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.**

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]

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.

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

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)?

The null hypothesis states that the fertilizer is not associated with tomato plant yield, and is the equivalent of saying that $\hat{\beta}_{FERTILIZER} = 0$.

The alternative hypothesis states that the fertilizer is associated with tomato plant yield, and is the equivalent of saying that $\hat{\beta}_{FERTILIZER} \neq 0$.

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

~~$8.09 - 2.80(0) - (8.09 - 2.80(1)) = 2.80$~~

~~Adjusting for manure, the expected yield is 2.80 units higher when tomatoes are picked weekly compared to when they are picked daily.~~

$8.09 - 2.80(1) - (8.09 - 2.80(0)) = -2.80$

Adjusting for manure, the expected yield is 2.80 units lower when tomatoes are picked weekly compared to when they are 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?

$$\big(8.09 + 6.92(1) - \ -2.80(0)\big) - \big(8.09 + 6.92(0) - \ -2.80(1)\big) = 6.92$$

The expected yield for a plant treated with manure under Condition A is 6.92 units higher than the expected yield for a control plant under Condition B.

$$\big(8.09 + 6.92(1) - 2.80(0)\big)$$
$$-$$
$$\big(8.09 + 6.92(0) \ - 2.80(1)\big)$$

$$= \ 6.92 - (-2.80)$$
$$= 9.72$$

The expected yield for a plant treated with manure under Condition A is 9.72 units higher than the expected yield for a control plant under Condition B.

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, if X_{COND} was instead coded: 1=Condition A, 0=Condition B, then I would expect the coefficient to be multiplied by negative one, so in this case the coefficient would become 2.80. This is because the relationship between condition and yield would remain the same, so holding fertilizer constant, the expected yield for condition A must still be 2.80 units higher than for condition B. When 0 is inserted into the equation for condition B, the expected yield is 8.09 holding fertilizer constant, and when 1 is inserted into the equation for condition A, the expected yield increases by 2.80(1), or 2.80 units.

(A 1-unit change in X_{cond} reflects the comparison of Condition B vs Condition A. If X_{cond} was coded the other way, the sign of the coefficient would be reversed)

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?

Based on this output and study design, it is not possible to determine if fertilizer confounds the relationship between condition and yield. In order to determine if fertilizer confounds condition, we must determine how much the coefficient estimate for condition changes before and after adjusting for fertilizer. Since the coefficient estimate for condition alone is not known, it cannot be determined if fertilizer is a confounder.

(Additionally, it was mentioned in the study design that condition was randomly assigned within a plot, so by design it cannot be associated with fertilizer, and therefore cannot be a confounder).

B [20 points]

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.

	Unstandardized coefficients		Standardized coefficients		
Model	B	Std. Error	Beta	*t*	*p* Value
(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.

On average, it appears that Tai Chi did improve participants' blood pressure ($p<0.001$). Looking at the unstandardized coefficient, and controlling for covariates, for the "Groups" variable, those who did do Tai Chi are expected to have a 4.959 ~~unit~~ mmHg lower blood pressure than those who did.

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.

The unstandardized coefficient says that a one mg/dL increase in HDL-C is associated with a 0.727 ~~unit~~ mmHg decrease in SBP. The standardized coefficient says that a 1 standard deviation increase in HDL-C is associated with a 0.453 standard deviation decrease in SBP.

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

~~The group (whether one participated in Tai Chi or not) had the strongest effect on SBP. The absolute value of the coefficient for "Groups" was the highest amongst all variables (4.959), which indicates the strength of the relationship with the dependent variable. Additionally, the coefficient was found to be statistically significant (p<0.001).~~

HDL-C has the highest effect on SBP. The absolute value of the standardized coefficient for HDL-C was the highest amongst all variables (0.453), indicating the strongest magnitude of the relationship between dependent and independent variables. **The standardized variables are better at comparing magnitudes of different coefficients, as they are all in standard deviation units.**

