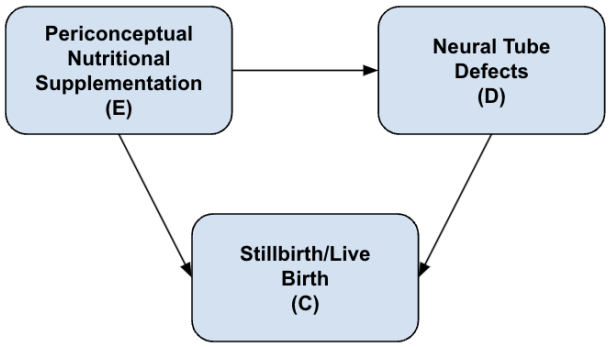
BIOST 536 Homework 3

Benjamin Stan

October 23, 2021

Question 1

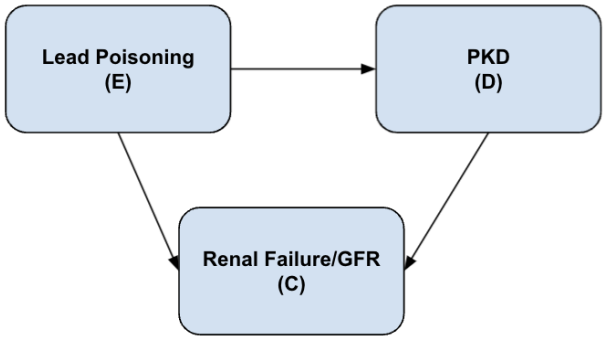
The DAG for the study is shown below:



Based on this causal model, limiting the study to live births could bias the results because it removes one of the effects of the outcome of interest (neural tube defects). This would filter down the types of cases in which the effect of E on D could be observed.

Question 2

The DAG for this study is shown below:



Based on this DAG, there is no need to adjust for GFR as a confounder. It is a collider on the path between the exposure (lead poisoning) and disease (PKD) of interest.

Question 3

The following questions pertain to this model

$$logit(p) = \beta_0 + \beta_1 * C + \beta_2 * E + \beta_3 * CE$$

a.) If the P(D) in the population is 10%, that means that 90% of the population does not have the disease. Under a 1:1 sampling pattern, the value of π is $0.9/0.1 = 9$. The expected value of $\hat{\beta}_0$ in terms of the population parameters π and β_0 is:

$$\begin{aligned} \hat{\beta}_0 &= log(\pi) + \beta_0 \\ &= log(9) + \beta_0 \end{aligned}$$

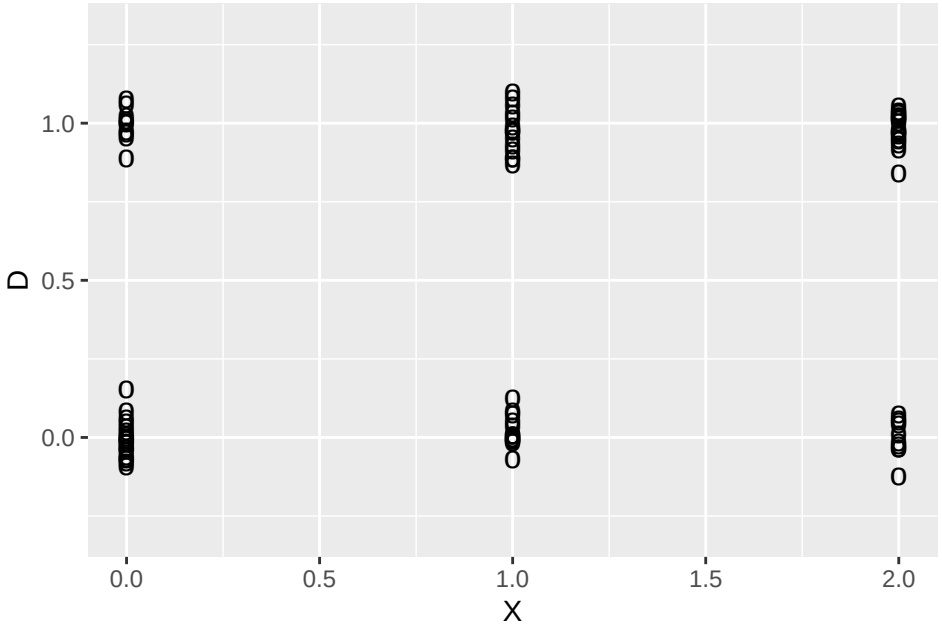
b.) Under a 1:9 sampling pattern, the value of π is 1. The expected value of $\hat{\beta}_0$ in terms of the population parameters π and β_0 is:

$$\begin{aligned} \hat{\beta}_0 &= log(\pi) + \beta_0 \\ &= log(1) + \beta_0 \\ &= \beta_0 \end{aligned}$$

This equivalence with the population parameter is expected because the sampling pattern matches the prevalence of the cases and controls in the population.

