**PM592: Regression Analysis for Data Science**

Name:
**Flemming Wu**

**HW5**

*Confounding, Interaction*

**Instructions**

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 3 questions and 30 points possible.

Researchers at Nittany University were interested in factors that influenced the satisfaction of individuals taking group exercise classes. They recruited 90 individuals who decided to enroll in one of three classes: 1) cardio, 2) strength, 3) flexibility. Individuals took group exercise classes every other day for two weeks. The data they collected is located in the "gx.csv" file.

The researchers were interested in the following:

- Do the rate of perceived exertion, instructor encouragement, participant control, and perceived competence relate to intrinsic satisfaction with the workouts?
- Do these effects vary based on the type of class the participant is engaged in?

**Data Dictionary**

| **Variable** | **Meaning** | **Coding** |
|---|---|---|
| classtype | Type of class (randomization condition) | 1 = Cardio<br>2 = Strength<br>3 = Flexibility |
| age | Age of participant (years) collected at baseline | |
| bmi | Body mass index of participant collected at baseline | |
| rpe | Mean of rate of perceived exertion across all workouts | 1-10 scale, higher score represents more exertion |
| encourage | Encouragement scale: "The instructor encouraged me" | 1-7 scale, 1 = strongly disagree, 7 = strongly agree |
| control | Control scale: "The instructor made me do things their way" | 1-7 scale, 1 = strongly agree, 7 = strongly disagree |
| perc_comp | Perceived Competence scale: "I believe I completed the exercises today the way they should be done" | 1-7 scale, 1 = strongly disagree, 7 = strongly agree |
| satisfaction | Satisfaction scale: a measure of how much intrinsic satisfaction participants had from the program | 6-21 scale, 6 = highly dissatisfied, 21 = highly satisfied |

Note: it may be helpful to convert "classtype" to a factor variable before you begin your analysis.

## Question 1 [8 points]

Perform a preliminary set of multivariable linear regressions to address the research questions.

| Variable | Model 1<br>Estimate (SE)<br>Unadjusted | Model 2<br>Estimate (SE)<br>Age-Adjusted | Model 3<br>Estimate (SE)<br>BMI-Adjusted | Model 4<br>Estimate (SE)<br>Age & BMI Adjusted |
|---|---|---|---|---|
| **Intercept** | 4.44 (SE=0.92) (p=6.04e-06***) | 2.48 (SE=1.12) (p=0.03*) | 3.87 (SE=1.49) (p=0.011*) | 2.14 (SE=1.56) (p=0.18) |
| **Class Type (ref: cardio)** | -4.12 (vs. strength) (SE=0.36) (p<2e-16***)<br>---<br>-0.92 (vs. flexibility) (SE=0.43) (p=0.035*) | -4.18 (vs. strength) (SE=0.35) (p<2e-16***)<br>---<br>-1.04 (vs. flexibility) (SE=0.41) (p=0.015*) | -4.12 (vs. strength) (SE=0.36) (p<2e-16***)<br>---<br>-0.92 (vs. flexibility) (SE=0.43) (p=0.037*) | -4.17 (vs. strength) (SE=0.35) (p<2e16***)<br>---<br>-1.03 (vs. flexibility) (SE=0.42) (p=0.016*) |
| **Perceived Exertion** | 0.26 (SE=0.11) (p=0.018*) | 0.28 (SE=0.10) (p=0.0096**) | 0.26 (SE=0.11) (p=0.019*) | 0.28 (SE=0.10) (p=0.01*) |
| **Encouragement** | 0.52 (SE=0.15) (p=0.00058***) | 0.48 (SE=0.14) (p=0.0011**) | 0.51 (SE=0.15) (p=0.0013**) | 0.47 (SE=0.15) (p=0.002**) |
| **Control** | 0.39 (SE=0.14) (p=0.0077**) | 0.37 (SE=0.14) (p=0.0089**) | 0.41 (SE=0.15) (p=0.0073**) | 0.39 (SE=0.15) (p=0.0096**) |
| **Competence** | 0.41 (SE=0.12) (p=0.00087***) | 0.44 (SE=0.12) (p=0.0003***) | 0.40 (SE=0.12) (p=0.0016**) | 0.43 (SE=0.12) (p=0.000545***) |

1a. [4 points] Construct a preliminary table of parameter estimates for 4 models: 1) unadjusted, 2) age-adjusted, 3) bmi-adjusted, and 4) age & bmi adjusted. Use the above table as a template. Note: you will have to figure out how to present the estimates in the table for "class type", clearly conveying information about the reference group.

