JOHN SCHOOF

Biostat/Epi 536 2020

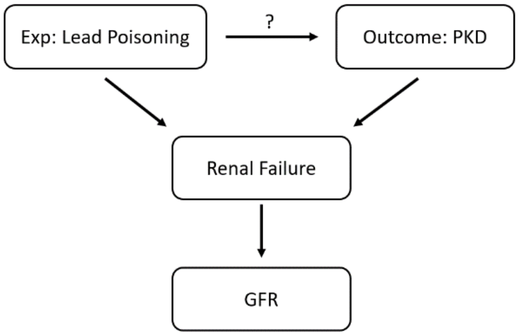
HW 3 (6 problems)

1. An investigator is planning a study of the effects of periconceptual nutritional supplementation (E) on neural tube defects (D). Fetuses developing with neural tube defects are more likely to result in spontaneous abortion or stillbirth (C=0) instead of live birth (C=1) compared to fetuses developing without these defects. Periconceptual nutritional supplementation is also believed to reduce the risk of spontaneous abortion or still birth through pathways independent of neural tube defects.

For convenience, the investigator is considering limiting his study to live births. Given the causal model described above, why might this bias the results?

This would cause index event bias, a type of collider stratification bias. Stillbirth is a collider of nutrition (exposure) and birth defects (outcome) and therefore we should not adjust for it. Live births are less likely to experience the outcome. This would result in spuriously low measure of excess risk.

2. Before it was understood that polycystic kidney disease (PKD) is a genetic disorder, Dr. Ott hypothesized that lead poisoning was a cause of PKD. In planning a study to collect evidence to study a possible effect of lead poisoning on PKD, Dr. Ott wonders whether glomerular filtration rate (GFR) is a confounder because prior work showed that GFR is associated with both lead poisoning and PKD. Suppose, in truth, lead poisoning is a cause of renal failure (the kidneys don't work as well as they should), affecting GFR. Similarly, PKD is a cause of kidney failure. Draw a DAG summarizing the information presented. Should Dr. Ott treat GFR as a confounder?



There is no need the adjust for GFR in this model. Renal failure is a collider, not GFR. And we do not want to adjust for Renal Failure because that would induce a relationship between lead poisoning and PKD whether one exists in reality or not.

3. Assume that P(D) in a population is 10%. Investigators plan to conduct a case-control study (unmatched) to study associations between D and a binary exposure E and also collect data on a binary covariate C. Their analysis model will be $\text{logit}(p) = \beta_0 + \beta_1 C + \beta_E E + \beta^* C \times E$.

A. The investigators will sample an equal number of cases and controls – 1:1 sampling. What is π, the ratio of sampling probabilities? What is the expected value of $\widehat{\beta_0}$? Write the expected value in terms of population parameters.

Pi = 1/(1/9) = 9
Beta0hat = beta0 + log(pi)

B. For every sampled case, investigators will sample 9 controls – 1:9 sampling. What is π, the ratio of sampling probabilities? What is the expected value of $\widehat{\beta_0}$? Write the expected value in terms of population parameters. Comment on whether this expected value surprises you, or if you can make sense of the difference from A.

Pi = 1/(8/9) = 1.13
Beta0hat = beta0 + log(pi)
The expected value is that same in both A and B. This is because beta0hat is not a population quantity. Its interpretation depends on the sampling method. Therefore, even with the same population, it can have different estimates.

4. Suppose you have data for 90 people on a continuous explanatory variable X and a binary outcome D, summarized in the following table:

X	**D**	**Frequency (number of people)**
0	no	20
0	yes	10
1	no	15
1	yes	15
2	no	10
2	yes	20

A. (part A is optional – you are encouraged to do Part A, but you can receive full credit without submitting part A) Although it is not very informative, make a scatterplot of X and D (coded as no=0 and yes=1). Do your best to make the scatterplot informative (for example, "jitter" to show overlapping points). Also, make the plot comparable with the plot you will make for Q6A (e.g. same axes scales).

B. Fit a simple logistic regression model of D on X. Write the fitted model.

$$\text{Logit }(\text{P}[D = 1|X = x]) = -0.693 + 0.693 * x$$

C. According to the fitted model, what is the probability of D for individuals with X=0? Why does this make sense?

$\text{Logit }(\text{P}[D = 1|X = 0]) = \beta_0 + \beta_1 * 0$
$log(odds) = -0.693 + 0.693 * 0 = -0.693$

$odds = e^{-0.693} = 0.5$
$probability = odds/(1 + odds) = 0.5/(1 + 0.5) = \mathbf{0.333}$

The probability of D for individuals with X=0 is 0.333. This makes sense because we see in our 2x3 table that 10 out of 30 people in the X=1 group are diseased. Therefore, the probability is also 0.333 if calculated using our table. The exact same as with using our logistic regression model.

D. According to the fitted model, what is the probability of D for individuals with X=1? Why does this make sense?

$\text{Logit}\ (\text{P}[D = 1|X = 1]) = \beta_0 + \beta_1 * 1$
$log(odds) = -0.693 + 0.693 * 1 = -0.693 + 0.693$
$odds = e^{-0.693+0.693} = 1$
$probability = odds/(1 + odds) = 1/(1 + 1) = 0.5$
The probability of D for individuals with X=1 is 0.5. This makes sense because we see in our 2x3 table that 15 out of 30 people in the X=1 group are diseased. Therefore the probability is also 0.5 if calculated using our table. The exact same as with using our logistic regression model.

