**Individual Project: Using Regression in a Selection Context**

Due date:

Part 2: April 14, 2022 (before class)

Description: HR analytics is a growing field that leverages statistical modeling and analysis in an effort to facilitate or improve traditional HR practice. As a field, HR is becoming less knowledge-based and more data-based. To maximize marketability, HR practitioners must hone their analytical thinking skills and improve their statistical skillset. Knowledge of core concepts will always be necessary for success in HR, but the ability to discover and interpret trends in data will provide additional opportunities in this field.

Objective: This project is intended to simulate a regression-based modeling approach to HR analytics in a selection context. You should be comfortable with based statistical analysis such as single and multiple regression. Relying on empirical evidence is a key component of evidence-based decision-making. Relying on objective, quantitative information can strengthen arguments, and it is important to understand when and how to use data to support desired courses of action.

For this project, you will be required to “clean” data, make decisions and recommendations based on statistical analyses, and integrate HR analytics with traditional HR concepts. In doing so, it is often helpful to consider the desired final product and related outcomes from a practical perspective. It is important to find the optimal solution, but it is also important to consider legal and ethical constraints. This project is designed to provide you with an opportunity to practice justifying decisions using a combination of empirical evidence and persuasive arguments.

* Please take some time to research careless responding and outliers. Key concepts include random responding, straightlining, and illegal values.
* Please take some time to research regression analysis. Key concepts include regression equations, intercepts, regression weights, and R2.
* Please take some time to consider the legal and ethical issues in HR. Key concepts include bias, discrimination, protected classes, and disparate impact.
* All analyses can be conducted in Microsoft Excel, but you are free to use other programs if desired.
* The instructor is available to answer questions and discuss or clarify any of these topics, but it is expected that you will first research them on your own.

Deliverables:

1. [Due date: October 7, 2021] A “cleaned and scored” dataset that (a) calculates the final 11 variable scores for all participants and (b) specifies the ID numbers of any individuals “screened out” prior to data analysis as well as the reason(s) they were screened. The accuracy of this dataset will determine 50% of your project grade.
2. [Due date: November 11, 2021] A short report (no more than 2 pages) addressing the two analyses (and associated questions) listed on the following pages. This report will determine 50% of your project grade.

Scenario: You have been tasked with designing a selection system for your organization. Specifically, you have been provided access to data on 9 predictor variables and 3 criterion variables for 1000 current employees of your organization. Details as follows:

*Predictor list (range of indicator scores listed in parentheses)*

1. Age in years
2. Aptitude
   1. Four-indicator scale (0 = Low aptitude, 100 = High aptitude)
3. Biological sex (0 = female, 1 = male)
4. Conscientiousness
   1. 10-indicator scale (1 = Low conscientiousness, 7 = High conscientiousness)
5. Grade point average (GPA)
6. Integrity
   1. 10-indicator scale (1 = Low integrity, 7 = High integrity)
7. Intelligence (55 = Low intelligence, 145 = High intelligence)
8. Interview scores
   1. Rated by 5 interviewers (0 = Bad interview, 10 = Good interview)
9. Years of work experience

*Criterion list (range of indicator scores listed in parentheses)*

1. Counterproductive work behaviors (CWBs)
   1. 5-indicator scale (1 = Low CWBs, 7 = High CWBs)
2. Intentions to quit
   1. 3-indicator scale (1 = Low intention to quit, 5 = High intention to quit)
3. Job performance
   1. Rated by 3 supervisors (0 = Low performance, 100 = High performance)

*Additional variables*

1. ID number (variable name “ID")
2. Instructed item (variable name “IItem”)
   1. “Please indicate response option 2 for this item”
3. Survey time in seconds (variable name “Time”)

*Potentially Useful Concepts*

* Correlation
* Regression (both single and multiple regression)
* R2

*Potentially Useful Microsoft Excel Concepts and Functions*

* Absolute cell referencing (the $ symbol)
* =AVERAGE()
* =CHOOSE()
* =CORREL()
* =IF()
* =SUM()
* =VLOOKUP()

*Tasks for Deliverable #2*:

Analysis 1:

1. Compute, report, and interpret bivariate correlations between the following variables:
   1. Intelligence and job performance
   2. Integrity and CWB
   3. Conscientiousness and intentions to quit
2. How can knowing these correlations inform selection procedures?

Analysis 2:

1. You have been tasked with designing a selection system for your organization. Explore combinations of variables in an effort to maximize the prediction of job performance. Specifically, job performance will serve as your criterion measure. ***You may only use one predictor from each of the following categories*** (thus, four or fewer total predictors):
   1. Ability data (Aptitude or Intelligence)
   2. Biodata (GPA, Interview, or Work Experience)
   3. Demographic data (Age or Sex)
   4. Personality data (Conscientiousness or Integrity)
2. Describe why you selected each predictor variable (and why you did not select others).
   1. Recommendation: describe your conceptual, theoretical, legal, and/or empirical reasons for each variable you included in your final predictive model.
3. Describe how you determined that the final set of predictors you chose resulted in maximization of the prediction of job performance (i.e., an acceptably high R2).

*What to turn in*: A Word document (two pages or less) containing (and explaining) the results of the two analyses above. **This write-up must \*explicitly report\* (a) the three correlations, (b) the set of predictors in your final model, and (c) the R2 value for your final model.**

*How it will be graded*: 30% accuracy of bivariate correlations, 30% maximization of prediction of job performance, 30% explanation of procedures and decisions, 10% spelling and grammar.

*The practice Excel worksheet (available on Blackboard) can provide you some practice with computing correlations and multiple regressions using Excel*.

\*Note: If using Excel for these analyses, you may need to modify the data to conduct analyses. For example, you may need to sort or rearrange the respondents (rows) of variables (columns).

*Common Excel error messages (and how to address them)*:

* “Regression - Input range contains non-numeric data.”
  + Issue: Some of your cells have data that are not numeric (e.g., empty or "blank" cells, formulas, spaces, or letters).
    - Fix #1: Make sure you did not select the variable label (e.g., "Criterion" in cell R1) alongside your numeric data.
    - Fix #2: Sort your rows so all rows with blank cells are at the top (or bottom), then only use respondents with non-blank cells.
    - Fix #3: Delete the rows that contain blank cells.
* “Regression - Input range must be a contiguous reference.”
  + Issue: You selected two or more separate ranges for your predictor variables
    - Fix: Rearrange (i.e., cut/paste) your predictor variables so they are adjacent, then select them all at once**.**