

PM592: Regression Analysis for Data Science

Exam 1 – 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 submit the exam late, you will be penalized 4 points for each minute (or fraction thereof) past the due time.**

A

[25 points]

C. Covington was studying ways to increase the yield of tomato plants in her garden. In June she planted 40 Gold Medal tomato plants in her backyard. Half of these were located in soil that received Steer's Pride fertilizer and half were located in control soil that did not receive fertilizer. Additionally, for each plot she randomly assigned 10 plants to Condition A (tomatoes picked daily) and 10 plants to Condition B (tomatoes picked weekly). The experiment lasted for 3 months, from June until the end of August.

Y_i = The combined yield (weight in pounds) of all tomatoes picked on plant i between June and August.

$$X_{FERTILIZER} = \begin{cases} 1, \text{treated} \\ 0, \text{control} \end{cases} \quad X_{COND} = \begin{cases} 1, \text{Condition B} \\ 0, \text{Condition A} \end{cases}$$

She fit the following model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_{FERTILIZER}X_{FERTILIZER} + \hat{\beta}_{COND}X_{COND}$$

With corresponding parameter estimates:

$\hat{\beta}_0$	$\hat{\beta}_{FERTILIZER}$	$\hat{\beta}_{COND}$
8.09	6.92	-2.80

A1. In notation, what is the null and alternative hypothesis that would test if Steer's Pride fertilizer is associated with the yield of tomato plants, controlling for the amount of tomato picking (condition)?

$$H_0: \beta_{FERTILIZER} = 0$$

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

A2. What is the effect of picking plants weekly (vs. daily) basis on expected yield, adjusting for manure?

$\hat{\beta}_{COND} = -2.80$; Plants picked weekly have an expected overall yield 2.80 pounds less than plants picked daily.

A3. What is the difference in expected yield for a plant treated with manure under Condition A vs. a control plant under Condition B?

$$\hat{Y}_1 = 8.09 + 6.92(1) - 2.80(0)$$

$$\hat{Y}_2 = 8.09 + 6.92(0) - 2.80(1)$$

$$\hat{Y}_2 - \hat{Y}_1 = 6.92 - (-2.80) = 9.72$$

A4. If X_{COND} was instead coded: 1=Condition A, 0=Condition B, then would you expect $\hat{\beta}_{COND}$ to change? Why or why not?

Yes. In the statement above, a 1-unit change in X_{COND} reflects the comparison of Condition B vs. Condition A. If X_{COND} were coded the other way – Condition A vs. Condition B – then we would expect the sign to be reversed ($-\hat{\beta}_{COND}$).

A5. Based on this output and based on the study design, can we determine whether fertilizer confounds the relationship between condition and yield? Why or why not?

No. Note either of the following:

- For fertilizer to sensibly be a confounder, it must be associated with condition and yield. Condition was randomly assigned within plot, so by design it cannot be associated with fertilizer. Therefore, it cannot sensibly be a confounder.
- To assess confounding numerically, we would have to compare the fertilizer-adjusted $\hat{\beta}_{COND}$ to the unadjusted value. The unadjusted value is not provided, so we cannot assess confounding numerically.

Pan et al. (2014) examined the effect of Tai Chi activity on blood pressure. Participants had been diagnosed with Phase I or II hypertension and chose which group they wanted to participate in: 1) the Tai Chi exercise group, in which they were instructed to participate in 60 minutes of Tai Chi per day for 6 days per week, or 2) the control group which did not perform Tai Chi. At baseline, the average SBP for participants in the Tai Chi group (158 mmHg) was similar to that of the control group (157 mmHg). The study lasted for 12 weeks.

The following table reflects the effect of several demographic and lab values, as well as the effect of the Tai Chi program ("Groups": 1=Tai Chi, 0=Control), on systolic blood pressure measured at the end of the study (SBP; mmHg).

Note that the "standardized coefficients" reflect the regression relationships using z-scores for the independent variables and the outcome instead of the raw scores.

