1. **(20 points) Do exercise 19.8.8 on page 475**

Exercise 19.8.8. A study was conducted to examine the association between consumption of caffeinated coffee and nonfatal myocardial infarction among males between the ages of 21 and 54 years. The dataset *coffee* contains information on a sample of 2496 adult males [292]. A binary variable indicating whether a study participant drinks caffeinated coffee is saved under the variable name *coffee*, with the value 1 indicating that he does drink coffee and 0 that he does not. The response occurrence of myocardial infarction is saved under the name *mi*, with 1 indicating that this event did occur and 0 that it did not.

$$In\left(\frac{\hat{p}}{1-\hat{p}}\right) = \widehat{\beta_0} + \widehat{\beta_1} x_1$$

(a) Fit a logistic regression model of the form where the response is occurrence of myocardial infarction, and x represents coffee consumption. Interpret the odds ratio of myocardial infarction for males who drink caffeinated coffee versus those who do not.

(b) Suppose you are concerned that smoking status is a confounder in the relationship between coffee consumption and myocardial infarction. The variable smoke takes the value 1 if a male is a self-reported smoker and 0 otherwise. Fit separate logistic regression models for smokers and nonsmokers. Within each subgroup, interpret the odds ratio of myocardial infarction for males who drink coffee versus those who do not.

(c) Using the entire sample, fit a logistic regression model to examine the relationship between coffee consumption and the occurrence of myocardial infarction, adjusting for smoking status. Interpret the odds ratio associated with coffee consumption.

(d) At the 0.05 level of significance, test the null hypothesis that the odds ratio associated with coffee consumption is equal to 1. What do you conclude?

(e) Construct a 95% confidence interval for the odds ratio of occurrence of myocardial infarction for coffee drinkers versus nondrinkers, adjusting for smoking status.

Solution:

(a)

```
> # ------------(a)------------
> # Fit a logistic regression model
> model <- glm(mi ~ coffee, data = coffee, family = binomial(link = "logit"))
> summary(model)

Call:
glm(formula = mi ~ coffee, family = binomial(link = "logit"),
    data = coffee)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4464  -1.4464   0.9304   0.9304   1.3106

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3079     0.1086  -2.834   0.0046 **
coffee        0.9211     0.1177   7.828 4.95e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3321.3  on 2495  degrees of freedom
Residual deviance: 3259.2  on 2494  degrees of freedom
AIC: 3263.2

Number of Fisher Scoring iterations: 4

> exp(model$coefficients[2])
  coffee
2.512051
>
```

From the model summary, we can see that the coefficient for coffee consumption is 0.9211, which means that coffee consumption is positively associated with the occurrence of myocardial infarction. The odds ratio for coffee consumption is exp(0.9211) = 2.512052, which means that the odds of myocardial infarction are 2.512052 times greater for males who drink caffeinated coffee compared to those who do not.

(b)

```
> # ------------(b)------------
> # Fit separate logistic regression models for smokers and nonsmokers
> model_smokers <- glm(mi ~ coffee, data = subset(coffee, smoke == 1), family = binomial(link = "logit"))
> summary(model_smokers)

Call:
glm(formula = mi ~ coffee, family = binomial(link = "logit"),
    data = subset(coffee, smoke == 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5993  -1.3991   0.8078   0.8078   1.1560

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.05064    0.15916   0.318     0.75
coffee       0.90190    0.16996   5.307 1.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1903.5  on 1558  degrees of freedom
Residual deviance: 1876.1  on 1557  degrees of freedom
AIC: 1880.1

Number of Fisher Scoring iterations: 4

> exp(model_smokers$coefficients[2])
  coffee
2.464292
> model_nonsmokers <- glm(mi ~ coffee, data = subset(coffee, smoke == 0), family = binomial(link = "logit"))
> summary(model_nonsmokers)

Call:
glm(formula = mi ~ coffee, family = binomial(link = "logit"),
    data = subset(coffee, smoke == 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1979  -1.1979  -0.9269   1.1570   1.4506

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6225     0.1526  -4.080 4.50e-05 ***
coffee        0.6707     0.1692   3.964 7.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1297.3  on 936  degrees of freedom
Residual deviance: 1281.1  on 935  degrees of freedom
AIC: 1285.1

Number of Fisher Scoring iterations: 4

> exp(model_nonsmokers$coefficients[2])
  coffee
1.955542
```

From the model summaries, we can see that in both the smoker and nonsmoker subgroups, the coefficient for coffee consumption is positive, indicating a positive association between coffee consumption and the occurrence of myocardial infarction. The odds ratios are exp(0.902) = 2.464 for smokers and exp(0.671) = 1.956 for nonsmokers. This means that the odds of myocardial infarction are higher for coffee drinkers in both subgroups, but the effect size is larger for smokers than for nonsmokers.

(c)

```
> # ------------(c)------------
> # Fit a logistic regression model adjusting for smoking status
> model_adj <- glm(mi ~ coffee + smoke, data = coffee, family = binomial(link = "logit"))
> summary(model_adj)

Call:
glm(formula = mi ~ coffee + smoke, family = binomial(link = "logit"),
    data = coffee)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5929  -1.2071   0.8126   0.8126   1.4930

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.71688    0.11886  -6.031 1.63e-09 ***
coffee       0.78639    0.12078   6.511 7.46e-11 ***
smoke        0.86890    0.08679  10.012  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3321.3  on 2495  degrees of freedom
Residual deviance: 3158.1  on 2493  degrees of freedom
AIC: 3164.1

Number of Fisher Scoring iterations: 4

> exp(model_adj$coefficients[2])
  coffee
2.195462
```

From the model summary, we can see that the coefficient for coffee consumption is still positive, indicating a positive association between coffee consumption and the occurrence of myocardial infarction, after adjusting for smoking status. The odds ratio for coffee consumption is exp(-0.7864) = 2.1955, which means that the odds of myocardial infarction are 2.1955 times higher for males who drink caffeinated coffee compared to those who do not, after adjusting for smoking status.

(d)

$$H_0: Odds\ Ratio\ for\ Coffee\ consumption\ =\ 1$$
$$H_a: Odds\ Ratio\ for\ Coffee\ consumption\ \neq\ 1$$

```
> # ------------(d)------------
> # Test the null hypothesis that the odds ratio associated with coffee consumption is equal to 1
> summary(model)

Call:
glm(formula = mi ~ coffee, family = binomial(link = "logit"),
    data = coffee)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4464  -1.4464   0.9304   0.9304   1.3106

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3079     0.1086  -2.834   0.0046 **
coffee        0.9211     0.1177   7.828 4.95e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3321.3  on 2495  degrees of freedom
Residual deviance: 3259.2  on 2494  degrees of freedom
AIC: 3263.2

Number of Fisher Scoring iterations: 4

> test <- summary(model)$coefficients["coffee", "Pr(>|z|)"]
> test
[1] 4.945685e-15
```

The coefficient estimate for coffee is 0.9211, indicating that coffee consumption is associated with an increase in the odds of myocardial infarction. The z-statistic for the hypothesis test of odds ratio = 1 is 7.828, and the corresponding p-value is 4.95e-15. Since the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that there is evidence of a significant association between coffee consumption and the odds of myocardial infarction.

(e)

```
> # ------------(e)------------
> # 95% confidence interval for odds ratio
> confint(model_adj, level = 0.95)
Waiting for profiling to be done...
                 2.5 %     97.5 %
(Intercept) -0.9517215 -0.4854285
coffee       0.5503456  1.0241008
smoke        0.6990682  1.0393346
```

The 95% confidence interval for the odds ratio of coffee drinkers versus nondrinkers is [0.55, 1.02], adjusting for smoking status. This indicates that, holding smoking status constant, coffee drinkers have lower odds of myocardial infarction than nondrinkers, and the true odds ratio is likely to be within this interval with 95% confidence.