The entries for class type use the "cardio" factor as the reference group. So, an entry "-4.1 (vs. strength)" indicates that the estimated satisfaction score for an individual who took a cardio class is 4.1 points lower than the estimated satisfaction score for an individual who took a strength class, holding all other variables constant.

1b. [2 point] For each independent variable in the table, state whether age and BMI appear to confound the relationship between that variable and satisfaction score, and why.

At first glance, looking at the coefficient estimates for each of the independent variables in the table, it appears that the estimates changed when age and BMI were added into the model compared to the unadjusted model. However, looking at the model in which only BMI was added, the coefficients did not change much from the unadjusted model. This leads me to think that age is the confounding variable for the original independent variables. In the code snippet below, I calculate the percent change in coefficient estimates before and after adjusting for age in my model:

```
> beta_pct_change <- function(adj, unadj) {
+   return((abs((unadj - adj)/unadj))*100)
+ }
> beta_pct_change(4.1245, 4.17708) # strength
[1] 1.258774
> beta_pct_change(.9213, 1.03705) # flexibility
[1] 11.16147
> beta_pct_change(.2610, .27592) # perceived exertion
[1] 5.407364
> beta_pct_change(.5246, .47857) # encouragement
[1] 9.618238
> beta_pct_change(.3947, .43744) # control
[1] 9.770483
> beta_pct_change(.4158, .42932) # competence
[1] 3.149166
```

In order for a variable to be a confounder, the variable must change the slope parameter by 10-20% and it must sensibly simultaneously cause the dependent and independent variables in the model.

Using a threshold of 10% for the change in the slope/beta parameter of each of the variables, it appears that class type(strength), perceived exertion, control, and competence were all not confounded by age and BMI. The class type(flexibility) had a change in slope of 11.16%. However, the data dictionary indicates that class type was randomized. Given this information, I would expect the randomization to mitigate confounding effects of age and BMI on class type, and therefore I do not consider class type to be confounded by age and BMI.

1c. [2 points] Based on your answer to (1b), which model do you feel comfortable proceeding with? Justify your answer.

In part 1b, I had found that age had changed the coefficient estimate for class type at a considerable level, 11%, but ultimately decided that it is not a confounding variable because of the study design. The class type variable was a randomized condition, which I believe would mitigates any causal relationship age would have on the class type an individual enrolled in, making it not a confounding variable. In addition to the lack of compelling evidence for age to be a confounding variable, leaving it out would keep the model parsimonious. For these reasons, I will proceed with the unadjusted model in which there are no additional confounding variables added into the model.

## Question 2 [13 points]

Add complexity to your model by testing whether model effects vary by class type.

2a. [2 points] Do any of the main independent variables (in the table in Question 1) interact with classtype in their association with satisfaction? Provide the p-values you used to test these interactions.

(use extra SS test to determine interaction effects for dummy variable sets)

```
> anova(
+   lm(satisfac ~ classtype.f + rpe + encourage + control + perc_comp, data=gx),
+   lm(satisfac ~ rpe*classtype.f + encourage + control + perc_comp, data=gx)
+ )
Analysis of Variance Table

Model 1: satisfac ~ classtype.f + rpe + encourage + control + perc_comp
Model 2: satisfac ~ rpe * classtype.f + encourage + control + perc_comp
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     83 148.03
2     81 127.09  2     20.94 6.6727 0.002077 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(
+   lm(satisfac ~ classtype.f + rpe + encourage + control + perc_comp, data=gx),
+   lm(satisfac ~ rpe + encourage*classtype.f + control + perc_comp, data=gx)
+ )
Analysis of Variance Table

Model 1: satisfac ~ classtype.f + rpe + encourage + control + perc_comp
Model 2: satisfac ~ rpe + encourage * classtype.f + control + perc_comp
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     83 148.03
2     81 143.47  2    4.5603 1.2873 0.2816
> anova(
+   lm(satisfac ~ classtype.f + rpe + encourage + control + perc_comp, data=gx),
+   lm(satisfac ~ rpe + encourage + control*classtype.f + perc_comp, data=gx)
+ )
Analysis of Variance Table

Model 1: satisfac ~ classtype.f + rpe + encourage + control + perc_comp
Model 2: satisfac ~ rpe + encourage + control * classtype.f + perc_comp
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     83 148.03
2     81 144.58  2    3.4542 0.9676 0.3843
> anova(
+   lm(satisfac ~ classtype.f + rpe + encourage + control + perc_comp, data=gx),
+   lm(satisfac ~ rpe + encourage + control + perc_comp*classtype.f, data=gx)
+ )
Analysis of Variance Table

Model 1: satisfac ~ classtype.f + rpe + encourage + control + perc_comp
Model 2: satisfac ~ rpe + encourage + control + perc_comp * classtype.f
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     83 148.03
2     81 144.95  2    3.0869 0.8625 0.4259

>
```