Table 3. Multiple linear regression model for the related influencing factors of SBP.

Model	Unstandardized coefficients		Standardized coefficients Beta	<i>t</i>	<i>p</i> Value
	B	Std. Error			
(Constant)	88.213	26.059		3.385	0.001
Age	0.235	0.120	0.058	1.950	0.053
Gender	-1.316	0.946	-0.042	-1.390	0.166
BMI	-0.301	0.273	-0.033	-1.105	0.271
Fasting glucose	1.654	0.965	0.054	1.715	0.088
Triglycerides	4.865	3.172	0.048	1.534	0.127
Total cholesterol	0.043	0.035	0.037	1.249	0.214
HDL-C	-0.727	0.098	-0.453	-7.385	<0.001
LDL-C	0.148	0.057	0.107	2.612	0.010
Trait anxiety	1.441	0.264	0.214	5.453	<0.001
State anxiety	0.066	0.301	0.008	0.220	0.826
Observation time	1.747	0.565	0.094	3.094	0.002
Groups	-4.959	0.855	-0.271	-5.802	<0.001

B1. On average, did Tai Chi improve participants' blood pressure? Explain the effect and provide a p-value to support your conclusion.

Yes. The groups had similar SBP at the start of the study and the beta coefficient for the effect of group is -4.959 ($p < 0.001$). Therefore, controlling for covariates, Tai Chi is on average associated with 4.959 mmHg lower blood pressure.

B2. HDL-C was measured in mg/dL. In one sentence each, describe the effect of HDL-C on SBP using 1) the unstandardized coefficient and 2) the standardized coefficient.

- 1) Adjusting for covariates, a 1 mg/dL increase in HDL-C is associated with a 0.727 mmHg decrease in SBP.
- 2) Adjusting for covariates, a 1 SD increase in HDL-C is associated with a 0.453 SD decrease in SBP.

B3. Which variable had the strongest effect on SBP? Briefly justify your response.

HDL-C because it has the largest magnitude standardized coefficient. Note that the unstandardized coefficient is affected by the unit of measurement while the standardized coefficients are all on SD units, which are more comparable across variables.

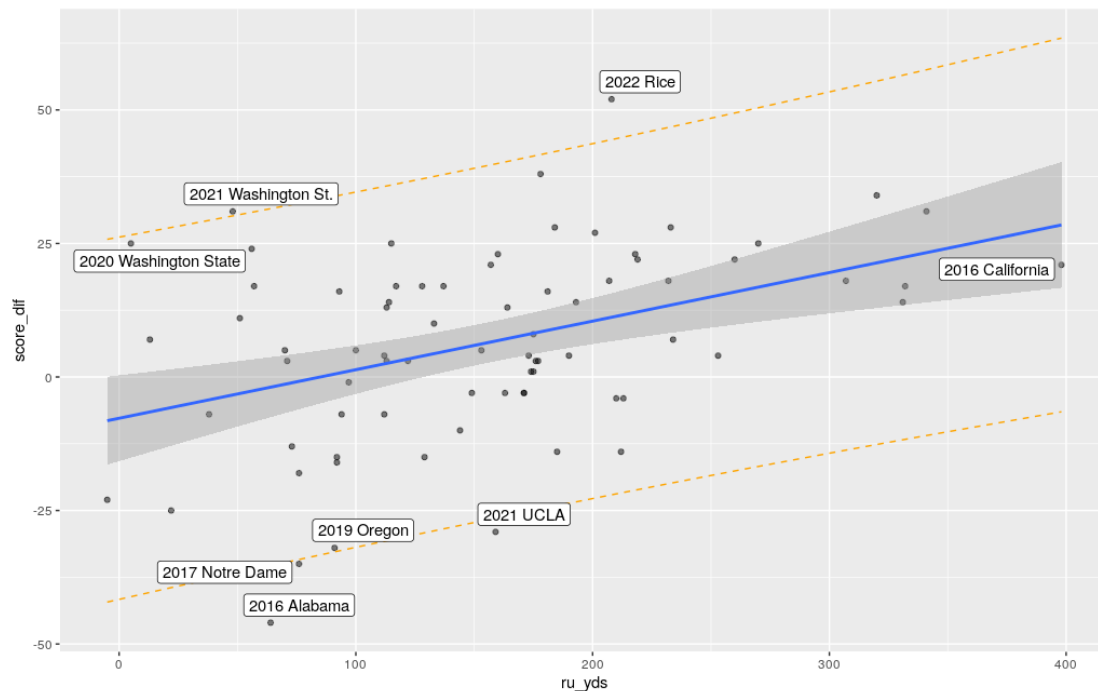
B4. What type of study design did the authors use? Briefly justify your response.

