PM592: Regression Analysis for Data Science
Exam 2 – Fall 2022

Instructions

- Answer questions directly on the exam sheet and show all work.
- You may use your class notes, R software, and a calculator.
- You may **not** consult with any resources that are not a part of this class, including obtaining outside help through websites or talking to others about this exam.
- You may not discuss this exam with classmates until after the final due date.
- Unless otherwise stated, use α = .05 when testing statistical hypotheses.
- You have 180 minutes to submit the exam after accessing it. Plan ahead as the submission process may take longer than expected. If you encounter difficulties uploading the exam, e-mail a copy to tpickeri@usc.edu.
- **If you submit the exam late, you will be penalized 4 points for each minute (or fraction thereof) past the due time.**

Statement of Academic Integrity

For this exam, I affirm the following:

- ✓ This exam reflects only my own work. I did not receive assistance from any other individual, nor did I provide assistance to any other student taking this exam.
- ✓ While I may use my own notes, I did not refer to any online source during the exam.
- ✓ I understand that acts of academic dishonesty may be penalized in accordance with Section 13 of the University of Southern California Community Standards, including possible "F" in the course, notation on transcript, and/or dismissal from academic programs (https://sjacs.usc.edu/students/academic-integrity/).

I affirm by typing my name below.

____

Name Date

A [25 points]

Diop et al. (2022) examined willingness of Qatari citizens to volunteer at the FIFA World Cup. They interviewed 6,071 Qatari citizens and ascertained variables of interest, including whether the participant was willing to volunteer at this year's World Cup.

Y = Willing to volunteer (1=Yes, 0=No).

$X_{INTEREST} = \begin{cases} 1, \text{interested in soccer} \\ 0, \text{not interested in soccer} \end{cases}$ $X_{AGE40} = \begin{cases} 1, \text{Age} > 40 \text{ years} \\ 0, \text{Age} \leq 40 \text{ years} \end{cases}$

They fit the following model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_{INTEREST} X_{INTEREST} + \hat{\beta}_{AGE40} X_{AGE40}$$

With corresponding parameter estimates:

$\hat{\beta}_0$	$\hat{\beta}_{INTEREST}$	$\hat{\beta}_{AGE40}$
0.02	0.09	-0.13

A1. In notation, what is the null and alternative hypothesis that would test if interest in soccer is associated with being willing to volunteer, controlling for age?

$H_0: \beta_{INTEREST} = 0$

$H_A: \beta_{INTEREST} \neq 0$

A2. According to this model, what is the odds ratio for the effect of interest in soccer on being willing to volunteer, adjusting for age?

$OR = e^{0.09-0} = 1.09$

A3. According to this model, what is the odds ratio of being willing to volunteer for a 45-year-old who is interested in soccer compared to a 50-year-old who is not interested in soccer?

$OR = e^{((0.09-0.13)-(0-(-0.13)))} = e^{0.09} = 1.09$

A4. According to this model, what is the odds ratio of being willing to volunteer for a 35-year-old compared to a 25-year-old, adjusting for interest in soccer?

