Stat 431 Assignment 2 Spring 2021

Due by 4:00pm EDT on Friday June 18, 2021 via Crowdmark

Notes for Submission:

Upload your assignment directly to Crowdmark via the link you receive by email. It is your responsibility to make sure your solution to each question is submitted in the correct section, that the pages are rotated correctly, and that everything is legible. *Be sure to give yourself ample time to upload your solutions before the deadline.*

- Typed solutions are preferred. The R markdown file used to generate the assignment will be provided for students who wish to use this as a starting point for their solution.
- Organization and comprehensibility are part of a full solution. Consequently, points will be deducted for solutions that are not organized or are incomprehensible.
- Be sure to show your work and include all R code and relevant output for each question (where applicable). You should use R comments to document your code but do not embed your solution/interpretation within the R comments.

Question 1 [12 marks] Adapted from Problem 9.4 of Dunn & Smyth (2018)

In Topic 2e and in Sect. 9.3 (p. 336), the probit binomial glm was developed as a threshold model. Here consider using the *logistic distribution* with mean μ and variance σ^2 (standard deviation σ) as the tolerance distribution. The logistic distribution has the probability density function (PDF):

$$f(y;\mu,\sigma^2) = \frac{\pi \exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}{\sigma\sqrt{3}\left[1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}\right]^2}$$

for $-\infty < y < \infty$, $-\infty < \mu < \infty$, and $\sigma > 0$.

(a) **[2 marks]** Show that the logistic distribution is not a member of the exponential family.

$$
\begin{aligned}
f(y;\mu,\sigma^2) &= \frac{\pi \exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}{\sigma\sqrt{3}\left[1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}\right]^2} \\
&= explog(\frac{\pi \exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}{\sigma\sqrt{3}\left[1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}\right]^2}) \\
&= \exp\{log(\pi \exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}) - log(\sigma\sqrt{3}\left[1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}\right]^2)\} \\
&= \exp\{log\pi - (y-\mu)\pi/(\sigma\sqrt{3}) - log(\sigma\sqrt{3}) - 2log([1+\exp\{-(y-\mu)\pi/(\sigma\sqrt{3})\}])\}
\end{aligned}
$$

Therefore, we will have $\theta = -\frac{\pi}{\sigma\sqrt{3}}, \phi = 1, a(\phi) = 1$

$$
\begin{aligned}
b(\theta) &= \mu(-\frac{\pi}{\sigma\sqrt{3}}) = \mu\theta \\
c(y;\phi) &= log\pi - log(\sigma\sqrt{3}) - 2log(1+\exp\{-(y-\mu)\pi/(\sigma\sqrt{3})\}) \\
&= log\pi - log(\frac{-\pi}{\theta}) - 2log(1+\exp\{(y-\mu)\theta\})
\end{aligned}
$$

