PHS 2000B Problem Set 3

Interaction (34 pts)

Due Thursday, February 20, 2025

Key concepts (15 pt)

There is considerable evidence that smoking and exposure to air pollution both are independent causes of lung cancer. However, in Turner et al. 2014[1], the authors were interested in whether there may be a joint effect of smoking and air pollution on lung cancer mortality. Below is a modified reproduction of their Table 2 showing the raw data on rates of lung cancer death within strata of smoking status (current vs. never smoker) and exposure to air pollution ($PM_{2.5}$ exposure above 75th percentile vs. below 25th percentile).

	Never Smoker		Current Smoker	
$PM_{2.5}$ exposure	Deaths	No. Subjects	Deaths	No. Subjects
Low ($\leq$ 25th percentile)	63	76,025	346	31,486
High ($\geq$ 75th percentile)	81	73,592	447	33,789

1. Calculate an appropriate measure of interaction on the additive scale and **interpret**. (2 pt)

Answer:

2. Calculate an appropriate measure of interaction on the multiplicative scale and **interpret**. (2 pt)

Answer:

3. Are your estimates of 1 and 2 similar or different? In general, when do you expect them to be similar and when do you expect them to be different? (2 pt)

Answer:

4. Calculate and **interpret** the RERI. (2 pt)

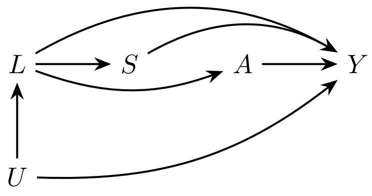
Answer:

[1] Michelle C. Turner, Aaron Cohen, Michael Jerrett, Susan M. Gapstur, W. Ryan Diver, C. Arden Pope, Daniel Krewski, Bernardo S. Beckerman, Jonathan M. Samet, Interactions Between Cigarette Smoking and Fine Particulate Matter in the Risk of Lung Cancer Mortality in Cancer Prevention Study II, American Journal of Epidemiology, Volume 180, Issue 12, 15 December 2014, Pages 1145–1149, https://doi.org/10.1093/aje/kwu275

5. Using counterfactual notation, write out the causal effect implied when the authors mention interest in a joint effect of intervening on smoking and air pollution. Please define how we can see if there is additive or multiplicative interaction. (3 pt)

Answer:

6. Turner and colleagues collected new data with the following underlying causal structure. L is a binary measured variable and U are unmeasured variables. What assumptions do you need to make to identify the joint effect of smoking S and air pollution A on lung cancer mortality Y with this new data? Show how you *could* estimate whether there is a causal interaction on the **multiplicative scale** using data and relevant assumptions (Hint: use the g-formula/standardization, following the steps shown in the lecture or lab). (4 pt)



Answer:

Modeling interactions (8 pt)

You are interested in a hypothesized interaction between two exposures, arsenic and tobacco smoke[2], on the incidence of skin lesions. Based on the literature you are fairly certain that exposure to arsenic in drinking water is a necessary cause of skin lesions, but believe that there may be significant interaction with other carcinogenic exposures like smoking. This data is found in the `arsenic.csv` file. At baseline you collect data on arsenic exposure (`arsenic`) from individual drinking sources and categorize them into "high" and "low" exposure categories. Likewise you ask participants whether they currently smoke (`smoker`) and record their current age (`age`). You then prospectively follow them and record who develops skin lesions (`lesions`).

The assumption that the outcome is rare is reasonable here.

1. Test your hypothesis by fitting a statistical model with a multiplicative interaction involving arsenic and smoking exposure and adjusting for age.

CODE:

```r
# Make sure eval = TRUE when knitting the assignment if using Rmd

arsenic <- read.csv("arsenic.csv")

m1 <- glm(lesions ~ arsenic + smoker + arsenic:smoker + age, data = arsenic, family = binomial)

summary(m1)
```

```
##
## Call:
## glm(formula = lesions ~ arsenic + smoker + arsenic:smoker + age,
##     family = binomial, data = arsenic)
##
## Coefficients:
##                Estimate Std. Error z value    Pr(>|z|)
## (Intercept)    -1.90677    0.35202   -5.42 0.000000061 ***
## arsenic         0.08676    0.34584    0.25       0.802
## smoker          0.45127    0.21661    2.08       0.037 *
## age            -0.00882    0.00859   -1.03       0.305
## arsenic:smoker  0.87110    0.44061    1.98       0.048 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 810.83  on 1027  degrees of freedom
## Residual deviance: 784.92  on 1023  degrees of freedom
## AIC: 794.9
##
## Number of Fisher Scoring iterations: 4
```

[2] While the data used in this problem set are fake, a more rigorous overview of this issue can be found here: Chen, Y., Graziano, J. H., Parvez, F., Hussain, I., Momotaj, H., Van Geen, A., ... & Ahsan, H. (2006). Modification of risk of arsenic-induced skin lesions by sunlight exposure, smoking, and occupational exposures in Bangladesh. Epidemiology, 459-467.