$OR = e^{0-0} = 1$

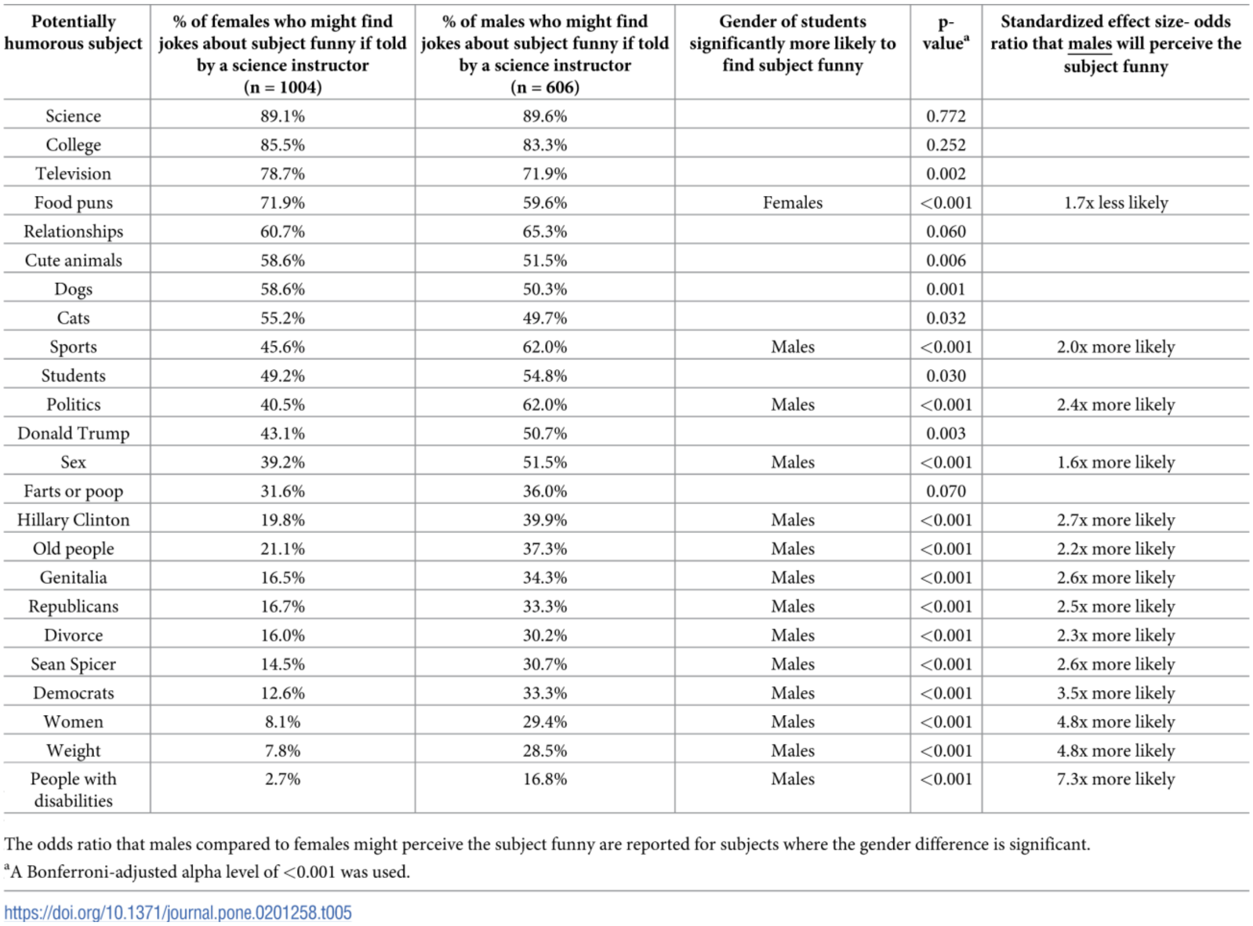
A5. What is the predicted probability of being willing to volunteer for a person who is ≤40 years old and not interested in soccer?

$$\hat{\pi} = \frac{e^{0.02+0+0}}{1 + e^{0.02+0+0}} = 0.505$$

B [20 points]

Cooper et al. (2018) studied the use of humor to improve learning in college science courses. They compiled a list of potentially humorous subjects and subsequently surveyed 1,610 students at Arizona State University about which topics they would potentially find funny in the context of science instruction. The proportions of students who said they would find each topic funny, stratified by gender, are listed below. P-values and odds ratios for this relationship are presented as well.

(Note: their phrasing of “more likely” is questionable—I would have used “as likely.”)



Potentially humorous subject	% of females who might find jokes about subject funny if told by a science instructor (n = 1004)	% of males who might find jokes about subject funny if told by a science instructor (n = 606)	Gender of students significantly more likely to find subject funny	p-value[a]	Standardized effect size- odds ratio that males will perceive the subject funny
Science	89.1%	89.6%		0.772	
College	85.5%	83.3%		0.252	
Television	78.7%	71.9%		0.002	
Food puns	71.9%	59.6%	Females	<0.001	1.7x less likely
Relationships	60.7%	65.3%		0.060	
Cute animals	58.6%	51.5%		0.006	
Dogs	58.6%	50.3%		0.001	
Cats	55.2%	49.7%		0.032	
Sports	45.6%	62.0%	Males	<0.001	2.0x more likely
Students	49.2%	54.8%		0.030	
Politics	40.5%	62.0%	Males	<0.001	2.4x more likely
Donald Trump	43.1%	50.7%		0.003	
Sex	39.2%	51.5%	Males	<0.001	1.6x more likely
Farts or poop	31.6%	36.0%		0.070	
Hillary Clinton	19.8%	39.9%	Males	<0.001	2.7x more likely
Old people	21.1%	37.3%	Males	<0.001	2.2x more likely
Genitalia	16.5%	34.3%	Males	<0.001	2.6x more likely
Republicans	16.7%	33.3%	Males	<0.001	2.5x more likely
Divorce	16.0%	30.2%	Males	<0.001	2.3x more likely
Sean Spicer	14.5%	30.7%	Males	<0.001	2.6x more likely
Democrats	12.6%	33.3%	Males	<0.001	3.5x more likely
Women	8.1%	29.4%	Males	<0.001	4.8x more likely
Weight	7.8%	28.5%	Males	<0.001	4.8x more likely
People with disabilities	2.7%	16.8%	Males	<0.001	7.3x more likely