Quasi-experimental. While there were two “treatment conditions,” participants self-selected into their condition. Therefore this is a cohort design.

C**[25 points]**

I downloaded data on USC football games from 2020 until present (usctrojans.com/sports/football). For each game, the *score difference* was computed as USC's score minus their opponent's score. I regressed the score difference on several variables, and found that score difference was significantly related to rushing yards.

The following graph displays the relationship between the score difference and rushing yards, with labels containing the year and opponent.



Call:

```
lm(formula = score_dif ~ ru_yds, data = fb)
```

Residuals:

Min	1Q	Median	3Q	Max
-44.093	-10.277	-0.276	11.588	40.814

Coefficients:

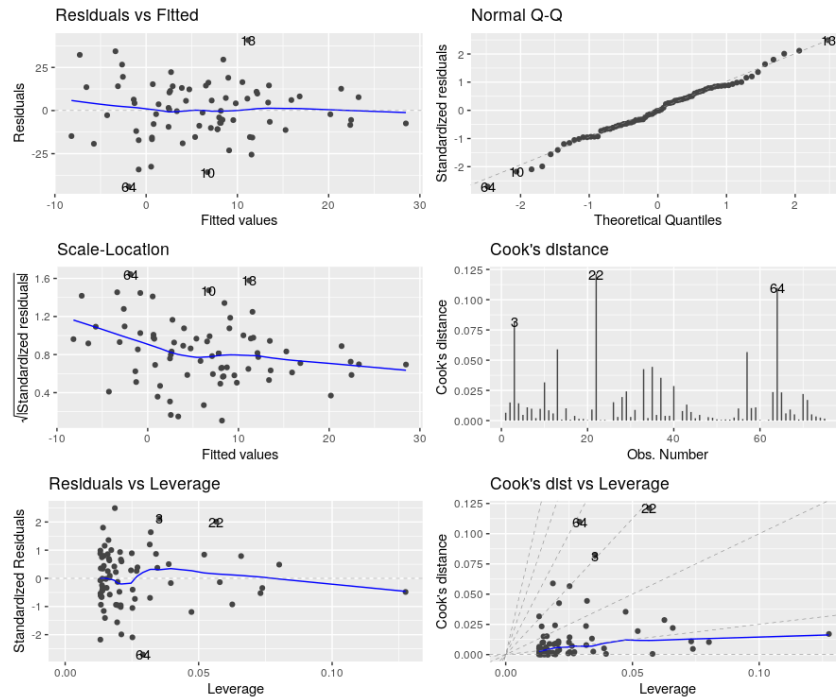
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-7.72675	4.03494	-1.915	0.059363 .
ru_yds	0.09093	0.02300	3.954	0.000174 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 74 degrees of freedom

