Multiple Regression
Second Exam

BeyChella Edition

Question	Points	Max Points
1		10
2		25
3		35
4		10
5		10
6		10
Total		100

Instructions: The examination lasts **80** minutes and all books are closed. You may use a calculator and a single 8.5 by 11 page of notes, front and back. Cell phones may not be used at any point. No interaction with anyone except the instructor is allowed. Show all of your work clearly!
In case of potential errors or ambiguity on the exam, please note them and state your assumptions.

Honor Code Statement: Smith College expects all students to be honest and committed to the principles of academic and intellectual integrity in their preparation and submission of course work and examinations.

Students and faculty at Smith are part of an academic community defined by its commitment to scholarship, which depends on scrupulous and attentive acknowledgement of all sources of information, and honest and respectful use of college resources.

Dishonest Examination Behavior: The unauthorized giving or receiving of information during examinations or quizzes (this applies to all types, such as written, oral, lab or take-home) is dishonest examination behavior.

Signature: I have read the above instructions and agree to abide by the Honor Code in taking this exam.

(printed name)

(signature)

1. **(10 points)** Question 1 includes 5 multiple choice / short answer questions about 'R' (1.A-1.E) worth **2 points** each.

   Questions 1.A-1.B refer to the code below:

```
> library(ipums)
> usa_ddi <- read_ipums_ddi("usa_00001.xml")
> usa_data <- read_ipums_micro(usa_ddi, verbose = FALSE)
```

   (a) What is the `usa_00001.XML` file?
      i. the data
      ii. **the data dictionary**
      iii. your variable search results from IPUMS
      iv. a description of how the study data were originally collected

   (b) What does the second line of code do?
      i. **reads in the zipped data file into 'R'**
      ii. reads in the data dictionary
      iii. imports the micro search results
      iv. generates a help file for your dataset in 'R'

   (c) What function in 'R' do we use to keep a specified list of variables?
      i. 'mutate()'
      ii. **'select()'**
      iii. 'filter()'
      iv. 'group_by()'

   (d) What function in 'R' do we use to keep a specified set of observations?
      i. 'mutate()'
      ii. 'select()'
      iii. **'filter()'**
      iv. 'group_by()'

   (e) If you had a labeled factor variable ('classyear': "First Year", "Sophomore", "Junior", "Senior"), which of the following would generate an indicator variable of whether or not someone was a sophomore. Select all that apply.
      i. `mutate(as_factor(if_else(classyear=="Sophomore", Sophomore, Not Sophomore)))`
      ii. `mutate(as_factor(if_else(classyear=="Sophomore",1,0)))`

Over the past two weekends, Beyoncé performed the headlining set on Saturday night at Coacella, a large, three day (Fri-Sun) music festival in Southern California. Her performance drew on her impressive catalog of music as well as many African-American musical and cultural traditions. It quickly generated considerable praise and critical attention – so much so that people quickly redubbed the festival BeyChella.

2. **25 points** Before Coachella, Beyoncé's mom, Ms. Tina Knowles-Lawson, expressed concern that White audience members may not understand all of the references, including the salience of "Lift Every Voice and Sing". To test that hypothesis, the *House of Deréon Foundation* funded a survey that asked a sample of 2,843 people for their race (Black/Hispanic/Other/White) and whether they knew "Lift Every Voice and Sing" was considered the Black National Anthem (yes/no).

```
     race
song  White Other Hispanic Black
  No   1116   142      148    66
  Yes   664   107       78   522
```

   (a) **(3 points)**. What were the odds that White attendees knew "Lift Every Voice and Sing"?
   $\frac{664}{1116} = 0.5949821$

   (b) **(3 points)**. What were the odds that Black attendees knew "Lift Every Voice and Sing"?
   $\frac{522}{66} = 7.909091$

```
Call:
glm(formula = song ~ race, family = binomial, data = levas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0915  -0.9663  -0.9201   1.4043   1.4587

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.51922    0.04901 -10.594   <2e-16 ***
raceOther     0.23623    0.13708   1.723   0.0848 .
raceHispanic -0.12128    0.14825  -0.818   0.4133
raceBlack     2.58724    0.13953  18.543   <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: 3937.6  on 2842  degrees of freedom
Residual deviance: 3396.1  on 2839  degrees of freedom
AIC: 3404.1

Number of Fisher Scoring iterations: 4
```

(c) **(6 points)**. When Blue Ivy (the foundation's Data Science Intern) analyzed the data, she decided to fit a logistic regression; her raw output is below. Show numerically how your answers above correspond to her output above. Briefly explain the relationship in a sentence.

OR: $\frac{7.909091}{0.5949821} = 13.29$

OR: $e^{\hat{\beta}_{raceBlack}} = e^{2.58724} = 13.29$

You can calculate the odds ratios of someone of black race knowing the song compared to the odds of white race knowing the song from either the table or by exponentiating the $\hat{\beta}$ coefficient. Either approach should yield the same results.

(d) **(3 points)**. Write the hypothetical and fitted regression equation (in logit form) for Blue Ivy's model.