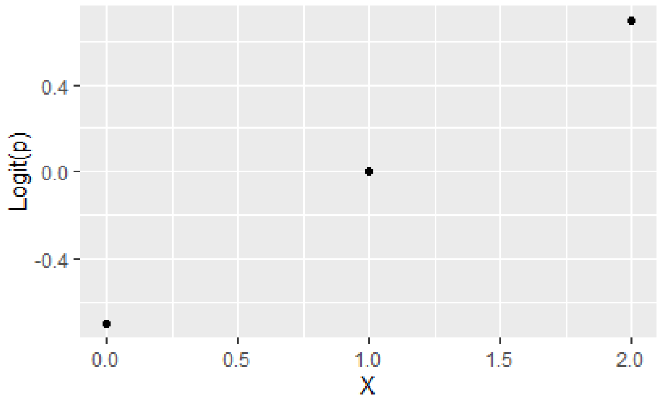
E. According to the fitted model, what is the probability of D for individuals with X=2? Why does this make sense?

$\text{Logit}\ (\text{P}[D = 1|X = 2]) = \beta_0 + \beta_1 * 2$
$log(odds) = -0.693 + 0.693 * 2 = -0.693 + 1.386$
$odds = e^{-0.693+1.386} = 2$
$probability = odds/(1 + odds) = 2/(1 + 2) = 0.667$
The probability of D for individuals with X=2 is 0.666. This makes sense because we see in our 2x3 table that 20 out of 30 people in the X=1 group are diseased. Therefore the probability is also 0.666 if calculated using our table. The exact same as with using our logistic regression model.

F. For each of X=0,1,2, estimate p, the probability of D, using the table above (not the logistic model). Complete the following table:

X	**p, probability of D**	**odds of D**	**log odds of D = logit(p)**
0	1/3	1/2	-0.69
1	½	1	0
2	2/3	2	0.69

G. Plot logit(p) against X. What do you notice?



The relationship between these points appears linear with an upward trend.

5. Suppose you have data for 900 people on a continuous explanatory variable X and a binary outcome D, summarized in the following table:

X	D	Frequency (number of people)
0	no	200
0	yes	100
1	no	150
1	yes	150
2	no	100
2	yes	200

Fit a simple logistic model to these data. Compare and contrast the results with your results from Q4. Your comparison should include regression parameter estimates, standard errors, and confidence intervals.

$$\text{Logit}\ (\text{P}[D = 1|X = x]) = -0.693 + 0.693 * x$$

The Q5 unadjusted logistic regression model that fit is shown above. Note that the point estimates for the coefficients are exactly the same as in the Q4 model. The standard errors, confidence intervals, and p values differ in the two models. In the Q5 model, β_1 is 0.693 with a 95% confidence interval (calculated using robust standard errors) of 0.693-0.523, and our β_1 in the Q4 model has a 95% confidence interval of 0.693-0.156. The Q5 model has much smaller standard errors, resulitng in the narrower confidence intervals that were just mentioned. The p value for the β_1 estimate in the Q5 model is also much smaller than that of the Q4 model. The p value for β_1 in the Q4 model is 0.011 compared to $1.2 * 10^{15}$.

6. Suppose you have data for 150 people on an explanatory, continuous variable X and a binary outcome D, summarized in the following table:

X	D	Frequency (number of people)
-1	no	30

0	no	20
0	yes	10
1	no	15
1	yes	15
2	no	10
2	yes	20
3	yes	30

A. (part A is optional – you are encouraged to do Part A, but you can receive full credit without submitting part A) Although it is not very informative, make a scatterplot of X and D (using standard 0/1 coding as in Q4). Do your best to make the scatterplot informative (for example, "jitter" to show overlapping points). Also, make your plot comparable with the plot from Q4A.

B. Fit a simple logistic regression model of D on X. Write the fitted model.

$$\text{Logit}\ (\text{P}[D = 1|X = x]) = -1.346 + 1.346 * x$$

C. According to the fitted model, what is the probability of D for individuals with X=0? Why is this different from Q3C?

$\text{Logit}\ (\text{P}[D = 1|X = 0]) = \beta_0 + \beta_1 * 0$

$log(odds) = -1.346 + 1.346 * 0 = -1.346$

$odds = e^{-1.346} = 0.26$

$probability = odds/(1 + odds) = 0.26/(1 + 0.26) = 0.206$

This model actually estimates a coefficient that results in a probability of 0.206. This slightly less than 1/3, our answer from Q4c. This makes sense because the model is fitting a logistic curve based on all of the data and with new data saying that the -1s have a probability of zero of being diseased the curve we fit will be pulled down at lower values of x. It is basically saying that lower values of x are associated with less disease.

D. According to the fitted model, what is the probability of D for individuals with X=1? How does this compare with Q3D? Why does this make sense?

$\text{Logit}\ (\text{P}[D = 1|X = 1]) = \beta_0 + \beta_1 * 1$

$log(odds) = -1.346 + 1.346 * 1 = -1.346 + 1.346$

$odds = e^{-1.346+1.346} = 1$

$probability = odds/(1 + odds) = 1/(1 + 1) = 0.5$

The probability is 0.5, exactly the same as in Q4d. This makes sense because the data for X=-1 and X=3 does not change the center of our logistic curve. The logistic curve still passed through this point.