The odds ratio that males compared to females might perceive the subject funny are reported for subjects where the gender difference is significant.

[a] A Bonferroni-adjusted alpha level of <0.001 was used.

https://doi.org/10.1371/journal.pone.0201258.t005

B1. They reported that males were 2.0 times as likely as females to find sport jokes to be funny. Create the 2-by-2 contingency table that this odds ratio would have been based on.

There are 1004 females and 45.6% of them find sports jokes funny. So, 458 females find sports jokes funny and 546 do not find sports jokes funny.

There are 606 males and 62.0% of them find sports jokes funny. So, 376 males find sports jokes funny and 230 do not find sports jokes funny.

The table is therefore:

	Funny	Not Funny
Male	376	230
Female	458	546

Note we can compute the OR = (376*546)/(458*230) = 1.95 (approximately 2.0)

B2. Using the 2x2 table you created in [B1], compute the 95% confidence interval on the odds ratio.

$$SE(\ln(OR)) = \sqrt{\frac{1}{376} + \frac{1}{230} + \frac{1}{458} + \frac{1}{546}} = 0.105$$

$$\ln(OR) = \ln(2) = 0.693$$

So the 95% CI on the ln(OR) = 0.693 +/- 1.96*0.105 = (0.487, 0.899)

And the 95% CI on the OR = (exp(0.487), exp(0.899) = (1.61, 2.46)

B3. Suppose you fit a logistic regression equation for the effect of gender on finding jokes about sports funny. What would be the fit value of the slope coefficient for gender?

The beta coefficient for a variable reflects the ln(OR) for a 1-unit increase in X. From B2 we computed the ln(OR) = 0.693, so we would expect the slope coefficient to be 0.693.

B4. Based only on your response in [B2], is the effect of gender on finding sports jokes funny statistically significant? Justify your response.

Yes. The 95% CI on the OR does not contain 1.0, so we would expect this effect to be statistically significant at alpha=0.05.

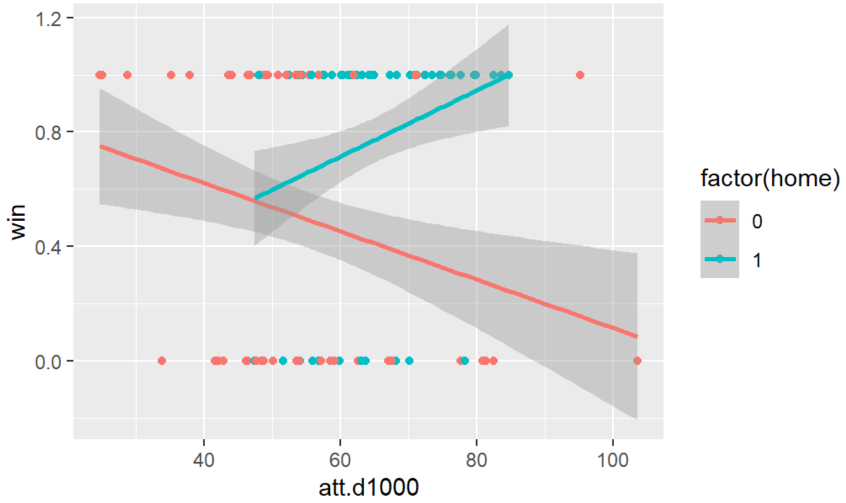
C [25 points]

More football! I downloaded data on USC football games from 2016 until present (usctrojans.com/sports/football). I wanted to know whether attendance at the football game was related to probability of winning the game, and whether this effect was the same for home vs. away games. Here are the variables:

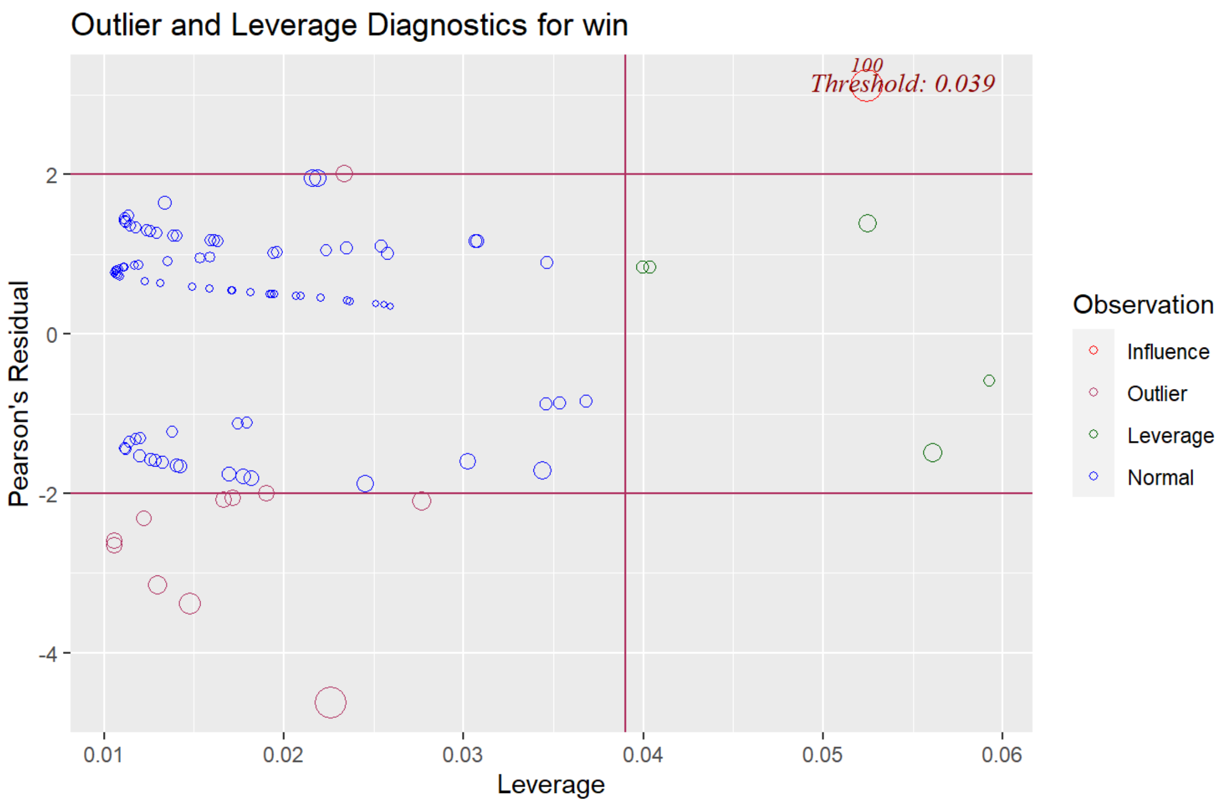
win: 1=USC won the game, 0=USC lost the game

home: 1=game played at USC, 0=game played away

att.d1000: attendance at the game, divided by 1000 (e.g., a "60" represents 60,000 people)



```
Call:
glm(formula = win ~ home * att.d1000, family = binomial, data = usc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2067  -1.1067   0.5274   0.9121   1.8573

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.01623    1.10777   1.820   0.0687 .
home           -5.50544    2.74597  -2.005   0.0450 *
att.d1000      -0.03726    0.01954  -1.907   0.0565 .
home:att.d1000  0.11174    0.04482   2.493   0.0127 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 133.62  on 101  degrees of freedom
Residual deviance: 116.49  on  98  degrees of freedom
AIC: 124.49

Number of Fisher Scoring iterations: 4
```



C1. Is the interaction between attendance and game location (home vs. away) statistically significant? Provide the test statistic and p-value.

Yes, we can look at the interaction term in the output. The test statistic is z=2.493, p=0.0127.

C2. Briefly (1-2 sentences) describe how attendance is related to the probability of winning when USC plays a home game.

