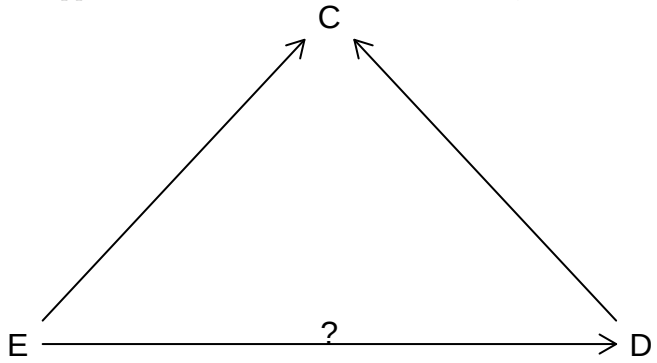
BIOST536 Homework 3

Ivy Zhang

10/19/2021

1

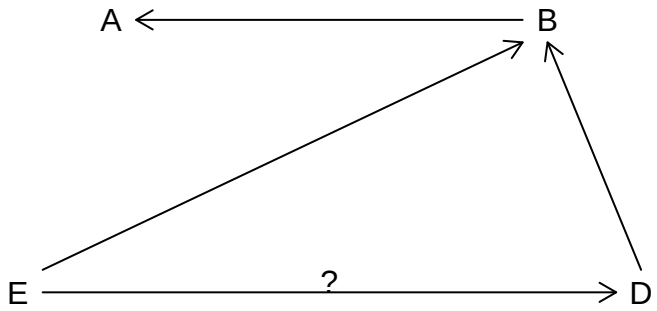
The following DAG tells us about the causal model based on the question explaination. As we can see birth status is a collider. If we stratifying or adjusting for birth status which is a collider , we will induce association between periconceptual nutritions supplementation and neutral tube defects, and lead to confounding. After we delete all arrows emanating from E, because of the association between periconceptual nutritions supplementation and neutral tube defects, we will have an unblocked path, therefore confounding.



1	E	Periconceptual nutritions supplementation
2	C	Birth status
3	D	Neural tube defects

2

The following graph is the DAG I draw to summarize the information. I don't think Dr.Ott should treat GFR as a confounder. If we delete all arrows emanating from E, we cannot find any backdoor path from E to D. GFR is on the casual pathway between lead poisoning and polycystic kidney disease. Therefore, Dr.Ott should not treat GFR as a confounder.



1 E Lead poisoning
2 A Glomerular filtration rate
3 B Kidney Failure
4 D Polycystic kidney disease

3

(A) π equals to 9 based on the description.

$$\pi = \frac{P[Z=1|D=1]}{P[Z=1|D=0]} = \frac{1}{1/9} = 9$$

$$\beta_0 = log\pi + logitP[D|E=0, C=0]$$

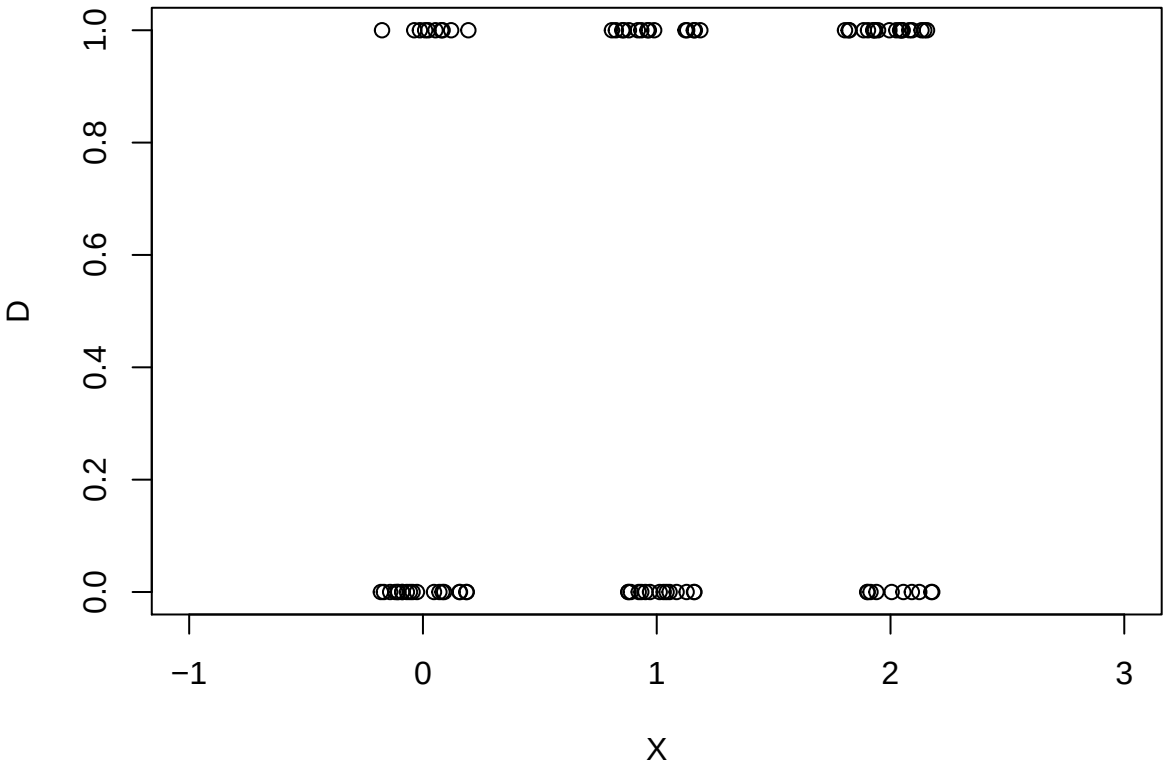
(B) π equals to 1 based on the description.

