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 $\alpha = .05$ 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).

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

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: $\hat{\beta}_{INTEREST} = 0$

H_A: $\hat{\beta}_{INTEREST} \neq 0$

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?

Odds ratio: $e^{\hat{\beta}_{INTEREST}} = e^{0.09} = 1.094$

Having an interest in soccer increases the odds of willingness to volunteer by 1.094 or 9.4%, adjusting for age.

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?

The equation of the best fit line: $\hat{Y} = 0.02 + 0.09X_{INTEREST} - 0.13X_{AGE40}$

~~$\hat{Y} = [0.02 + 0.09(1) - 0.13(1)] - [0.02 + 0.09(0) - 0.13(0)] = -0.02 - 0.02 = -0.04$~~

$OR = e^{[0.02+0.09(1)-0.13(0)]-[0.02+0.09(0)-0.13(0)]} = e^{0.09} = 1.09$

The odds ratio is given by: $e^{0.09} = 1.09$. A 45-year old who is interested in soccer is associated with 1.09 times the odds of willingness to volunteer, or a 9% increase, compared to a 50-year old who isn't interested in soccer.

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?

$$\hat{Y} = [0.02 - 0.13(0)] - [0.02 - 0.13(0)] = 0$$

The odds ratio is given by $e^0 = 1$. The odds ratio of being willing to volunteer for a 35-year old compared to a 25-year old adjusting for interest in soccer is 1, or in other words they have the same odds of volunteering.

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

$$\hat{Y} = 0.02 + 0.09(0) - 0.13(0) = 0.02$$

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

The predicted probability of being willing to volunteer for a person who is less than or equal to 40 years old and not interested in soccer is given by $\frac{e^{0.02}}{1+e^{0.02}} = 0.55$, or a 55% probability of willingness to volunteer.

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.

	Funny	Not Funny	Total
Male	376	230	606
Female	458	546	1004
Total	834	776	1610

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

The odds ratio = $\frac{(376)(546)}{(230)(458)} = 1.95$

The 95% CI of the log odds ratio is given by $\ln(OR) \pm 1.96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$

$$\ln(1.95) \pm 1.96\sqrt{\frac{1}{376}+\frac{1}{230}+\frac{1}{458}+\frac{1}{546}} = (0.462, 0.873)$$

So, to get the 95% CI for the odds ratio, exponentiate the lower and upper bounds: $(e^{0.462}, e^{0.873}) = (1.59,\ 2.40)$

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 fit value for the coefficient for the effect of gender on finding jokes about sports funny would be equal to the log odds ratio: $\ln(1.95) = 0.668$

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

Yes, because the 95% confidence interval for the odds ratio does not include 1.0. Therefore, at $\alpha = 0.05$, we would expect to find the effect to be statistically significant. An odds ratio of 1.0 would imply that there is no effect of gender on finding sports jokes funny.

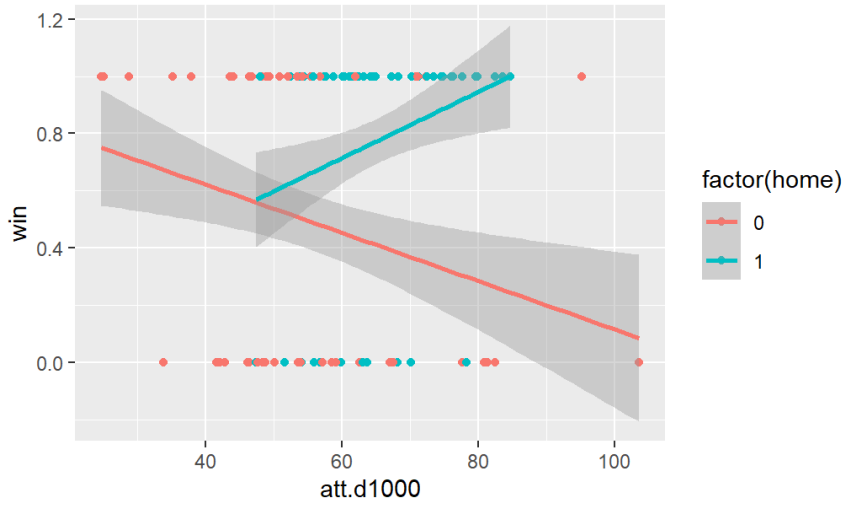
C [25 points]

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:

