BIOST 536: Homework 3

Department of Biostatistics @ University of Washington

October 19, 2024

Designing a Descriptive Table

1. Your team has conducted a case-control study to examine the evidence that a (binary) exposure affects a disease, and you have analyzed study data. You are beginning to write a scientific article on your findings. The first table in your article will describe study participants by sharing descriptive statistics for key variables. How would you construct your table? Specifically, would you compute and report descriptive statistics on the whole sample (no stratification)? Or would you stratify on some variable and, if so, what variable? Write a short paragraph to explain your choice.

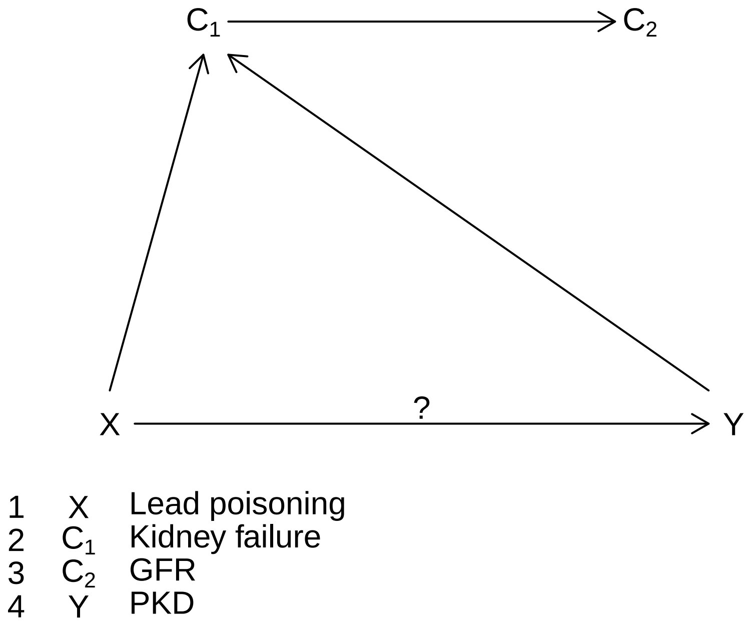
   If it helps you to be more concrete, assume that D is a type of cancer, the exposure is a genetic variant, and the hypothesis is that having one or more copies of the genetic variant increases a person's risk of the cancer.

Characteristic	**Overall**, N = 10,000	**Disease**, N = 5,000	**No disease**, N = 5,000
Exposure, E			
Exposed	3,548 (35%)	2,283 (46%)	1,265 (25%)
Unexposed	6,452 (65%)	2,717 (54%)	3,735 (75%)
Covariate, C	0.66 (0.11, 1.16)	0.94 (0.48, 1.35)	0.33 (-0.08, 0.83)

Designing a descriptive table in this example, I would provide overall and stratified variable summaries by cases and controls. Stratification serves to illustrate characteristics of the two outcome populations, while overall summaries allow researchers to briefly inspect how the case and control populations are reflected in the whole population. See Table 1 for an example.

Designing a DAG

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?



I am unsure, but I believe we would not need to adjust for confounding because GFR is not a cause of PKD.

Case-Control Parameterization

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 $logit(p) = \beta_0 + \beta_1 \cdot C + \beta_E \cdot E + \beta^* \cdot C \times E$

(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 $\hat{\beta}_0$? Write the expected value in terms of population parameters.

In a case-control setting, let D denote a case $(D = 1)$ and control $(D = 0)$, and Z denote a sampled $(Z = 1)$

and unsampled ($Z = 0$) event. Then, π defines a measure of how much more likely a case is to be sampled than a control:

$$\pi = \frac{P(Z{=}1|D{=}1)}{P(Z{=}1|D{=}0)} = \frac{\frac{P(Z=1\cap D=1)}{P(D=1)}}{\frac{P(Z=1\cap D=0)}{P(D=0)}} = \frac{|Z{=}1\cap D{=}1|}{|Z{=}1\cap D{=}0|} \cdot \frac{P(D{=}0)}{P(D{=}1)}$$

And, from the logistic regression equation above, β_0 is estimated by $\hat{\beta}_0 = log(\pi) + \beta_0^*$, where β_0^* is the population log odds of disease for unexposed with covariate $C = 0$.

