PM592: Regression Analysis for Data Science
Exam 1 – Fall 2023

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 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.

Flemming Wu _____ October 16, 2023

Name Date

A [25 points]

Dr. Heckmann studied the effect of a new medication on weight loss in patients who had undergone osteoarthritis surgery. Weight was measured at baseline (time=0) and weight loss (as a percent of baseline weight) was assessed for each participant at intervals across 5 years.

Y = Amount of weight change, as a percent of baseline weight.

$X_{TIME} = \text{time since study start, in months}$

They weren't sure about the form of the relationship between weight change and time, so they fit the following four models:

(I) $\hat{Y} = \hat{\beta}_{TIME}\sqrt{X_{TIME}}$

(II) $\hat{Y} = \hat{\beta}_{TIME}\ln(X_{TIME})$

(III) $\hat{Y} = \hat{\beta}_{TIME}e^{X_{TIME}}$

(IV) $\hat{Y} = \hat{\beta}_{TIME}X_{TIME}$

With corresponding beta parameter estimates, p-values for beta, and model r-squared:

Model	$\hat{\beta}_{TIME}$	P	R^2
I	-1.94	<0.001	0.27
II	-2.60	<0.001	0.22
III	-5.19	<0.001	0.11
IV	-0.69	<0.001	0.19

A1. It is not typical to run a model without an intercept. In this situation, why might the authors have not included an intercept term in any of the models that were examined?

The intercept term represents the expected Y value for when X is equal to 0, and removing the intercept from a regression model means that the authors expected the starting value of the Y variable to be 0. In this case, Y represents the amount of weight change since the study began. Removing the intercept from the model makes sense given the study design, as the amount of weight change at the time the study begins is equal to 0, indicating no change.

A2. The authors decided to use Model I as their final model. What is the best rationale for using Model I to describe the relationship between time and weight change?

Considering that the p-values for the coefficient estimates in all four models are statistically significant (p<0.001), Model I makes the best case for the final model because its R^2 value is the highest of all models ($R^2 = 0.27$). This means that modeling wight change as a function of the square root of time in months is able to explain the highest amount of variance in weight change, 27%.

A3. What is the estimated weight change score ($\hat{Y}$) for someone at the beginning of the study (i.e., when $X_{TIME} = 0$)?

The estimated weight change score for someone at the beginning of the study is $\hat{Y} = -1.94\sqrt{0} = \mathbf{0}$

A4. What is the estimated weight change score ($\hat{Y}$) for someone after 9 months of follow-up?

The estimated weight change score for someone after 9 months of follow-up is $\hat{Y} = -1.94\sqrt{9} = -1.94(3) = \mathbf{-5.82}$

A5. What is the interpretation of the parameter estimate $\hat{\beta}_{TIME} = -1.94$?

The parameter estimate $\hat{\beta}_{TIME} = -1.94$ indicates that a one-unit increase in the square root of time in months since the study began, the expected weight score decreases by 1.94 units.

B [20 points]

Lee et. al (2022) studied the determinants of internalized sexual stigma (ISS) in a sample of 1,000 lesbian, gay, and bisexual Taiwanese young adults. There are three components to ISS (social discomfort, sexuality, and identity), but we will focus just on the *identity stigma* component (reflecting the propensity to have a negative self-attitude as a member of a sexual minority).

Y = The ISS identity score on a scale ranging from 1 - 5, with higher values reflecting more negative self-attitude toward one's identity.

$$X_{GEN} = \begin{cases} 1, \text{male} \\ 0, \text{female} \end{cases} \qquad X_{SO} = \begin{cases} 1, \text{homosexual identity} \\ 0, \text{bisexual identity} \end{cases}$$

They fit the following model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_{GEN}X_{GEN} + \hat{\beta}_{SO}X_{SO} + \hat{\beta}_{INX}X_{GEN}X_{SO}$$

And found the model fit well with the corresponding parameter estimates:

$\hat{\beta}_0$	$\hat{\beta}_{GEN}$	$\hat{\beta}_{SO}$	$\hat{\beta}_{INX}$
1.23	3.56	-0.13	-2.46

(equation: $\hat{Y} = 1.23 + 3.56X_{GEN} - 0.13X_{SO} - 2.46X_{GEN}X_{SO}$)

B1. What is the predicted ISS identity score for a person who identifies as bisexual female?

The predicted ISS identity score for a person who identifies as bisexual female is 1.23

$$\hat{Y} = 1.23 + 3.56(0) - 0.13(0) - 2.46(0)(0) = \mathbf{1.23}$$

B2. What is the predicted ISS identity score for a person who identifies as bisexual male?