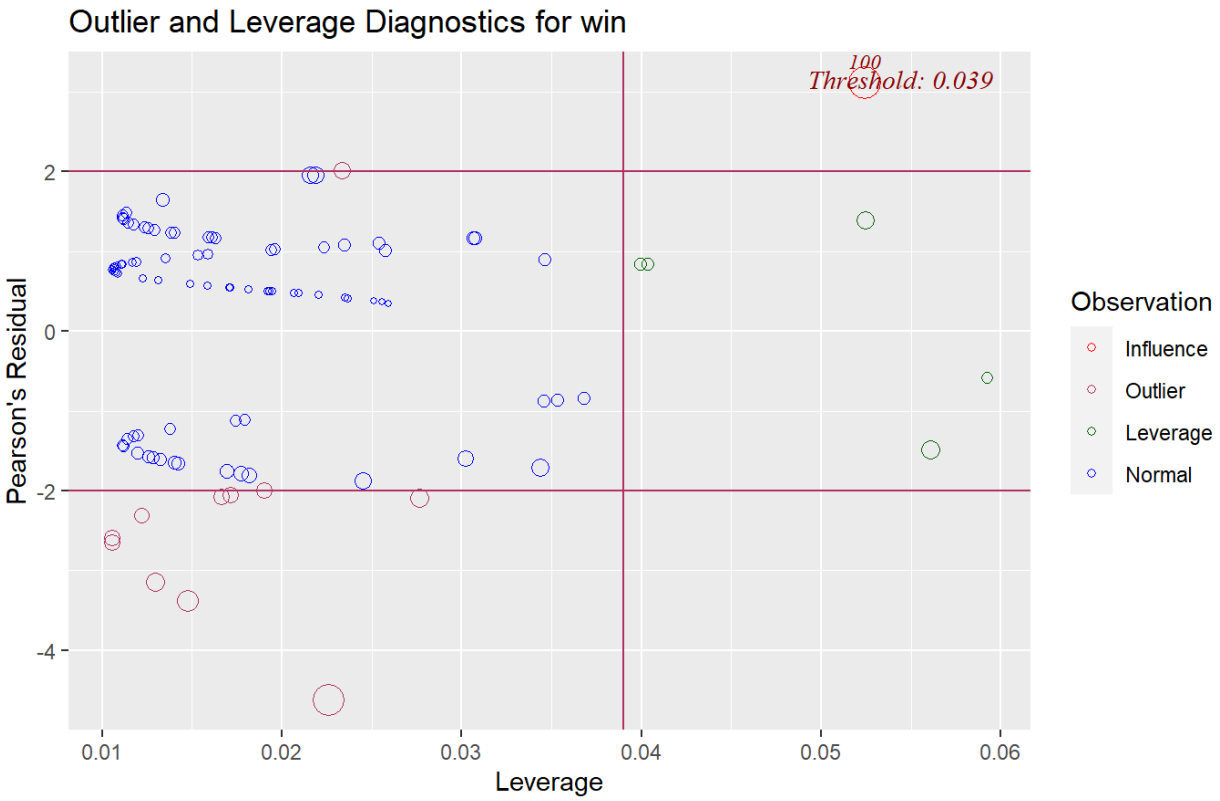
win: 1=USC won the game, 0=USC lost the game

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.

Since there is only 1 interaction term (between home and attendance), the p-value for the interaction term can be used to determine if it is significant. According to the model output, the interaction between attendance and game location is statistically significant (z=2.493, p=0.0127).

If there was more than one interaction term, would have to use the likelihood ratio test to determine its significance.

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

The estimated model parameters are: $\hat{Y} = 2.016 - 5.505X_{home} - 0.037X_{attendance} + 0.112X_{home}X_{attendance}$

Plugging in 1 for X_{home}, the equation becomes: $\hat{Y} = 2.016 - 5.505 - 0.037X_{attendance} + 0.112X_{attendance} = -3.489 + 0.075X_{attendance}$

The association between attendance and probability of winning when USC plays a home game is $e^{0.075} = 1.078$, or in other words a 1,000 person increase in attendance at a home game is associated with a 7.8% increase in the probability of USC winning the game.

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

Plugging in 0 for X_{home}, the equation becomes: $\hat{Y} = 2.016 - 0.037X_{attendance}$
The association between attendance and probability of winning when USC plays an away game is $e^{-0.037} = 0.964$, or in other words a 1,000 person increase in attendance at an away game is associated with a 3.6% decrease in the probability of USC winning the game.

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

$$R^2_{McFadden} = 1 - \frac{D_1}{D_0}$$
$$= 1 - \frac{116.49}{133.62} = 0.128$$

This pseudo R-Squared value compares the deviance of the model under consideration to the null model with no coefficients. The null model is considered the worst possible model, and will therefore have the highest deviance. The addition of coefficients should decrease the deviance, and the pseudo R-Squared is the ratio of the model's deviance to the null model's deviance. In theory a good model will have low deviance, and would decrease the ratio in the equation, and would thus cause R-Squared to approach 1. Roughly 12% of the variance in the outcome (winning vs. losing) is explained by whether USC played at home or away.

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

There are a few outliers that can be seen in the outlier and leverage diagnostics plot, especially negative outliers. Six points appear to have high leverage. However, only observation 100 has both high leverage and is an outlier and should possibly be assessed as to why this is. Overall though, the model appears to fit quite well.

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: _265_

The analysis looked at the effect of the ERAS protocol on recovery time of patients who underwent spine surgery, whether the effect varied by ethnicity, and controlled for age. First, the distributions of each of the covariates were assessed. Then, the functional form for age was determined using the fractional polynomials method. From the output, it appears that age was best encoded as a linear variable. Then, three models were fit: a linear regression, a Poisson regression, and a negative binomial regression. Since the distribution for length of stay was revealed to be right skewed, a linear regression would not be the appropriate modelling approach, as the assumptions would be violated. A dispersion test was also performed, which revealed that the dispersion parameter was 3.48 (p<0.001). Given this information, the negative binomial model is the best modelling approach and I will consider it to be the final model. The final model output reveals that the interaction between age and ERAS is statistically significant at a relaxed alpha level of 0.15 (p=0.054). According to the simple slopes output from the negative binomial model, the ERAS protocol decreases the expected length of stay by $e^{-0.20} = 0.819$ *days compared to those in the standard care group for non-White patients, but the association is not statistically significant (p=0.27). For white patients, the expected length of stay is decreased by* $e^{-0.64} = 0.52$ *days compared to those in the standard care group, and this association is statistically significant (p<0.001). The next steps that need to be done are to assess the GOF statistics and residuals to see if there are any influential points.*

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
```