|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 2 = Strength 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**  **Estimate (SE) Unadjusted** | **Model 2**  **Estimate (SE) Age-Adjusted** | **Model 3**  **Estimate (SE) BMI-Adjusted** | **Model 4**  **Estimate (SE) 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\*\*\*)  **--------------**  -0.92 (vs. flexibility) (SE=0.43) (p=0.035\*) | -4.18 (vs. strength) (SE=0.35) (p<2e-16\*\*\*)  --------------  -1.04 (vs. flexibility) (SE=0.41) (p=0.015\*) | -4.12 (vs. strength) (SE=0.36) (p<2e-16\*\*\*)  --------------  -0.92 (vs. flexibility) (SE=0.43) (p=0.037\*) | -4.17 (vs. strength) (SE=0.35) (p<2e16\*\*\*)  --------------  -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. |

The following function calculates the percent change in the beta estimate for each of the variables used in the regression model before and after adjusting for age and BMI:

> beta\_pct\_change <- function(adj, unadj) {

+ return((abs((unadj - adj)/unadj))\*100)

+ }

> beta\_pct\_change(4.1245, 4.17838) # strength

[1] 1.289495

> beta\_pct\_change(.9213, 1.03237) # flexibility

[1] 10.75874

> beta\_pct\_change(.2610, .27563) # perceived exertion

[1] 5.30784

> beta\_pct\_change(.5246, .46771) # encouragement

[1] 12.16352

> beta\_pct\_change(.3947, .38527) # control

[1] 2.447634

> 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 10.75%. 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. Lastly, the slope estimate for encouragement changed by 12.16%, meeting the threshold criteria. Additionally, BMI could potentially be a cause of how much encouragement an individual receives; those with a BMI further from the overall mean BMI of individuals could receive more encouragement from workout instructors than those who have a BMI that is closer to the mean BMI. Age could also sensibly confound encouragement, as younger individuals tend to be more physically fit than older individuals. This can lead to instructors giving more encouragement to older individuals who are less likely to be physically fit.

|  |  |
| --- | --- |
|  | 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 and BMI to be confounders of the relationship between encouragement received from instructor and satisfaction with the workout program. When accounted for, age and BMI changed the slope parameter estimate for encouragement by over 10%. However, upon closer inspection of the change in the coefficient for between age-adjusted, BMI-adjusted, and age and BMI-adjusted models, it appears that the change occurred only when age is added to the model

Conceptually, an individual who has a BMI closer to the mean BMI, and who is younger (and therefore more likely to be in good physical condition) could sensibly receive less encouragement from instructors than those who have higher or lower BMIs and/or are older and less physically active (i.e. less likely to be in good physical condition). Individuals that are more physically fit could also potentially feel less satisfaction with the workout class, as they may not find it as challenging as those who are less physically fit. With that, I would feel most comfortable proceeding with the model that adjusts for age and BMI.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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. |

> lm(satisfac ~ (rpe + encourage + control + perc\_comp)\*classtype.f + age + bmi, data=gx) %>% summary()

Call:

lm(formula = satisfac ~ (rpe + encourage + control + perc\_comp) \*

classtype.f + age + bmi, data = gx)

Residuals:

Min 1Q Median 3Q Max

-2.5378 -0.6837 -0.1210 0.7430 2.9070

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.870263 1.960974 1.464 0.14757

rpe 0.327892 0.174695 1.877 0.06452 .

encourage 0.358899 0.214771 1.671 0.09899 .

control 0.263678 0.236862 1.113 0.26927

perc\_comp 0.332607 0.181474 1.833 0.07091 .

classtype.fstrength 1.041886 1.981322 0.526 0.60058

classtype.fflexibility 2.459505 1.997980 1.231 0.22228

age 0.059999 0.019733 3.041 0.00328 \*\*

bmi 0.008864 0.042481 0.209 0.83529

rpe:classtype.fstrength 0.349475 0.249705 1.400 0.16588

rpe:classtype.fflexibility -0.505513 0.257581 -1.963 0.05351 .

encourage:classtype.fstrength -0.176989 0.341746 -0.518 0.60610

encourage:classtype.fflexibility 0.327246 0.325826 1.004 0.31852

control:classtype.fstrength 0.426015 0.316227 1.347 0.18209

control:classtype.fflexibility -0.125560 0.344520 -0.364 0.71658

perc\_comp:classtype.fstrength -0.031913 0.278111 -0.115 0.90896

perc\_comp:classtype.fflexibility -0.012630 0.270272 -0.047 0.96285

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.207 on 73 degrees of freedom

Multiple R-squared: 0.8513, Adjusted R-squared: 0.8187

F-statistic: 26.11 on 16 and 73 DF, p-value: < 2.2e-16

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=0.054).

None of the other p-values for the interaction terms were found to be statistically significant at the alpha level cutoff (𝛼=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, plus age and BMI, and include an interaction term between class type and rpe. The following are the parameter estimates, standard errors, and p-values of all of the coefficients of the final model:

> lm(satisfac ~ age + bmi + rpe\*classtype.f + encourage + control + perc\_comp, data=gx) %>% summary()

Call:

lm(formula = satisfac ~ age + bmi + rpe \* classtype.f + encourage +

control + perc\_comp, data = gx)

Residuals:

Min 1Q Median 3Q Max

-2.85649 -0.76159 -0.07984 0.79601 3.11695

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.29360 1.66847 1.375 0.17312

age 0.05898 0.01879 3.140 0.00238 \*\*

bmi 0.01174 0.04088 0.287 0.77464

rpe 0.31570 0.16708 1.890 0.06249 .

classtype.fstrength 1.74767 1.54747 1.129 0.26216

classtype.fflexibility 3.22660 1.25062 2.580 0.01173 \*

encourage 0.38982 0.13972 2.790 0.00660 \*\*

control 0.40600 0.13488 3.010 0.00350 \*\*

perc\_comp 0.33046 0.11346 2.913 0.00466 \*\*

rpe:classtype.fstrength 0.37416 0.23602 1.585 0.11689

rpe:classtype.fflexibility -0.51285 0.22611 -2.268 0.02605 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

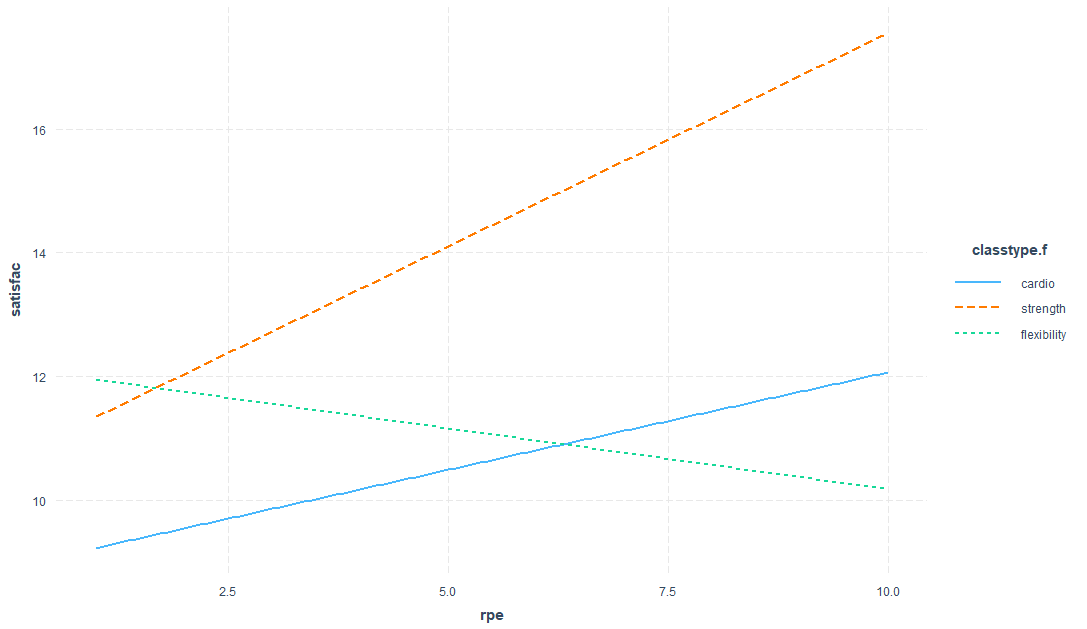
Residual standard error: 1.194 on 79 degrees of freedom

Multiple R-squared: 0.8423, Adjusted R-squared: 0.8224

F-statistic: 42.2 on 10 and 79 DF, p-value: < 2.2e-16

|  |  |
| --- | --- |
|  | 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 rpe 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.