The predicted ISS identity score for a person who identifies as bisexual male is 4.79

$$\hat{Y} = 1.23 + 3.56(1) - 0.13(0) - 2.46(1)(0) = \mathbf{4.79}$$

B3. Among males, what is the estimated difference in ISS identity score for a person who identifies as homosexual vs. someone who identifies as bisexual?

Among males, the estimated difference in ISS identity score for a person who identifies as homosexual vs. someone who identifies as bisexual is:

$$\hat{Y} = 1.23 + 3.56(1) - 0.13(1) - 2.46(1)(1)$$

$-$

$$\hat{Y} = 1.23 + 3.56(1) - 0.13(0) - 2.46(1)(0)$$

$$= -0.13 - 2.46 = \mathbf{-2.59}$$

2.59 units lower for males that identify as homosexual vs. males that identify as bisexual.

B4. Among females, what is the estimated difference in ISS identity score for a person who identifies as homosexual vs. someone who identifies as bisexual?

Among females, the estimated difference in ISS identity score for a person who identifies as homosexual vs. someone who identifies as bisexual is:

$$\hat{Y} = 1.23 + 3.56(0) - 0.13(1) - 2.46(0)(1)$$

$-$

$$\hat{Y} = 1.23 + 3.56(0) - 0.13(0) - 2.46(0)(0)$$

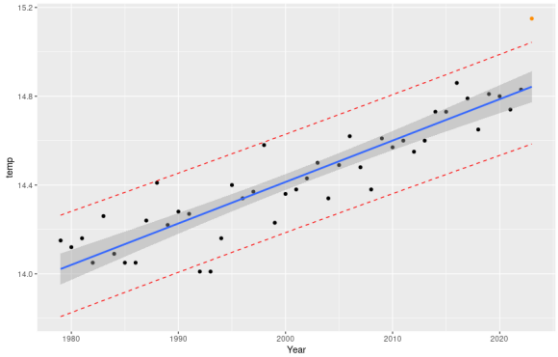
$$= \mathbf{-0.13}$$

0.13 units lower for females that identify as homosexual vs. females that identify as bisexual.

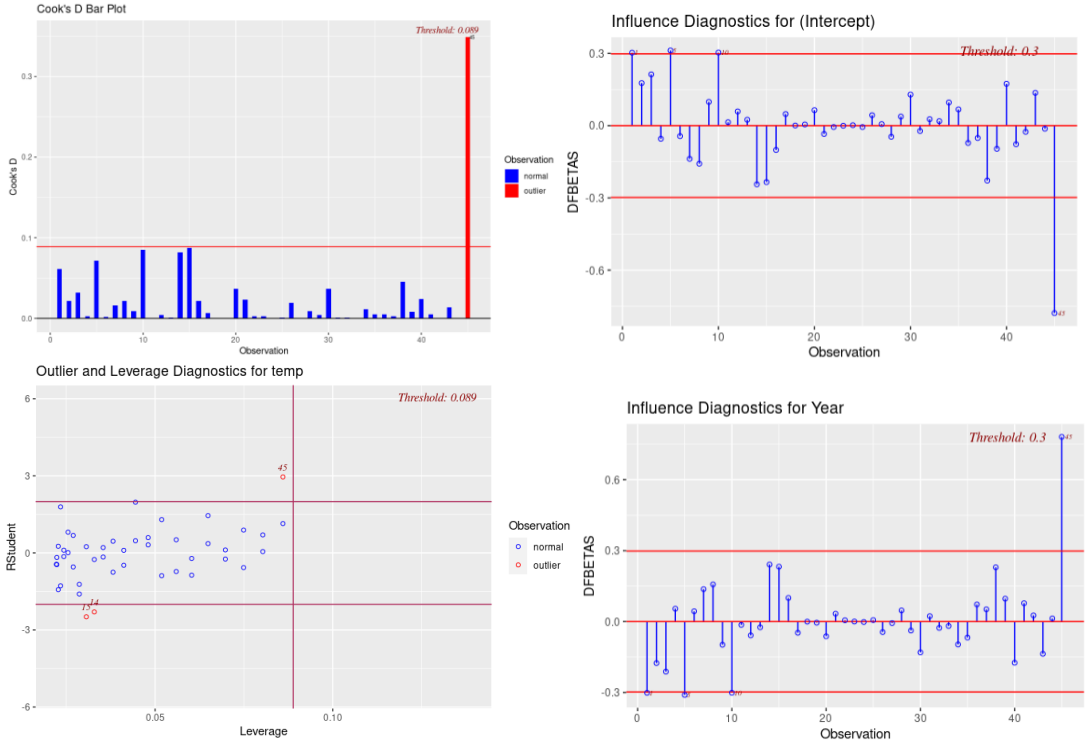
C [25 points]

Zeke Hausfather, from the Berkeley Earth climate data project, was recently in the news for describing a temperature trend as being "gobsmackingly bananas." He fit a linear trend for the average September global surface temperature from 1979 to 2022. Then he added the data point for September 2023 and examined how the fit of the line changed.