$$\pi = \frac{P[Z=1|D=1]}{P[Z=1|D=0]} = \frac{1}{9/9} = 1$$

$$\beta_0 = log\pi + logitP[D|E=0, C=0] = 0 + logitP[D|E=0, C=0] = logitP[D|E=0, C=0]$$

4

A



(B) We fitted the model as:

$$logit(P[D = 1|X]) = -0.693 + 0.693 \times X$$

(C)

$$P[D = 1|X = 0] = \frac{e^{-0.693}}{1 + e^{-0.693}} = 0.333$$

We estimated the probability of D for individuals with X = 0 equals to 0.5. Because for 30 individuals who have X = 1, one-third of them have D = 1.0.333 is very similar to one-third.

(D)

$$P[D = 1|X = 1] = \frac{e^{0}}{1 + e^{0}} = 0.5$$

We estimated the probability of D for individuals with X = 1 equals to 0.5. Because for 30 individuals who have X = 1, half of them have D = 1.

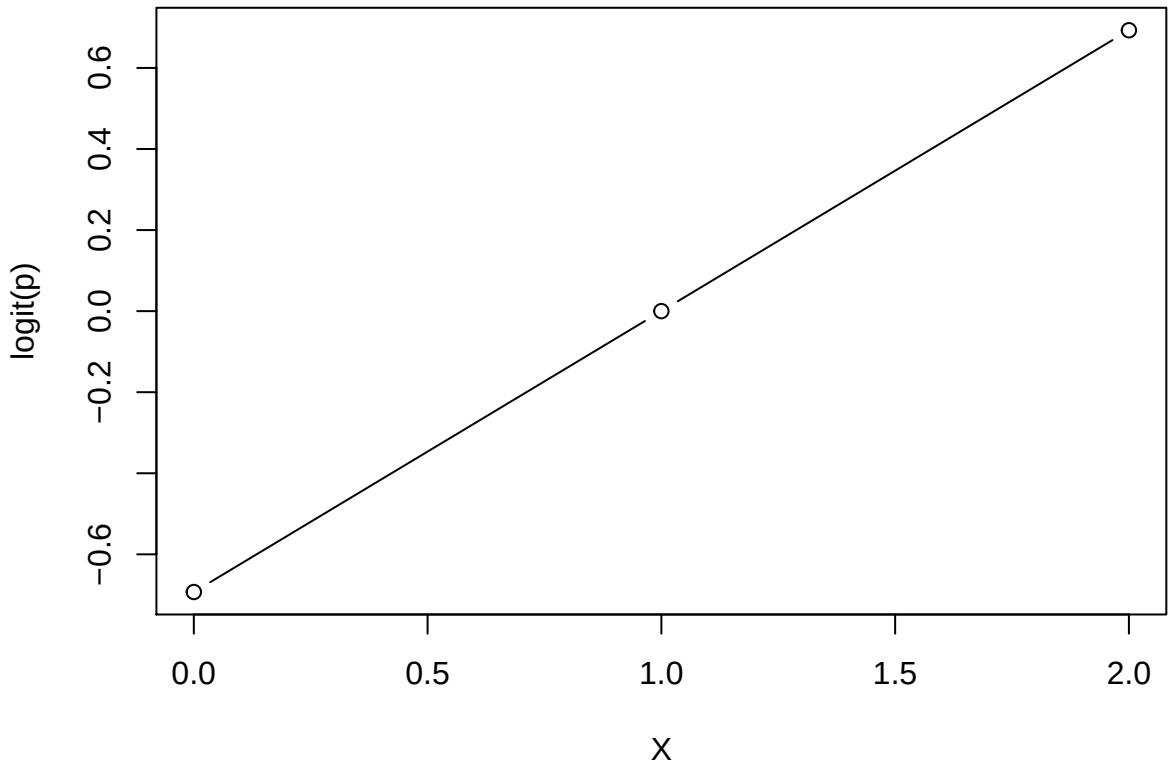
(E)

