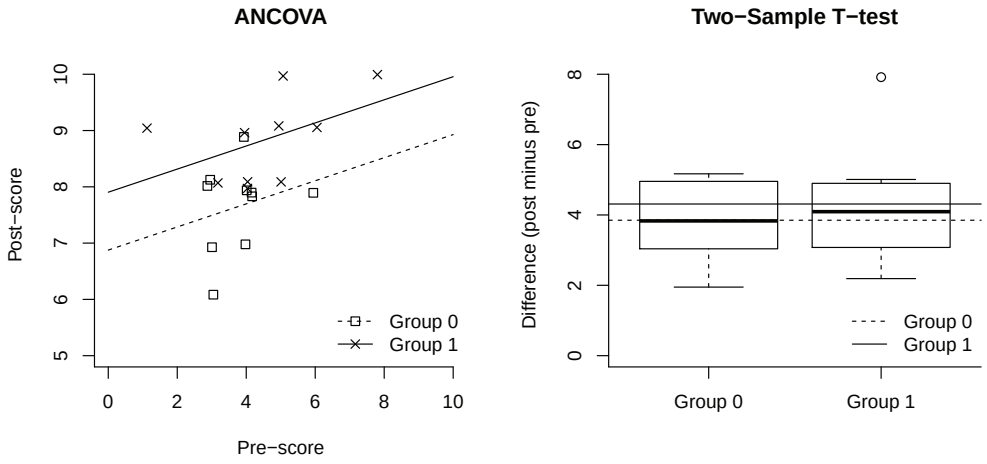

Figure 5. Data comparing traditional lecture (Group 0) and active-based learning (Group 1)

6. Discussion

Lord's Paradox has been known for a long time, and it has been explained graphically in literature, but it has received less attention analytically. Using the covariance of linearly combined random variables, we can show that the parameter β_1 in the two-sample t-test and the parameter γ_1 in the ANCOVA model are different by the magnitude of $(\gamma_2 - 1)\kappa_1$, where κ_1 is the difference in the average baseline score, comparing Group 1 (treatment) to Group 0 (control). In practice, it is difficult to have $(\gamma_2 - 1)\kappa_1 = 0$ in observational studies. This article can be summarized by the three scenarios in terms of the educational research scenarios presented in Section 4.

- When students with high baseline scores belong to the treatment group, which means $\kappa_1 > 0$, we have $\beta_1 < \gamma_1$. In an extreme case, we may have the opposite signs $\gamma_1 > 0$ and $\beta_1 < 0$.
- When students with low baseline scores belong to the treatment group, which means $\kappa_1 < 0$, we have $\beta_1 > \gamma_1$. When the treatment has no effect at all (i.e., H_0: $\gamma_1 = 0$ is true), there is a good chance of rejecting H_0: $\beta_1 = 0$ in favor of H_1: $\beta_1 > 0$ under the two-sample t-test with a large sample size.
- When students are randomized so that the average baseline score is same in the two groups, which means $\kappa_1 = 0$, we have $\beta_1 = \gamma_1$. In most practical situations, where pre- and post-scores are positively correlated in both groups, statistical power to conclude H_1: $\gamma_1 > 0$ in the ANCOVA model is greater than statistical power to conclude H_1: $\beta_1 > 0$ in the two-sample t-test as heuristically explained in Appendix 1.

The proposition in Section 3.1 is mentioned in most introductory statistical theory courses, and students can have deeper understanding of the two-sample t-test and the ANCOVA model through the examples.

In observational studies, we sometimes consider the propensity score, the conditional probability of assignment to a particular group (i.e., control or treatment) as a function of other variables, say $(W_1, \ldots, W_k)$ (Rosebaum & Rubin, 1983). The association between $(W_1, \ldots, W_k)$ and X_i does not necessarily imply the association between $(W_1, \ldots, W_k)$ and Y_i. In general, the difference between β_1 in the two-sample t-test and γ_1 in the multiple linear regression $Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i$ can be quantified as $\beta_1 - \gamma_1 = (\gamma_2 - 1)\kappa_1 + \sum_{j=1}^{k} \alpha_j \nu_j$, where $\nu_j \equiv E(W_{j,i} \mid X_i = 1) - E(W_{j,i} \mid X_i = 0)$. See Appendix 2 for detail. If $W_{j,i}$ is not associated with Y_i given all other covariates (i.e., $\alpha_j = 0$), it does not contribute to the difference between β_1 and γ_1. The same argument holds for the use of a scalar propensity score, say S_i. The role of propensity score depends on the linear relationship between S_i and Y_i and $E(S_i \mid X_i = 1) - E(S_i \mid X_i = 0)$. Without any association between S_i and Y_i, the propensity score does not play any role in the difference between β_1 and γ_1.

References

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology, 34*, 215-220. https://doi.org/10.1093/ije/dyh299

Fox, J., & Weisberg, S. (2011). An R companion to applied regression (2nd edition). Thousand Oaks, CA: Sage.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemilogy, 162*, 267-278. https://doi.org/10.1093/aje/kwi187

Holland, P., & Rubin, D. (1983). On Lord's Paradox. In Principals of Modern Psychological Measurement (H. Wainer and S. Messick, eds.). Hillsdale, NJ: Erlbaum.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*, 304-305. http://dx.doi.org/10.1037/h0025105

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72*, 337-338. http://dx.doi.org/10.1037/h0028108

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analysing data: A model comparison perspective (2nd edition). Mahwah, NJ: Erlbaum.