The following plot shows the relationship between global average surface temperature in September (degrees C) and year, for 1979 to 2022. The best fit linear regression line is shown in blue, with confidence intervals on the line in grey and 95% prediction intervals in dashed red. The temperature data point for 2023 was added in orange.



Diagnostics for a model that includes all data from 1979 to 2023 is presented here. Of note, Observation 45 refers to the data point from year 2023.



C1. Using the 95% prediction intervals, provide a 1-sentence description of whether the data point for 2023 is poorly fit. In your response, use the meaning of the 95% prediction interval to explain why or why not.

The 95% prediction interval represents the interval in which we are 95% confident that a new value sampled from the same distribution would lie according to the regression model, and since observation 45 lies outside of the 95% prediction interval, I argue that it is poorly fit by the model.

C2. Using the Cook's distance, provide a 1-sentence description of whether the data point for 2023 is poorly fit. In your response, use the meaning of Cook's distance to explain why or why not.

Cook's distance measures how much all the slope and intercept estimates change with the deletion of each observation, and as can be seen in both of the Cook's D bar plots, the intercept and slope both change a lot with the deletion of observation 45 compared to all of the other points, making observation 45 poorly fit.

C3. Using the residuals and leverage plot, provide a 1-sentence description of whether the data point for 2023 is poorly fit. In your response, use the meaning of the residuals and leverage to explain why or why not.

According to the residuals and leverage plot, observation 45 has a high residual, meaning the difference between the expected and actual value is high, and has more leverage than almost all of the other points, meaning that its distance to the mean of all X values is far, both of which point to observation 45 being poorly fit by the model.

C4. Based on the DFBETAS, does the inclusion of the data point from 2023 increase or decrease the estimated value of i) the intercept and ii) the slope?

According to the DFBETAS plots, the inclusion of observation 45 decreases the estimated value of the intercept, since it has a negative DFBETAS value for the intercept, and increases the estimated value of the slope, since it has a positive DFBETAS value for the slope.

D [30 points]

Dr. Oh studied mental health outcomes in college students. He wanted to know whether certain racial/ethnic groups experience more or less impairment due to depression.

Dr. Oh's research assistant kept changing their mind about whether they wanted to work on this specific project. In the interim, they had produced a data analysis and sent the output. Interpret this output to form a cohesive report on what was performed. The main research question is whether impairment due to depression varies by race/ethnicity. Dr. Oh had said to control for level of depression, and also possibly gender and age as confounders.

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: _____318_____

The analysis performed was a multivariate linear regression of impairment due to depression on ethnicity. First, the analyst checked the distributions of all of the variables. Then, an initial unadjusted model was created with only the ethnicity variable. Next, the analyst adjusted for severity of the depression as a confounding variable, which did seem to change the parameter estimates considerably, so depression severity was kept in the model. In the next two models, the analyst adjusted for age and gender simultaneously, first with gender and age as-is, and then with age encoded as a $\frac{1}{x^2} + \frac{1}{x}$ *polynomial term. It is not clear whether the hierarchical or fractional polynomials approach was taken to achieve the best high-order term for the age variable and the p-values for the higher-order terms are missing from the output. The analyst appears to have included gender and polynomial encoding of age as confounders in the final model and assessed the assumptions of linear regression. However, neither of the models that adjusted for age and gender seemed to change the slope estimates for the variables of interest considerably (i.e. over 10%). For this reason, I would argue that model 2 should be the final model. My final model (model 2) explains 45.8% of the variance in impairment (F=1.815e+04, p<2.2e-16). According to the final model, adjusting for depression severity, the interpretations of the parameters are as follows. The expected impairment score for Asian students is 0.11 units lower compared to White students (p<2e-16). The expected impairment score for Black students is 0.053 units lower compared to White students (p=1.35e-15). The expected impairment score for Hispanic students is 0.039 units lower compared to White students (p=1.53e-7). The expected impairment score for multiracial students is 0.026 units higher compared to white students (p=7.82e-5). Next steps should be to assess the assumptions of linear regression for my final model, and check for influential points using Cook's Distance, DFFITS, and DFBETAS.*