$log(odds) = \beta_0 + \beta_1 raceOther + \beta_2 raceHispanic + \beta_3 raceBlack$

$log(odds) = \hat{\beta}_0 + \hat{\beta}_1 raceOther + \hat{\beta}_2 raceHispanic + \hat{\beta}_3 raceBlack$ or $log(odds) = -0.51922 + 0.23623 raceOther + -0.12128 raceHispanic + 2.58724 raceBlack$

```
Call:
glm(formula = song ~ race, family = binomial, data = levas)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0915  -0.9663  -0.9201   1.4043   1.4587

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.51922    0.04901 -10.594   <2e-16 ***
raceOther     0.23623    0.13708   1.723   0.0848 .
raceHispanic -0.12128    0.14825  -0.818   0.4133
raceBlack     2.58724    0.13953  18.543   <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: 3937.6  on 2842  degrees of freedom
Residual deviance: 3396.1  on 2839  degrees of freedom
AIC: 3404.1

Number of Fisher Scoring iterations: 4
```

(e) **(10 points)** Is there a significant difference in knowing this song by race? Use the evidence from Blue Ivy's model to test this question.

i. State the formula and related hypothesis for the test.
$H_0 : \beta_i = 0$
$H_A : \beta_i \neq 0$
$G = -2log(Likelihood)_{null} - -2log(Likelihood)_{residual}$

ii. Conduct the test and state your conclusion from the test in a sentence. *[χ^2 critical values for 1 d.f.=3.84, 2 d.f.=5.991, 3 d.f.=7.815, 4 d.f.=9.488]*
3937.6-3396.1 = 541.5
Since $541.5 > 7.815$, we can reject the null hypothesis and conclude that there is a difference in the odds of knowing the song by race.

Questions 3-5 use the same data, described here.

Singing along at concerts has been associated with concertgoers' vocal strain and temporary voice loss. Beyoncé had a set that was >1 hour long and included 48 songs (including samples; some songs were longer than others). She finished her set with "Love on Top", a song known for its key-changes, where the song modulates to a higher musical key such that the same tune is being sung at increasingly higher notes (the song has 5 total key changes).

After attending weekend 1 and noticing how many people were missing their voices on Sunday (the day after Beyoncé's set), Cristina Yang, MD, PhD advised Coachella to gather data on all 125,000 attendees the day after Beyoncé's set on the number of Beyoncé songs they sang along to during her set, whether or not they sang along to "Love on Top", how many of the "Love on Top" key 5 changes they sang, and whether or not they lost their voice (yes vs. no [reference]).

3. **(35 points)**. Bring the beat* in. (*coefficients) The following tables have rows of the beta coefficients and standard errors in parentheses; numbered columns reflect each of the four regression models that were fit.

Log-Odds of Losing Voice at #Beychella

	Dependent variable:			
	lostvoice			
	(1)	(2)	(3)	(4)
numsongs	0.050***	0.051***	0.052***	0.052***
	(0.002)	(0.002)	(0.002)	(0.002)
loveontop		1.756***	1.796***	0.888***
		(0.046)	(0.047)	(0.070)
keychanges			0.814***	0.661***
			(0.021)	(0.022)
loveontop:keychanges				1.046***
				(0.074)
Constant	2.784***	2.060***	0.771***	0.963***
	(0.031)	(0.033)	(0.042)	(0.045)
Observations	125,000	125,000	125,000	125,000
Log Likelihood	-12,975.580	-12,066.120	-11,107.180	-10,972.360
Akaike Inf. Crit.	25,955.160	24,138.250	22,222.360	21,954.720
Note:				*p<0.1; **p<0.05; ***p<0.01

(a) **(5 points)** Based on model 1, calculate and interpret the estimated odds of voice loss for two individuals who sang 36 and 37 songs, respectively, from Model 1. Use this information to calculate the odds ratio of voice loss between singing 36 and 37 songs.

$$Odds_{36} = e^{2.784+(0.050*36)} = 97.90523$$

The odds of losing your voice after singing 36 songs was 97.9 (to 1).

$$Odds_{37} = e^{2.784+(0.050*37)} = 102.9249$$

The odds of losing your voice after singing 37 songs was 102.9 (to 1).

$$\frac{102.9249}{97.90523} = 1.051271$$

The odds of losing your voice for every additional song sung is 1.05 times higher than someone who sang one fewer songs, on average.

(b) **(3 points)** Based on Model 1, calculate the odds ratio of voice loss for 'numsongs' and interpret it in a sentence.

$$OR = e^{0.050)} = 1.05$$

The odds of losing your voice is 1.05 higher for every additional song sung, on average.

(c) **(3 points)** Based on model 2, calculate the odds ratio of voice loss for 'loveontop' and interpret it in a sentence.

$$OR = e^{1.756)} = 5.79$$

The odds of losing your voice is 5.79 times higher, on average, among those who sang Love on Top compared to those who did not sing Love on Top, adjusted for the number of songs sung.

(d) **(3 points)** Based on model 3, calculate the 95% CI for the odds ratio for voiceloss for 'keychanges'. $z^* = 1.96$

$$OR = e^{0.814 \pm 1.96 \times 0.021)} = 2.17, 2.35$$

We are 95% confident that the true odds of losing your voice for each additional key change sung was between 2.17 and 2.35 times singing one fewer key change, adjusted for singing love on top and total number of songs sung.

