STA2101 Assignment 5*

Ahmed Nawaz Amanullah †

February 2021

1

1.1 Problem

If two events have equal probability, the odds ratio equals _____.

1.2 Solution

1

2

2.1 Problem

For a multiple logistic regression model, if the value of the kth explanatory variable is increased by c units and everything else remains the same, the odds of $Y = 1$ are _____ times as great. Prove your answer.

---

†with help from the Overleaf team

2.2 Solution

$e^{\beta_k c}$. To get this, take the odds of $Y = 1$ for any data point,

$$e^{\beta_0+\beta_1 x_{i,1}+...+\beta_k x_{i,k}+...\beta_{p-1}x_{i,p-1}} \tag{1}$$

Then, if we increase the the value of the kth explanatory variable by c units, while keeping everything else same, the changed odds of $Y = 1$ is

$$e^{\beta_0+\beta_1 x_{i,1}+...+\beta_k (x_{i,k}+c)+...\beta_{p-1}x_{i,p-1}} \tag{2}$$

Therefore the odds of $Y = 1$ changes by

$$\frac{e^{\beta_0+\beta_1 x_{i,1}+...+\beta_k x_{i,k}+...\beta_{p-1}x_{i,p-1}}}{e^{\beta_0+\beta_1 x_{i,1}+...+\beta_k (x_{i,k}+c)+...\beta_{p-1}x_{i,p-1}}} \tag{3}$$

$$= \frac{e^{\beta_0}\prod_{j=1}^{p-1} e^{\beta_j x_{i,j}}}{e^{\beta_k c}e^{\beta_0}\prod_{j=1}^{p-1} e^{\beta_j x_{i,j}}} \tag{4}$$

$$e^{\beta_k c} \tag{5}$$

which is our answer, as required.

3

3.1 Problem

For a multiple logistic regression model, let $P(Y_i = 1|x_{i,1}, ..., x_{i,p-1}) = \pi(\mathbf{x}_i)$. Show that a linear model for the log odds is equivalent to

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T\boldsymbol{\beta}}} \tag{6}$$

3.2 Solution

The log odds of $Y_i = 1$ is $\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right)$. So a linear model for the log odds is $\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1}x_{i,p-1} = \mathbf{x}_i^T\boldsymbol{\beta}$. Then, solving for $\pi(\mathbf{x}_i)$,

$$\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} = \mathbf{x}_i^T \boldsymbol{\beta} \tag{7}$$

$$\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = e^{\mathbf{x}_i^T \boldsymbol{\beta}} \tag{8}$$

$$\pi(\mathbf{x}_i) = (1-\pi(\mathbf{x}_i))e^{\mathbf{x}_i^T \boldsymbol{\beta}} \tag{9}$$

$$\left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)\pi(\mathbf{x}_i) = e^{\mathbf{x}_i^T \boldsymbol{\beta}} \tag{10}$$

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \tag{11}$$

we get what we wanted to show, that a linear model for the log odds, $\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} = \mathbf{x}_i^T \boldsymbol{\beta}$ is equivalent to $\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}$, as required.

4

4.1 Problem

Write the log likelihood for a general logistic regression model, and simplify it as much as possible. Of course use the result of the last question.

4.2 Solution

$$L(\beta_0, ..., \beta_{p-1}) = \prod_{i=1}^{n} f(x_i; \beta_0, ..., \beta_{p-1}) \tag{12}$$

$$= \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i} \tag{13}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right)^{y_i} \left(1 - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right)^{1-y_i} \tag{14}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right)^{y_i} \left(\frac{1}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right)^{1-y_i} \tag{15}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{y_i \mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) \tag{16}$$

$$= \frac{e^{\sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta}}}{\prod_{i=1}^{n} \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)} \tag{17}$$

So the log likelihood expression is

$$l(\beta_0, ..., \beta_{p-1}) = \log L(\beta_0, ..., \beta_{p-1}) \tag{18}$$

$$= \log \left(\frac{e^{\sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta}}}{\prod_{i=1}^{n} \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right)} \right) \tag{19}$$

$$= \sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right) \tag{20}$$

$$= \sum_{i=1}^{n} \left(y_i \mathbf{x}_i^T \boldsymbol{\beta} - \log \left(1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}\right) \right) \tag{21}$$

5

5.1 Problem

A logistic regression model with no explanatory variables has just one parameter, β_0. It('s ?) also the same probability $\pi = P(Y = 1)$ for each case.

