Homework 8

Denis Ostroushko

2022-11-08

16.1

In this problem we will estimate the effect of consuming Brussels sprouts on the DIM (a metabolite of I3C) measurements in urine of the test participants. We will refer to this as simply 'treatment' later in the homeowrk.

```
## [1] 11.30053
```

For a crossover design, will use the following model:

$$Model : Measurement = \alpha * Treatment + \beta * Order + Other$$

We will use the data provided to estimate the parameter α, with an estimator $\hat{a}$. In order to get the estimator, a treatment effect, we need the average of the two quantities. Treatment effect is given by:

$$a = \frac{\bar{x}_1 + \bar{x}_2}{2}$$

Where $\bar{X}_1 = A1 - B2$, i.e. the treatment effect for a group of people who consumed Brussels sprouts and then a placebo, some cabbage.

Where $\bar{X}_2 = A2 - B1$, i.e. the treatment effect for a group of people who consumed placebo and then Brussels sprouts.

We get $X1$ and $X2$ using the code below:

```
X1 <- group_1_data[group_1_data$period == 1, ]$response -
                    group_1_data[group_1_data$period == 2, ]$response
X2 <- group_2_data[group_2_data$period == 2, ]$response -
                    group_2_data[group_2_data$period == 1, ]$response
```

The average of the two quantities is the overall treatment effect, $\hat{a} = 8.7245$, which matches the output in SAS.

We can use a t-test to determine if the treatment is indeed effective, by making sure that the effect is statistically different from 0. Before stating the test, we need to get a standard error for the estimate $\hat{a}$.

We obtain pooled standard error from the two samples using this formula:

$$s^2 = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$$

We then obtain the standard error for the estimate by diving s^2 by 2, as stated on slide 24.

Code for this calculation is given below:

```
pooled_var <- (sd(X1)^2 / (length(X1)) + sd(X2)^2 / (length(X2)))
Standard_error <- sqrt(pooled_var / 4)
```

Thus, we obtain the estimate for the standard error equal to 1.6808. I notice that the estimate is different by that provided in the SAS output. I tried to obtain an estimate using multiple formulas for pooled standard error, and the closest estimate I could get was 1.585. However, 1.6808 is obtained through the more straightforward formula.

Now we can state the formal T-test:

- $H_0 : \hat{a} = 0$
- $H_a : \hat{a} \neq 0$
- T-test statistic: 5.1906
- Critical cutoff value T^*: 2.0687
- $P(T^* > T) = 0$
- Comment on the estimates: Because the value of the pooled standard error does not match that in the SAS output, none of the test statistics and p-value match the output in SAS. However, I consider these as within an acceptable margin of error, especially given that the estimate for the treatment effect is spot on. I will understand and accept any verdict regarding the grading and comments.
- Conclusion: Results are statistically significant, we reject the null hypothesis and conclude that consumption of Brussels sprouts increases the measurement of biomarker of interest in urine.

17.1

Before fitting the logistic regression model, we assess the balance of covariates in the two groups using a two-sample t-tests for averages and proportions between Arm 5 and Arm 6.

Table below provides sample averages for Age and FTND, and the proportions for Gender, Education, and Income binary variables.

Variable	Arm 5	Arm 6	P-value
Averages			
age	44.6388889	45.1710526	0.8347021
FTND	3.1666667	3.6184211	0.1015770
Proportions			
gender	0.4722222	0.4736842	1.0000000
Education	0.6666667	0.6184211	0.7756448
Income	0.2777778	0.3684211	0.4638255

Table comments

1. It looks like ages are fairly balanced between the two arms. T-test p-value is 0.84, which is a pretty big p-value, so the chance that the two average values are not statistically different is pretty high
2. Averages of FTND are not statistically different at the $\alpha = 0.05$ level, but it is pretty close. We might need to balance the samples using a propensity score model
3. The proportion and men and women in the sample is essentially identical in the two samples
4. Percent of people with college or higher education in the two samples is similar
5. Percent of people who earn above $30,000 per year is similar

We fit the logistic regression model and provide a summary of the model below:

Predictor	Estiamte	Standard Error	Z Value	P value
(Intercept)	-0.470954	1.029277	-0.457558	0.647270
age	0.005682	0.016598	0.342313	0.732115
FTND	0.291828	0.166754	1.750058	0.080108
gender	-0.017048	0.422011	-0.040396	0.967777
educ2	-0.276705	0.438114	-0.631581	0.527661
income30	0.501308	0.452928	1.106815	0.268374

No notable changes appear in the model summary. All p-values are pretty similar to the individual appropriate tests. Standard errors are pretty high for all estimated coefficients, which is only interesting for the observation point of view. We will not use this model for inferences on coefficients.

To get the treatment effect, we need to estimate the probability of being in Arm 6 for each participant, denote as $\hat{\pi}_i$, and calculate log odds, given by:

$$X_i = ln(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i})$$

We then compare the average log-odds between members of Arm 5 and 6 to obtain the effect size:

Arm	Mean	Std. Error
Arm 5	0.6535316	0.0818532
Arm 6	0.8471550	0.0496438

T-test statement and conclusion is given below:

- $H_0 : \bar{X}_{arm\ 5} = \bar{X}_{arm\ 6}$
- $H_a : \bar{X}_{arm\ 5} \neq \bar{X}_{arm\ 6}$
- T-test statistic: 2.0226
- Cutoff value T^*: 1.9818
- $P(T^* > T) = 0.0475$
- Conclusion: we reject the null hypothesis and conclude that the effect size is different for the two groups. Therefore, we can conclude that the two samples are not identical, there is confounding present, and we need to use propensity score-based matching to estimate the difference in measurements of interest between the two samples.