Since $c(y;\phi)$ contains θ in the formula, which is not allowed

Therefore, the logistic distribution is not a member of the exponential family

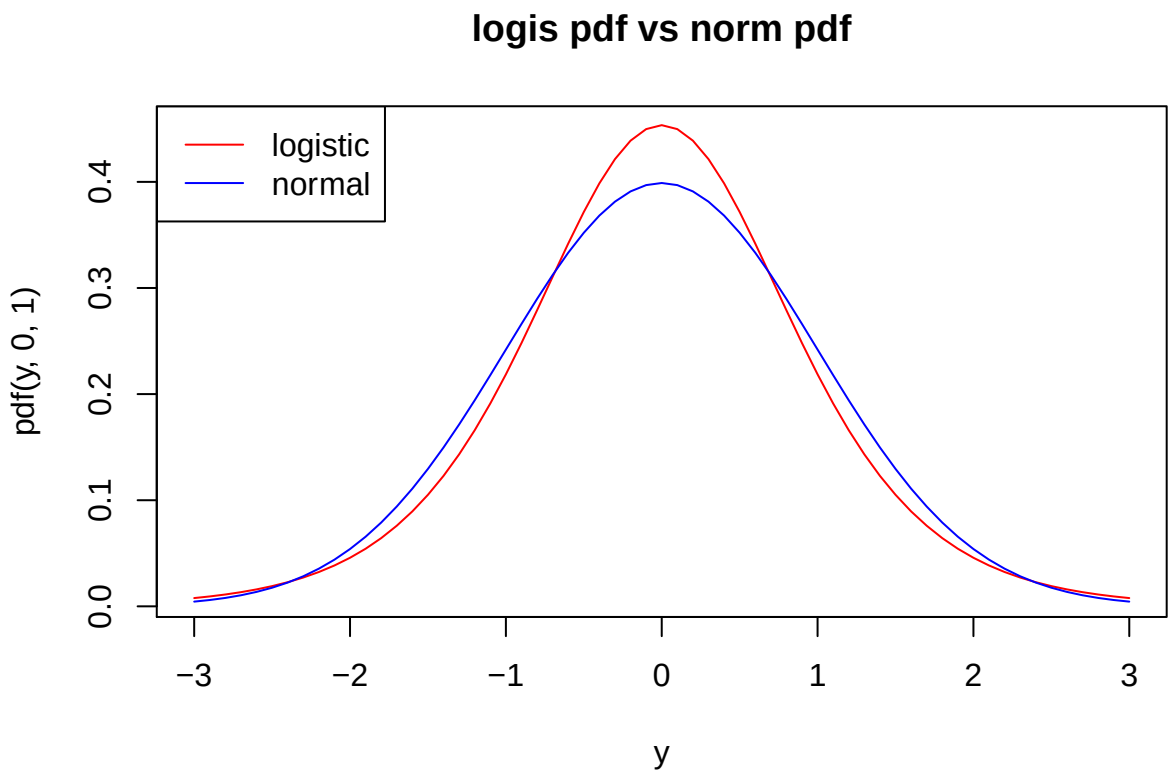
(b) **[3 marks]** Determine the cumulative distribution function (CDF) for the logistic distribution.

$$
\begin{aligned}
F(y) &= \int_{-\infty}^{y} \frac{\pi \exp\left\{-(s-\mu)\pi/(\sigma\sqrt{3})\right\}}{\sigma\sqrt{3}\left[1+\exp\left\{-(s-\mu)\pi/(\sigma\sqrt{3})\right\}\right]^2} ds \\
\text{Let } u &= 1+\exp\left\{-(s-\mu)\pi/(\sigma\sqrt{3})\right\}, du = -\frac{\pi}{\sigma\sqrt{3}}\exp\left\{-(s-\mu)\pi/(\sigma\sqrt{3})\right\} ds \\
\frac{\pi}{\sigma\sqrt{3}} ds &= -\frac{1}{u-1} du, \text{ sub this into the original we get:} \\
F(y) &= \int_{\infty}^{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}} \frac{-(u-1)}{(u-1)u^2} du \\
&= \int_{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}^{\infty} \frac{(u-1)}{(u-1)u^2} du \\
&= \int_{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}^{\infty} \frac{1}{u^2} du \\
&= \int_{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}^{\infty} \frac{1}{u^2} du \\
&= -\frac{1}{u}|_{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}^{\infty} \\
&= \frac{1}{1+\exp\left\{-(y-\mu)\pi/(\sigma\sqrt{3})\right\}}
\end{aligned}
$$

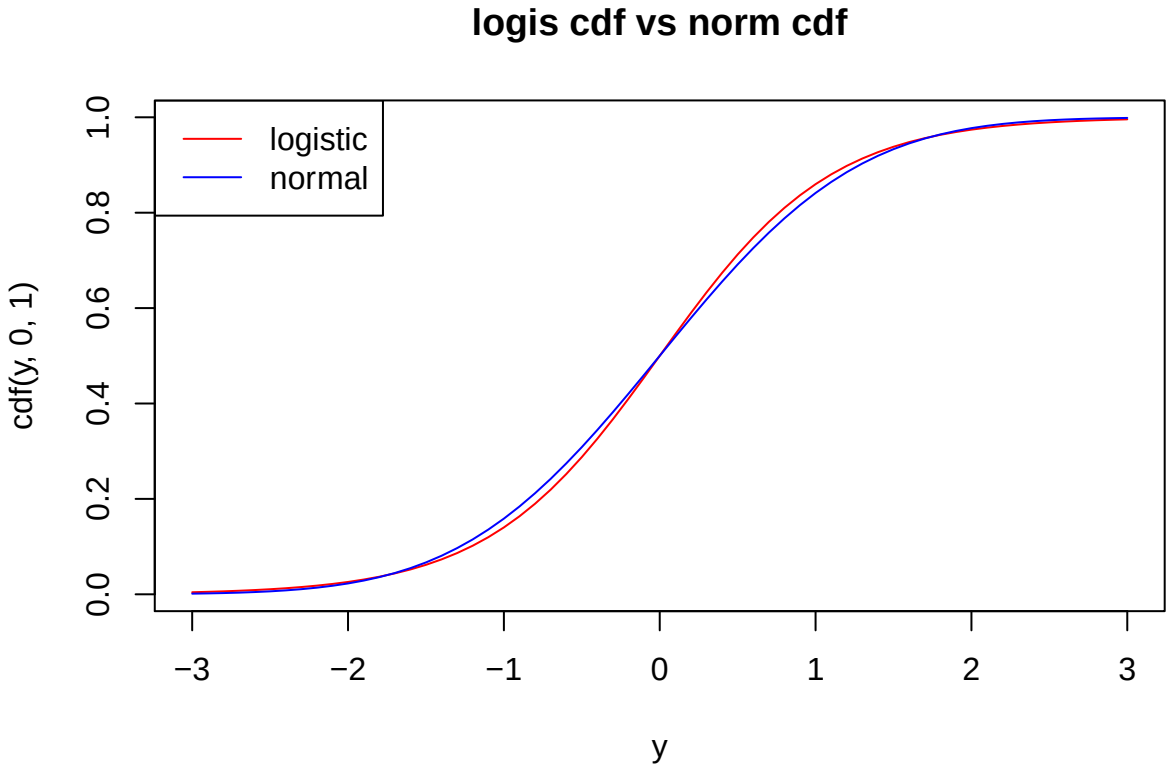
(c) **[4 marks]** Plot the PDF and CDF for the logistic distribution with mean 0 and variance 1. Also plot the same graphs for the normal distribution with mean 0 and variance 1. Comment on the similarities and differences between the two sets of functions.

```
> pdf <- function(y, m, s) {
+     pi * exp(-(y - m) * pi/(s * sqrt(3)))/((s * sqrt(3)) * (1 + exp(-(y - m) * pi/(s * sqrt(3))))^2)
+ }
>
```

```
> cdf <- function(y, m, s) {
+     1/(1 + exp(-(y - m) * pi/(s * sqrt(3))))
+ }
>
> y <- seq(-3, 3, 0.1)
>
> plot(y, pdf(y, 0, 1), type = "l", col = "red", main = "logis pdf vs norm pdf")
> lines(y, dnorm(y, 0, 1), col = "blue")
> legend("topleft", legend = c("logistic", "normal"), col = c("red", "blue"), lty = 1)
```




```
> plot(y, cdf(y, 0, 1), type = "l", col = "red", main = "logis cdf vs norm cdf")
> lines(y, pnorm(y, 0, 1), col = "blue")
> legend("topleft", legend = c("logistic", "normal"), col = c("red", "blue"), lty = 1)
```



Comments: From two graphs, we can see that both distributions have nearly the same tails but a variation in the middle. In the middle part in pdf plot, logistic distribution increases faster than normal distribution and logistic distribution has a higher value. From cdf plot, we see that from -1 to 0, logistic distribution has a higher value but from 0 to 1, normal distribution has a higher value.

(d) **[3 marks]** Show that if we use the logistic distribution as the tolerance distribution then this corresponds to a binomial glm with a logistic link function. What is the relationship between (β_0, β_1) the parameters of a simple logistic regression model and (μ, σ) the mean and standard deviation of the logistic tolerance distribution?

Let $\Pi(x) = F(x) = \dfrac{1}{1+\exp\left\{-(x-\mu)\pi/(\sigma\sqrt{3})\right\}}$

Then $g(\Pi) = logit(\Pi) = log(\dfrac{\Pi}{1-\Pi})$

$$= log(\frac{\frac{1}{1+\exp\left\{-(x-\mu)\pi/(\sigma\sqrt{3})\right\}}}{1-\frac{1}{1+\exp\left\{-(x-\mu)\pi/(\sigma\sqrt{3})\right\}}})$$

$$= log(\frac{1}{\exp\left\{-(x-\mu)\pi/(\sigma\sqrt{3})\right\}})$$

$$= (x-\mu)\pi/(\sigma\sqrt{3}) = \beta_0 + \beta_1 x$$

Therefore, $\beta_0 = -\dfrac{\mu\pi}{\sigma\sqrt{3}}, \beta_1 = \dfrac{\pi}{\sigma\sqrt{3}}$

Sub these two into $\Pi(x)$ we will get:

$$\Pi(x) = \frac{1}{1+\exp\left\{-(x-\mu)\pi/(\sigma\sqrt{3})\right\}}$$

$$= \frac{1}{1+\exp\{-\beta_0-\beta_1 x\}}$$

$$= \frac{\exp\{\beta_0+\beta_1 x\}}{1+\exp\{\beta_0+\beta_1 x\}}$$

, which is a binomial glm with a logistic link function as required

Question 2 [12 marks] Adapted from Problems 9.12 of Dunn & Smyth (2018)

Chromosome aberration assays are used to determine whether or not a substance induces structural changes in chromosomes. One study (Williams (1988)) compared the results of two substances at various doses (Table 9.12). A large number of cells were sampled at each dose to see how many were aberrant.