The equation describing the outcome is:

$$logit(\hat{\pi}) = 2.02 - 5.51X_{HOME} - 0.04X_{ATT1000} + 0.11(X_{HOME})(X_{ATT1000})$$

When USC plays a home game this equation becomes:

$$logit(\hat{\pi}) = 2.02 - 5.51(1) - 0.04X_{ATT1000} + 0.11(1)(X_{ATT1000}) = -3.49 + 0.07X_{ATT1000}$$

Higher attendance is associated with greater odds of winning. (Namely, each 1000 person increase in attendance is associated with exp(0.07) = 1.07 times the odds of winning.)

C3. Briefly (1-2 sentences) describe how attendance is related to the probability of winning when USC plays an away game.

The equation describing the outcome is:

$$logit(\hat{\pi}) = 2.02 - 5.51X_{HOME} - 0.04X_{ATT1000} + 0.11(X_{HOME})(X_{ATT1000})$$

When USC plays an away game this equation becomes:

$$logit(\hat{\pi}) = 2.02 - 5.51(0) - 0.04X_{ATT1000} + 0.11(0)(X_{ATT1000}) = 2.02 - 0.04X_{ATT1000}$$

Higher attendance is associated with lower odds of winning. (Namely, each 1000 person increase in attendance is associated with exp(-0.04) = 0.96 times the odds of winning.)

C4. Compute McFadden's R-squared for this model. (Hint: Lab 8, 3.3.2) Interpret this value.

$$R^2_{McFadden} = 1 - \frac{D_1}{D_0} = 1 - \left(\frac{116.49}{133.62}\right) = 0.128.$$

This is a pseudo R-squared value. Roughly: 12.8% of the variance in the outcome (winning vs. losing) is explained by where USC played (home vs. away) and the attendance.

C5. Briefly describe what the outlier and leverage diagnostics plot says about the fit of the model.

Most observations fit the model well. There is one observation (100) that may be influential and should be checked.

D [30 points]

Dr. Buser was studying the recovery time of patients who underwent spine surgery. Her team wanted to implement a new protocol – Enhanced Recovery After Surgery (ERAS) – which had been implemented in other fields to improve patient care. Patients were randomized into either the ERAS group or standard care group (n=104 per group).

They also suspected that ERAS may have different efficacy depending on the patient's ethnicity (white vs. non-white).

Dr. Buser moved away to New York and left her team with the output her statistician had given her. They're calling on you to interpret this output to form a cohesive report on what was performed. The main research question is whether participants in the ERAS group had shorter recovery time in the hospital, and whether this effect differed depending on patient ethnicity.

Based only on the output in the appendix, write brief report detailing the methods, results, and conclusions from the available analyses. Your report must be in paragraph format (i.e., no bullet points). Any text that appears after 350 words will be deleted and not graded.

You should comment on:

- The type of analysis performed
- The steps involved in building and selecting the best model
- Which final model(s) you chose to address the research question
- An interpretation of the parameters of interest in final model, including relevant coefficients and p-values
- Information about how well the final model fits (if provided)
- Any missing or next steps that may be appropriate

Please state your word count here: ________________

[+3] We examined the functional form of age; it was linearly related to the outcome
[+9] We examined 3 models: linear, Poisson, and Negative Binomial
[+3] The Poisson model was overdispersed
[+3] Use the NB model since LOS is a skewed outcome
[+3] There is an interaction between ERAS and race (p=0.05)
[+3] ERAS reduces length of stay for white patients, but not for non-white patients
[+6] For nonwhite patients, ERAS is associated with exp(-0.20) = 0.81 times the LOS (p=0.27). For white patients, ERAS is associated with exp(-0.64) = 0.53 times the LOS (p<0.01).
[+3] Need to examine GOF statistics and residuals

