**Jon Doretti**

**DSC 324/424: Assignment #1**

**Due: Sunday, July 24, 2022 by 11:59PM (by midnight)**

**Total: 50 points**

**Problem 1(5 points –** Due **Monday, July 18th, 2022** at 5PM) Introduce yourself on D2L by posting to the Class Introductions forum on D2L. Include a bit of information about yourself including some of the following. Note, this

- Name

- Undergraduate Degree

- Major/Degree Program (Concentration)/Time in Program (e.g. 3rd quarter, 2nd yr, graduating this quarter)

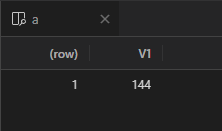
- Position at Work, if applicable

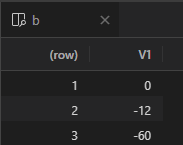
- What is your experience with R? Have you used it for any courses? For work? - What interests you about Advanced Data Analysis?

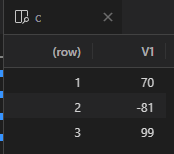
- Field(s) of Interest and/data

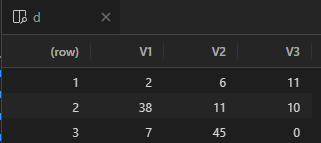
- Hobbies

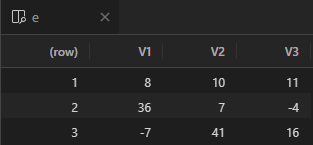
**Problem 2 (10 points)** Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

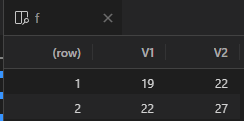
a) 

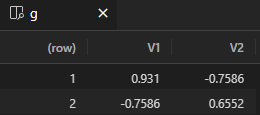
b)

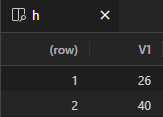
c)

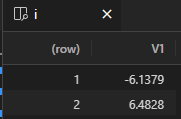
d)

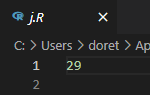
e)

f)

g)

h)

i)

j)

**Problem 3 (10 points –other types of regression models):** There are other types of regression models outside of linear and logistic regression. **Using Google Scholar**, locate **a journal article**, which utilizes **one** of the **types of regressions** listed below or another regression outside of linear/logistic that interests you. **Write a summary** of the journal article and how it utilizes the regression model in **two to three paragraphs**. **Cite the paper in APA format. Choose one of the following regressions:**

3. Elastic Net Regression

Within machine learning, there is a subset that focuses on the use for algorithmic trading. The use of data and data visualization to build, predict and execute successful trades within any financial market is on the rise. Various techniques shed light on the many aspects of algorithmic trading. The elastic net regression can be used in areas such as portfolio optimization. In Shen et. al’sDoubly regularized portfolio with Risk Minimization, the researcher team used elastic net regression to esure “ (1) only a small set of assets are selected to invest in each period; (2) portfolios in consecutive trading periods are similar” (Shen et al, abstract). This is optimal for their portfolio as their goal is to account for capital tax gains and transaction costs while also maintaining a minimal risk portfolio. However risk is being calculated through different models other than the elastic net regression.

The use of the elastic net regression is to weigh two coefficients: (1) structured sparsity and (2) turnover. Structured sparsity is weighted to select a small set of assets as stated above and turnover is to weigh the similarity of portfolios in consecutive trading periods. Using these coefficients along with standard risk minimization models is how Shen et al derive the doubly regularized portfolio. Using their equation 6 they are able to use the elastic net regression or elastic net penalty as they referred to is as. The elastic net regression has also been used, according to Shen et al, to derive the doubly regularized support vector machine in research done by Wang, Zhu, and Zou 2006. The elastic net regression was used to calculate a unique minimum of assets and a minimum of trades to be made of these assets in respect to the coefficients.

The results showed that the double regularized portfolio beats many benchmarks and backtests. However, one flaw of many trading algorithms is that they are over-fitted. Therefore data is skewed. Shen et al claim that the double regularized portfolio “consistently achieves higher risk-adjusted returns with much lower implied cost measured by turnover rates” (1291, Conclusion). The double regularized portfolio grows more steadily and with less volatility through most investment periods.