$$P[D = 1|X = 2] = \frac{e^{0.693}}{1 + e^{0.693}} = 0.667$$

We estimated the probability of D for individuals with X = 1 equals to 0.667. Because for 30 individuals who have X = 1, two-third of them have D = 1.0.677 is very similar to two-third.

(F)

X	p, probabilit of D	odds of D	log odds of D = logit(p)
0	0.333	0.5	-0.693
1	0.500	1.0	0.000
2	0.333	2.0	0.693



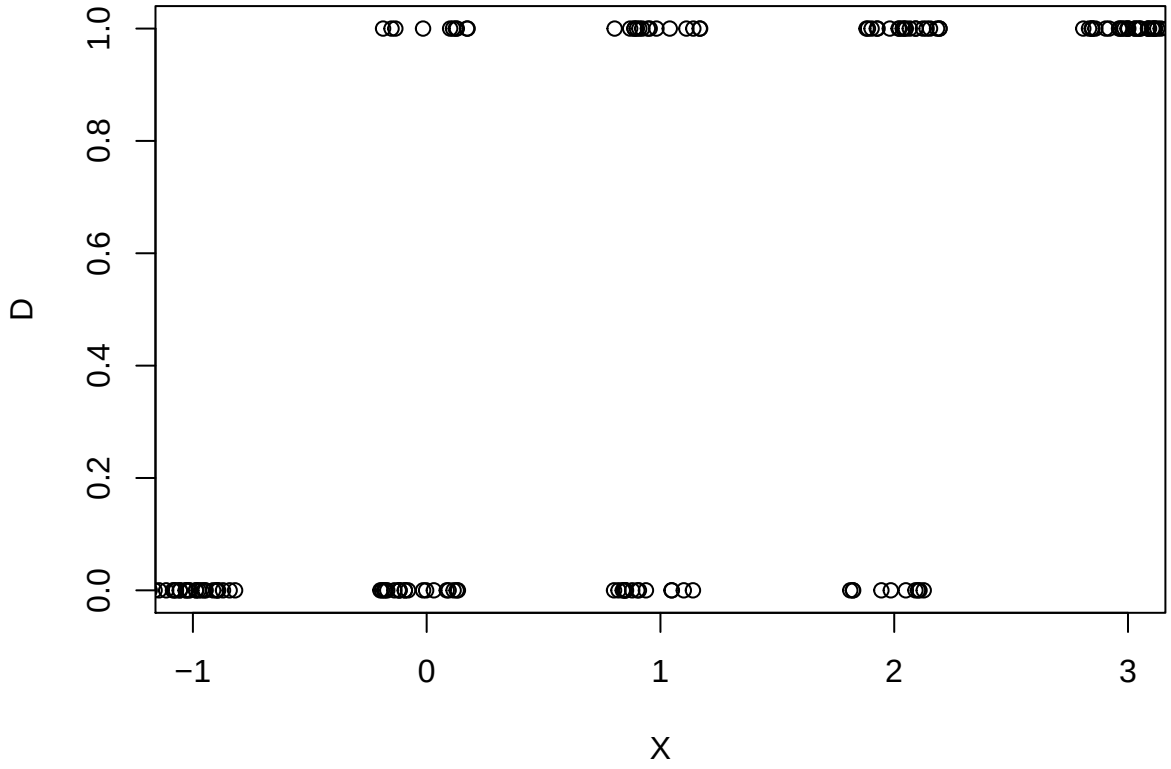
(G)

From the previous plot, we can see a clearly linare trend of logit(P) over X. When X becomes larger, logit(P) becomes larger.

5

From the simple logistic model, we notice that the estimated coeffiecients of the parameter from question 5 model are exactly the same as question 4. However,both the roubust standard error and the naive standard error seems to be much smaller in question 5 model compared to question 4 model. The confidence intervals of the two estiamted coeffecients are narrower compared to the question 4 model.