Appendix

```
dep_impa: impairment due to depression, on a 0-5 scale with higher values reflecting more impairment
dep_cat: severity of depression, on a 0-5 scale, with higher values reflecting more severity
age: age, in years
male: 1=male, 0=female
raceth2.f = race/ethnicity
```

```
> dat15109 %>% skimr::skim(age, male, dep_cat, dep_impa, raceth2.f)

── Variable type: factor ─────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 raceth2.f             0             1 FALSE          5 Whi: 66778, Asi: 12772, Two: 10157, Bla: 9894

── Variable type: numeric ────────────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean    sd p0 p25 p50 p75 p100 hist
1 age                   0             1 23.5   7.00  18  19  21  25   80 ▇▂▁▁
2 male                  0             1  0.277 0.448  0   0   0   1    1 ▇▁▁▂
3 dep_cat               0             1  1.49  1.24   0   1   1   2    4 ▇▇▃▂▂
4 dep_impa              0             1  2.13  0.837  1   2   2   3    4 ▃▇▁▃▁

> dat15109 %>% count(raceth2.f)
# A tibble: 5 × 2
  raceth2.f       n
  <fct>       <int>
1 White       66778
2 Asian       12772
3 Black        9894
4 Hispanic     7771
5 Two or More 10157

> summary(model1)

Call:
lm(formula = dep_impa ~ raceth2.f, data = dat15109)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2458 -0.2458 -0.1442  0.7542  2.0020

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.144239   0.003230 663.903  < 2e-16 ***
raceth2.fAsian    -0.146275   0.008060 -18.147  < 2e-16 ***
raceth2.fBlack    -0.054791   0.008991  -6.094  1.1e-09 ***
raceth2.fHispanic  0.014942   0.010003   1.494    0.135
raceth2.fTwo or More  0.101601   0.008889  11.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8346 on 107367 degrees of freedom
Multiple R-squared:  0.00525,    Adjusted R-squared:  0.005213
F-statistic: 141.7 on 4 and 107367 DF,  p-value: < 2.2e-16

> summary(model2)

Call:
lm(formula = dep_impa ~ raceth2.f + dep_cat, data = dat15109)

Residuals:
     Min       1Q   Median       3Q      Max
-2.31692 -0.41873  0.07343  0.52824  2.63920

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)           1.471760   0.003275 449.450  < 2e-16 ***
raceth2.fAsian       -0.110961   0.005950 -18.648  < 2e-16 ***
raceth2.fBlack       -0.053032   0.006636  -7.991 1.35e-15 ***
raceth2.fHispanic    -0.038773   0.007386  -5.250 1.53e-07 ***
raceth2.fTwo or More  0.025933   0.006566   3.950 7.82e-05 ***
dep_cat               0.454808   0.001518 299.537  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.616 on 107366 degrees of freedom
Multiple R-squared:  0.4581,     Adjusted R-squared:  0.4581
F-statistic: 1.815e+04 on 5 and 107366 DF,  p-value: < 2.2e-16
```

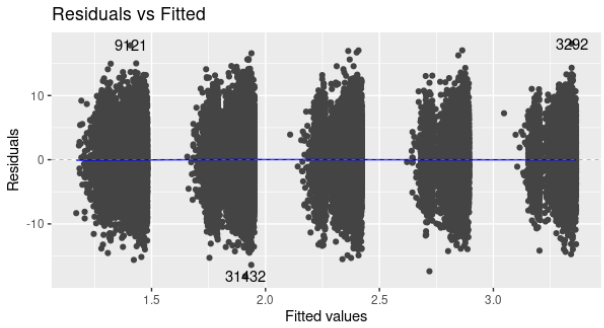
> summary(model3)

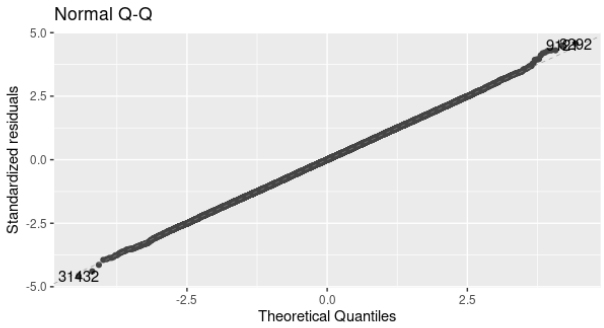
```
Call:
lm(formula = dep_impa ~ raceth2.f + dep_cat + age + male, data = dat15109)

Residuals:
     Min       1Q   Median       3Q      Max
-2.33922 -0.41308  0.04693  0.49781  2.65388

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.5661438  0.0074942 208.982  < 2e-16 ***
raceth2.fAsian       -0.1084927  0.0059452 -18.249  < 2e-16 ***
raceth2.fBlack       -0.0529712  0.0066401  -7.977 1.51e-15 ***
raceth2.fHispanic    -0.0384712  0.0073753  -5.216 1.83e-07 ***
raceth2.fTwo or More  0.0243920  0.0065566   3.720 0.000199 ***
dep_cat               0.4508757  0.0015321 294.287  < 2e-16 ***
age                  -0.0030454  0.0002707 -11.251  < 2e-16 ***
male                 -0.0616645  0.0042153 -14.629  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.615 on 107364 degrees of freedom
Multiple R-squared:  0.4598,     Adjusted R-squared:  0.4598
F-statistic: 1.306e+04 on 7 and 107364 DF,  p-value: < 2.2e-16
```

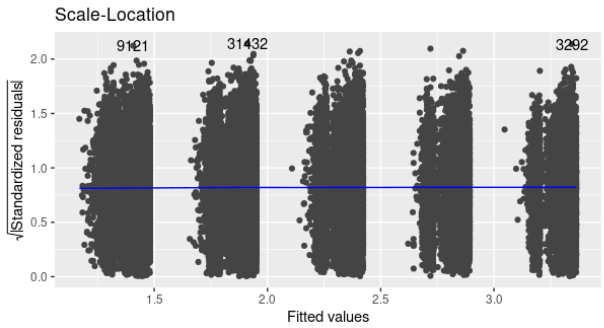
> summary(model4)

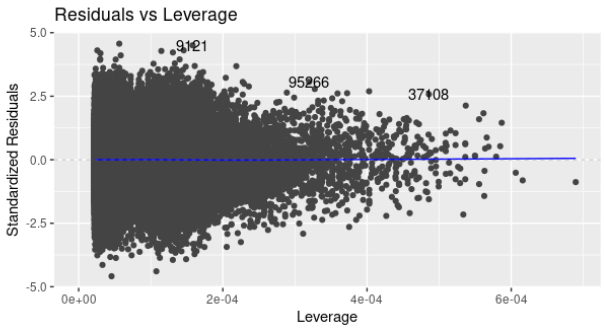
```
Call:
lm(formula = dep_impa ~ raceth2.f + dep_cat + I(age^-2) + I(age^-1) +
    male, data = dat15109)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35220 -0.41730  0.04474  0.48936  2.71384

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           9.328e-01  3.128e-02  29.824  < 2e-16 ***
raceth2.fAsian       -1.131e-01  5.950e-03 -19.014  < 2e-16 ***
raceth2.fBlack       -5.139e-02  6.636e-03  -7.744 9.69e-15 ***
raceth2.fHispanic    -3.986e-02  7.370e-03  -5.409 6.36e-08 ***
raceth2.fTwo or More  2.500e-02  6.550e-03   3.817 0.000135 ***
dep_cat               4.508e-01  1.531e-03 294.425  < 2e-16 ***
I(age^-2)            -3.322e+02  1.970e+01 -16.860  < 2e-16 ***
I(age^-1)             2.804e+01  1.598e+00  17.553  < 2e-16 ***
male                 -6.309e-02  4.213e-03 -14.977  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6145 on 107363 degrees of freedom
Multiple R-squared:  0.4608,     Adjusted R-squared:  0.4608
F-statistic: 1.147e+04 on 8 and 107363 DF,  p-value: < 2.2e-16
```

> autoplot(model4)


Residuals vs Fitted
Residuals
Fitted values
9121
3292
31432
10
0
-10
1.5
2.0
2.5
3.0


Normal Q-Q
Standardized residuals
Theoretical Quantiles
31432
5.0
2.5
0.0
-2.5
-5.0


Scale-Location
√|Standardized residuals|
Fitted values
9121
31432
3292
2.0
1.5
1.0
0.5
0.0


Residuals vs Leverage
Standardized Residuals
Leverage
9121
95266
37408
5.0
2.5
0.0
-2.5
-5.0
0e+00
2e-04
4e-04
6e-04