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

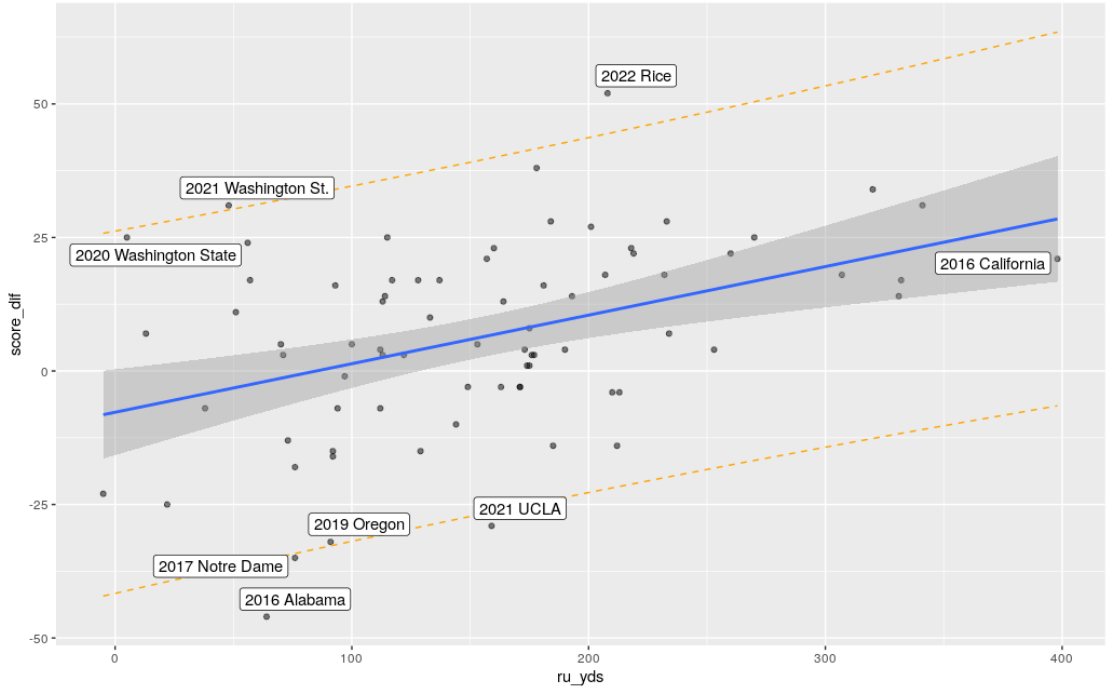
~~Since the participants chose whether they wanted to participate in Tai Chi or not, the type of study design the authors used can be classified as an observational study. There was no randomization in the assignment of groups.~~

The type of study design used was quasi-experimental. The subjects were separated into two groups/conditions but were allowed to choose the group to participate in, making this a cohort study.

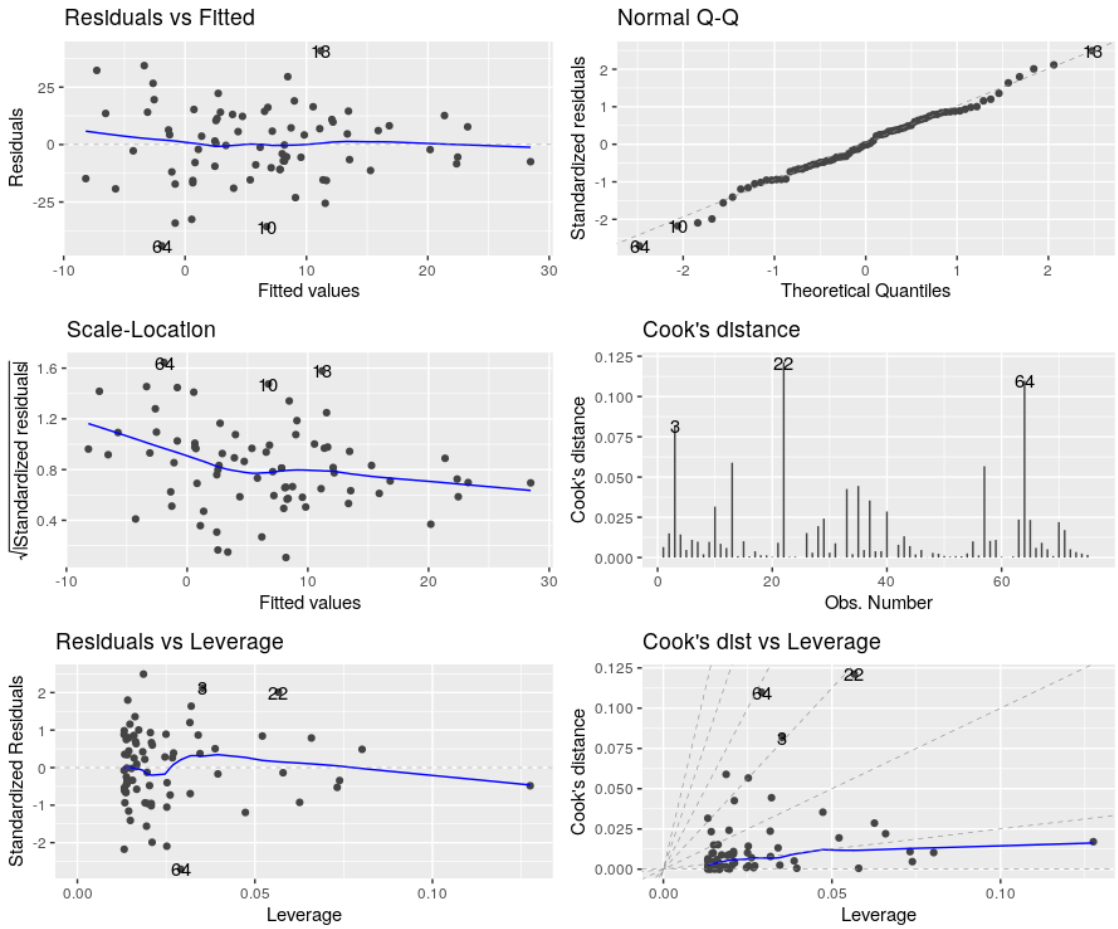
C [25 points]

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 model fit the data relatively well (F=15.63, p=0.00017), the data seem to follow a relatively linear pattern ~~and the assumptions of linearity and normality are met~~. 17.44% of the variance in score difference is explained by rushing yards. ~~However, the model could be improved with a transformation because the Scale-Location plot shows some degree of heteroscedasticity.~~ The assumptions of linear regression appear to be met, suggesting that the model is valid.

