Jiayuan Shi
MA684 HW8

1A. Which variables are significant predictors of not using vitamins during pregnancy? Explain.
Solution: Marital status, age, education and smoking status are significant predictors of not using vitamins during pregnancy. If there are no differences in the odds in the two groups, the odds ratio would be 1.0. Since 1.0 is only contained in the 95% CI of American Indian in the race variable, there is no significant difference in the odds of a American Indian for those using vs. not using vitamins, and the other variables like marital status, age, education and smoking status are significant.

1B. Describe the association between education and not using vitamins during pregnancy.
Solution: The association between education and not using vitamins during pregnancy is significant, because the confidence intervals for its odds ratios of Education Year<20 and Education Year=2 groups both do not contain the null value of 1.0.

Those receive less than 12 years of education have 1.75 times the odds of not using vitamins during pregnancy as those receive more than 12 years of education.

Those receive 12 years of education have 1.59 times the odds of not using vitamins during pregnancy as those receive more than 12 years of education.

1C. Describe the association between smoking status and not using vitamins during pregnancy.
Solution: The association between smoking status and not using vitamins during pregnancy is significant, because the confidence interval for its odds ratio of smoker group does not contain the null value of 1.0.

Smokers are less likely to not use vitamins during pregnancy than non-smokers. Smokers have 0.9 times the odds of not using vitamins during pregnancy as those non-smokers.

2A. Complete the above tables.
Solution:

Variable	Coefficient	Standard Error	Wald Chi-Square	p-value
Intercept	-1.354	---	---	---
Sex Female	-0.673	0.163	17.047	<0.001
Grade 10	0.117	0.245	0.228	0.633
Grade 11	0.240	0.240	1	0.317
Grade 12	0.614	0.235	6.827	0.009

Variable	Odds Ratio	95% CI for OR
Intercept	---	---
Sex Female	0.510	(0.371, 0.702)
Grade 10	1.124	(0.695, 1.817)

Grade 11	1.271	(0.794, 2.035)
Grade 12	1.848	(1.166, 2.929)

2B. Describe differences in smoking for females vs. males, based on the odds ratios from the above tables.
Solution: 95% CI for OR of the variable Sex Female is (0.371, 0.702), does contain null value of 1.0, indicating smoking significantly differs between females and males. Thus, controlling for grade, females have 0.51 times the odds of smoking compared to males. Female students are significantly less likely to smoke than male students.

2C. Describe the differences in smoking across grades, based on the odds ratios from the above tables.
Solution: Students in Grade 10 are more likely to smoke than students in Grade 9. Controlling for other variables, students in Grade 10 have 1.124 times the odds of smoking compared to students in Grade 9. 95% CI for OR of students in Grade 10 is (0.695, 1.817), contains null value of 1.0, indicating this difference of smoking is not significant.

Students in Grade 11 are more likely to smoke than students in Grade 9. Controlling for other variables, students in Grade 11 have 1.271 times the odds of smoking compared to students in Grade 9. 95% CI for OR of students in Grade 11 is (0.794, 2.035), contains null value of 1.0, indicating this difference of smoking is not significant.

Students in Grade 12 are more likely to smoke than students in Grade 9. Controlling for other variables, students in Grade 12 have 1.848 times the odds of smoking compared to students in Grade 9. 95% CI for OR of students in Grade 12 is (1.166, 2.929), does not contain null value of 1.0, indicating this difference of smoking is significant.

2D. What is the predicted probability of smoking (calculated using the slopes from the above logistic regression), for:
- a 9th grade girl
- a 9th grade boy
- a 12th grade boy?
Solution: Predicted probabilities:

$$\hat{p}_x = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}$$

For,

- Predicted probability of a 9th grade girl: $\frac{e^{-1.354-0.673}}{1+e^{-1.354-0.673}}$=0.116
- Predicted probability of a 9th grade boy: $\frac{e^{-1.354}}{1+e^{-1.354}}$=0.205
- Predicted probability of a 12th grade boy: $\frac{e^{-1.354+0.614}}{1+e^{-1.354+0.614}}$=0.323

Based on this analysis, 11.6% of the 9th grade girls are expected to smoke, 20.5% of the 9th grade boys are expected to smoke, and 32.3% of the 12th grade boys are expected to smoke.

3A. As a preliminary step in the analysis we need to account for the missing data on voting - re-code the voting variable to account for missing data:

Solution:

```
> vote[vote==9] <- NA
> table(vote)
vote
  0   1
151 293
```

3B. What percent of the sample voted (this percent should be based on the number with non-missing data, not on the total sample size)? Give a 95% confidence interval for the percent of registered voters who voted:

Solution: Percent of the sample voted = 293/(293+151) = 0.6599 = 65.99%

95% confidence interval for the percent of registered voters who voted is,

$$0.6599 \pm 1.96 * \sqrt{\frac{0.6599 * (1 - 0.6599)}{151 + 293}} = (0.616, 0.704)$$

Check on the n's and values to make sure those who didn't vote are excluded:

```
> length(which(is.na(vote)==TRUE))
[1] 56
> length(vote)
[1] 500
```

N = 500-56 = 151+293 = 444, so those who didn't vote are excluded.

3C. Does the percent voting differ between males and females? Perform a chi-square test to examine this question, reporting the percent voting for males and for females along with the chi-square statistic and p-value.

```
> table(vote,sexf)
    sexf
vote   0   1
   0  82  69
   1 161 132
> chisq.test(table(vote,sexf),correct=TRUE)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(vote, sexf)
X-squared = 0.00081549, df = 1, p-value = 0.9772
```

Solution:

Percent voting for males = 161/(161+82) = 0.6626 = 66.26 %

Percent voting for females = 132/(69+132) = 0.6567 = 65.67 %

χ^2 (1 d.f.) = 0.00081549, p = 0.9772. We fail to reject the hull hypothesis, which is there is no significant difference between male and females on the percent with voting. Hence, with 95% confidence, we can say that the percent voting does not differ significantly between males and females.

3D. What is the odds ratio describing the odds of voting for females compared to males? Give the 95% confidence interval for this odds ratio.

```
> oddsratio.wald(table(vote,sexf))
$data
      sexf
vote     0   1 Total
  0     82  69   151
  1    161 132   293
  Total 243 201   444

$measure
    odds ratio with 95% C.I.
vote  estimate     lower    upper
   0 1.0000000        NA       NA
   1 0.9743451 0.6568483 1.445308

$p.value
    two-sided
vote midp.exact fisher.exact chi.square
   0         NA           NA         NA
   1  0.8969736    0.9200573  0.8972111

$correction
[1] FALSE

attr(,"method")
[1] "Unconditional MLE & normal approximation (Wald) CI"
```

Solution:
The output gives the odds ratio as 0.974, which means females have 0.974 times the odds of voting compared to males.
The 95% confidence interval for the odds ratio is (0.657, 1.445). Since 1.0 is contained in the confidence interval given above, there is no significant difference in the odds of voting for females vs. males.

3E. Perform a multiple logistic regression predicting whether someone votes from their age, sex, and political party.

```
> log.out <- glm(vote ~ age + sexf + relevel(factor(party),ref='1'),
+ family=binomial(link=logit))
> summary(log.out)

Call:
glm(formula = vote ~ age + sexf + relevel(factor(party), ref = "1"),
    family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0726  -1.2643   0.7476   0.9422   1.2498
```

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.500928   0.348212  -1.439 0.150272
age                              0.019342   0.006374   3.035 0.002408
sexf                            -0.035287   0.208538  -0.169 0.865631
relevel(factor(party), ref = "1")2  0.231287   0.220786   1.048 0.294839
relevel(factor(party), ref = "1")3  1.286830   0.333551   3.858 0.000114

(Intercept)
age                             **
sexf
relevel(factor(party), ref = "1")2
relevel(factor(party), ref = "1")3 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 569.29  on 443  degrees of freedom
Residual deviance: 544.38  on 439  degrees of freedom
  (56 observations deleted due to missingness)
AIC: 554.38

Number of Fisher Scoring iterations: 4

> exp(coef(log.out))
                  (Intercept)                           age                          sexf
                    0.6059683                     1.0195305                     0.9653284
relevel(factor(party), ref = "1")2 relevel(factor(party), ref = "1")3
                    1.2602211                     3.6212881

> exp(confint(log.out))
                                2.5 %   97.5 %
(Intercept)                     0.3049058 1.197272
age                             1.0069671 1.032482
sexf                            0.6415614 1.454408
relevel(factor(party), ref = "1")2 0.8181280 1.946015
relevel(factor(party), ref = "1")3 1.9306921 7.193526
```

Solution:

Results of a logistic regression predicting voting

Variable	Odds Ratio	p-value	95% CI
Age	1.0195	0.0024	(1.0070, 1.0325)
Sex (F vs. M)	0.9653	0.8656	(0.6416, 1.4544)
Political Party*			
Republicans	1.2602	0.2948	(0.8181, 1.9460)
Independents	3.6213	0.0001	(1.9307, 7.1935)

* Democrats are the reference category

Which of the variables in the model are significantly associated with the chance that someone votes? Describe the significant associations.
Solution: Given the 95% CI for the odds ratios, Age($p=0.0024<0.05$) and Independents party($p=0.0001<0.05$) in the Political Party are significantly associated with the chance that someone votes. Also, 95% CI for OR of Age is (1.0070, 1.0325), does not contain null value of 1.0, indicating voting significantly differs with age. 95% CI for OR of Independents party is (1.9307, 7.1935), does not contain null value of 1.0, indicating voting significantly differs with Independents party.

Controlling for sex and political party, people one year older have 1.0195 times the odds of voting as people one year younger.

Controlling for age and sex, Independents have 3.6213 times the odds of voting as Democrats.