Log-Odds of Losing Voice at #Beychella

	Dependent variable:			
	lostvoice			
	(1)	(2)	(3)	(4)
numsongs	0.050***	0.051***	0.052***	0.052***
	(0.002)	(0.002)	(0.002)	(0.002)
loveontop		1.756***	1.796***	0.888***
		(0.046)	(0.047)	(0.070)
keychanges			0.814***	0.661***
			(0.021)	(0.022)
loveontop:keychanges				1.046***
				(0.074)
Constant	2.784***	2.060***	0.771***	0.963***
	(0.031)	(0.033)	(0.042)	(0.045)
Observations	125,000	125,000	125,000	125,000
Log Likelihood	-12,975.580	-12,066.120	-11,107.180	-10,972.360
Akaike Inf. Crit.	25,955.160	24,138.250	22,222.360	21,954.720
Note:				*p<0.1; **p<0.05; ***p<0.01

(e) **(3 points)** Based on model 4, calculate the odds ratio of voice loss for 'keychanges' and interpret it in a sentence.

$$OR = e^{0.814)} = 2.26$$

The odds of losing your voice is 2.26 times higher, on average, for every additional key change someone sang, adjusted for singing Love on Top and number of songs sung.

(f) **(4 points)** Based on model 4, calculate the probability of losing their voice for someone who sang 36 songs and didn't sing "Love on Top".

odds: $e^{0.963+(0.052*36)} = 17.0304$

probability: $\frac{17.0304}{18.0304} = 0.944$

(g) **(4 points)** Based on model 4, calculate the probability of losing their voice for someone who sang 36 songs and sang all 5 key changes in "Love on Top".

odds: $e^{0.963+(0.052*36)+(0.888)+(0.661*5)+(1.046*5)} = 210659.8$

probability: $\frac{210659.8}{210660.8} = 0.999$

(h) **(10 points)** Is Model 4 better than Model 2? State your hypotheses, conduct the test, and intrepret your conclusion in a sentence in the context of this scenario. *[χ^2 critical values for 1 d.f.=3.84, 2 d.f.=5.991, 3 d.f.=7.815, 4 d.f.=9.488]*

$H_0 : \beta_3 = \beta_4 = 0$

$H_A : \beta_i \neq 0$

$G = -2log(Likelihood)_{nested} - -2log(Likelihood)_{full}$

$G = (-2 * -12,066.120) - (-2 * -10,972.360)$

$G = 2187.52$

Since 2187.52>5.99, we can reject the null hypothesis that both keychanges and the interaction between keychanges and Love on Top are both 0 and conclude that at least one of them is not 0. Thus, model 4 is better.

Log-Odds of Losing Voice at #Beychella

	Dependent variable:			
	lostvoice			
	(1)	(2)	(3)	(4)
numsongs	0.050***	0.051***	0.052***	0.052***
	(0.002)	(0.002)	(0.002)	(0.002)
loveontop		1.756***	1.796***	0.888***
		(0.046)	(0.047)	(0.070)
keychanges			0.814***	0.661***
			(0.021)	(0.022)
loveontop:keychanges				1.046***
				(0.074)
Constant	2.784***	2.060***	0.771***	0.963***
	(0.031)	(0.033)	(0.042)	(0.045)
Observations	125,000	125,000	125,000	125,000
Log Likelihood	-12,975.580	-12,066.120	-11,107.180	-10,972.360
Akaike Inf. Crit.	25,955.160	24,138.250	22,222.360	21,954.720
Note:				*p<0.1; **p<0.05; ***p<0.01

4. **(10 points total)** Visually depict each of these four models on the log-odds scale, with 'numsongs' on the x-axis. Don't worry about the exact numbers on the axes, just illustrate the general direction and magnitude the coefficients. Clearly label where each term in the model is being represented. State any assumptions you have to make.

   (a) Model 1 Simple logistic regression (looks linear on the log-odds scale) - one intercept and one slope

   (b) Model 2 Logistic regression with parallel slopes: two intercept and one slope

   (c) Model 3 Multiple logistic regression with parallel slopes: two intercept and one slope (now in 3D, so planes not lines)

   (d) Model 4 Multiple logistic regression with different slopes: two intercept and two slopes (still in 3D)

5. **10 points**. What are the assumptions of these regression models and are they met? Briefly describe each assumption and state whether you believe they are met. If you cannot evaluate a given assumptions from the data provided, clearly state all steps you would take to evaluate that assumption.

   Independence: Whether the observations are independent from each other. You need to know something about the data to evaluate. It's met in this case if individuals only surveyed once.

   Randomness: Whether the explanatory variable or associated between explanatory and response variable were randomly assigned or if sampling was random. again, you have to know something about the data. Probably not met – singing along wasn't randomly assigned.

   Linearity: Divide the sample into equal sized groups (i.e., groups of people based on how many songs they sang), estimate the mean value of losing their voice for each group, plot the log odds versus the mean value of how many songs that group sang. we can't do that with the information given here.