Table 9.12 The number of aberrant cells for different doses of two substances

Substance	Dose (in mg/ml)	No. cell samples	No. cells aberrant	Substance	Dose (in mg/ml)	No. cell samples	No. cells aberrant
A	0	400	3	B	0.0	400	5
A	20	200	5	B	62.5	200	2
A	100	200	14	B	125.0	200	2
A	200	200	4	B	250.0	200	4
				B	500.0	200	7

(a) **[3 marks]** Fit one or more binomial glm(s) with the logistic link to determine if there is evidence of a difference between the two substances. Justify your conclusions.

```
> s <- factor(c("A", "A", "A", "A", "B", "B", "B", "B", "B"))
> d <- c(0, 20, 100, 200, 0, 62.5, 125, 250, 500)
> c <- c(400, 200, 200, 200, 400, 200, 200, 200, 200)
> a <- c(3, 5, 14, 4, 5, 2, 2, 4, 7)
> y <- cbind(a, c - a)
>
> m1 <- glm(y ~ s + d, family = binomial(link = logit))
> summary(m1)

Call:
glm(formula = y ~ s + d, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2300  -0.4616  -0.1080   0.2164   3.0396

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.816884   0.214839 -17.766  < 2e-16 ***
sB          -0.802928   0.347343  -2.312  0.02080 *
d            0.002689   0.001013   2.655  0.00793 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.619  on 8  degrees of freedom
Residual deviance: 16.629  on 6  degrees of freedom
AIC: 52.423

Number of Fisher Scoring iterations: 5
```

Comment: Consider the null hypothesis H_0=0 where A and B are indifferent. we see that the p-value of β_1 is 0.02080<0.05, there is an evidence against the null hypothesis. Therefore, we reject H_0 and conclude that there is a difference between substances.

(b) **[3 marks]** Use the dose and the logarithm of dose as an explanatory variable in separate glms, and compare. Which is better, and why?

```
> d <- d + 1/1e+10  #since d contains zero
> m1 <- glm(y ~ s + d, family = binomial(link = logit))
> summary(m1)

Call:
glm(formula = y ~ s + d, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2300  -0.4616  -0.1080   0.2164   3.0396

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.816884   0.214839 -17.766  < 2e-16 ***
sB          -0.802928   0.347343  -2.312  0.02080 *
d            0.002689   0.001013   2.655  0.00793 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.619  on 8  degrees of freedom
Residual deviance: 16.629  on 6  degrees of freedom
AIC: 52.423

Number of Fisher Scoring iterations: 5
> m2 <- glm(y ~ s + log(d), family = binomial(link = logit))
> summary(m2)

Call:
glm(formula = y ~ s + log(d), family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3520  -1.1668  -0.6905   1.1295   2.3386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.48243    0.19968 -17.440  < 2e-16 ***
sB          -0.54310    0.30251  -1.795  0.07261 .
log(d)       0.03903    0.01425   2.739  0.00617 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25.619  on 8  degrees of freedom
Residual deviance: 14.094  on 6  degrees of freedom
AIC: 49.889

Number of Fisher Scoring iterations: 4
```

Comment: Then the $\hat{\beta}_{2a}$ is 0.002689 and $\hat{\beta}_{2b}$=0.04005. Since 1 unit increase in dose in model 1 increases the responses by 0.002689 and 1 unit increase in dose in model 2 by 0.04005. Therefore, the model with log of dose has a mroe significant impact on the responses. We will say that the second model is better than the first one.

(c) **[3 marks]** Compute the 95% confidence interval for the dose regression parameter, and interpret the regression parameter.

Since, $\hat{\beta}_2 = 0.002689$ and $se(\hat{\beta}_2) = 0.001013$, we will obtain the confidence interval as:

```
> c(0.002689 - 1.96 * 0.001013, 0.002689 + 1.96 * 0.001013)
[1] 0.00070352 0.00467448
```

β_0 is the log odds of aberrant cells where the substance A and 0 mg/ml dose are used. β_1 is the log odds of aberrant cells using substance A versus substance B. β_2 is the log odds of aberrant cells does up by 0.002689 with one unit increase in doses.

(d) **[3 marks]** Williams (1988) states that the spontaneous aberration rate is typically around 2%. Using the best fitting logistic regression model, determine the dose of each substance at which we expect an aberration rate greater than 2%.

```
> log(0.02/(1 - 0.02))
[1] -3.89182
```

Therefore, $\beta_0+\beta_1 x_1+\beta_2 x_2=-3.816884 - 0.802928x_1 + 0.002689x_2 \geq -3.89182$

For substance A, x_1=0, we will have the doses:

```
> (-3.89182 + 3.816884)/0.002689
[1] -27.86761
```

For substance B, x_1=1, we will have the doses:

```
> (-3.89182 + 3.816884 + 0.802928)/0.002689
[1] 270.7296
```

Comment: Therefore, even we don't use doses, we will get aberrant rate greater than or equal to 2%. But if we are using substance B, we must use over 270.7296 mg/ml to get the aberrant rate over 2%.

Question 3 [16 marks]

Browning et al. (2021) conducted a survey of university students in the United States from mid-March to early-May 2020 to assess the psychological impacts COVID-19 on students. We will use a subset of the data from this survey to address one of the study objectives: "to evaluate potential sociodemographic, lifestyle-related, and awareness of people infected with COVID-19 risk factors that could make students more likely to experience these [psychological] impacts.'' The analysis that you will conduct in this assignment question is simpler than that used by the authors and therefore you should not expect your results exactly match those in the published manuscript.

Download the original data from https://doi.org/10.1371/journal.pone.0245327.s010 and save it in your R Working Directory. The following R code reads in the original data file and prepares it for our analysis. We will restrict our analysis to the representative sample of students taken from North Carolina State University with complete data. You must run this code before proceeding the answer the questions below.

```
> # Save the original .csv file in your R Working Directory and then run this code block to
> # input the data and prepare it for our analysis.
> COVIDdata = read.csv("journal.pone.0245327.s010.csv")
>
> # Limit the data to students from NCSU and a restricted set of explanatory variables
> COVIDdata_NCSU = COVIDdata[(!is.na(COVIDdata$Source) & (COVIDdata$Source == "NCState")), names(COVIDdata) %in%
+     c("Health_General", "Hrs_Screen", "Hrs_Outdoor", "Hrs_Exercise", "Class_Self", "Infected_Any",
+         "Female", "BMI", "Educ_College_Grad", "Age", "Classification_High", "Ethnoracial_Group_White1_Asian2",
+         "Age_18to25")]
>
> # Remove observations with missing Ethnoracial data (all other variable are complete)
> COVIDdata_NCSU = COVIDdata_NCSU[!is.na(COVIDdata_NCSU$Ethnoracial_Group_White1_Asian2), ]
>
> # Create factor variables where necessary
> COVIDdata_NCSU$Infected_Any = factor(COVIDdata_NCSU$Infected_Any)
> COVIDdata_NCSU$Educ_College_Grad = factor(COVIDdata_NCSU$Educ_College_Grad)
> COVIDdata_NCSU$Ethnoracial_Group_White1_Asian2 = factor(COVIDdata_NCSU$Ethnoracial_Group_White1_Asian2)
> COVIDdata_NCSU$Age_18to25 = factor(COVIDdata_NCSU$Age_18to25)
>
> # str(COVIDdata_NCSU) # Display data set structure, commented out to save space
```

The following variables are included in our analysis data set:

- `Health_General`: Numeric measure of self-reported general health on range from (1=poor to 5=excellent)
- `Hrs_Screen`: Numeric measure of hours of screen time over the previous 24 hours
- `Hrs_Outdoor`: Numeric measure of hours spent outdoors over the previous 24 hours
- `Hrs_Exercise`: Numeric measure of hours spent exercising over the previous 24 hours
- `Class_Self`: Numeric measure of self-reported socioeconomic class (1=working class to 5=upper class)
- `Infected_Any`: Binary indicator for knowing someone infected with COVID-19
- `Female`: Numeric measure gender (=1 female to =0 male, includes some non-integer values)
- `BMI`: Numeric measure of self-reported Body Mass Index (kg/m^2)
- `Educ_College_Grad`: Binary variable =1 for graduate students and =0 for undergraduate students
- `Age`: Categorical/factor variables at five levels: "18 to 24","25 to 32", "33 to 44", ...
- `Classification_High`: Binary indicator that the student has a high COVID-19 psychological impact profile (the response of interest)
- `Ethnoracial_Group_White1_Asian2`: Categorical/factor variables at three levels: 0="Black or Hispanic", 1="Non-Hispanic White", 2="Non-Hispanic Asian"
- `Age_18to25`: Binary variable =1 for Age 18 to 24 and =0 otherwise (yes, this appears to have been incorrectly labeled 18to25 instead of 18to24)