The following output provides the class type stratum-specific estimates of the relationships between rpe and satisfaction:

> sim\_slopes(m, pred = rpe, modx = classtype.f)

**SIMPLE SLOPES ANALYSIS**

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

Est. S.E. t val. p

------- ------ -------- ------

-0.20 0.16 -1.22 0.23

*Slope of rpe when classtype.f = strength:*

Est. S.E. t val. p

------ ------ -------- ------

0.69 0.16 4.21 0.00

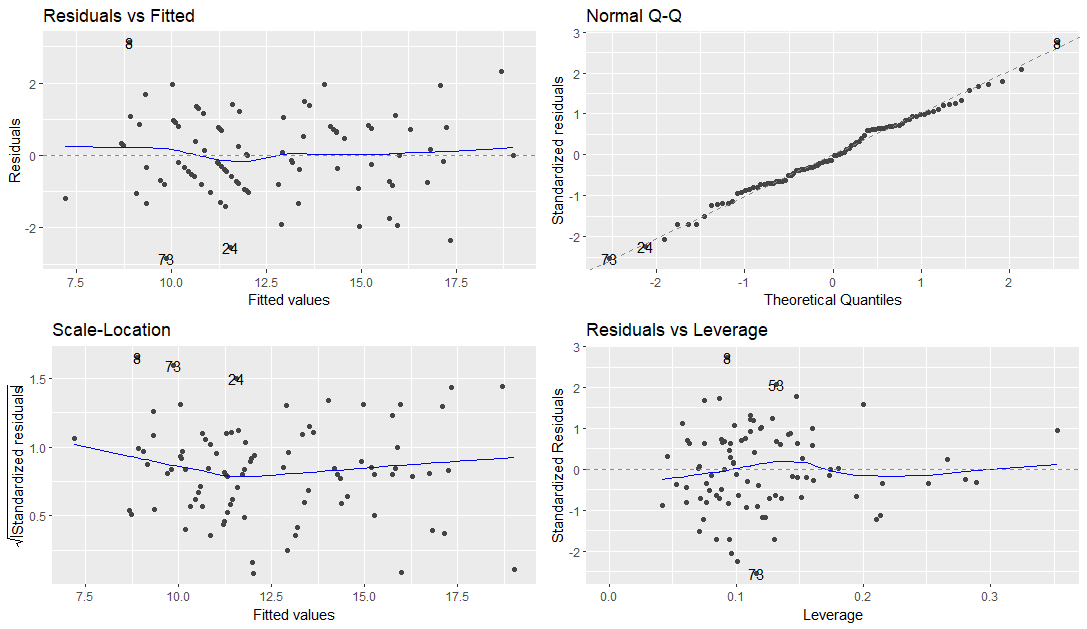
*Slope of rpe when classtype.f = cardio:*

Est. S.E. t val. p

------ ------ -------- ------

0.32 0.17 1.89 0.06

|  |  |
| --- | --- |
|  | 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 rate of perceived exertion, instructor 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 they did affect the coefficient estimates of instructor encouragement considerably, and reasoned that these are plausible confounding variables. 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 the interaction between class type and perceived exertion was significant at an alpha level of 0.15. So, in my final model I included the original variables of interest, plus age, BMI, and the interactions between class type and perceived exertion. The equation of my final model is:

The final model was checked for the assumptions of linear regression and no outliers with high leverage were found.

The interpretations of the coefficients are as follows:

(Holding all other variables not mentioned in each statement constant):

* A one-year increase in age is associated with a 0.06 unit predicted increase in satisfaction score (p=0.002)
* A one-unit increase in BMI is associated with a 0.01 unit predicted increase in satisfaction score (p=0.77)
* A one-unit increase in rate of perceived exertion is associated with a 0.32 unit predicted increase in satisfaction score (p=0.06)
* An individual who took the strength class is predicted to have a 1.75-unit higher mean than an individual who took the cardio class (p=0.26)
* An individual who took the strength class is predicted to have a 3.23-unit higher mean than an individual who took the cardio class (p=0.01)
* A one-unit increase in instructor encouragement is associated with a 0.40 unit predicted increase in satisfaction score (p=0.007)
* A one-unit increase in participant control is associated with a 0.41-unit predicted increase in satisfaction score(p=0.004)
* A one-unit increase in perceived competence is associated with a 0.33-unit predicted increase in satisfaction score(p=0.005)
* For an individual who took the strength class, the slope associated with rate of perceived instruction increases by 0.37 compared to an individual who took the cardio class
* For an individual who took the flexibility class, the slope associated with rate of perceived instruction decreases by 0.51 compared to an individual who took the cardio class

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

My model explains 84.2% of the variation in satisfaction scores (=0.8423).