Multiple R-squared: 0.1744, Adjusted R-squared: 0.1633

F-statistic: 15.63 on 1 and 74 DF, p-value: 0.0001743



```
# A tibble: 10 × 5
  obs year opponent      dffit cook.d
  <int> <dbl> <chr>      <dbl> <dbl>
1    10  2021 UCLA      -0.258  0.0316
2    13  2022 Rice       0.356  0.0589
3    22  2020 Washington State 0.503  0.121
4    46  2018 Oregon State -0.0963 0.00470
5    51  2017 Stanford  -0.0336 0.000572
6    58  2017 Arizona State  0.143  0.0103
7    59  2017 Arizona    -0.147  0.0109
8    64  2016 Alabama    -0.491  0.110
9    70  2016 Arizona     0.209  0.0220
10   71  2016 California  -0.184  0.0171
```

C1. Briefly (1-2 sentences) explain how well this model fits the data and why.

The overall model is statistically significant ($F_{1,74} = 15.63, p < 0.001$) and fits moderately well; 17.44% of the variation in score difference is explained by rushing yards. The assumptions of linear regression appear to hold, suggesting this model is valid.

C2. Observation 64 represents a 2016 game in which USC was badly beat by Alabama (the final score was 52 to 6). Is this (observation 64) an influential point? Why? Are there any other points that are just as influential?

Observation 64 appears to have a large standardized residual, Cook's distance, and dffit; it does appear influential. Observation 22 (2020 Washington State) appears even more influential, having values for all these metrics that are higher than Observation 64.

C3. If we removed observation 64, how would you expect the estimate of the intercept and slope to change? Why?

Observation 64 has a large negative residual and is toward the “left” side of the line. If we removed it, then the line would shift higher on the left. This would in effect increase the value of the intercept and decrease the slope.

C4. Considering the number of rushing yards USC had against California in 2016, was the score difference in that game abnormal? Explain why or why not.

No, the residual for this game is relatively small in comparison to others in the data.

An additional model was run with both passing yards (pa_yds) and rushing yards (ru_yds). The ANOVA table is displayed below.

```
> m4 <- lm(score_dif ~ ru_yds + pa_yds, data=fb)
> aov(m4)
Call:
aov(formula = m4)
```

Terms:

	ru_yds	pa_yds	Residuals
Sum of Squares	4273.979	1830.415	18399.013
Deg. of Freedom	1	1	73

Residual standard error: 15.87581

C5. How much of the variation in outcome is explained by the model with both rushing yards and passing yards?

Rushing yards and passing yards collectively explain $(4273.979 + 1830.415)$ SS. There is a total of $(4273.979 + 1830.415 + 18399.013)$ SS. Therefore these two variables explain $(4273.979 + 1830.415) / (4273.979 + 1830.415 + 18399.013) = 24.9\%$ of the variation in the outcome.

Dr. Kim studied the physical activity patterns of adolescents. She wanted to know whether children who had been diagnosed with congenital adrenal hyperplasia (CAH) had different levels of physical activity through adolescence.

Dr. Kim's data analyst silent-quit on her, and she didn't know what to make of all the output she had been given. Interpret this output to form a cohesive report on what was performed. The main research question is whether levels of physical activity differ between CAH and non-CAH children at different ages. Dr. Kim thought that ethnicity might be a potential confounder.

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: _____

[+2] Linear regression
[+2] Distribution of MVPA was not assessed
[+4] Interaction model for age tertile and CAH; p-value not assessed
[+2] Model also adjusting for ethnicity
[+4] Ethnicity appears to confound the relationship between age and MVPA
[+4] Should perform the Extra SS test to determine the significance of the age x CAH interaction.
[+4] CAH is not associated with MVPA in the highest two tertiles.
[+4] CAH IS associated with MVPA in the youngest tertile ($p=0.02$).
[+4] In the youngest tertile, adjusting for ethnicity, CAH have 13.4 minutes lower MVPA on average compared to controls.
[+4] We did not examine model diagnostics – would want to look at LINE assumptions and influential points.