(a) Write π as a function of β_0; show your work.

(b) The *invariance principle* of maximum likelihood estimation says the MLE of a function of the parameter is that function of the MLE. It is very handy. Now, still considering a logistic regression model with no explanatory variables,

i. Suppose $\bar{y}$ (the sample proportion of $Y = 1$ cases) is 0.57. What is $\hat{\beta}_0$? Your answer is a number.

ii. Suppose $\hat{\beta}_0 = -0.79$. What is $\bar{y}$? Your answer is a number.

5.2 Solution

5.2.1 (a)

$$\log \frac{\pi}{1-\pi} = \beta_0 \tag{22}$$

$$\frac{\pi}{1-\pi} = e^{\beta_0} \tag{23}$$

$$\pi = (1-\pi)e^{\beta_0} \tag{24}$$

$$\left(1 + e^{\beta_0}\right)\pi = e^{\beta_0} \tag{25}$$

$$\pi = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{26}$$

5.2.2 (b)

i. Assuming a Bernoulli model with π as the probability of success (i.e. $Y = 1$), then recall from the slides that the MLE is the sample proportion $\hat{\pi} = \bar{y} = 0.57$. Then by the invariance principle,

$$\hat{\beta}_0 = \log \frac{\hat{\pi}}{1-\hat{\pi}} = \log \frac{0.57}{1-0.57} \approx 0.28 \tag{27}$$

ii. Again, by the invariance principle,

$$\bar{y} = \hat{\pi} = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}} = \frac{e^{-0.79}}{1 + e^{-0.79}} \approx 0.31 \tag{28}$$

6

6.1 Problem

Consider a logistic regression in which the cases are newly married couples with both people from the same religion. The explanatory variables are total family income and religion. Religion is coded A, B, C and None (let's call "None" a religion), and the response variable is whether the marriage lasted 5 years (1=Yes, 0=No).

(a) Write a linear model for the log odds of a successful[1] marriage. You

[1] I agree, this may be a modest definition of success.

do not have to say how the dummy variables are defined. You will do that is the next part.

(b) Make a table with four rows, showing how you would set up indicator dummy variables for Religion, with None as the reference category.

(c) Add a column showing the odds of the marriage lasting 5 years. The *symbols* for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row. But of course your answer contains β values. Denote income by x.

(d) For a constant value of income, what is the ratio of the odds of a marriage lasting 5 years or more for Religion C to the odds of lasting 5 years or more for No Religion? Answer in terms of the β symbols of your model.

(e) Holding income constant, what is the ratio of the odds of lasting 5 years or more for religion A to the odds of lasting 5 years or more for religion B? Answer in terms of the β symbols of your model.

(f) You want to test whether controlling for income, Religion is related to whether the marriage lasts 5 years. State the null hypothesis in terms of one or more β values.

(g) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from religion C, allowing for income. State the null hypothesis in terms of one or more β values.

(h) You want to test whether marriages between people of No Religion with an average income have a 50-50 chance of lasting 5 years. State the null hypothesis in symbols. To hold income to an "average" value, just set $x = \bar{x}$.

6.2 Solution

6.2.1 (a)

$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4}$, where $\pi_i = \pi(\mathbf{x}_i) = P(Y_i = 1|x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$ is the conditional probability of a successful marriage, $x_{i,1}$ is the total family income, and $x_{i,2}$, $x_{i,3}$ and $x_{i,4}$ are dummy variables for the religion.

6.2.2 (b)

Religion	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
A	1	0	0
B	0	1	0
C	0	0	1
None	0	0	0

6.2.3 (c)

Religion	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	Odds of Successful Marriage
A	1	0	0	$e^{\beta_0}e^{\beta_1 x}e^{\beta_2}$
B	0	1	0	$e^{\beta_0}e^{\beta_1 x}e^{\beta_3}$
C	0	0	1	$e^{\beta_0}e^{\beta_1 x}e^{\beta_4}$
None	0	0	0	$e^{\beta_0}e^{\beta_1 x}$

6.2.4 (d)

$$\frac{e^{\beta_0}e^{\beta_1 x}e^{\beta_4}}{e^{\beta_0}e^{\beta_1 x}} = e^{\beta_4} \tag{29}$$