(mention R-squared, model diagnostics, and test statistic and p-value)

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?

~~I would not classify observation 64 as an influential point, its cook's distance value is 0.110, which is below the 0.5 threshold that I would use to identify influential points worth investigating. Other observations with similar cook's distance values are 22 and 8.~~

Observation 64 is an influential point because it has a large standardized residual (from the Residuals vs Leverage plot), a large cook's distance, and high DFFIT value. Additionally, observation 22 appears to be even more influential, having higher values for the aforementioned metrics.

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

Given that observation 64 is a negative outlier on the ~~bottom~~"left" end of the regression line, I would expect that its removal would cause the slope to decrease and intercept to increase. This is because the regression line fits the data to minimized the sum of square residuals of all points, so without having to minimize observation 64's residual, the line ~~and intercept can be higher and fit the rest of the data better.~~ would shift higher on its left side.

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.

Considering the number of rushing yards USC had against California in 2016, the score difference in that game was not abnormal. The point for that observation lies very close to the regression line and is within the standard error of the line boundaries, making the point well-approximated by the regression. (The residual is very small compared to others)

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?

The variation in outcome explained by the model with both rushing yards and passing yards is (4273.979 + 1830.415) / (4273.979 + 1830.415 + 18399.013) = 24.9% of the variation.

D [30 points]

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: ____183_____

The analysis performed was a linear regression of congenital (CAH) on physical activity (MVPA). First, the analyst checked the distributions of all of the variables except MVPA, which is missing from the output. The model was then built with an interaction term between age, which was converted into a categorical variable (tertile), and CAH. The significance of the interaction term was not assessed with an Extra Sums of Squares test. Additionally, the final model included ethnicity as a confounder, and it does appear to change the coefficient estimates for age considerably. Since an interaction term was added to the final model, the association between CAH and MVPA was assessed for each age tertile. Based on the output, it seems that CAH is significantly associated with MVPA only for the youngest tertile, and not for the other two. For the youngest tertile, non-CAH children are expected to have 13.4 minutes higher MVPA per day than CAH children. Lastly, the output does not show whether the final model was checked for the assumptions of linear regression or if any influential points were identified and addressed.

Mention the following:

- *Type of regression (linear regression; univariate, multivariate)*
- *Look at distributions of dependent and independent variables*
- *Interaction terms (need p-values; has to be from Extra SS Test if interaction is a set of dummy variables)*
- *Confounding (do the beta values for variables of interest change?)*
- *If interaction terms significant (or included in final model) when is the association between X and Y significant?*

- *Make sure to interpret coefficients correctly (take into account transformations and units of measurement)*
- *Check assumptions of linear regression*
- *Check 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
```

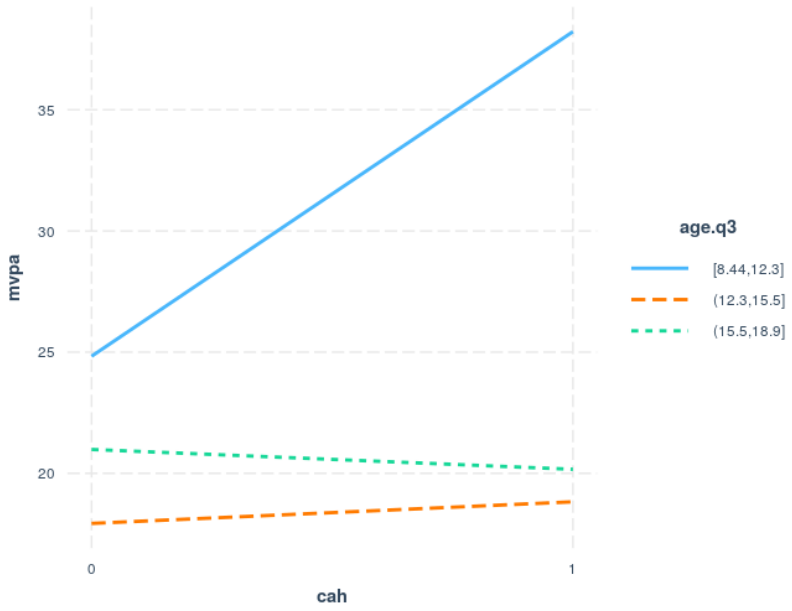
> summary(mvpa_age.fit.2)

```
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
```

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


35
30
25
20
mvpa
0
1
cah
age.q3
[8.44,12.3]
(12.3,15.5]
(15.5,18.9]