(a) **[4 marks]** The R code below fits a logistic regression model to this data. Provide the estimate and interpretation of the *exponentiated* regression parameters associated with: `Hrs_Screen`, `Infected_Any`,

and both of the `Ethnoracial_Group_White1_Asian2` variables.

```
> # Fit a main effects logistic regression model
> model1 = glm(Classification_High ~ Female + Age + Ethnoracial_Group_White1_Asian2 + Class_Self +
+     Health_General + BMI + Hrs_Screen + Hrs_Outdoor + Hrs_Exercise + Educ_College_Grad + Infected_Any,
+     family = binomial(link = "logit"), data = COVIDdata_NCSU)
```

```
> summary(model1)

Call:
glm(formula = Classification_High ~ Female + Age + Ethnoracial_Group_White1_Asian2 +
    Class_Self + Health_General + BMI + Hrs_Screen + Hrs_Outdoor +
    Hrs_Exercise + Educ_College_Grad + Infected_Any, family = binomial(link = "logit"),
    data = COVIDdata_NCSU)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7050  -1.0757  -0.7634   1.1723   1.8212

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.137093   0.532321  -0.258 0.796763
Female                           0.641159   0.123443   5.194 2.06e-07 ***
Age25 to 32                      0.247117   0.163376   1.513 0.130392
Age33 to 44                      0.674348   0.335705   2.009 0.044563 *
Age45 to 54                      0.668704   0.558473   1.197 0.231158
Age55 to 64                    -12.303896 324.743746  -0.038 0.969777
Ethnoracial_Group_White1_Asian21   0.280955   0.201063   1.397 0.162309
Ethnoracial_Group_White1_Asian22   0.550986   0.236807   2.327 0.019980 *
Class_Self                      -0.166599   0.062700  -2.657 0.007883 **
Health_General                  -0.225893   0.059626  -3.789 0.000152 ***
BMI                              0.004515   0.013459   0.335 0.737301
Hrs_Screen                       0.034977   0.022164   1.578 0.114541
Hrs_Outdoor                     -0.051784   0.049113  -1.054 0.291706
Hrs_Exercise                     0.032813   0.073759   0.445 0.656420
Educ_College_Grad1               0.021507   0.146708   0.147 0.883452
Infected_Any1                    0.376658   0.140636   2.678 0.007401 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1809.2  on 1311  degrees of freedom
Residual deviance: 1726.1  on 1296  degrees of freedom
AIC: 1758.1

Number of Fisher Scoring iterations: 11
```

Estimate:

```
> exp(0.034977)  # Hrs_Screen
[1] 1.035596
> exp(0.376658)  # Infected_Any
```

```
[1] 1.457406
> exp(0.280955)  # Ethnoracial_Group_White1_Asian21
[1] 1.324394
> exp(0.550986)  # Ethnoracial_Group_White1_Asian22
[1] 1.734963
```

Hrs_screen: the odds ratio of the response of interest increases by 1.035596 with one unit increases in numeric measure of hours of screen time over the previous 24 hours, while other variables remain fixed.

Infected_Any: the odds ratio of the response of interest with knowing someone infected with COVID-19 versus not-known while keeping other variables fixed.

Ethnoracial_Group_White1_Asian21: the odds ratio of the response of interest who are Non-Hispanic White versus who are Black or Hispanic while keeping other variables fixed.

Ethnoracial_Group_White1_Asian22: the odds ratio of the response of interest who are Non-Hispanic Asian versus who are Black or Hispanic while keeping other variables fixed.