Question 4

a.) The scatterplot of D vs. X is shown below:



b.) The fitted model for the logistic regression of D on X is shown below

$$logit\left(\widehat{P(D=1|X)}\right) = -0.69 + 0.69 * X$$

c.) According to the fitted model, the probability of D for individuals with X=0 is

$$\frac{exp(-0.69)}{1+exp(-0.69)} = \frac{0.502}{1.502} = 0.33$$

This estimate of 0.33 makes sense because one-third of subjects with X=0 have a value of D=1.

d.) According to the fitted model, the probability of D for individuals with X=1 is

$$\frac{exp(0)}{1+exp(0)} = \frac{1}{2} = 0.5$$

This estimate of 0.5 makes sense because one-half of subjects with X=1 have a value of D=1.

e.) According to the fitted model, the probability of D for individuals with X=2 is

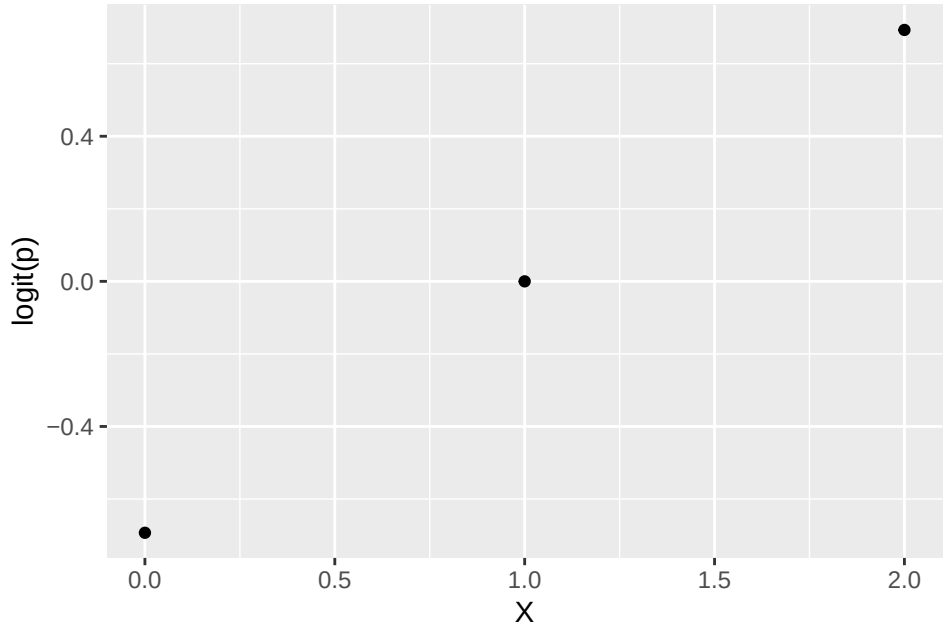
$$\frac{exp(0.69)}{1+exp(0.69)} = \frac{1.99}{2.99} = 0.67$$

This estimate of 0.67 makes sense because two-thirds of subjects with X=2 have a value of D=1.