In this study, $\pi = 1 \cdot \frac{0.9}{0.1} = 9$; therefore, $\hat{\beta}_0 = 0.954 + \beta_0^*$.

(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 $\hat{\beta}_0$? 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).

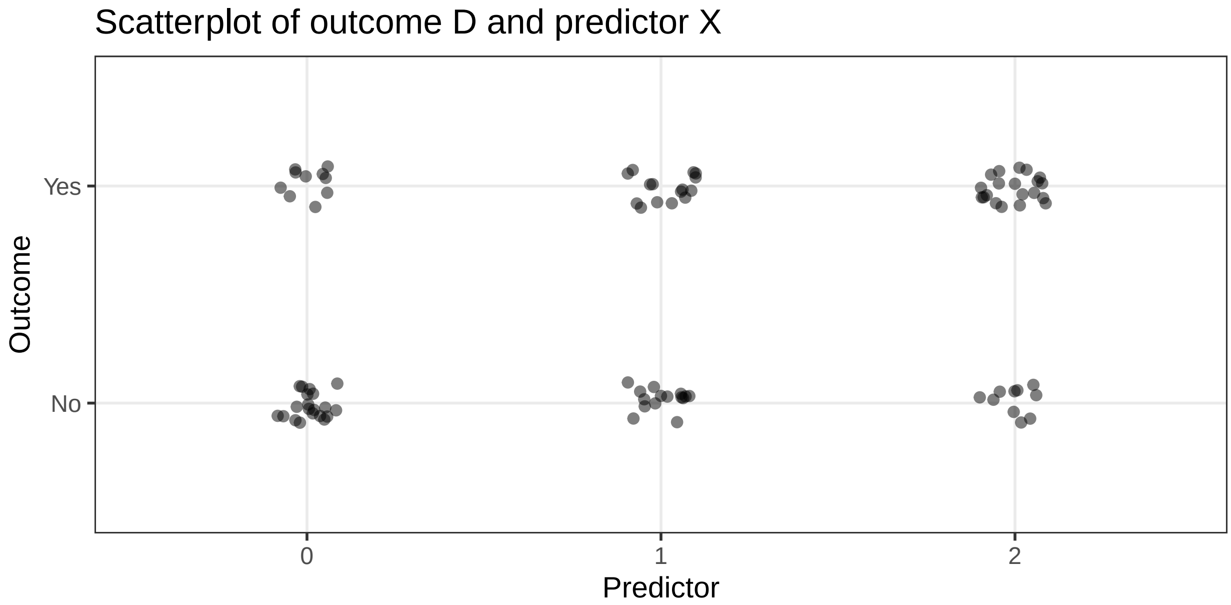
In this study, $\pi = \frac{1}{9} \cdot \frac{0.9}{0.1} = 1$; therefore, $\hat{\beta}_0 = 0 + \beta_0^*$. This is expected because we no longer over-representing cases in our study and are able to estimate parameters of the population.

Logistic Regression Practice

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

(a) (Optional) 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 Question 6a (e.g. same axes scales).



(b) Fit a simple logistic regression model of D on X. Write the fitted model.

$lo\hat{g}it(p) =$ -0.693 \$ + \$ 0.693 $\cdot X$ is our model of the log odds of disease from some continuous variable X.

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

From $x = logit(p) = log\frac{p}{1-p}$ we derive $p = expit(x) = \frac{exp(x)}{1+exp(x)}$ where p denotes $P(D = 1|X = x)$.

$\hat{P}(D = 1|X = 0) = 0.333$. Low odds of disease within this group is expected because the majority of their sample did not have disease.

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

$\hat{P}(D = 1|X = 1) = 0.5$. This is reasonable because this subsample was evenly split diseased and not diseased.