Shen, W., Wang, J., & Ma, S. (2014). Doubly regularized portfolio with Risk Minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(1). https://doi.org/10.1609/aaai.v28i1.8906

**Problem 4 (10 points-Data Ethics or Data Integrity): Using Google Scholar**, locate **a journal article**, which discusses data ethics or data integrity in terms of big data in your field of interest. **Write a summary** of the journal article and how it utilizes data ethics or data integrity in **two to three paragraphs**. **Cite the paper in APA format.**

Algorithmic trading is more than building portfolios using machine learning. Algorithmic trading also uses high frequency trading. High frequency trading helps price discovery in financial markets. However, in the short term high frequency trading can cause volatility and high price fluctuations. High frequency trading is not impermeable to manipulation as there are many strategies to manipulate asset value and access to data according to Merli et al. For example, “Another manipulative strategy consists in placing a large number of orders in the market (quote stuffing). Quote stuffing may affect slow traders market access” (Merli et al, 185, HFT and manipulation). As a result there have been many fail-safes implemented into the financial markets. For example the circuit breaker:

Circuit Breaks: there are interruptions of trading mechanisms. They are used to facilitate the management of momentary orders imbalances that can cause sudden price movements (Poirer 2012). Pausing markets can be a good way to allow market participants to recalibrate their strategies and to reset their own algorithms parameters (IOSCO 2011). (Merli et al, 187, Policy Responses).

These manipulation tactics are highly unethical and majority of the time are carried out by large hedge funds. This is because the majority, everyday investors do not control the necessary capital in order to carry out these manipulation strategies. It is big corporate investors that are hiding the data from the regular or retail investors, manipulating them in order to make money off their decisions. In order to combat these types of manipulation, governmental bodies have required transparency in the markets:

Transparency: Firms should periodically produce reliable and clear reports in order to inform stakeholders about their performances. They must ensure a regular update on the state of the art of their procedures. In addition, companies should keep detailed records of information about key decisions, system properties, testing methodologies to enable competent authorities to carry out their monitoring role (ESMA 2012). (Merli et al, 189, HFT & Ethics)

Merli et al concluded that “strategies and automated trading systems should plan their trading decisions process basing it on fairness, prudence and diligence principles. These principles are the basis of good governance” (190). However, the last sentence is the key, basis of good governance. With no enforcement of the rules, those who wish to take advantage will without repercussion and manipulate data and asset values.

Merli, R., Massa, I., & Lucchetti, M. (2014). HIGH FREQUENCY TRADING: TECHNOLOGY, REGULATION AND ETHICAL ISSUES. *– Achievements and challenges of commodity science in the age of globalization, Polish Society of Commodity Science, Cracow, pp. 179-192. ISBN: 978-83-938909-0-3*

**Problem 5: (15 pts – regression analysis, visualization, and interpretation):** The insurance\_dataset.csv dataset contains 1338 observations (rows) and 7 features (columns). The insurance data contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

We are interested in which independent variables are significant for **predicting the expenses** by the other predictors.

a. (5 points) Before running any regressions make sure to check for

multicollinearity. How did you check for multicollinearity? If there is

multicollinearity, how do you plan to resolve it? Are there any other issues with the dataset we must consider before running the regressions?

VIF is a method of checking for multicollinearity. A VIF < 5 means there is no multicollinearity. To solve multicollinearity you can remove the variable with the highest VIF from the data set. After, another multicollinearity should be run to ensure there are no more variables that have multicollinearity.

b. Run a multiple regression of price on the variables listed above.

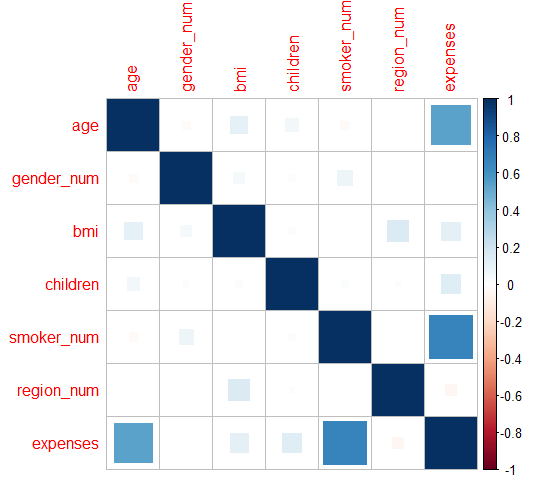
i. (5 points) Run the model **using an automatic method** (i.e., stepwise, forward, backward). Explain why you chose the method. Comment on

the overall significance of the regression fit. Which predictors have

coefficients that are significantly different from zero at the .05 level?

I chose to use the forward selection model because after calculating the VIF on the first expenses model I noticed that there was no multicollinearity between any of the variables and expenses. This made me want to know the variable(s) that had the most effect on expenses. I found that all variables are significant on the expenses. The overall significance of the regression fit could have been better. There is not much variance between the R(.7507) and adjusted-R(.7498).

ii. (5 points) Using the variables above, **create a visualization**, which will provide an interesting story or insight within this data.



The above graph shows the correlation between all variables with BMI influencing the most outside of expenses. Expenses has the greatest correlation between variables and it is interesting to see as the VIF did not return a value greater than 5 for any variable.