1a. Write out the model you have used. (1 pt)

Answer:

1b. Calculate and **interpret** the estimate and a 95% confidence interval for the relevant interaction term. (2 pt)

Answer:

1c. Does your analysis support this hypothesis? (1 pt)

Answer:

2. Calculate and **interpret** estimates and associated 95% CIs for the RERI and attributable proportion (i.e., the proportion of the outcome in the doubly exposed that is due to the interaction). (4 pt)

Hint: Run the code in the `interaction_code.R` file (you can also use `source("interaction_code.R")` if you've saved the `interaction_code.R` file in your working directory) and then use the `additive_interactions()` function on your model result to get the these measures.

Code:

```
source("interaction_code.R")

# now you'll just need to call the function and interpret the results
# see the OSF repository for documentation and vignettes: https://osf.io/7ccpp/
```

```
m <- glm(lesions ~ arsenic + smoker + arsenic:smoker + age, data = arsenic, family = binomial)
rs <- additive_interactions(m, "lesions")
```

```
##              Stat     Est    CI.lo  CI.hi     p.val.0 p.val.epi p.val.suff.cause
## 1            RERI 2.43147  0.35486 4.5081 0.010869596    0.3419          0.08834
## 2              AP 0.59414  0.31610 0.8722 0.000014054        NA               NA
## 3         arsenic 0.02931 -0.20485 0.2635 0.403101177        NA               NA
## 4          smoker 0.18442 -0.01498 0.3838 0.034937146        NA               NA
## 5 arsenic:smoker 0.78627  0.44675 1.1258 0.000002826        NA               NA
```

Answer:

Subgroup analysis (11 pt)

In 2008, the state of Oregon expanded its Medicaid program to cover low-income residents who were previously uninsured. At the time, the waitlist to join the program was much larger than the number of available slots that the state could fund so officials held a lottery to determine which individuals on the waitlist would be offered slots. Selected adults won the opportunity to apply for Medicaid and to enroll if they met eligibility requirements. This lottery presented an opportunity to study the impact of Medicaid. In what became known as the "Oregon Health Insurance Experiment"[3], researchers used the random assignment during the expansion to study the causal effects of Medicaid on financial and health outcomes.

In this section, you will be using a subset of the actual replication data from the OHIE to examine whether there are important or interesting interactions between Medicaid and other baseline variables.

1. Load the `ohie.csv` dataset and look over the codebook (`ohie_codebook.xlsx`). Identify an outcome (highlighted in green) that interests you and at least one baseline variable (highlighted in blue) that you think could have an important statistical interaction with Medicaid assignment, the main exposure of interest (`treatment`). Run a regression to determine whether this is the case (*Note: you will be ignoring the significance level of the interaction term so you do not need to base your variable selection on the p-value*). Consider the type of outcome when specifying your regression model and use whatever scale (multiplicative or additive) you think is most relevant for the variables you chose.

```
# insert your R code here to run the model
# use summary(model) to see the estimated coefficients of your model
# use confint(model) or the standard error provided in summary(model)
# to generate 95% confidence intervals for estimated coefficients
```

1a. Write out the model you used and report the estimate and 95% confidence interval for the interaction term. (1 pt)

Answer:

1b. **Intepret** the interaction term. *Note: If selecting a categorical variable for interaction, provide interpretations all interaction coefficients. If using a logistic or log-linear model, you may interpret on the log-scale or exponentiated.* (1 pt)

Answer:

2. Do you think the relationship between Medicaid assignment and your chosen variable is a causal interaction or a heterogeneous effect? Explain which assumptions are necessary for it to be a causal interaction or a heterogeneous effect? Please define the assumptions using notation. Which is more likely? (5 pt)

Answer:

3. Ignoring the statistical significance of the interaction term (i.e. assuming there is some sort of effect), draw a DAG showing this hypothesized relationship between Medicaid assignment, your outcome, and your baseline variable. Make sure to include any other relevant variables, measured or unmeasured. (2 pt)

Answer:

[3] Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J., Newhouse, J., ... Finkelstein, A. (2013). The Oregon Experiment — Effects of Medicaid on Clinical Outcomes. The New England Journal of Medicine, 368(18), 1713-1722.

4. If you were planning a follow up based on your result, would you target a different subpopulation or would you potentially attempt a more comprehensive program to intervene on both interaction variables? Why? (2 pt)

Answer: