Homework 6 - Drew Davison

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
New names:
Rows: 147 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (8): ...1, female, age, highstatus, yrsmoke, cigsday, bird, cancer

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Data: Association Between Bird-Keeping and Risk of Lung Cancer

A 1972-1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague. They identified 49 cases of lung cancer among patients who were registered with a general practice, who were age 65 or younger, and who had resided in the city since 1965. Each patient (case) with cancer was matched with two control subjects (without cancer) by age and sex. Further details can be found in Holst, Kromhout, and Brand (1988).

Age, sex, and smoking history are all known to be associated with lung cancer incidence. Thus, researchers wished to determine after age, sex, socioeconomic status, and smoking have

been controlled for, is an additional risk associated with birdkeeping? The data (Ramsey and Schafer 2002) is found in [`birdkeeping.csv`(data/birdkeeping.csv)].[1]

The paper that this exercise is based upon can be found here and please read it before completing the assignment. (https://www.bmj.com/content/bmj/297/6659/1319.full.pdf)

Exercise 1

Part a

Create a segmented bar chart and appropriate table of proportions showing the relationship between birdkeeping and cancer diagnosis. Summarize the relationship in 1 - 2 sentences.

Note

```
# Create a table of counts
count_table <- table(birds$bird, birds$cancer)

# Convert to proportions by row (birdkeeping)
prop_table <- prop.table(count_table, margin = 1)
print("Proportion Table:")
```

```
[1] "Proportion Table:"
```

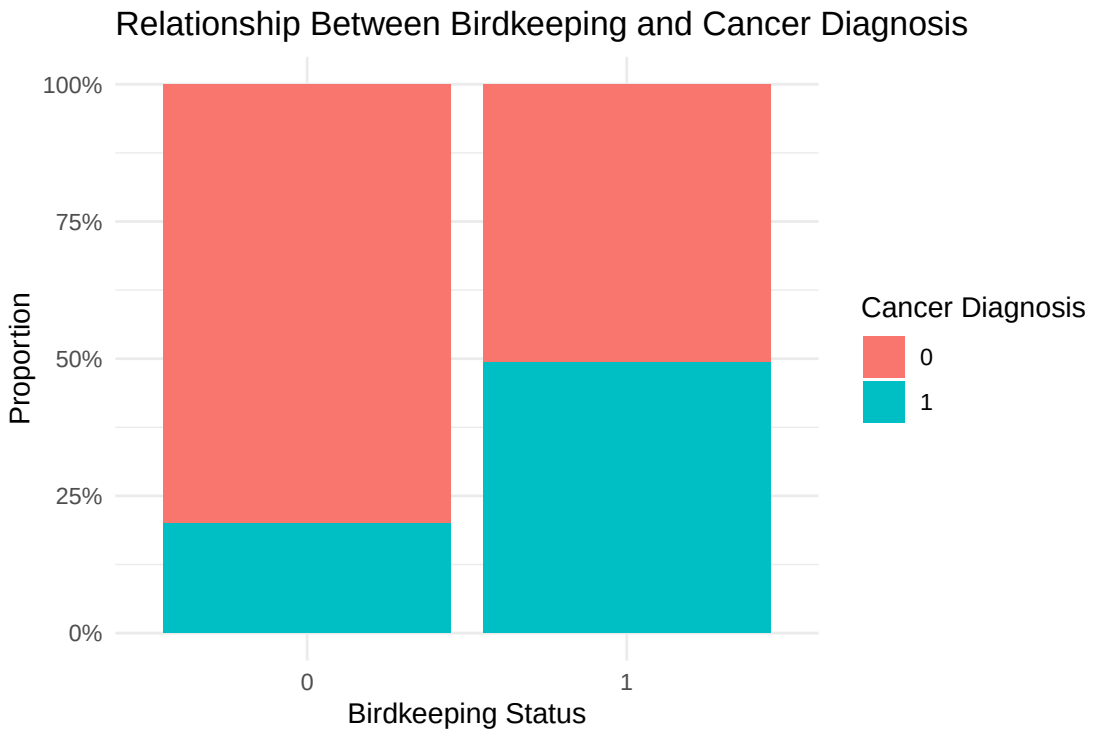
```
print(round(prop_table, 2))
```

```
     0    1
0 0.80 0.20
1 0.51 0.49
```