Appendix

mvpa: minutes moderate-to-vigorous physical activity per day
CAH: 1=CAH, 0=Control
Age: age, in years
Age.q3 = age tertile

```
> dat14849 %>% skim(mvpa, age)
```

```
— Variable type: numeric —
  skim_variable n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist
1 mvpa          0                1 18.7 11.6 2.32 9.79 16.4 26.0 59.6 
2 age          0                1 14.0 2.85 9.07 12.1 13.5 16.3 18.9 
```

```
> dat14849 %>% group_by(age.q3) %>% skim(age)
```

```
— Variable type: numeric —
  skim_variable age.q3 n_missing complete_rate mean sd p0 p25 p50 p75 p100 hist
1 age [8.44,12.3] 0 1 10.8 1.19 9.07 9.91 10.5 11.8 12.3 
2 age (12.3,15.5] 0 1 13.3 0.936 12.3 12.6 13.2 13.9 15.5 
3 age (15.5,18.9] 0 1 17.1 1.08 15.7 16.0 17.3 17.7 18.9 
```

```
> dat14849 %>% count(age.q3)
```

```
# A tibble: 3 × 2
  age.q3 n
  <fct> <int>
1 [8.44,12.3] 11
2 (12.3,15.5] 12
3 (15.5,18.9] 14
```

```
> dat14849 %>% count(cah)
```

```
# A tibble: 2 × 2
  cah.f n
  <int> <int>
1 1 17
2 0 20
```

```
> dat14849 %>% count(ethnicity.f)
```

```
# A tibble: 2 × 2
  ethnicity.f n
  <fct> <int>
1 Non-Hispanic 15
2 Hispanic 22
```

```
> summary(mvpa_age.fit.1)
```

Call:

```
lm(formula = mvpa ~ age.q3 * cah, data = dat14849)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.6250	-6.1510	0.2656	3.7778	24.2222

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.444	3.973	5.649	3.34e-06 ***
age.q3(12.3,15.5]	-8.806	5.893	-1.494	0.145
age.q3(15.5,18.9]	-7.424	5.619	-1.321	0.196
cah	-12.917	5.893	2.192	0.036 *
age.q3(12.3,15.5]:cah	12.843	8.197	-1.567	0.127
age.q3(15.5,18.9]:cah	12.244	7.896	-1.551	0.131

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.732 on 31 degrees of freedom

Multiple R-squared: 0.3964, Adjusted R-squared: 0.299
F-statistic: 4.071 on 5 and 31 DF, p-value: 0.005886

> summary(mvpa_age.fit.2)

Call:

lm(formula = mvpa ~ age.q3 * cah + ethnicity.f, data = dat14849)

Residuals:

Min	1Q	Median	3Q	Max
-17.4321	-4.8554	-0.8284	3.5728	21.3596

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	24.830	3.907	6.356	5.18e-07 ***
age.q3(12.3,15.5]	-6.897	5.632	-1.225	0.2302
age.q3(15.5,18.9]	-3.846	5.552	-0.693	0.4938
cah	-13.394	5.568	2.405	0.0225 *
ethnicity.fHispanic	-7.157	3.275	-2.185	0.0368 *
age.q3(12.3,15.5]:cah	12.502	7.741	-1.615	0.1168
age.q3(15.5,18.9]:cah	14.212	7.509	-1.893	0.0681 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.188 on 30 degrees of freedom

Multiple R-squared: 0.4793, Adjusted R-squared: 0.3751

F-statistic: 4.602 on 6 and 30 DF, p-value: 0.002008

> sim_slopes(mvpa_age.fit.2, cah, modx="age.q3")

SIMPLE SLOPES ANALYSIS

Slope of cah when age.q3 = (15.5,18.9]:

Est.	S.E.	t val.	p
0.82	5.01	-0.16	0.87

Slope of cah when age.q3 = (12.3,15.5]:

Est.	S.E.	t val.	p
-0.89	5.39	0.17	0.87

Slope of cah when age.q3 = [8.44,12.3]:

Est.	S.E.	t val.	p
-13.39	5.57	2.41	0.02