Appendix

```
los: length of stay in hospital recovery (in days)
age: age of participant (in years)
eras: 1=ERAS, 0=Control (Standard Care)
white.f: factor variable indicating white vs. nonwhite ethnicity

> dat14786 %>% count(eras.f)
# A tibble: 2 × 2
  eras.f       n
  <fct>    <int>
1 Non-ERAS   104
2 ERAS       104

> dat14786 %>% count(white.f)
# A tibble: 2 × 2
  white.f               n
  <fct>             <int>
1 White               123
2 Non-White    85

> dat14786 %$% summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.00   56.00   66.50   63.68   72.25   93.00 

> dat14786 %$% summary(los)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   3.000   4.212   6.000  32.000 

> mfp(los ~ fp(age)+eras, family=poisson, data=dat14786)
Call:
mfp(formula = los ~ fp(age) + eras, data = dat14786, family = poisson)


Deviance table:
                 Resid. Dev
Null model       641.0544
Linear model     592.1554
Final model      592.1554

Fractional polynomials:
           df.initial select alpha df.final power1 power2
eras                1      1  0.05        1      1      .
age                 4      1  0.05        1      1      .


Transformations of covariates:
              formula
age    I((age/100)^1)
eras             eras

Rescaled coefficients:
   Intercept        eras.f         age.1
    1.357321     -0.461153      0.004439

Degrees of Freedom: 207 Total (i.e. Null);  205 Residual
Null Deviance:      641.1
Residual Deviance: 592.2         AIC: 1218


> summary(model.1)

Call:
glm(formula = los ~ eras * white.f + age, data = dat14786)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
```

```
-4.765  -2.346  -1.381   1.433  26.675  

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          4.25555    1.47410   2.887 0.004312 ** 
eras                                -2.65809    0.72620  -3.660 0.000321 ***
white.fNon-White                    -0.98422    0.79621  -1.236 0.217837    
age                                  0.02096    0.02159   0.971 0.332818    
eras.fERAS:white.fNon-White          1.83254    1.13999   1.608 0.109498    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 16.17932)

    Null deviance: 3530.7  on 207  degrees of freedom
Residual deviance: 3284.4  on 203  degrees of freedom
AIC: 1176.2

Number of Fisher Scoring iterations: 2

> summary(model.2)

Call:
glm(formula = los ~ eras * white.f + age, family = poisson, 
    data = dat14786)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4731  -1.3719  -0.6743   0.6670   7.8984  

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          1.395144   0.181393   7.691 1.46e-14 ***
eras                                -0.642987   0.091248  -7.047 1.83e-12 ***
white.fNon-White                    -0.194243   0.088556  -2.193  0.02828 *  
age                                  0.005084   0.002676   1.900  0.05743 .  
eras:white.fNon-White                0.444616   0.141318   3.146  0.00165 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 641.05  on 207  degrees of freedom
Residual deviance: 582.18  on 203  degrees of freedom
AIC: 1211.6

Number of Fisher Scoring iterations: 5

> summary(model.3)

Call:
glm.nb(formula = los ~ eras * white.f + age, data = dat14786, 
    init.theta = 2.37639126, link = log)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5704  -0.9765  -0.4340   0.4036   3.4774  

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                          1.421167   0.300810   4.724 2.31e-06 ***
eras                                -0.643151   0.148471  -4.332 1.48e-05 ***
white.fNon-White                    -0.202430   0.156113  -1.297   0.1947    
age                                  0.004723   0.004426   1.067   0.2859    
eras:white.fNon-White                0.447461   0.232067   1.928   0.0538 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.3764) family taken to be 1)
```

```
    Null deviance: 220.87  on 207  degrees of freedom
Residual deviance: 199.54  on 203  degrees of freedom
AIC: 1022.2

Number of Fisher Scoring iterations: 1


              Theta:  2.376
          Std. Err.:  0.351

 2 x log-likelihood:  -1010.175

> AER::dispersiontest(model.2)

        Overdispersion test

data:  model.2
z = 3.3838, p-value = 0.0003574
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  3.476044

> sim_slopes(model.1, pred=eras, modx=white.f)
SIMPLE SLOPES ANALYSIS

Slope of eras when white.f = Non-White:

   Est.   S.E.   t val.      p
------- ------ -------- ------
  -0.83   0.88    -0.94   0.35

Slope of eras when white.f = White:

   Est.   S.E.   t val.      p
------- ------ -------- ------
  -2.66   0.73    -3.66   0.00

> sim_slopes(model.2, pred=eras, modx=white.f)
SIMPLE SLOPES ANALYSIS

Slope of eras when white.f = Non-White:

   Est.   S.E.   z val.      p
------- ------ -------- ------
  -0.20   0.11    -1.84   0.07

Slope of eras when white.f = White:

   Est.   S.E.   z val.      p
------- ------ -------- ------
  -0.64   0.09    -7.05   0.00

> sim_slopes(model.3, pred=eras, modx=white.f)
SIMPLE SLOPES ANALYSIS

Slope of eras when white.f = Non-White:

   Est.   S.E.   z val.      p
------- ------ -------- ------
  -0.20   0.18    -1.10   0.27

Slope of eras when white.f = White:

   Est.   S.E.   z val.      p
------- ------ -------- ------
  -0.64   0.15    -4.33   0.00
```