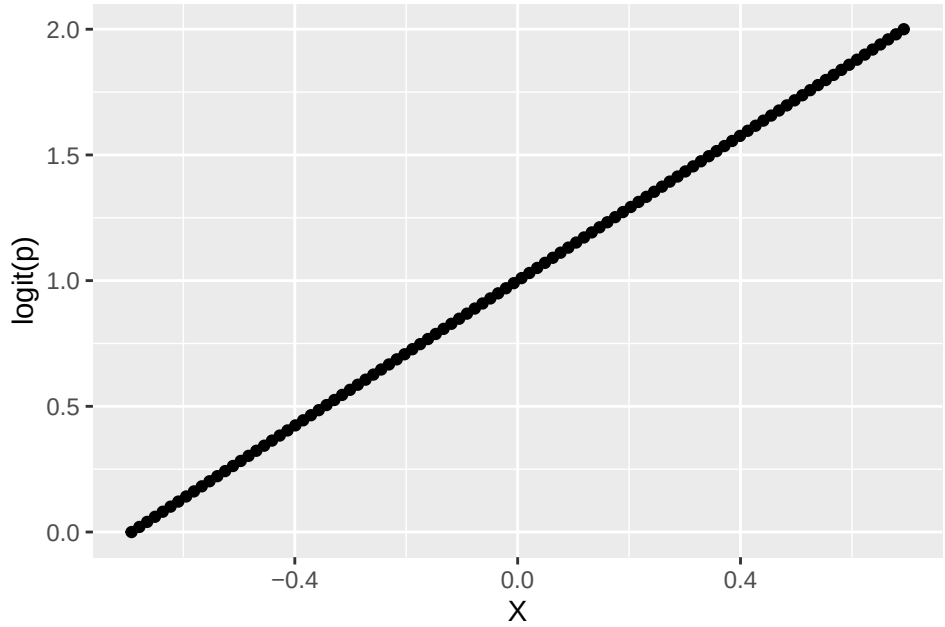
f.) The completed table is below:

X	P(D=1)	Odds of D	log-odds of D
0	0.333	0.5	-0.693
1	0.5	1	0
2	0.666	2	0.693

g.) The graph of logit(p) vs. X is shown below:



It can be noted that the graph of logit(p) vs. X is a straight line. To confirm this, a sequence of 100 values between 0 and 2 was used for X and the fitted values are shown below.



Question 5

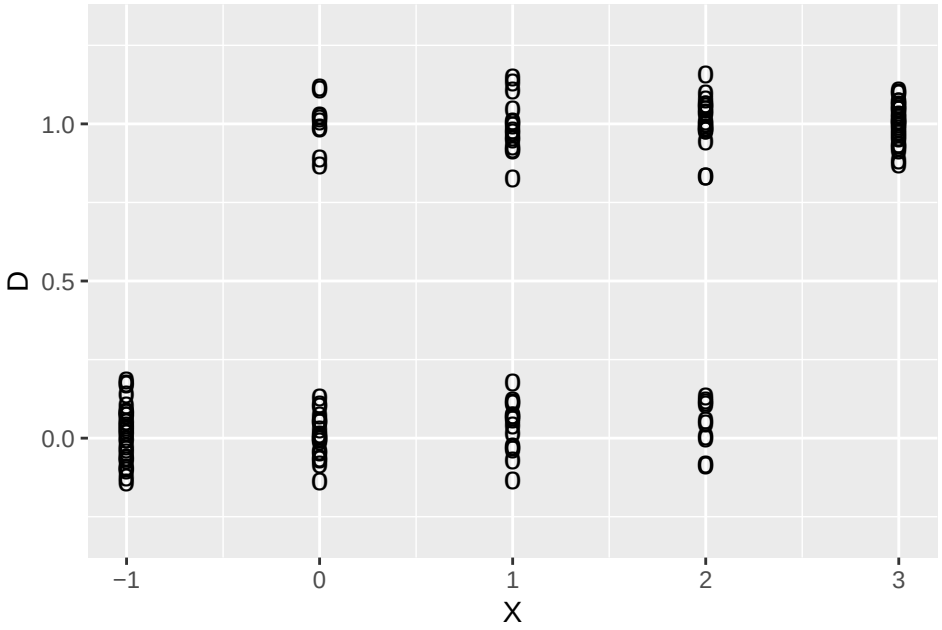
The fitted model for the logistic regression of D on X is shown below

$$logit\left(\widehat{P(D=1|X)}\right) = -0.69 + 0.69 * X$$

Comparing these results to those from Q4, the logistic regression coefficients are the same. This is because the proportion of values in each X/D bucket were the same, but the number of samples increased by a factor of 10. However, the standard error for the coefficient on X in Q4 is 0.27, compared to 0.087 in this model. This is to be expected given the 10x increase in samples, which would lead to lower standard errors. Similarly, this model has narrower confidence intervals. The Q4 coefficient on X is -0.69 with 95% robust CI of (-1.38,-0.0058), while this model has a coefficient on X of -0.69 with a 95% robust CI of (-0.91,-0.48).