6.2.5 (e)

$$\frac{e^{\beta_0}e^{\beta_1 x}e^{\beta_2}}{e^{\beta_0}e^{\beta_1 x}e^{\beta_3}} = e^{\beta_2-\beta_3} \tag{30}$$

6.2.6 (f)

$H_0 : \beta_2 = \beta_3 = \beta_4 = 0.$

6.2.7 (g)

$H_0 : \beta_2 - \beta_4 = 0.$

6.2.8 (h)

$H_0 : \beta_0 + \beta_1\bar{x} = \log\frac{0.5}{1-0.5} = 0.$

7

7.1 Problem

People who raise large numbers of birds inhale potentially dangerous material, especially tiny fragments of feathers. Can this be a risk factor for lung cancer, controlling for other possible risk factors? Which of those other possible risk factors are important? Here are the variables in the file http://www.utstat.utoronto.ca/~brunner/data/illegal/birdlung.data.txt. These data are from a textbook called the *Statistical Sleuth* by Ramsey and Schafer, and are used without permission.

Variable	**Values**
Lung Cancer	1=Yes, 0=No
Gender	1=Female, 0=Male
Socioeconomic Status	1=High, 0=Low
Birdkeeping	1=Yes, 0=No
Age	
Years smoked	
Cigarettes per day	

If you look at **help(colnames)**, you can see how to add variable names to a data frame. It's a good idea, because if you can't remember which variables are which during a quiz, you're out of luck.
First, make tables of the binary variables using **table**, Use **prop.table** to find out the percentages. What proportion of the sample had cancer. Any comments?
There is one primary issue in this study: Controlling for all other variables, is birdkeeping significantly related to the chance of getting lung cancer? Carry out a likelihood ratio test to answer the question.

(a) In symbols, what is the null hypothesis?

(b) What is the value of the likelihood ratio test statistic G^2? The answer is a number.

(c) What are the degrees of freedom for the test? The answer is a number.

(d) What is the p-value? The answer is a number.

(e) What do you conclude? Presence of a relationship is not enough. Say what happened.

(f) For a non-smoking, bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.

(g) Obtain a 95% confidence interval for that last probability. Your answer is a pair of numbers. There is an easy way and a hard way. Do it the easy way.

(h) Your answer to the last question made you uncomfortable. Why? Another approach is to start with a confidence interval for the log odds, and then use the fact that the function $p(x) = \frac{e^x}{1+e^x}$ is strictly increasing in x. Get the confidence interval this way. Again, your answer is a pair of numbers. Which confidence interval do you like more?

(i) Naturally, you should be able to interpret all the Z-tests too. Which one is comparable to the main likelihood ratio test you have just done?

(j) Controlling for all other variables, are the chances of cancer different for men and women?

(k) Also, are *any* of the explanatory variables related to getting lung cancer? Carry out a single likelihood ratio test. You could do it from the default output with a calculator, but use R. Get the p-value, too.

(l) Now please do the same as the last item, but with a Wald test. Of course you should display the value of W_n, the degrees of freedom and the p-value.

(m) Finally and just for practice, fit a simple logistic regression model in which the single explanatory variable is number of cigarettes per day.

   i. When a person from this population smokes ten more cigarettes per day, the odds of lung cancer are multiplied by r (odds ratio). Give a point estimate of r. Your answer is a number.

   ii. Using the **vcov** function and the delta method, give an estimate of the asymptotic variance of r. Your answer is a number.

Please bring your R printout for this question to the quiz. Also, this question requires some paper and pencil work, and it would be fair to ask for something like that on the quiz too.