According to my test for whether class type has an effect on any of the main independent variables, it appears that class type has a statistically significant interaction with rpe (perceived exertion) (p=0.002). None of the other p-values for the interaction terms were found to be statistically significant at the alpha level cutoff ($\alpha$=0.15).

2b. [3 points] Based on your analyses so far, and considering interactions and confounding (you may reassess confounding after making a decision on 2a), decide on your preliminary final model—the model that best describes the researchers'' questions. Provide the parameter estimates, standard errors, and p-values of the coefficients.

My final model will include all of the main variables, and include interaction terms between class type and rpe and class type. The following are the parameter estimates, standard errors, and p-values of all of the coefficients of the final model:

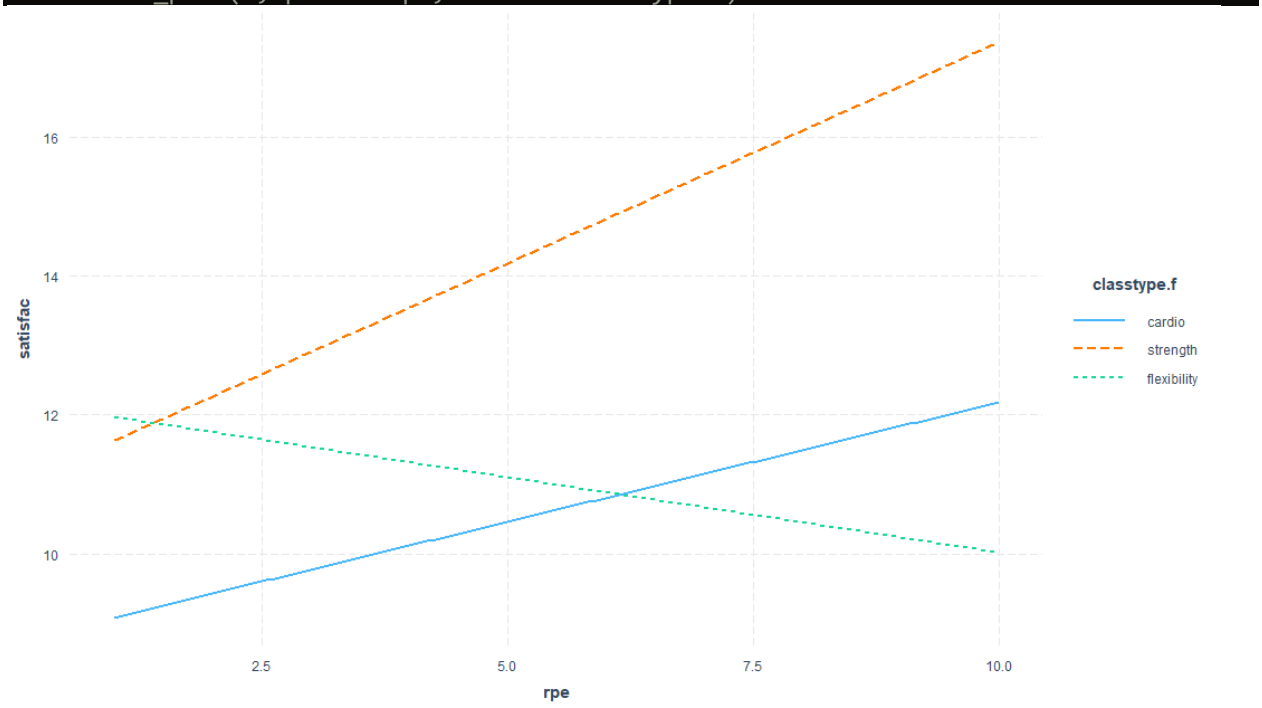
```
> m <- lm(satisfac ~ classtype.f*rpe + encourage + control + perc_comp, data=gx)
> summary(m)

Call:
lm(formula = satisfac ~ classtype.f * rpe + encourage + control +
    perc_comp, data = gx)

Residuals:
     Min       1Q   Median       3Q      Max
-3.10993 -0.81659 -0.07467  0.81297  3.11786

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.2092     1.2740   3.304  0.00142 **
classtype.fstrength         2.4639     1.6042   1.536  0.12845
classtype.fflexibility      3.5471     1.3074   2.713  0.00814 **
rpe                         0.3660     0.1744   2.098  0.03900 *
encourage                   0.4630     0.1414   3.274  0.00156 **
control                     0.4121     0.1363   3.024  0.00334 **
perc_comp                   0.3150     0.1162   2.710  0.00821 **
classtype.fstrength:rpe     0.2537     0.2441   1.039  0.30172
classtype.fflexibility:rpe -0.5849     0.2359  -2.479  0.01524 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.253 on 81 degrees of freedom
Multiple R-squared:  0.8221,     Adjusted R-squared:  0.8045
F-statistic: 46.79 on 8 and 81 DF,  p-value: < 2.2e-16
```