(b) **[3 marks]** The `Age` variable has five levels and is represented in `model1` by four binary variables. Fit a new model with the same linear predictor as `model1` but without the `Age` variable (call this `model2`). Use this model to conduct a deviance test of the null hypothesis that `Age` in not an important explanatory variable. Be sure to carefully state the null and alternative hypotheses in terms of the regression coefficients (be explicit about which model you are referring to) and give the formula of the test statistic and its asymptotic distribution under the null hypothesis. What is the conclusion of the test?

```
> model2 = glm(Classification_High ~ Female + Ethnoracial_Group_White1_Asian2 + Class_Self +
+     Health_General + BMI + Hrs_Screen + Hrs_Outdoor + Hrs_Exercise + Educ_College_Grad + Infected_Any
+     family = binomial(link = "logit"), data = COVIDdata_NCSU)
>
> summary(model2)

Call:
glm(formula = Classification_High ~ Female + Ethnoracial_Group_White1_Asian2 +
    Class_Self + Health_General + BMI + Hrs_Screen + Hrs_Outdoor +
    Hrs_Exercise + Educ_College_Grad + Infected_Any, family = binomial(link = "logit"),
    data = COVIDdata_NCSU)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6642  -1.0822  -0.7749   1.1747   1.7912

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -0.179258   0.530312  -0.338 0.735345
Female                           0.631456   0.122806   5.142 2.72e-07 ***
Ethnoracial_Group_White1_Asian21  0.253006   0.200339   1.263 0.206630
Ethnoracial_Group_White1_Asian22  0.548086   0.235880   2.324 0.020148 *
Class_Self                      -0.175670   0.062060  -2.831 0.004645 **
Health_General                  -0.213182   0.059191  -3.602 0.000316 ***
BMI                              0.009376   0.013291   0.705 0.480563
Hrs_Screen                       0.035006   0.022090   1.585 0.113036
Hrs_Outdoor                     -0.047407   0.048671  -0.974 0.330044
```

```
Hrs_Exercise                      0.024605   0.073020   0.337 0.736142
Educ_College_Grad1                0.072762   0.143110   0.508 0.611151
Infected_Any1                     0.349305   0.139804   2.499 0.012471 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1809.2  on 1311  degrees of freedom
Residual deviance: 1734.3  on 1300  degrees of freedom
AIC: 1758.3

Number of Fisher Scoring iterations: 4
```

The full model model1 is $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} + \beta_{15} x_{i15}$ where x_{i1} represents `Female`, x_{i2} represents `Age25 to 32`, x_{i3} represents `Age33 to 44`, x_{i4} represents `Age45 to 54`, x_{i5} represents `Age55 to 64`, x_{i6} represents `Ethnoracial_Group_White1_Asian21`, x_{i7} represents `Ethnoracial_Group_White1_Asian22`, x_{i8} represents `Class_Self`, x_{i9} represents `Health_General`, x_{i10} represents `BMI`, x_{i11} represents `Hrs_Screen`, x_{i12} represents `Hrs_Outdoor`, x_{i13} represents `Hrs_Exercise`, x_{i14} represents `Educ_College_Grad1`, x_{i15} represents `Infected_Any`.

Since in the reduced model, all ages are gone, we will have a null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ and $H_A : \beta_2 \neq 0 or \beta_3 \neq 0 or \beta_4 \neq 0 or \beta_5 \neq 0$.

Then, since the Deviance test statistic is $D = -2log[R(\hat{\pi}) = -2log(\frac{L(\tilde{\pi})}{L(\hat{\pi})})] = -2[l(\tilde{\pi}) - l(\hat{\pi})] \sim X^2_{(n-p)}$ where $\tilde{\pi}$ is model1 and $\hat{\pi}$ is model2, $\Delta D = D_0 - D_A = -2[l(\tilde{\pi}) - l(\hat{\pi})] \sim X^2_{(n-p)}$

```
> 1 - pchisq(model2$deviance - model1$deviance, 1300 - 1296)
[1] 0.08657205
```

Since 0.08657205>0.05, there's no evidence against the null hypothesis. We say that the reduced model 2 is adequate comparing to full model 1.

(c) **[3 marks]** An alternative to dropping `Age` from `model1` is to replace it with the binary `Age_18to25` variable. Fit a new model with the same linear predictor as `model1` but with `Age_18to25` instead of `Age` (call this `model3`). Use this model to conduct a deviance test of the null hypothesis that this reparameterization is adequate. Be sure to carefully state the null and alternative hypotheses in terms of the regression coefficients (be explicit about which model you are referring to) and give the formula of the test statistic and its asymptotic distribution under the null hypothesis. What is the conclusion of the test?

```
> model3 = glm(Classification_High ~ Female + Ethnoracial_Group_White1_Asian2 + Age_18to25 +
+     Class_Self + Health_General + BMI + Hrs_Screen + Hrs_Outdoor + Hrs_Exercise + Educ_College_Grad +
+     Infected_Any, family = binomial(link = "logit"), data = COVIDdata_NCSU)
> summary(model3)

Call:
glm(formula = Classification_High ~ Female + Ethnoracial_Group_White1_Asian2 +
    Age_18to25 + Class_Self + Health_General + BMI + Hrs_Screen +
    Hrs_Outdoor + Hrs_Exercise + Educ_College_Grad + Infected_Any,
    family = binomial(link = "logit"), data = COVIDdata_NCSU)
```

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.733  -1.078  -0.765   1.173   1.811

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      0.161778   0.553504   0.292 0.770073
Female                           0.641802   0.123211   5.209  1.9e-07 ***
Ethnoracial_Group_White1_Asian21 0.269104   0.200818   1.340 0.180234
Ethnoracial_Group_White1_Asian22 0.533151   0.236277   2.256 0.024042 *
Age_18to251                     -0.328629   0.149089  -2.204 0.027507 *
Class_Self                      -0.160614   0.062524  -2.569 0.010204 *
Health_General                  -0.224257   0.059544  -3.766 0.000166 ***
BMI                              0.005671   0.013430   0.422 0.672814
Hrs_Screen                       0.033679   0.022133   1.522 0.128099
Hrs_Outdoor                     -0.049882   0.048827  -1.022 0.306962
Hrs_Exercise                     0.033545   0.073409   0.457 0.647700
Educ_College_Grad1               0.012689   0.146094   0.087 0.930786
Infected_Any1                    0.371860   0.140436   2.648 0.008099 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1809.2  on 1311  degrees of freedom
Residual deviance: 1729.4  on 1299  degrees of freedom
AIC: 1755.4

Number of Fisher Scoring iterations: 4
```

$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5$ vs H_A : one of the coefficient does not equal to others.

$\Delta D = D_0 - D_A = -2[l(\hat{\pi}) - l(\tilde{\pi})] \sim X^2_{(n-p)}$ where $\tilde{\pi}$ is model 1 and $\hat{\pi}$ is model 3.

```
> 1 - pchisq(model3$deviance - model1$deviance, 1299 - 1296)
[1] 0.3513265
```

Since 0.3513265>0.05, we can not reject the null hypothesis. Therefore, the reduced model 3 is adequate comparing to model 1.

(d) **[2 marks]** Regardless of your conclusions from part (b) and (c) use `model3` for parts (d) to (f). Calculate the model-based probability a NCSU student with the same personal characteristics as you (in April 2020) would have a high COVID-19 psychological impact profile. Or if you do not wish to use your own personal characteristics you may instead make-up the profile of a hypothetical student.

```
> x <- c(1, 0, 0, 1, 1, 1, 4, 21, 8, 10, 0, 0, 0)
> exp(x %*% summary(model3)$coefficients[1:13])/(1 + exp(x %*% summary(model3)$coefficients[1:13]))
          [,1]
[1,] 0.3096753
```

I will have 30.96753% probability of having a high COVID-19 psychological impact profile.

(e) **[2 marks]** Calculate a 95% confidence interval for your estimate from part (d).

```
> tmp <- summary.glm(model3)
>
> v <- tmp$cov.unscaled
>
> x <- as.matrix(x, 13, 1)
>
> t(x) %*% v %*% x
          [,1]
[1,] 0.2707391
>
> c(0.3096753 - 1.96 * sqrt(0.2707391), 0.3096753 + 1.96 * sqrt(0.2707391))
[1] -0.7101636  1.3295142
```

Therefore, the 95% confidence interval is (-0.7101636, 1.3295142)

(f) **[2 marks]** Write one paragraph summarizing the results and highlighting the important associations identified by this logistic regression model.

As we can see, a normal student like me are less likely to be identified as a high COVID-19 psychological impact profile because I have only 30.96753% probability. As we can see in the model, "female" has the lowest p-value with only 1.9e-07, this implies that "female" is an important association with the model. Another one is "health general" with p-value of 0.000166. Even though it has a much higher p-value than "female", it's still a main factor of psychological impact, which makes sense in the real life. "Educ_College_Grad1" has a p-value of 0.930786. This means that the logistic regression model not really depends on the education level.