6



(A)

(B) We fitted the model as:

$$logit(P[D = 1|X]) = -1.35 + 1.35 \times X$$

(C)

$$P[D = 1|X = 0] = \frac{e^{-1.35}}{1 + e^{-1.35}} = 0.206$$

We estimated the probability of D for individuals with X = 0 equals to 0.206. It is different beacuase now the data contains more levels for X. For level with X = -1, there are 30 observations with all no diseases. Therefore, when we treat x as a continuous variable, our estiamtion of logit(P) will be dragged down by these extreme observations. **(D)**

$$P[D = 1|X = 1] = \frac{e^{0}}{1 + e^{0}} = 0.5$$

We estimated the probability of D for individuals with X = 1 equals to 0.5. Because for 30 individuals who have X = 1, half of them have D = 1. Although we have 30 no-disease observatiions when X = -1, we also have 30 observations who have disease with X = 3. The effect of 60 observations will conteract each other when X = 1 when we are fitting the linear models in logit(P).

Code Appendix

```
#---------------Set Up-------------------
knitr::opts_chunk$set(echo =FALSE,results= FALSE, warning = FALSE, message = FALSE, fig.height =5)
library(dplyr)
library(dagR)
library(uwIntroStats)
library(reshape)
library(knitr)
```

```r
options(digits = 3)
#--------------------Q1 Dag---------------------------
dag = dag.init(outcome = NULL, exposure = NULL,
               covs = 1,
               arcs = c(0,1,
                        -1,1),
               x.name = "Periconceptual nutritions supplementation", y.name = "Neural tube defects",
               cov.names = "Birth status",
               symbols = c("E","C","D"))
dag.draw(dag)
#--------------------------Q2 Dag---------------------------
dag2 = dag.init(outcome = NULL, exposure = NULL,
                covs = c(1,2),
                arcs = c(2,1,
                         0,2,
                         -1,2),
                x.name = "Lead poisoning", y.name = "Polycystic kidney disease",
                cov.names = c("Glomerular filtration rate","Kidney Failure"),
                symbols = c("E","A","B","D"))
dag.draw(dag2)


#-------------------------4A Scatter Plot----------------------
q4 = matrix(NA, ncol = 2, nrow = 90)
q4[1:30,1] = 0
q4[31:60,1] = 1
q4[61:90,1] = 2
q4[1:20,2] = 0
q4[21:30,2] = 1
q4[31:45,2] = 0
q4[46:60,2] = 1
q4[61:70,2] = 0
q4[71:90,2] = 1
plot(q4[,2]~jitter(q4[,1]),xlab = "X", ylab = "D", xlim = c(-1,3))
colnames(q4) = c("X","D")
q4 = data.frame(q4)
#--------------------4B Logistic Regression-------------------
logmol =  regress("odds", D~X, data = q4)
coef(logmol)
#----------------------Question 4 Table---------------------------
table4 = matrix(NA, ncol = 4, nrow = 3)
table4[,1] = c(0,1,2)
table4[,2] = c(0.333,0.500,0.333)
table4[,3] = c(0.5,1,2)
table4[,4] = c(-0.693,0,0.693)
colnames(table4) = c("X", "p, probabilit of D", "odds of D",
                     "log odds of D = logit(p)")
kable(table4)

#------------------Q4 plot----------------------------
plot(table4[,4]~X, data = table4, type="b",ylab = "logit(p)")
#-------------------Question 5--------------
dat5 = data.frame(D = c(0,1,0,1,0,1),
```

```
                  X = c(0,0,1,1,2,2),
                  freq = c(200,100,150,150,100,200))
dat5_exp = untable(dat5[,c("X","D")], num = dat5$freq)
dat5_mod = regress("odds", D~X, data = dat5_exp)
coef(dat5_mod)
coef(logmol)
#-----------------Question 6----------------------
dat6 =data.frame(X = c(-1, rep(c(0,1,2),each = 2), 3),
                 D = c(0, rep(c(0,1),3),1),
                 freq = c(30,20,10,15,15,10,20,30))
dat6_exp = untable(dat6[,c("X","D")], num = dat6$freq)
plot(D~jitter(X),xlab = "X", ylab = "D", xlim = c(-1,3),data = dat6_exp)
#--------------------Question6B Logistic Regression-------------------
q6mol =  regress("odds", D~X, data = dat6_exp)
coef(q6mol)
```