Question 6

a.) The scatterplot of D vs. X is shown below:



b.) The fitted model for the logistic regression of D on X is shown below

$$logit\left(\widehat{P(D=1|X)}\right) = -1.35 + 1.35 * X$$

c.) According to the fitted model, the probability of D for individuals with X=0 is

$$\frac{exp(-1.35)}{1+exp(-1.35)} = \frac{0.259}{1.259} = 0.206$$

This estimate of 0.206 differs from that in Q4c because the data that is used to fit the model differs. The new data does not allow the fitted model to exactly estimate the proportion of subjects with a value of D=1 at X=0. This change is due to the new covariate values at X=-1 and X=3. The model is not able to accurately predict the proportion of subject with D=1 for any value aside from X=1, the midpoint of the data which happens to correspond to a P(D|X=1) of 0.5.

d.) According to the fitted model, the probability of D for individuals with X=1 is

$$\frac{exp(0)}{1+exp(0)} = \frac{1}{2} = 0.5$$

This exactly the matches the result from Q4d. This estimate of 0.5 makes sense because one-half of subjects with X=1 have a value of D=1. The model is able to properly estimate this value because the value of X=1 is the midpoint in the data and has a P(D|X=1) of 0.5.

Appendix

```
## Question 1
setwd("/Users/bstan/Documents/UW/Courses/BIOST 536")
rm(list = ls())
```

```r
library(tidyverse)
library(tidyr)
library(tinytex)
library(sandwich)
knitr::include_graphics("figs/hw3_q1_dag.png")

## Question 2
knitr::include_graphics("figs/hw3_q2_dag.png")

## Question 4a
df <- data.frame(rbind(t(replicate(n=20,c(0,0))),
                       t(replicate(n=10,c(0,1))),
                       t(replicate(n=15,c(1,0))),
                       t(replicate(n=15,c(1,1))),
                       t(replicate(n=10,c(2,0))),
                       t(replicate(n=20,c(2,1)))
                       ))
colnames(df) <- c("X","D")
set.seed(44)
df$D2 <- df$D+rnorm(nrow(df), mean=0, sd=0.05)
ggplot(data = df, aes(X,D2)) +geom_point(size=4,shape='o') +
  xlab("X") +
  ylab("D") +
  ylim(-0.3,1.3)

## Question 4b
glm <- glm(D~X, data = df, family = "binomial")
coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[1] + c(0, qnorm(c(0.025, 0.975))) * rob_se[1]
conf_int
summary(glm)

## Question 4g
### Plot values from table
df_plot <- data.frame(cbind(c(0,1,2),c(-0.693,0,0.693)))
colnames(df_plot) <- c("x","y")
ggplot(df_plot,aes(x,y)) + geom_point() +
  xlab("X") +
  ylab("logit(p)")

### Plot sequence of values
df_plot <- data.frame(seq(0,2,length.out=100),-0.693+0.693*seq(0,2,length.out=100))
colnames(df_plot) <- c("x","y")
ggplot(df_plot,aes(y,x)) + geom_point() +
  xlab("X") +
  ylab("logit(p)")

## Question 5
df <- data.frame(rbind(t(replicate(n=200,c(0,0))),
                       t(replicate(n=100,c(0,1))),
                       t(replicate(n=150,c(1,0))),
```

```
                        t(replicate(n=150,c(1,1))),
                        t(replicate(n=100,c(2,0))),
                        t(replicate(n=200,c(2,1)))
                        ))
colnames(df) <- c("X","D")

glm <- glm(D~X, data = df, family = "binomial")
coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[1] + c(0, qnorm(c(0.025, 0.975))) * rob_se[1]
conf_int
summary(glm)

## Question 6a
df <- data.frame(rbind(t(replicate(n=30,c(-1,0))),
                        t(replicate(n=20,c(0,0))),
                        t(replicate(n=10,c(0,1))),
                        t(replicate(n=15,c(1,0))),
                        t(replicate(n=15,c(1,1))),
                        t(replicate(n=10,c(2,0))),
                        t(replicate(n=20,c(2,1))),
                        t(replicate(n=30,c(3,1)))
                        ))
colnames(df) <- c("X","D")
set.seed(41)
df$D2 <- df$D+rnorm(nrow(df), mean=0, sd=0.08)
ggplot(data = df, aes(X,D2)) +geom_point(size=4,shape='o') +
  xlab("X") +
  ylab("D") +
  ylim(-0.3,1.3)

## Question 6b
glm <- glm(D~X, data = df, family = "binomial")
coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int <- coef[1] + c(0, qnorm(c(0.025, 0.975))) * rob_se[1]
conf_int
summary(glm)
```