Standardization

Part 1: Non-parametric standardization

Suppose the NHEFS data is from an ideal randomized trial to compute the effect of smoking cessation A (`qsmk`) on weight gain Y (`wt82_71`). Participants were randomized to quit smoking (*A=1*) or not quit smoking (*A=2*) conditional on their exercise level (`exercise`) at baseline.

This means that the treated and untreated are *exchangeable* conditional on exercise. That is, we have conditional exchangeability.

One approach to estimating the effect of smoking on weight gain while adjusting for other variables is standardization. Recall that standardization will give us an estimate of the marginal average causal effect. Standardization entails estimating the distribution of L and the mean outcome conditional on treatment and L:

$$\mathrm{E}[Y^a] = \sum_{l}^{L} \Pr[L = l] \times \mathrm{E}[Y|A = a, L = l]$$

To start, we need to import the NHEFS dataset into R:

```
nhefs <- read.csv('nhefs_nomiss.csv')
```

1. **Complete the following table. Sample R code is provided to generate the estimates in the first row of the table. You will need to modify this code to complete the remaining cells of the table.**

```
# Distribution of L:
table(nhefs["exercise"])
prop.table(table(nhefs["exercise"]))

# Conditional means for the outcome

  # L=0, A=1
  mean(nhefs[which(nhefs$exercise==0 & nhefs$qsmk==1),]$wt82_71)

  # L=0, A=0
  mean(nhefs[which(nhefs$exercise==0 & nhefs$qsmk==0),]$wt82_71)
```

Exercise level (L)	**Estimates for $\Pr[L = l]$**	**Estimates for $\mathrm{E}[Y \mid A = 1, L = l]$**	**Estimates for $\mathrm{E}[Y \mid A = 0, L = l]$**
Much exercise $\mathbf{L = 0}$	0.192	4.574	2.728
Moderate exercise $\mathbf{L = 1}$			
Little or no exercise $\mathbf{L = 2}$			

2. **Using the numbers that you calculated in Question 1, calculate an estimate for $\mathrm{E}[Y^{a=1}]$.**

Answer:

3. **Using the numbers that you calculated in Question 2, calculate an estimate for $\mathrm{E}[Y^{a=0}]$.**

Answer:

4. **What is the marginal average causal effect of smoking cessation on weight gain, that is, $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$?**

Answer:

5. **The stratum-specific estimates for the effect of smoking cessation on weight gain are as follows:**
   L=0: 1.85 kg
   L=1: 2.96 kg
   L=2: 2.47 kg
   How do the stratum-specific effects compare against the marginal effect?

Answer:

Part 2: Parametric standardization for a continuous outcome

In reality, NHEFS is not an ideal randomized trial. Rather, it is an observational study. As such, we will probably need to adjust for many confounders in order to compute the effect of smoking cessation A (`qsmk`) on weight gain Y (`wt82_71`).

Suppose the treated and untreated are exchangeable conditional on the following covariates:

- Exercise (`exercise`)
- Age (`age`)
- Age^2
- Sex (`sex`)
- Education (`education`)

With many covariates, we'll need to perform standardization using models. Recall that there are four steps to standardization when using statistical software:

1. Expansion of the dataset
2. Outcome modelling
3. Prediction
4. Standardization by averaging

Expansion of the dataset. We copy the dataset twice. In the first copy, we set all of the values for treatment to 1; in the second copy, we set all of the values for treatment to 0.

```
# Copy 1: set all values of treatment to 1
nhefs1 <- nhefs
nhefs1$qsmk <- 1

# Copy 2: set all values of treatment to 0
nhefs0 <- nhefs
nhefs0$qsmk <- 0
```

Outcome modelling. We fit a model for the outcome conditional on treatment and the covariates using our original dataset (*nhefs*).

```
lin.mod <- lm(wt82_71~qsmk + as.factor(exercise) + age + I(age^2) +
  sex + as.factor(education), data=nhefs)
summary(lin.mod)
```

Prediction. Using our model, we obtain predicted values for the outcome in our two copies of the dataset. In the first copy of the dataset, the outcomes are predicted based on everyone having a treatment value of 1; in the second copy of the dataset, the outcomes are predicted based on everyone having a treatment value of 0.

```
nhefs1$predict.y <- predict(lin.mod, newdata=nhefs1, type="response")
nhefs0$predict.y <- predict(lin.mod, newdata=nhefs0, type="response")
```

Standardization by averaging: We take the predicted values obtained in the previous step, and average them. This gives us $E[Y^{a=1}]$ and $E[Y^{a=0}]$.

6. **What is the marginal average causal effect of smoking cessation on weight gain, that is, $\mathbf{E}[Y^{a=1}] - \mathbf{E}[Y^{a=0}]$?**

Answer:

7. **Provide a causal interpretation for the number that you calculated in Question 7.**

Answer:

Part 3: Parametric standardization for a dichotomous outcome

Now, suppose we are interested in the effect of smoking cessation *A* (`qsmk`) on death (`death`). We'll use the same set of covariates as in Part 2:

- Exercise (`exercise`)
- Age (`age`)
- Age^2
- Sex (`sex`)
- Education (`education`)

Again, we'll have to perform standardization using models. The four steps are exactly the same:

1. Expansion of the dataset
2. Outcome modelling
3. Prediction
4. Standardization by averaging

The only difference is that we will have to use logistic regression in step 2 (rather than linear regression).

Modify the code provided in Part 2 to try this yourself!

8. First, you will need to expand the dataset. Provide your R code below:

Answer:

9. **Second, you will need to fit a model for the outcome. Remember you'll need to use the glm function and specify family=binomial in order to fit a logistic regression model (rather than a linear regression model). Provide your R code below:**

   Hint: Your output should look something like this:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9196  -0.5681  -0.3261  -0.1980   2.8977

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -4.4746565  1.4241948  -3.142 0.001679 **
qsmk                   -0.0369671  0.1688389  -0.219 0.826690
as.factor(exercise)1   -0.0550115  0.2245934  -0.245 0.806504
as.factor(exercise)2    0.1653622  0.2234229   0.740 0.459221
age                     0.0387405  0.0577544   0.671 0.502361
I(age^2)                0.0007379  0.0005818   1.268 0.204680
sex                    -0.5981422  0.1576533  -3.794 0.000148 ***
as.factor(education)2 -0.3724349  0.2128112  -1.750 0.080106 .
as.factor(education)3 -0.6944548  0.1971156  -3.523 0.000427 ***
as.factor(education)4 -0.4597077  0.3477331  -1.322 0.186164
as.factor(education)5 -0.6328623  0.2915904  -2.170 0.029978 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1503.7  on 1565  degrees of freedom
Residual deviance: 1123.1  on 1555  degrees of freedom
AIC: 1145.1

Number of Fisher Scoring iterations: 6
```

Answer:

10. **Third, obtain the predicted values for the outcome based on the model that you fit in the previous step. Provide your R code below:**

Answer:

11. Fourth, find the average of the predicted outcomes in each of the copies of the dataset. Provide your R code below:

Answer:

12. What is the estimate for $\mathbf{E}\left[Y^{a=1}\right]$?

Answer:

13. What is the estimate for $\mathbf{E}\left[Y^{a=0}\right]$?

Answer:

14. Calculate the causal risk difference.

Answer:

15. Calculate the causal risk ratio.

Answer:

16. Provide a causal interpretation for the causal risk ratio.

Answer:

17. Calculate the causal odds ratio.

Answer: