STAT 426 Assignment 10

Due Tuesday, November 9, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

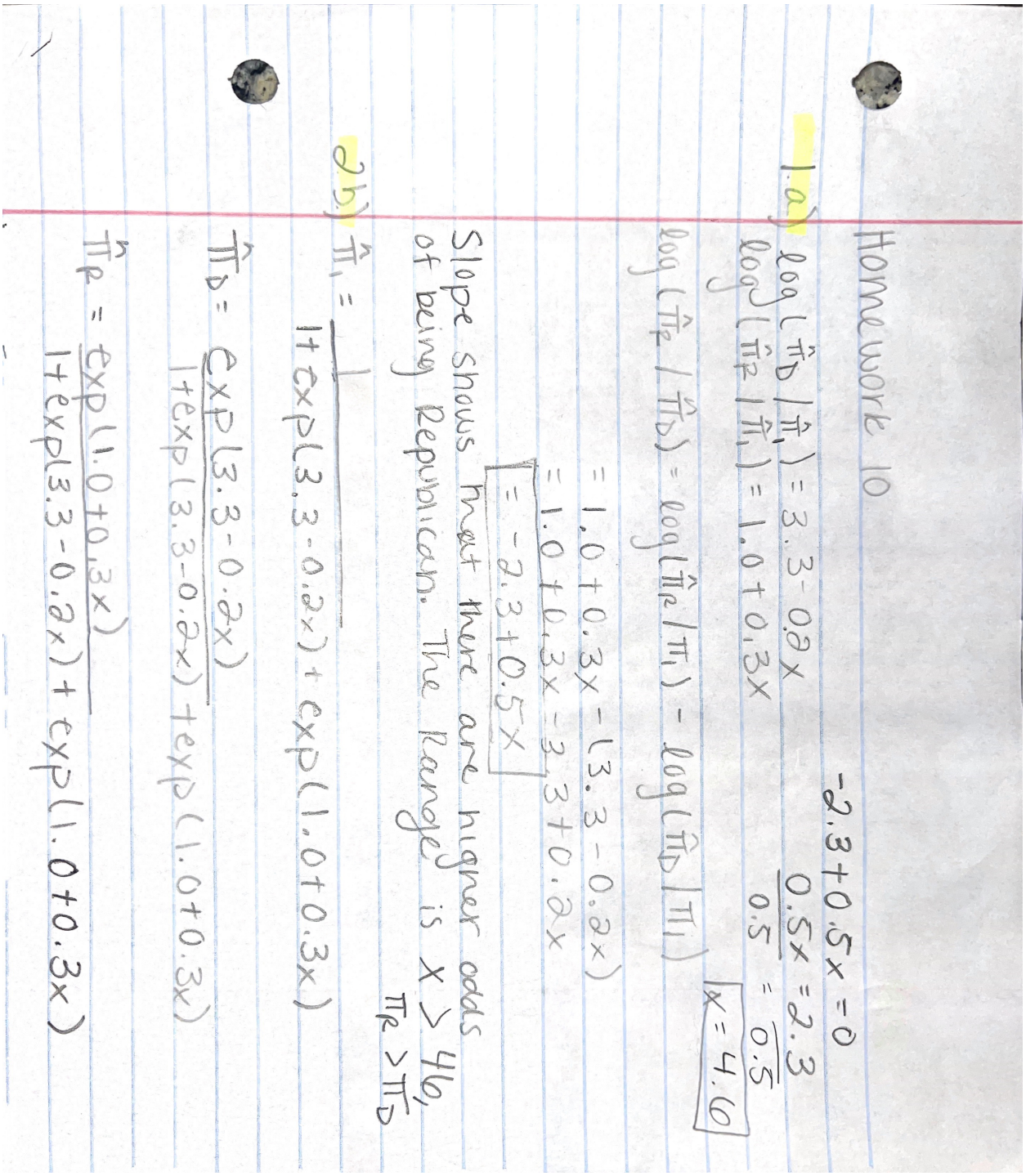
Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1. (6 pts) Interpreting a baseline category logistic model

(Agresti, Exercise 7.2) A model fit predicting preference for U.S. president (Democrat, Republican, Independent) using x = annual income (in \$10,000) gives the prediction equations $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$.

a) (2 pts) Find the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$ and interpret the slope. For what range of x is $\hat{\pi}_R > \hat{\pi}_D$?

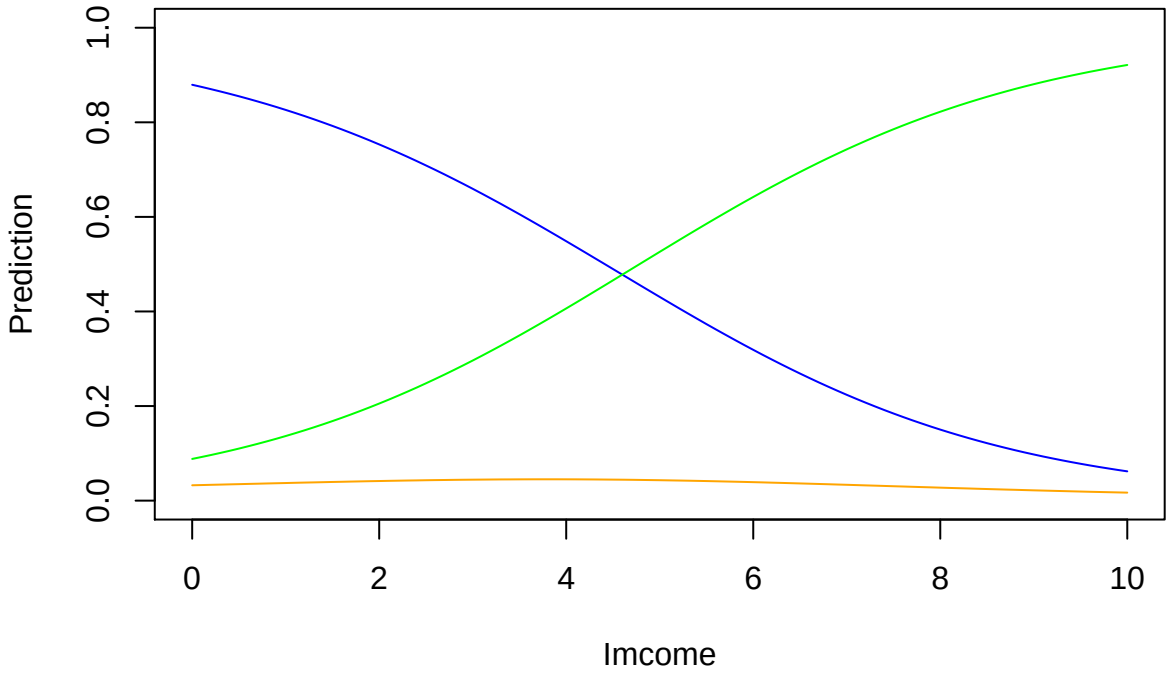
b) (2 pts) Find the prediction equations for $\hat{\pi}_I$, $\hat{\pi}_D$, and $\hat{\pi}_R$.



c) (2 pts) Plot $\hat{\pi}_D$, $\hat{\pi}_I$ and $\hat{\pi}_R$ on the same graph for x between 0 and 10 (recall how to plot functions using the `curve` function in R).

```
pi_i = function(x){1/ (1 + exp(3.3 - 0.2*x) + exp(1.0+0.3*x))}
pi_d = function(x){exp(3.3 - 0.2*x)/ (1 + exp(3.3 - 0.2*x) + exp(1.0+0.3*x))}
pi_r = function(x){exp(1.0 + 0.3*x)/ (1 + exp(3.3 - 0.2*x) + exp(1.0+0.3*x))}
```

```
curve(pi_i, xlim = c(0,10), ylim = c(0,1), col = "orange", xlab = "Imcome", ylab = "Pred
curve(pi_d, col = "Blue", xlab = "Imcome", ylab = "Prediction", add = TRUE)
curve(pi_r, col = "Green", xlab = "Imcome", ylab = "Prediction", add = TRUE)
```



Problem 2. (6 pts) Finding a good model

The data below refer to the dependence of Party Affiliation on Gender and Race.

```
political = data.frame(
  Gender = c("M", "M", "F", "F"),
  Race = c("W", "B", "W", "B"),
  Dem = c(132, 42, 172, 56),
  Rep = c(176, 6, 129, 4),
  Ind = c(127, 12, 130, 15)
)
political
```

```
##   Gender Race Dem Rep Ind
## 1      M    W 132 176 127
## 2      M    B  42   6  12
## 3      F    W 172 129 130
## 4      F    B  56   4  15
```

a) (2 pts) Find a baseline category model with party as the nominal response that is as simple as possible while still providing an adequate fit to the data. Show statistically that it fits well and that none of its variables should be dropped.

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
mod1 <- vglm(
cbind(Dem, Rep, Ind) ~ Gender+Race, family=multinomial, data= political)

mod2 <- vglm(
cbind(Dem, Rep, Ind) ~ Gender, family=multinomial, data= political)

mod3 <- vglm(
cbind(Dem, Rep, Ind) ~ Race, family=multinomial, data= political)

AIC(mod1)
```

```
## [1] 52.35686
```

```
AIC(mod2)
```

```
## [1] 125.0811
```

```
AIC(mod3)
```

```
## [1] 61.79868
```

```
#Based off of AIC, Mod1 with all variables is preferred.
```

b) (2 pts) What is the baseline category for party in your model?

- My baseline in my model is The Independent party.

c) (2 pts) Using your fitted model in a), explain how the log odds of Democrat versus Republican vary across Gender and Race.

```
summary(mod1)
```

```
##
## Call:
## vglm(formula = cbind(Dem, Rep, Ind) ~ Gender + Race, family = multinomial,
##     data = political)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1   1.3882     0.2296   6.045 1.49e-09 ***
## (Intercept):2  -1.1771     0.3807  -3.092  0.00199 **
## GenderM:1      -0.2202     0.1583  -1.391  0.16412
## GenderM:2       0.3526     0.1651   2.136  0.03271 *
## RaceW:1        -1.1183     0.2335  -4.789 1.68e-06 ***
## RaceW:2         1.1598     0.3801   3.051  0.00228 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 0.1982 on 2 degrees of freedom
##
## Log-likelihood: -20.1784 on 2 degrees of freedom
##
## Number of Fisher scoring iterations: 3
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  3  of the response
```

- The first odds ratio compares being Democratic to an Independent. The second compares being Republican to an Independent. A male in the first odds ratio is likely to be a Democrat compered to a female. A male in the second odds ratio is more likely to be a Republican compared to a female. For someone who is white in the first odds ratio is less likely to be a Democrat compared to someone who is in the second odds ratio who is black. However, those who are white in the second odds ratio more likely to be a Republican.

Problem 3 (8 pts) Working with an ordinal model

Consider an $I \times J$ contingency table with an ordinal column variable Y having levels 1,2,. . . , J, and row variable X having numerical scores $\{x_i = i\}$ for rows $i = 1, 2, \ldots, I$. We consider the model

$$\text{logit}\,[P(Y \leq j \,|\, X = x_i)] = \alpha_j + \beta x_i, \quad i = 1, \ldots, I;\ j = 1, \ldots, J-1. \tag{1}$$

a) (2 pts) Show that $\text{logit}\left[P(Y \leq j \mid X = x_{i+1})\right] - \text{logit}\left[P(Y \leq j \mid X = x_i)\right] = \beta$.

3a.) $\text{logit}[P(Y \leq j) \mid X = x_i)] = \alpha_j + \beta x_i$

$\text{logit}[P(Y \leq j) \mid X + x_{i+1}] = \alpha_j + \beta x_{i+1}$

$\text{logit}[P(Y \leq j) \mid X = x_i + 1)] - \text{logit}[P(Y \leq j) \mid X = x_i)] =$

$= \alpha_j + \beta x_{i+1} - (\alpha_j + \beta x_i) = \beta(x_{i+1} - x_i) = \beta$

$(x_{i+1} - x_i) = 1$

b) (2 pts) Consider the 2×2 table with expected cell counts represented by μ_{ab}, $a = 1, 2$; $b = 1, 2$.

	$Y_j \leq j$	$Y_j > j$
$X = x_{i+1}$	μ_{11}	μ_{12}
$X = x_i$	μ_{21}	μ_{22}

Show that the result in a) implies that e^{β} is the odds ratio for this table. Hint: Consider how to express conditional probabilities such as $P(Y \leq j \,|\, X = x_i)$ in terms of the expected cell counts μ_{ab}.

c) (2 pts) Assuming all IJ cells have nonzero counts, find the residual degrees of freedom for the model in Equation (1).

d) (2 pts) Show that if $\beta = 0$ then the model implies that $P(Y = j \,|\, X = x_j)$ does not depend on x_j, so X and Y are independent.

3b) $\text{logit}[P(Y=1|X=i)] = \log \frac{P(Y=1|X=i)}{P(Y=2|X=i)}$

$= \log \frac{\pi_{i1}}{\pi_{i2}} = \log \mu_{i1} - \log \mu_{i2}$

$= \beta(x_1 - x_2)$

When $p=1$, $x_1 - x_2 = 1$, the cumulative odds ratio is e^{β}

3c.) $(I-1)$ = categories of X

$(J-1)$ = categories of Y

df = #cell − #parameters

residual df = $(I-1)(J-1)$

3d) $P(Y=j|X=x_j) = P(Y \leq j | X=x_j) - P(Y \leq j-1 | X=x_j)$

$= \alpha_j + \beta x_i - \alpha_{j-1} - \beta x_i$ if $\beta = 0$

$= \alpha_j - \alpha_{j-1}$. → Does not depend on x_j

Therefore X & Y are independent