[1] This problem is adapted from Section 6.8.1, Ex 4.

```r
# Convert to a data frame for ggplot
df_plot <- as.data.frame(count_table)
colnames(df_plot) <- c("Birdkeeping", "Cancer", "Count")

# Plot segmented bar chart
ggplot(df_plot, aes(x = Birdkeeping, y = Count, fill = Cancer)) +
  geom_bar(stat = "identity", position = "fill") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Relationship Between Birdkeeping and Cancer Diagnosis",
    y = "Proportion",
    x = "Birdkeeping Status",
    fill = "Cancer Diagnosis"
  ) +
  theme_minimal()
```



The proportion table reveals:
0-0: 80% of participants without birdkeeping did not have lung cancer. 0-1: 20% of participants without birdkeeping did have lung cancer. 1-0: 51% of participants with birdkeeping did not have lung cancer. 1-1: 49% of participants with birdkeeping did have lung cancer.

It appears that the relationship between birdkeeping and lung cancer is, for the given sample, that there is a higher percentage of people with lung cancer that keep birds than for those who do not keep birds.

Part b

Calculate the unadjusted odds ratio of a lung cancer diagnosis comparing birdkeepers to non-birdkeepers. Interpret this odds ratio in context. (Note: an unadjusted odds ratio is found by not controlling for any other variables.)

Note

```
# Create a 2x2 table for birdkeeping and lung cancer
table_lung <- table(birds$bird, birds$cancer)

# View the table
print(table_lung)
```

```
   0  1
0 64 16
1 34 33
```

```
# Calculate the odds ratio manually
# OR = (a/c) / (b/d) = (exposed cases / exposed non-cases) / (unexposed cases / unexposed n

# Extract counts
a <- table_lung["1", "1"]        # birdkeepers with lung cancer
b <- table_lung["0", "1"]         # non-birdkeepers with lung cancer
c <- table_lung["1", "0"]    # birdkeepers without lung cancer
d <- table_lung["0", "0"]     # non-birdkeepers without lung cancer

# Calculate odds ratio
odds_ratio <- (a / c) / (b / d)
print(paste("Unadjusted Odds Ratio:", round(odds_ratio, 2)))
```

```
[1] "Unadjusted Odds Ratio: 3.88"
```

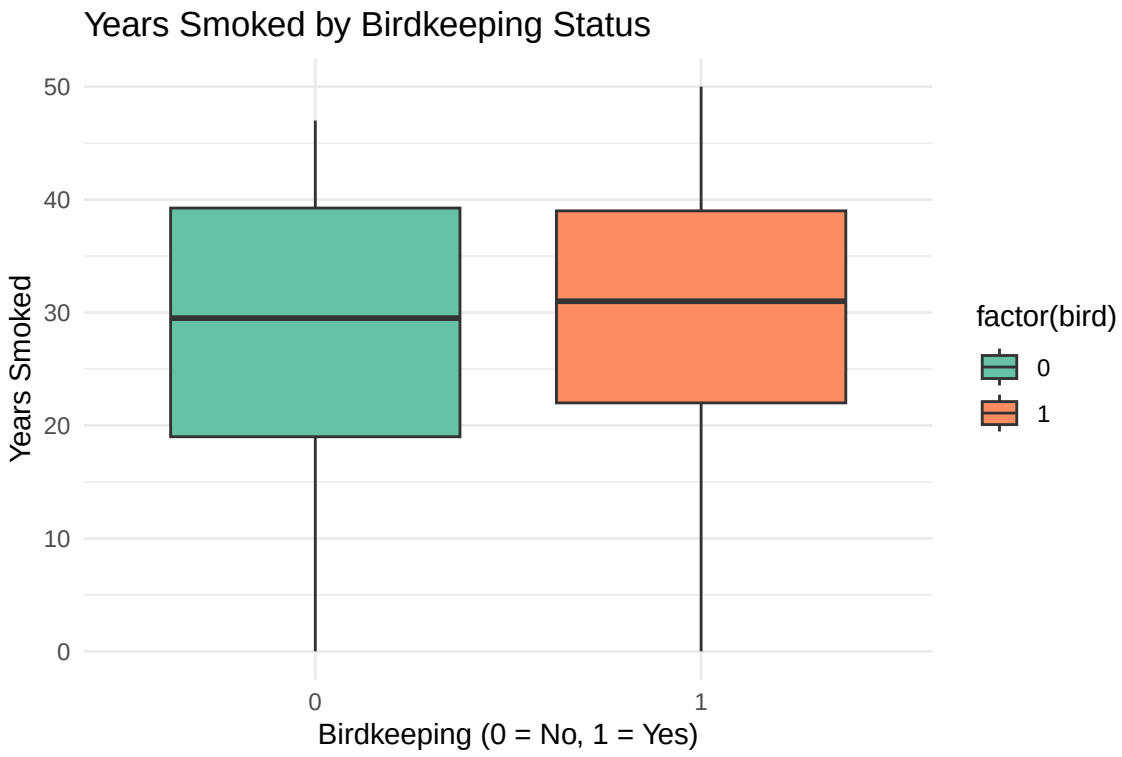
The unadjusted odds ratio of 3.88 indicates that birdkeepers had 3.88 times the odds of being diagnosed with lung cancer compared to non-birdkeepers. If the OR is greater than 1, this suggests a positive association between birdkeeping and lung cancer. If less

than 1, it suggests a negative association This is a crude measure and does not account for potential confounders such as smoking or age.

Part c

Does there appear to be an interaction between number of years smoked and whether the subject keeps a bird? Demonstrate with an appropriate plot and briefly explain your response.

```
ggplot(birds, aes(x = factor(bird), y = yrsmoke, fill = factor(bird))) +
  geom_boxplot() +
  labs(
    title = "Years Smoked by Birdkeeping Status",
    x = "Birdkeeping (0 = No, 1 = Yes)",
    y = "Years Smoked"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#66c2a5", "#fc8d62"))
```



The median smoking years for those who keep birds is slightly higher than those who do not,

however, the IQR is smaller for those who keep birds. The 25th percentile for those who keep birds is a greater number of years than those who do not keep birds. The 75th percentile for those who keep birds is a fewer number of years than those who do not keep birds.

In conclusion, it does not appear that there is a meaningful difference in years smoked based on if someone keeps birds.

Before answering the next questions, fit logistic regression models in R with cancer as the response and the following sets of explanatory variables:

- model1 = age, yrsmoke, cigsday, female, highstatus, bird
- model2 = yrsmoke, cigsday, highstatus, bird
- model3 = yrsmoke, bird
- model4 = yrsmoke, bird, yrsmoke:bird

```
# Model 1: age, yrsmoke, cigsday, female, highstatus, bird
model1 <- glm(cancer ~ age + yrsmoke + cigsday + female + highstatus + bird,
              data = birds,
              family = binomial)

# Model 2: yrsmoke, cigsday, highstatus, bird
model2 <- glm(cancer ~ yrsmoke + cigsday + highstatus + bird,
              data = birds,
              family = binomial)

# Model 3: yrsmoke, bird
model3 <- glm(cancer ~ yrsmoke + bird,
              data = birds,
              family = binomial)

# Model 4: yrsmoke, bird, and interaction between yrsmoke and bird
model4 <- glm(cancer ~ yrsmoke * bird,
              data = birds,
              family = binomial)

# Summary of all models
tidy(model1)
```

```
# A tibble: 7 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -1.94      1.80      -1.07  0.283
2 age          -0.0398    0.0355    -1.12  0.263
```

```
3 yrsmoke       0.0729    0.0265     2.75  0.00594
4 cigsday       0.0260    0.0255     1.02  0.308
5 female        0.561     0.531      1.06  0.291
6 highstatus    0.105     0.469      0.225 0.822
7 bird          1.36      0.411      3.31  0.000923
```

```
tidy(model2)
```

```
# A tibble: 5 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)  -3.38      0.708     -4.78  0.00000177
2 yrsmoke       0.0491    0.0188     2.62  0.00891
3 cigsday       0.0286    0.0244     1.17  0.241
4 highstatus   -0.0689    0.453     -0.152 0.879
5 bird          1.49      0.403      3.69  0.000223
```

```
tidy(model3)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)  -3.18      0.636     -5.00 0.000000582
2 yrsmoke       0.0582    0.0168     3.46 0.000544
3 bird          1.48      0.396      3.73 0.000194
```

```
tidy(model4)
```

```
# A tibble: 4 x 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -3.00       0.898     -3.34  0.000844
2 yrsmoke       0.0528     0.0256     2.06  0.0394
3 bird          1.18       1.15       1.03  0.304
4 yrsmoke:bird  0.00930    0.0340     0.274 0.784
```

Part d

Is there evidence that we can remove age and female from our model? Perform an appropriate test comparing model1 to model2; give a test statistic and p-value, and state a conclusion in context.

Note

```
# Perform likelihood ratio test comparing Model 1 and Model 2
anova(model2, model1, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: cancer ~ yrsmoke + cigsday + highstatus + bird
Model 2: cancer ~ age + yrsmoke + cigsday + female + highstatus + bird
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       142     156.72
2       140     154.20  2   2.5257   0.2828
```

The p-value of 0.2828 is greater than 0.05, which means we fail to reject the null hypothesis. This suggests that removing age and female from the model does not significantly worsen the model.

Part e

Carefully interpret each of the four model coefficients (including the intercept) in model4 in context.

Part f

If you replaced yrsmoke everywhere it appears in model4 with a mean-centered version of yrsmoke, tell what would change among these elements: the 4 coefficients, the 4 p-values for coefficients, and the residual deviance.

Part g

Observe that model3 is a potential final model based on this set of predictor variables. How does the adjusted odds ratio for birdkeeping from model3 compare with the unadjusted odds ratio you found in (b)? Is birdkeeping associated with a significant increase in the odds of developing lung cancer, even after adjusting for other factors?

Part h

Discuss the scope of inference in this study. Can we generalize our findings beyond the subjects in this study? Can we conclude that birdkeeping causes increased odds of developing lung cancer? Do you have other concerns with this study design or the analysis you carried out?

Exercise 2

(Ataman and Sarıyer 2021) use ordinal logistic regression to predict patient wait and treatment times in an emergency department (ED). The goal is to identify relevant factors that can be used to inform recommendations for reducing wait and treatment times, thus improving the quality of care in the ED.

The data include daily records for ED arrivals in August 2018 at a public hospital in Izmir, Turkey. The response variable is Wait time, a categorical variable with three levels:

- Patients who wait less than 10 minutes
- Patients whose waiting time is in the range of 10-60 minutes
- Patients who wait more than 60 minutes

Part a

Compare and contrast the proportional odds model with the multinomial logistic regression model. Write your response using 3 - 5 sentences. You can find a brief review of the proportional odds model here: https://library.virginia.edu/data/articles/fitting-and-interpreting-a-proportional-odds-model and https://online.stat.psu.edu/stat504/lesson/8/8.4

Part b

Table 5 in the paper contains the output for the wait time and treatment time models. Consider only the model for wait time. Describe the effect of arrival mode (ambulance, walk-in) on the waiting time. Note: walk-in is the baseline in the model. (A link to the paper can be found in the slides).

Part c

Consider output from both the wait time and treatment time models. Use the results from both models to describe the effect of triage level (red = urgent, green = non-urgent) on the wait and treatment times in the ED. Note: red is the baseline level.

Exercise 3

Ibanez and Roussel (2022) conducted an experiment to understand the impact of watching a nature documentary on pro-environmental behavior. The researchers randomly assigned the 113 participants to watch an video about architecture in NYC (control) or a video about Yellowstone National Park (treatment). As part of the experiment, participants played a game in which they had an opportunity to donate to an environmental organization. The data set is available in nature.csv in the data folder. We will use the following variables:

- donation_binary: 1 - participant donated to environmental organization versus 0 - participant did not donate.
- Age: age in years
- Gender: Participant's reported gender
- Treatment: Urban (T1)" - the control group versus "Nature (T2)" - the treatment group.
- NEP_high: 1 - score of 4 or higher on the New Ecological Paradigm (NEP) versus 0 - score of less than 4.

See the Introduction and Methods sections of Ibanez and Roussel (2022) for more details about the variables and see the class slides regarding the url for the paper.

```
nature <- read_csv("~/STA310/HW-6/STA310/Data/nature.csv", show_col_types = FALSE)
# https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0275806
nature = nature %>% select(c("donation_binary", "Age", "Gender", "Treatment", "nep_high", "D
# summary(nature)
```

Part a

Figure 2 on pg. 9 of the article visualizes the relationship between donation amount and treatment. Recreate this visualization using your own code. Use the visualization to describe the relationship between donating and the treatment.

Part b

Fit a probit regression model using age, gender, treatment, nep_high and the interaction between nep_high and treatment predict the likelihood of donating. (Note: Your model will be similar (but not exactly the same) as the "Likelihood" model in Table 5 on pg. 11.) Display the model.

Part c

Describe the effect of watching the documentary on the likelihood of donating.

Part d

Based on the model, what is the predicted probability of donating for a 20-year old female in the treatment group with a NEP score of 3?