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

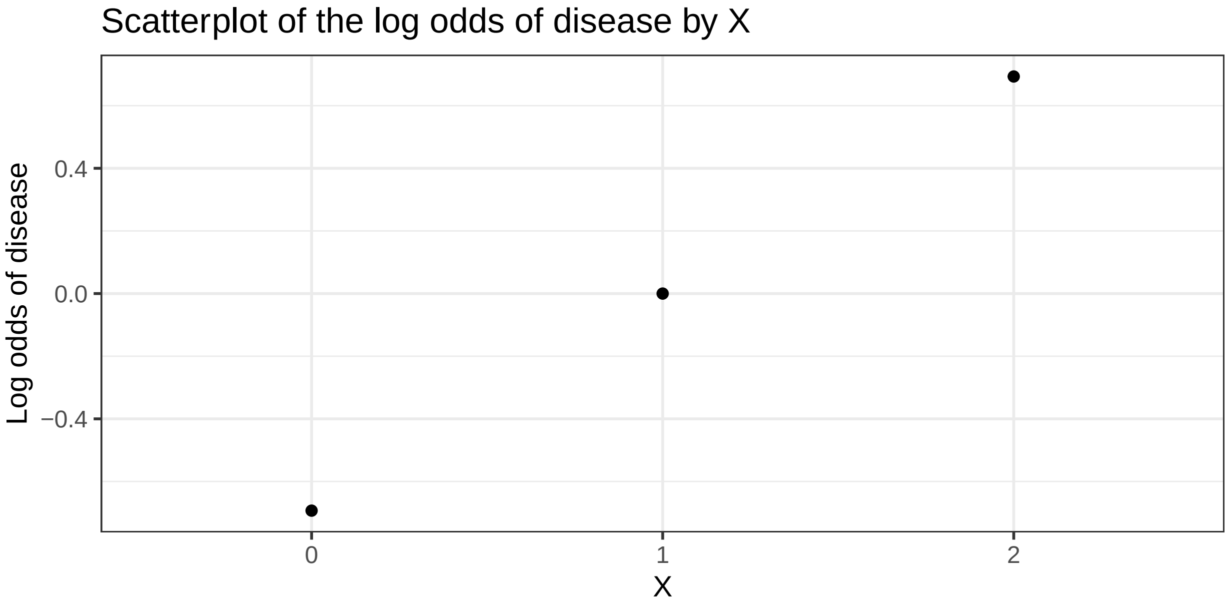
$\hat{P}(D = 1|X = 2) = 0.667$. This is reasonable because this subsample was mostly diseased.

(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:

Table 3: Estimation table

X	P(D)	odds(D)	log odds(D)
0	0.333	0.5	-0.693
1	0.500	1.0	0.000
2	0.667	2.0	0.693

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



The log odds across X are linear. This is because the X=0 and X=2 groups have inverted odds of disease and the X=1 group has 1:1 odds of disease.

5. Suppose you have data for 900 people on a continuous explanatory variable X and a binary outcome D, summarized in Table 5 below.

   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.

Table 4: Estimation table

X	D	Frequency
0	no	200
0	yes	100
1	no	150
1	yes	150
2	no	100
2	yes	200

X	P(D)	odds(D)	log odds(D)
0	0.333	0.5	-0.693
1	0.500	1.0	0.000
2	0.667	2.0	0.693

The models from Questions 4 and 5 were fit to very similar data, except that data from the latter was fit from duplicate sets of data from the first. Fitting a model with more of the same data will produce more precise but identical estimates. Then, it is reasonable to find in Table 6 that the two models have the same estimated odds of disease, with the latter model estimates producing smaller standard errors and more narrow confidence intervals.

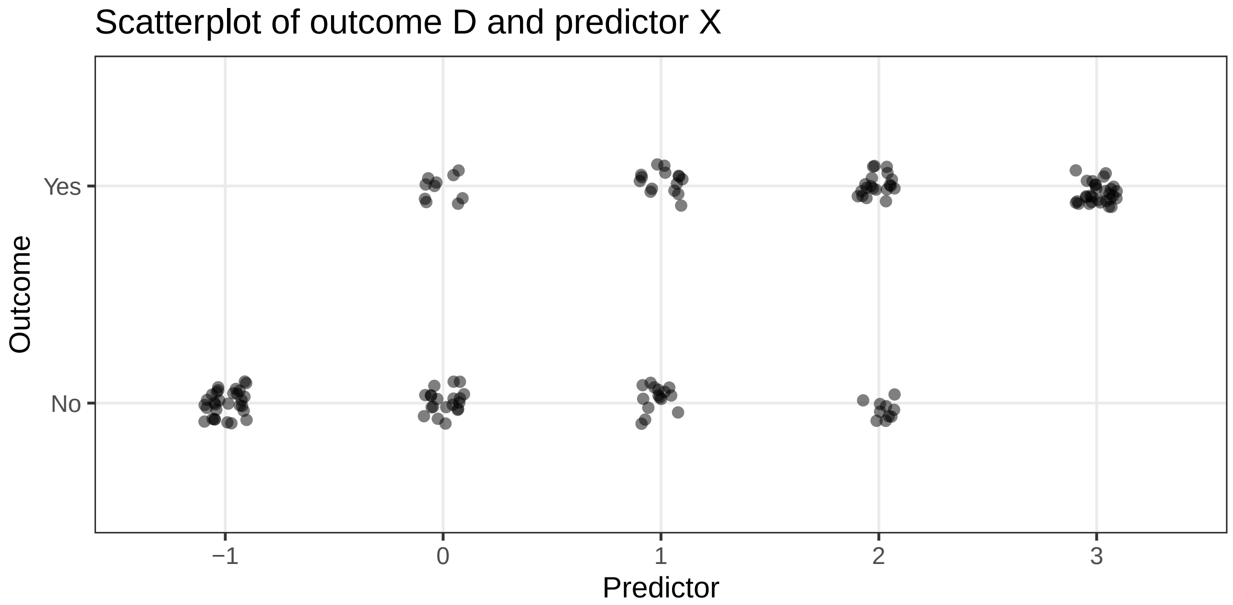
Table 6: Coefficient comparison table

Model	Term	Estimate	Std.Error	Conf.Low	Conf.High
Question 4	(Intercept)	0.5	0.351	0.245	0.978
	X	2.0	0.274	1.184	3.488
Question 5	(Intercept)	0.5	0.111	0.401	0.620
	X	2.0	0.087	1.690	2.374

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
-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) (Optional) 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.

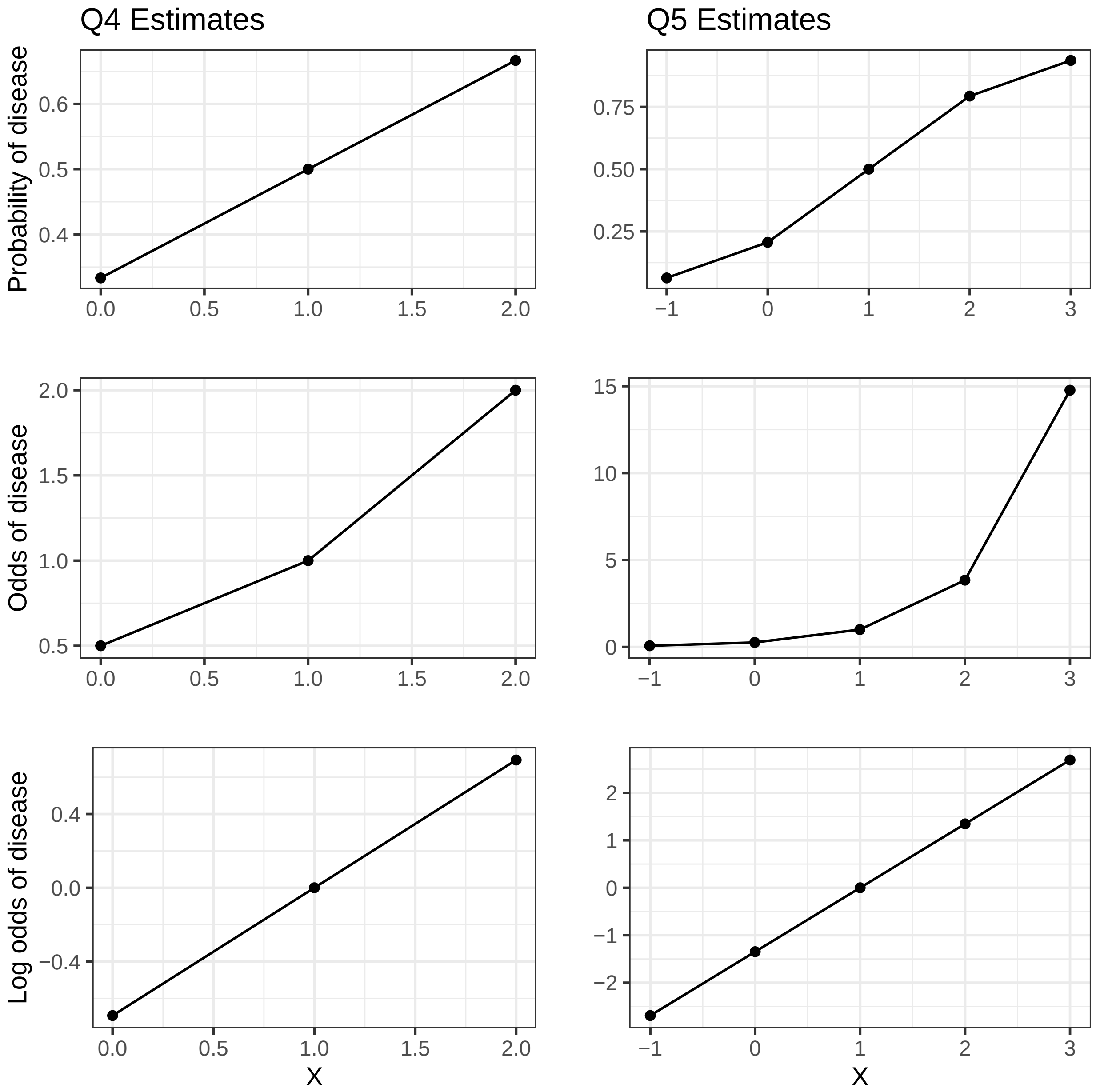
Table 8: Estimation table

	X	P(D)	odds(D)	log odds(D)
1	-1	0.063	0.068	-2.69
2	0	0.206	0.260	-1.35
4	1	0.500	1.000	0.00
6	2	0.794	3.843	1.35
8	3	0.937	14.767	2.69

$\hat{logit}(p) =$ -1.346 $ + $ 1.346 $\cdot X$ is our model of the log odds of disease from some continuous variable X.

(c) According to the fitted model, what is the probability of D for individuals with X=0? Why is this different from Question 4 (c)?

$\hat{P}(D = 1|X = 0) = 0.206$. Logistic regression is fit from maximum likelihood estimation (MLE), which tries to best estimate the odds of all subpopulations. See in Figure **this one** that the model best fits the probability, odds, and log odds of disease for each subpopulation.



(d) According to the fitted model, what is the probability of D for individuals with X=1? How does this compare with Q4D? Why does this make sense?

$\hat{P}(D = 1|X = 1) = 0.5$

Answer here

End of report. Code appendix begins on the next page.

Code Appendix

```r
# clear environment
rm(list=ls())

# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-')
ndigits = 3
options(digits = ndigits)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# load relevant packages
library(dplyr)      # data frame manipulation
library(ggplot2)    # plotting
library(gtsummary)  # table summaries
library(dagR)       # DAG simulation
library(broom)      # model coefficient table
library(knitr)      # pretty tables
library(tidyr)      # row replication
library(gridExtra)  # grid of plots
# compute the logit function for probability p
# returns log odds of p
logit <- function(p) {
  logit = log(p / (1-p))
  return(logit)
}

# compute expit function for x, the log odds of p
# returns probability p
expit <- function(x) {
  expit = exp(x) / (1 + exp(x))
  return(expit)
}

## ===============
## Question 1
## ===============
# simulate example cancer data
set.seed(1)
n <- 5000
pE <- 0.15  # (baseline) probability of exposure
# create example data
data <- data.frame(D = sample(rep(0:1, n)))   # 1:1 case-control
data$E <- rbinom(n, size=1, p = pE+(0.4*data$D))   # P(E|case) > P(E|control)
data$C <- rnorm(n, mean = pE+(1*data$D), sd = 0.5) # mean(C|case) > mean(E|ctrl)

data %>%
  dplyr::mutate(D = ifelse(D==1, "Disease", "No disease"),
                E = ifelse(E==1, "Exposed", "Unexposed"),) %>%
  gtsummary::tbl_summary(by = D,
                         label = list(E~"Exposure, E",
                                      C~"Covariate, C")) %>%
  add_overall() %>%
  bold_labels()

## ===============
## Question 2
## ===============
dag <- dagR::dag.init(covs = c(1,1), arcs = c(0, 1, 1, 2, 3, 1),
                      y.name = "PKD", x.name = "Lead poisoning",
                      cov.names = list("Kidney failure","GFR")) %>%
  dag.draw()

## ===============
## Question 4
## ===============
q4data <- data.frame(X = c(0,0,1,1,2,2),
                     D = rep(c("no","yes"),3),
                     Frequency = c(20,10,15,15,10,20))
knitr::kable(q4data)

q4data <- q4data %>% tidyr::uncount(Frequency)
q4data %>%
  mutate(Outcome = ifelse(D=="yes", "Yes", "No"),
         Predictor = as.factor(X)) %>%
  ggplot(mapping = aes(x = Predictor, y = Outcome)) +
  geom_jitter(height = 0.1, width = 0.1, alpha = 0.5) +
  labs(title = "Scatterplot of outcome D and predictor X") + theme_bw()
# fit simple logistic regression
mod.lr <- glm(as.factor(D)~X, family=binomial(), q4data)
coefs <- mod.lr$coefficients
# predictions of P(D) across values of X
q4_probs <- unique(mod.lr$fitted.values)
# create estimation table
q4res <- q4data %>%
  count(X, D, name="Frequency") %>%
  # group by X values, to make sum(Frequency) work in the next line
  group_by(X) %>%
  mutate(pD = Frequency/sum(Frequency),  # compute #(D=d,X=x) / #(X=x)
         oddsD = pD / (1-pD),
         logoddsD = log(oddsD)) %>%
  subset(D == "yes", select = -c(D, Frequency))

q4res %>% kable(col.names = c('X', 'P(D)', 'odds(D)', 'log odds(D)'),
                caption = "Estimation table")
ggplot(q4res, aes(as.factor(X), logoddsD)) +
  geom_point() +
  labs(title = "Scatterplot of the log odds of disease by X") +
  ylab("Log odds of disease") + xlab("X") + theme_bw()

## ===============
## Question 5
## ===============
q5data <- data.frame(X = c(0,0,1,1,2,2),
                     D = rep(c("no","yes"),3),
                     Frequency = 10*c(20,10,15,15,10,20))
kable(q5data)
q5data <- q5data %>% tidyr::uncount(Frequency)
# fit simple logistic regression
mod.lr2 <- glm(as.factor(D)~X, family=binomial(), q5data)
# predictions of P(D) across values of X
q5_probs <- unique(mod.lr2$fitted.values)

# create estimation table
q5res <- q5data %>%
  count(X, D, name="Frequency") %>%
  # group by X values, to make sum(Frequency) work in the next line
  group_by(X) %>%
  mutate(pD = Frequency/sum(Frequency),  # compute #(D=d,X=x) / #(X=x)
         oddsD = pD / (1-pD),
         logoddsD = log(oddsD)) %>%
  subset(D == "yes", select = -c(D, Frequency))

q5res %>% kable(col.names = c('X', 'P(D)', 'odds(D)', 'log odds(D)'),
                caption = "Estimation table")
# bind rows of models from Questions 4 and 5
rbind(
  mod.lr %>%
    broom::tidy(exponentiate = TRUE, conf.int = TRUE) %>%
    dplyr::select(-c(p.value, statistic)) %>%
    dplyr::mutate_if(is.numeric, round, ndigits),
  mod.lr2 %>%
    tidy(exponentiate = TRUE, conf.int = TRUE) %>%
    select(-c(p.value, statistic)) %>%
    mutate_if(is.numeric, round, ndigits)
  ) %>%
  # add a new column denoting model source
  mutate(model = c("Question 4", "", "Question 5", "")) %>%
  # place it first
  select(model, everything()) %>%
  # and rename columns for pretty print
  kable(caption = "Coefficient comparison table",
        col.names = c("Model", "Term", "Estimate",
                      "Std.Error", "Conf.Low", "Conf.High"))

## ===============
## Question 6
## ===============
q6data <- data.frame(X = c(-1,0,0,1,1,2,2,3),
                     D = c("no", rep(c("no","yes"),3), "yes"),
                     Frequency = c(30,20,10,15,15,10,20,30))
kable(q6data)
q6data <- q6data %>% tidyr::uncount(Frequency)
q6data %>%
  mutate(Outcome = ifelse(D=="yes", "Yes", "No"),
         Predictor = as.factor(X)) %>%
  ggplot(mapping = aes(x = Predictor, y = Outcome)) +
  geom_jitter(height = 0.1, width = 0.1, alpha = 0.5) +
  labs(title = "Scatterplot of outcome D and predictor X") + theme_bw()
# fit simple logistic regression
mod.lr3 <- glm(as.factor(D)~X, family=binomial(), q6data)
coefs3 <- mod.lr3$coefficients

# predictions of P(D) across values of X
q6_probs <- unique(mod.lr3$fitted.values)

# create estimation table
q6res <- q6data %>%
  count(X, D, name="Frequency") %>%  # collapse rows
  select(X) %>% unique() %>%  # select unique X rows
  # insert and calculate estimates
  mutate(pD = q6_probs,
         oddsD = pD / (1-pD),
         logoddsD = log(oddsD))

q6res %>% kable(col.names = c('X', 'P(D)', 'odds(D)', 'log odds(D)'),
                caption = "Estimation table")
gg <- ggplot(q4res, aes(X, pD)) +
  geom_point() + geom_line() +
  labs(title = "Q4 Estimates") + xlab("") +
  ylab("Probability of disease") + theme_bw()

gg2 <- ggplot(q4res, aes(X, oddsD)) +
  geom_point() + geom_line() +
  # labs(title = "Odds of disease by X") +
  ylab("Odds of disease") + xlab("") + theme_bw()

gg3 <- ggplot(q4res, aes(X, logoddsD)) +
  geom_point() + geom_line() +
  # labs(title = "Log odds of disease by X") +
  ylab("Log odds of disease") + xlab("X") + theme_bw()

gg4 <- ggplot(q6res, aes(X, pD)) +
  geom_point() + geom_line() +
  labs(title = "Q5 Estimates") +
  ylab("") +  xlab("") + theme_bw()

gg5 <- ggplot(q6res, aes(X, oddsD)) +
  geom_point() + geom_line() +
  # labs(title = "Odds of disease by X") +
  ylab("") +  xlab("") + theme_bw()

gg6 <- ggplot(q6res, aes(X, logoddsD)) +
  geom_point() + geom_line() +
  # labs(title = "Log odds of disease by X") +
  ylab("") +  xlab("X") + theme_bw()

gridExtra::grid.arrange(gg, gg4, gg2, gg5, gg3, gg6)
```

End of document.