7.2 Solution

The R code implemented to carry out the computations within these questions is disclosed below.

```
# STA2101 Assignment 5 Ques 7

# first read data
d_source <- "http://www.utstat.toronto.edu/~brunner/data/illegal/"
d_source <- paste(d_source, "birdlung.data.txt",sep="")
bl_data <- read.table(d_source)

# rename variables
cnames <- c("lc", "sex", "ses", "bk", "age", "yrs_s", "cig")
colnames(bl_data) <- cnames

# proportion of data investigation
# make tables for each binary variable
# Lung cancer
lc_freq <- table(bl_data$lc)
lc_prop <- prop.table(lc_freq)
# report proportion
sprintf("Lung cancer, No?: %.3f, Yes?: %.3f", lc_prop[1], lc_prop[2])

# Sex
sex_freq <- table(bl_data$sex)
sex_prop <- prop.table(sex_freq)
# report proportion
sprintf("Sex, Male: %.3f, Female: %.3f", sex_prop[1], sex_prop[2])

# Socioeconomic Status
ses_freq <- table(bl_data$ses)
ses_prop <- prop.table(ses_freq)
# report proportion
```

```
sprintf("SE Status , Low: %.3f , High: %.3f", ses_prop[1], ses_prop[2])

# Birdkeeping
bk_freq <- table(bl_data$bk)
bk_prop <- prop.table(bk_freq)
# report proportion
sprintf("Birdkeep , No?: %.3f , Yes?: %.3f", bk_prop[1], bk_prop[2])

# for part (b) want to compute G2 for null hypothesis test
#       (i.e all else considered, is bk related to chance of lc)
# fit logistic regression model
# fit full model
log_m1 <- glm(lc ~ sex+ses+bk+age+yrs_s+cig, data = bl_data,
family = binomial)
# check summary
summary(log_m1)

# fit reduced model (with just to control variables)
log_m2 <- glm(lc ~ sex+ses+age+yrs_s+cig, data = bl_data, family = binomial)
# check summary
summary(log_m2)

# G2
G2 <- log_m2$deviance - log_m1$deviance

# Report G2
sprintf("G2: %.3f", G2)

# for part (d) need to compute and report p-value
# degrees of freedom is 1 (only one equal sign in null hypothesis)
dof <- 1
p_val <- 1 - pchisq(G2, df = dof)
# Report p-value
sprintf("p-value: %.5f", p_val)

# what happens when anova is used?
anova(log_m2, log_m1, test = "Chisq")
```

```
# part e
# see if we reject
# significance level
alpha <- 0.05
reject <- 'no'
if(p_val<alpha){
        reject <- 'yes'
}
sprintf('Reject?: %s', reject)

# report exponential of bk estimated coefficient
odds_r <- exp(log_m1$coefficients["bk"])
sprintf("odds_r: %.3f", odds_r)

# part (f) requires the estimation of lung cancer probability
# for non-smoking (yrs_s=cig=0) women (sex=1) of avg age
#       (age = mean(age)) and low se status (ses = 0)
#   who birdkeeps (bk = 1)
# use predict to obtain estimate probability
age_bar <- mean(bl_data$age)
input_f <- data.frame(sex=1,ses=0,bk=1,age=age_bar,yrs_s=0,cig=0)
pre_f <- predict(log_m1,newdata=input_f,type="response", se.fit=T)
pre_f

# compare with manual process
x <- c(1, 1, 0, 1, age_bar, 0, 0)
xb <- sum(log_m1$coefficients*x)
phat <- exp(xb)/(1+exp(xb))
# report
sprintf("(f) estimated prob: %.3f", phat)

# part (g): need confidence interval for answer in (g)
# first try out delta method se computation just for comparison
# covariance matrix
Vhat <- vcov(log_m1)
# delta method
denom <- (1+exp(xb))^2
gdot <- x*exp(xb)/denom
```

```
gdot <- matrix(gdot, nrow = 1)
se_f_manual <- sqrt(gdot%*%Vhat%*%t(gdot))
# report
sprintf("(f) estimated se: %.4f", se_f_manual)

# Obtain and report CI
# bounds
lowerCL <- pre_f$fit - qnorm(0.975)*pre_f$se.fit
upperCL <- pre_f$fit + qnorm(0.975)*pre_f$se.fit
# now prediction
phat <- pre_f$fit
results <- cbind(phat, lowerCL, upperCL)
colnames(results) <- c("phat", "LowerCL", "UpperCL")
results

# part (h) - compute the 95% CI using different approach
#       this time we first obtain CI for log odds
# use predict without response this time...
# plan to use predict to get s.e as well
# note that alternatively could have gotten s.e from estimated
#       covariance matrix as sqrt(x*Vhat*t(x)) but since results are same
#       omitting for sake of brevity
pre_h <- predict(log_m1,newdata=input_f, se.fit=T); pre_h

# Obtain and report CI for log odds
# bounds
lowerCL <- pre_h$fit - qnorm(0.975)*pre_h$se.fit
upperCL <- pre_h$fit + qnorm(0.975)*pre_h$se.fit
# now prediction
log_odds <- pre_h$fit
results <- cbind(log_odds, lowerCL, upperCL)
colnames(results) <- c("log_odds", "LowerCL", "UpperCL")
results

# next want to use the fact that p(x) = exp(x)/(1+exp(x)) is increasing
#       to obtain the 95% CI for prob
# bounds
lowerCL <- exp(lowerCL)/(1+exp(lowerCL))
```

```
upperCL <- exp(upperCL)/(1+exp(upperCL))
# now prediction
phat <- exp(log_odds)/(1+exp(log_odds))
results <- cbind(phat, lowerCL, upperCL)
colnames(results) <- c("phat", "LowerCL", "UpperCL")
results

# part (j) LR test equivalent
# fit model for control variables
# fit reduced model (with just to control variables)
log_m3 <- glm(lc ~ ses+bk+age+yrs_s+cig,data = bl_data,
family = binomial)
# check summary
summary(log_m3)

# G2
G2 <- log_m3$deviance - log_m1$deviance

# Report G2
sprintf("G2: %.3f", G2)

# degrees of freedom is 1 (only one equal sign in null hypothesis)
dof <- 1
p_val <- 1 - pchisq(G2, df = dof)
# Report p-value
sprintf("p-value: %.5f", p_val)

# what happens when anova is used?
anova(log_m3, log_m1, test = "Chisq")

# part k
# Using LR test to see if any variable is related to chance of
#   lung cancer
# use null.deviance item of full model to get the negative LL of
#       the reduced model with just the intercept term
# LR test stat
G2 <- log_m1$null.deviance - log_m1$deviance
# p-value; degrees of freedom is 6 as there are 6 explanatory
```

```
#       variables whose coefficients are 0 under null hypothesis
pval <- 1-pchisq(G2, df = 6)
# report results
sprintf("G2: %.3f", G2)
sprintf("p-value: %.5f", pval)

# part l - repeat k, but using the Wald test
# Borrow Prof. Brunner's Wtest function
source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/Wtest.txt")
# Obtain full model coefficients
beta_hat1 <- log_m1$coefficients
# We will have 6 rows - 1 each for the 6 explanatory variables
# As under null hypothesis the regression coefficients for those are 0
L1 <- cbind(c(0,0,0,0,0,0), diag(6))
# covariance matrix
Vhat <- vcov(log_m1)
# Carry out Wald test
Wtest(L1, beta_hat1, Vhat)

# part m
# fit logistic regression model with just cig as explanatory variable
log_m4 <- glm(lc ~ cig,data = bl_data,family = binomial)
# check summary
summary(log_m4)

# obtain odds ratio for part (i)
beta_1 <- log_m4$coefficients["cig"]
# odds ratio
r <- exp(10*beta_1)
# report
sprintf("Odds multiplied by r: %.5f", r)

# part (ii) - obtain asymptotic vairance of r from (i)
# use vcov to get Vhat from model
Vhat <- vcov(log_m4)
# gdot for r
gdot <- c(0, 10*exp(10*beta_1))
gdot <- matrix(gdot, nrow = 1)
```

```
# get aymptotic variance by delta method
r_var <- gdot %*% Vhat %*% t(gdot)
# report asymptotic variance
sprintf("Asymptotic variance of r: %.5f", r_var)
```

We executed this code in R console and extracted the responses for this question from the output observed.

Firstly, approximately 33.3% of the sample had lung cancer. By doing a Google search on the people diagnosed with lung cancer within the US, we found a figure of about half a million, making it a small percentage of the total population (which is around 300 million). As a result, my humble thought is that perhaps this sample proportion is a consequence of the sampling process, as in, for example, the number of people with or without lung cancer was chosen beforehand (of course, it could also be that the sampling was done in a different country, a specific subset of the population was sampled, or birdkeeping really does impact the chance of getting lung cancer).

The response to the primary issue in the study (i.e. controlling for all other variables, is birdkeeping significantly related to the chance of getting lung cancer?) is within the responses to the sub-questions of this question, which are disclosed as follows.

(a) $H_0 : \beta_3 = 0$, where β_3 is the regression coefficient for the birdkeeping explanatory variable.

(b) $G^2 \approx 11.670$.

(c) 1 (number of non-redundant equalities or difference in number of betas (?) is the degrees of freedom).

(d) 0.00064.

(e) Since p-value is less than the default/ typical significance level of $\alpha = 0.05$, we say that there is evidence to suggest that there is association between birdkeeping and likelihood of getting diagnosed with lung cancer, specifically birdkeeping is associated with a higher chance of getting lung cancer (the regression coefficient in the full model is positive). Additionally, interpreting the regression coefficient, the odds of getting lung cancer for people who birdkeep a lot is about 3.906 times as great.

(f) Approximately 0.093 or 9.3%.

(g) The estimated probability is a function of the betas, so easy way is presumably to use delta method to compute the standard error of the estimated probability, then use it to calculated the confidence interval (seemingly from the R code/ console output, the **predict** function already returns the standard error). The confidence interval computed is approximately (-0.044, 0.230).

(h) The answer to the last question made me uncomfortable because the lower bound of the confidence interval is negative (and negative probability does not make sense). The confidence interval retrieved using the method proposed in this sub-question is (0.020, 0.341). I like this confidence interval more than the last one as the confidence interval bounds are within the range of possible probability values (i.e. between 0 and 1).

(i) The Z-test for testing the null hypothesis of regression coefficient for birdkeeping being equal to 0. It is already disclosed in the summary of the full model, specifically we observe test statistic value of $Z \approx 3.313$ and approximate p-value of 0.000923. Since p-value is less than the default/ typical significance level of $\alpha = 0.05$, we arrive at the same conclusion as the likelihood ratio test.

(j) From the summary of the full model, we have for the gender/sex variable the approximate test statistic and p-value of 1.057 and 0.290653 respectively. Since p-value is greater than the default/ typical significance level of $\alpha = 0.05$, we say that, controlling for all other variables, there is insufficient evidence to suggest that the chances of cancer are different for men and women.

(k) We have $G^2 \approx 32.937$ and approximate p-value of 0.00001 (and degrees of freedom of 6, as there are 6 explanatory variables). Since p-value is less than the default/ typical significance level of $\alpha = 0.05$, there is evidence to suggest that at least one of the explanatory variables is related to getting lung cancer.

(l) We have $W_n \approx 22.6$, degrees of freedom of 6 (as there are 6 explanatory variables) and approximate p-value of 0.00095. Since p-value is less

than the default/ typical significance level of $\alpha = 0.05$, we arrive at the same conclusion as the likelihood ratio test.

(m) i. 1.66745 (i.e. $e^{10\beta_1}$).

ii. 0.10452 (i.e. $\begin{pmatrix} 0 & 10e^{10\beta_1} \end{pmatrix} \hat{V} \begin{pmatrix} 0 \\ 10e^{10\beta_1} \end{pmatrix}$, where $\hat{V}$ is retrieved using R's **vcov** function).

8

8.1 Problem

In the usual multiple regression model, the X matrix is an $n \times p$ matrix of known constants. But in practice, the explanatory variables are often random and not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathcal{X} = X$. Here, the $n \times p$ matrix $\mathcal{X}$ is used to denote the matrix containing the random explanatory variables. It does not have to be *all* random. For example the first column might contain only ones if the model has an intercept.

8.1.1 (a)

Show that the least-squares estimator $(X^T X)^{-1} X^T \mathbf{y}$ is conditionally unbiased. You've done this before.

8.1.2 (b)

Show that $\hat{\beta} = (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \mathbf{y}$ is also unbiased unconditionally. If it helps, you may assume that the explanatory variables are discrete, so you can write a multiple sum. Or, you could do it using just expected values.

8.1.3 (c)

A similar calculation applies to the significant level of a hypothesis test. Let F be the test statistic (say for an F-test comparing full and reduced models),

and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory values. That is, $P(F > f_c|\mathcal{X} = X) = \alpha$. Find the *unconditional* probability of a Type I error. Again, you may assume that the explanatory variables are discrete if you wish. I used the so-called Law of Total Probability.

8.2 Solution

8.2.1 (a)

$$E((\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathbf{y}|\mathcal{X} = X) \tag{31}$$
$$= (X^TX)^{-1}X^TE(\mathbf{y}|\mathcal{X} = X) \tag{32}$$
$$= (X^TX)^{-1}X^TE(\mathcal{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}|\mathcal{X} = X) \tag{33}$$
$$= (X^TX)^{-1}X^T(X\boldsymbol{\beta} + 0) \tag{34}$$
$$= \boldsymbol{\beta} \tag{35}$$

8.2.2 (b)

$$E(\hat{\beta}) = E((\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathbf{y}) \tag{36}$$
$$= E(E((\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\mathbf{y}|\mathcal{X} = X)) \tag{37}$$
$$= E(\boldsymbol{\beta}) \tag{38}$$
$$= \boldsymbol{\beta} \tag{39}$$

8.2.3 (c)

$$P(F > f_c) = E(I(F > f_c)) \tag{40}$$
$$= E(E(I(F > f_c)|\mathcal{X} = X)) \tag{41}$$
$$= E(P(F > f_c|\mathcal{X} = X)) \tag{42}$$
$$= E(\alpha) \tag{43}$$
$$= \alpha \tag{44}$$

9

9.1 Problem

Ordinary linear regression is often applied to data sets where the explanatory variables are best modeled as random variables: write $y_i = \mathcal{X}_i^T \boldsymbol{\beta} + \epsilon_i$. In what way does the usual conditional linear regression model with normal errors imply that random explanatory variables must be independent of the error term? Hint: assume the random vector $\mathcal{X}_i$ and the scalar random variable ϵ_i are both continuous. What is the conditional distribution of ϵ_i given $\mathcal{X}_i = X_i$?

9.2 Solution

Starting with the hint, $\epsilon_i | \mathcal{X}_i = X_i \sim N(0, \sigma^2)$, where $\sigma^2 > 0$ is an unknown constant (this is a condition of the usual conditional linear regression model with normal errors). Now we want the same distribution for ϵ_i *unconditionally*, i.e. $\epsilon_i \sim N(0, \sigma^2)$, as a lot of our testing and inferential computations rely on this assumption. Now, we know $f(\epsilon_i | \mathcal{X}_i = X_i) = \frac{f(\epsilon_i, \mathcal{X}_i = X_i)}{f(\mathcal{X}_i = X_i)}$. Note that if we have that random explanatory variables are independent of the error term, then

$$
\begin{aligned}
& f(\epsilon_i | \mathcal{X}_i = X_i) && (45) \\
&= \frac{f(\epsilon_i, \mathcal{X}_i = X_i)}{f(\mathcal{X}_i = X_i)} && (46) \\
&= \frac{f(\epsilon_i) f(\mathcal{X}_i = X_i)}{f(\mathcal{X}_i = X_i)} && (47) \\
&= f(\epsilon_i) && (48)
\end{aligned}
$$

Which is what we want (i.e. $\epsilon_i \sim N(0, \sigma^2)$). Since we needed that random explanatory variables are independent of the error term, the usual conditional linear regression model with normal errors imply that random explanatory variables must be independent of the error term, as required.

N.B. In my previous response, I argued that if random explanatory variables are not independent of the error term, then we would have that $f(\epsilon_i | \mathcal{X}_i = X_i) \neq f(\epsilon_i)$ and $f(\epsilon_i | \mathcal{X}_i = X_i)$ would depend on the distribution of $\mathcal{X}_i$, which we do not know and would not be exactly normal. So the tests and

inferential tools we use in regular conditional linear regression, which relies on normal error assumption, are not valid anymore. So the usual conditional linear regression model with normal errors imply that random explanatory variables must be independent of the error term, as required.

10

10.1 Problem

For a model with just one (random) explanatory variable, show that $E(\epsilon_i|X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that a standard regression model without the normality assumption still implies zero covariance (though not necessarily independence) between the error term and explanatory variables.

10.2 Solution

Provided that $E(\epsilon_i|X_i = x_i) = 0$ for all x_i, then,

$$E(\epsilon_i) = E(E(\epsilon_i|X_i = x_i)) = E(0) = 0 \tag{49}$$

Then,

$$\begin{aligned} Cov(X_i, \epsilon_i) &= E((X_i - E(X_i))(\epsilon_i - E(\epsilon_i))) && (50)\\ &= E(\epsilon_i(X_i - E(X_i))) && (51)\\ &= E(\epsilon_i X_i) - E(X_i)E(\epsilon_i) && (52)\\ &= E(\epsilon_i X_i) && (53)\\ &= E(E(\epsilon_i X_i|X_i = x_i)) && (54)\\ &= E(x_i E(\epsilon_i|X_i = x_i)) && (55)\\ &= E(0) && (56)\\ &= 0 && (57) \end{aligned}$$

as required.