~~The equation of my final model is: $\hat{Y}_{satisf} = 5.00 + 0.345X_{rpe} + 1.81X_{classtype(str)} + 1.86X_{classtype(flexiblility)} + 0.309X_{encourage} + 0.407X_{control} + 0.305X_{perc\_comp} + 0.294X_{classtype(str)}X_{rpe} + -0.562X_{classtype(flexiblility)}X_{rpe} + 0.105X_{classtype(str)}X_{encourage} + 0.372X_{classtype(flexiblility)}X_{encourage}$~~

The equation of my final model is: $\hat{Y}_{satisf} = 4.2 + 0.37X_{rpe} + 2.46X_{classtype(str)} + 3.55X_{classtype(flexibility)} + 0.46X_{encourage} + 0.41X_{control} + 0.32X_{perc\_comp} + 0.25X_{classtype(str)}X_{rpe} + -0.58X_{classtype(flexibility)}X_{rpe}$

2c. [5 points] For the variables that have significant interaction terms, describe the nature of how class type interacts these variables, providing the stratum-specific estimates of the relationships between that variable and satisfaction. Provide a plot that illustrates the interaction.

```
> interact_plot(m, pred = rpe, modx = classtype.f)
```



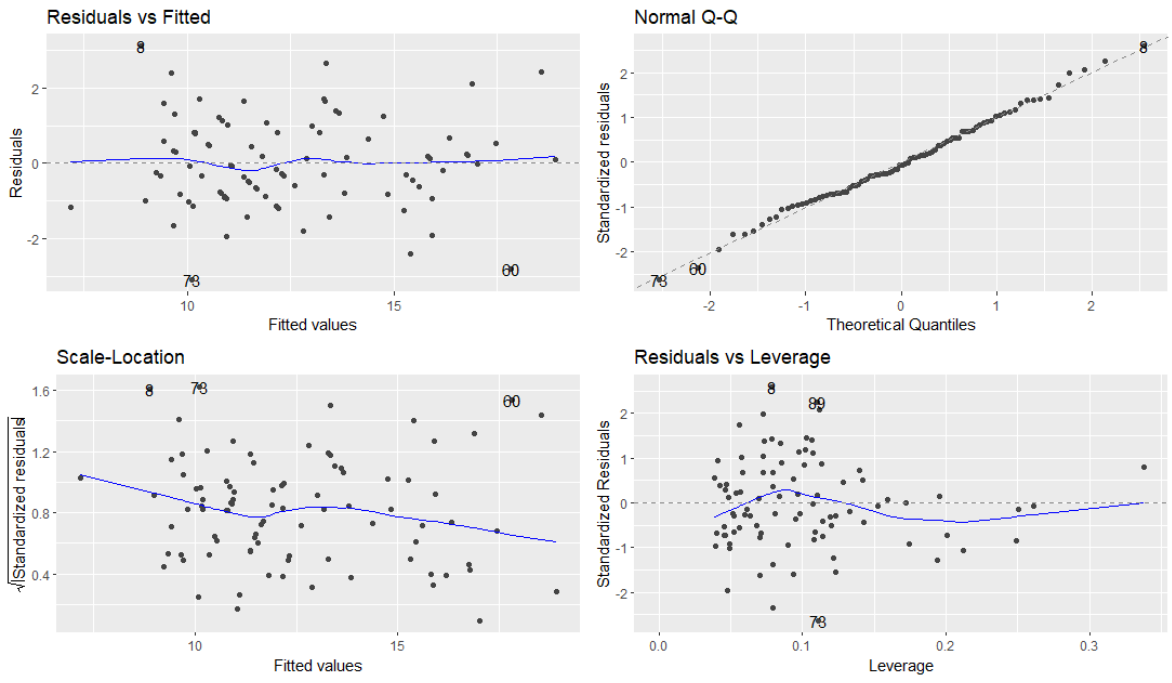

Holding all other variables constant, the slope and intercept estimate for rate of perceived exertion vs. satisfaction is highest for those who took the strength class, and both slope and intercept are lower for those who took the cardio class. Slope estimates for those who took the strength and cardio classes are positive. Furthermore, the intercept is highest and the slope becomes negative for those who took the flexibility class.

```
> sim_slopes(m, pred = rpe, modx = classtype.f)
SIMPLE SLOPES ANALYSIS
```

```
Slope of rpe when classtype.f = flexibility:

   Est.   S.E.   t val.      p
------- ------ -------- ------
  -0.22   0.17    -1.29   0.20

Slope of rpe when classtype.f = strength:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.62   0.17     3.64   0.00

Slope of rpe when classtype.f = cardio:

  Est.   S.E.   t val.      p
------ ------ -------- ------
  0.37   0.17     2.10   0.04
```

It appears that rpe is related to satisfaction only for when class type is strength or cardio.

2d. [3 points] Evaluate your model assumptions of linear regression.




Linearity: The residuals vs. fitted values plot shows a random scatter around a mean line of 0, the assumption of linearity is met.

Normality: The normal Q-Q plot shows the standardized residuals following a straight line, the assumption of normality is met.

Homoscedasticity: The scale-location plot shows that the standardized residuals are mostly scattered randomly around a mean line of zero, the standardized residuals at lower fitted values are very slightly higher, but there is no alarming violation, so the assumption of homoscedasticity is met.

## Question 3 [9 points]

Provide a conclusion.

3a. [7 points] Write a concluding methods/results paragraph. Keep your conclusion brief (only your first 400 words will be graded), providing only the most important aspects of your methods and results. Your conclusion should explain:

- The research question you attempted to address
- The steps you took to evaluate the research question, including your modeling approach