Pearl, J. (2016). Lord's Paradox revisited C (Oh Lord! Kumbaya!). *Journal of Causal Inference, 4*(2). https://doi.org/10.1515/jci-2016-0021

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosebaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Ross, S. (2008). A first course in probability (8th edition). Upper Saddle River, NJ: Prentice Hall.

Stigler, S. (1997). Regression toward the mean, historically considered. *Statistical Methods in Medical Research, 6*, 103-114. https://doi.org/10.1177/096228029700600202

Tu, Y., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's paradox, Lord's Paradox, and suppression effects are the same phenomenon – the reversal paradox. *Emerging Themes in Epidemiology, 5*(2). https://doi.org/10.1186/1742-7622-5-2

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). Mathematical statistics with applications (7th edition). Belmont, London: Thomson Brooks/Cole.

Wainer, H., & Brown, L. M. (2006). Three statistical paradoxes in the interpretation of group differences: illustrated with medical school admission and licensing data. *Handbook of Statistics, 26*, 893-918. https://doi.org/10.1016/S0169-7161(06)26028-0

Wright, D. B. (2003). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology, 73*, 123-136. https://doi.org/10.1348/000709903762869950

Wright, D. B. (2006). Comparing groups in a before-after design: when t test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*, 663-675. https://doi.org/10.1348/000709905X52210

Appendix 1

In some sense, the two-sample t-test and the ANCOVA model have a common structure:

$$D_i = \beta_0 + \beta_1 X_i + \epsilon_i ,$$
$$D_i^* = \gamma_0 + \gamma_1 X_i + \delta_i ,$$

where $D_i = Y_i - Z_i$ in the two-sample t-test and $D_i^* = Y_i - \gamma_2 Z_i$ in the ANCOVA model. In hypothesis testing, when $\beta_1 = \gamma_1$, we can gain statistical power by having a smaller standard error (SE), and a lower SE can be achieved by a smaller variance of the dependent variable, D_i and D_i^*, given X_i. Assume subjects are randomized so that X_i and Z_i are

uncorrelated. Using the proposition in Section 3.1, we can express $V(D_i^*)$ as

$$\begin{aligned} V(D_i^*) &= V(Y_i - \gamma_2 Z_i) \\ &= V(Y_i) + \gamma_2^2\, V(Z_i) - 2\gamma_2\, Cov(Y_i, Z_i) \\ &= [V(Y_i) + V(Z_i) - 2\, Cov(Y_i, Z_i)] + \left[\gamma_2^2\, V(Z_i) - 2\gamma_2\, Cov(Y_i, Z_i) - V(Z_i) + 2\, Cov(Y_i, Z_i)\right] \\ &= V(D_i) - \left[(1-\gamma_2^2)\, V(Z_i) - 2(1-\gamma_2)\, Cov(Y_i, Z_i)\right], \end{aligned}$$

where

$$\begin{aligned} Cov(Y_i, Z_i) &= Cov(\gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \delta_i, Z_i) \\ &= \gamma_1\, Cov(X_i, Z_i) + \gamma_2\, V(Z_i) \\ &= \gamma_2\, V(Z_i) \end{aligned}$$

because $Cov(X_i, Z_i) = 0$ by the randomization. Therefore, we can simplify

$$\begin{aligned} V(D_i^*) &= V(D_i) - \left[(1-\gamma_2^2)\, V(Z_i) - 2(1-\gamma_2)\,\gamma_2\, V(Z_i)\right] \\ &= V(D_i) - (1-\gamma_2)\, V(Z_i)\, \left[(1+\gamma_2) - 2\gamma_2\right] \\ &= V(D_i) - (1-\gamma_2)^2\, V(Z_i)\,. \end{aligned}$$

To this end, we have $V(D_i^*) < V(D_i)$.

Appendix 2

In the two-sample T-test, the parameter of interest is

$$\beta_1 = \frac{Cov(X_i, D_i)}{V(X_i)} = E(D_i \mid X_i = 1) - E(D_i \mid X_i = 0)\,, \tag{5}$$

where $D_i = Y_i - Z_i$. If the multiple linear regression model is given by

$$Y_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i\,,$$

we can write

$$D_i = \gamma_0 + \gamma_1 X_i + (\gamma_2 - 1)\, Z_i + \sum_{j=1}^{k} \alpha_j W_{j,i} + \delta_i\,.$$

Then the parameter of interest in the two-sample t-test is

$$\begin{aligned} \beta_1 &= \frac{\gamma_1 V(X_i) + (\gamma_2 - 1)\, Cov(X_i, Z_i) + \sum_{j=1}^{k} \alpha_j\, Cov(X_i, W_{j,i})}{V(X_i)} \\ &= \gamma_1 + (\gamma_2 - 1)\, \frac{Cov(X_i, Z_i)}{V(X_i)} + \sum_{j=1}^{k} \alpha_j\, \frac{Cov(X_i, W_{j,i})}{V(X_i)}\,. \end{aligned}$$

Since X_i is a Bernoulli random variable, as in Equation (5),

$$\begin{aligned} \kappa_1 &\equiv \frac{Cov(X_i, Z_i)}{V(X_i)} = E(Z_i \mid X_i = 1) - E(Z_i \mid X_i = 0)\,, \\ \nu_j &\equiv \frac{Cov(X_i, W_{j,i})}{V(X_i)} = E(W_{j,i} \mid X_i = 1) - E(W_{j,i} \mid X_i = 0) \end{aligned}$$

for $j = 1, \ldots, k$. Therefore,

$$\beta_1 = \gamma_1 + (\gamma_2 - 1)\,\kappa_1 + \sum_{j=1}^{k} \alpha_j \nu_j\,.$$