**Lab**

**Regression Methods for Causal Mediation Analysis**

**P1822 – Statistical Methods for Causal Inference**

**Linda Valeri**

Analyze a simulated data based on the case-control lung cancer study that was described in the lecture [data was simulated so that the distribution of outcomes, exposures and covariates were roughly similar to those in the original case control study]. Note that these simulated data are restricted to Caucasians and so population stratification is not likely an issue in the analyses you will run. The assumption that the outcome is rare is very reasonable here. The .csv dataset can be found on canvas. The data consists of the following variables:

Variable in Dataset Description

case lung cancer

snp 0 vs. 1/2 C alleles at rs8034191

smoking 1=ever smoker; 0=never smoker

sex 1=male 2=female

colgrad 1=college graduate 0=not college graduate

age in years

Let the exposure be the genetic variant, the mediator smoking status and the outcome lung cancer status.

Employ counterfactual approaches to mediation analysis and estimate direct effect and indirect effect.

1. Write a DAG that represents the mediation analysis of interest. Think about confounders you would like to adjust for.
2. Write the estimand for the controlled direct effect odds ratios had everyone been a smoker. Under which assumptions is this quantity identified? Which models you need to fit to estimate this quantity?
3. Write the estimand for the controlled direct effect odds ratios had everyone not been a smoker.
4. We can evaluate these and all other effects at the mean level of the covariates using the *CMAverse R package*.
5. Write the estimand for the natural direct effect and indirect effect odds ratios. Under which assumptions are these quantity identified? Which models you need to fit to estimate these quantities?

1. Calculate and interpret estimates of what the natural direct and indirect effect and total effect odds ratios would be, allowing for exposure-mediator interaction.