- How you addressed confounding and interactions and what you found
- Convince the reader you have a good model with respect to the assumptions and potential outliers
- Provide an interpretation of coefficients in your final model (with relevant p-values), keeping in mind how this interpretation relates to the research question

I attempted to address whether the exertion, encouragement, participant control, and perceived competence relate to workout satisfaction, and whether these effects vary based on the type of workout. I first checked if age and BMI confounded any of the variables of interest, and found that while they did affect the coefficient estimates of class type, I chose not to include these variables in my final model based on the class type randomized study design. Next, I checked for interactions to see if there were any strata-specific effects of the class type on the other main variables. I found that class type had interactions with exertion and encouragement at an alpha level of 0.15. So, in my final model I included the original variables of interest, plus the class type interactions with exertion and encouragement. The final model was checked for the assumptions of linear regression and no outliers with high leverage were found to have significant effects on the estimated parameters of the model.
The interpretations of the coefficients are as follows:
(Holding all other variables not mentioned in each statement constant):

- A one-unit increase in exertion is associated with a 0.34 unit predicted increase in satisfaction score (p=0.05)
- An individual who took the strength class is predicted to have a 1.8-unit higher mean satisfaction score than an individual who took the cardio class (p=0.32)
- An individual who took the flexibility class is predicted to have a 1.86-unit higher mean satisfaction score than an individual who took the cardio class (p=0.28)
- A one-unit increase in encouragement is associated with a 0.31 unit predicted increase in satisfaction score (p=0.099)
- A one-unit increase in control is associated with a 0.41-unit predicted increase in satisfaction score (p=0.004)
- A one-unit increase in competence is associated with a 0.31-unit predicted increase in satisfaction score (p=0.01)
- Compared to someone that did cardio, the slope associated with exertion increases by 0.29 for someone that did strength training (p=0.24)
- Compared to someone that did cardio, the slope associated with exertion decreases by 0.56 for someone that did flexibility training (p=0.02)

- Compared to someone that did cardio, the slope associated with encouragement increases by 0.10 for someone that did strength training (p=0.69)
- Compared to someone that did cardio, the slope associated with encouragement increases by 0.37 for someone that did flexibility training (p=0.12)

| 3b. [2 points] How much of the variation in satisfaction scores is explained by your model? |
|---|

My model explains 82.8% of the variation in satisfaction scores ($R^2$=0.8276).