***Fall 2020 6372 Midterm***

Directions: Please provide all of your answers in this document and submit it to 2DS. Note: By turning in a midterm, you promise to not have consulted with anyone about the content of this test during the administration of it through any media (chat, phone, email, etc.). Not adhering to the SMU honor code has serious consequences.

***ATTENTION: Please include your responses in a different color font.***

Every item/question is worth 3 points each. Feel free (not expected) to provide comments to your MTC answers for partial credit.

\_\_d \_\_ 1. If a model consistently under or over predicts the response for certain predictor values or combinations, the test error (ASE) will increase due to

1. High bias
2. High variance
3. Overfitting
4. Both high bias and high variance

\_\_\_d \_ 2. The perks of cross validation over a simple train/test set split is (are)

1. CV can be used when sample size is not very big
2. Variability in the train/test set split can be assessed for CV
3. CV will create the best model possible
4. All of the above
5. Only a and b

\_\_\_b\_\_ 3. The major advantage of a ***multiple linear regression model*** is that

1. They will not over fit data set sets
2. They are easy to interpret
3. They do not require EDA
4. It can forecast time series data

\_\_e \_ 4. Which model fit metric does not have any sort of control for overfitting?

1. AIC
2. BIC
3. Test ASE
4. Train ASE
5. Adjusted R-squared
6. All of the above

\_\_\_b\_\_ 5. When applied on the same set of predictors, a LASSO multiple linear regression fit with a large penalty value will yield (Note: MLR here is in reference to fitting the model on a set of predictors only)

1. A more complex model than an MLR model containing the same set
2. A less complex model than an MLR model containing the same set
3. A model with the same complexity as the MLR model containing the same set
4. There is no definitive call here

\_\_\_a\_\_ 6. Which ASE Plot below provides more evidence of overfitting? Include some thoughts on why? Just focus on the Train/Test plots.

1. The plot on the LEFT.
2. The plot on the RIGHT.
3. Neither show any signs of overfitting.

|  |  |
| --- | --- |
| LEFT | RIGHT |
| A screenshot of a social media post  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

**For each of the following scenarios, select which statistical model would be the most appropriate**

\_\_\_d\_\_ 7. *Scenario 1*: A medical researchers is studying two groups of patients (Healthy children and children with Lupus). For each of the two groups of kids, the medical researcher has data collected (a continuous response) on each of the patients prior to being administered a drug and then an additional measurement is taken one month after taking the drug.

1. Multiple Linear Regression
2. Two Way ANOVA
3. Time Series
4. Repeated measures
5. Something we haven’t covered yet.

\_\_\_d\_\_ 8. *Scenario 2*: Suppose we collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

1. Multiple Linear Regression
2. Two Way ANOVA
3. Time Series
4. Repeated measures
5. Something we haven’t covered yet.

\_\_e\_\_\_ 9. *Scenario 3*: We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

1. Multiple Linear Regression
2. Two Way ANOVA
3. Time Series
4. Repeated measures
5. Something we haven’t covered yet.

**SHORT ANSWERS (3pts each)**

10. A predictive modeler is not happy with his current multiple linear regression model’s prediction accuracy. He currently is using 6 continuous predictors and 2 categorical. These are the only variables he currently has to work within his data set and he is using them as is. All regression coefficients are both statistically and practically significant and assumptions are met. Provide some thoughts or a suggestion on what the modeler could do to try to improve his model for prediction.

The modeler could run pairs on R or plot side by side plots to examine which predictors are highly correlated to each other and remove ones he feels are irrelevant. You could also look at the VIF from SAS to give you an idea of which ones(high VIFs near or above 5) to take out.

11. State the major statistical issue that arises when conducting an analysis that assumes independent errors when the truth is the errors are actually correlated.

Multicollinearity withing features.

12. For the time series graph below, upon visual inspection, do you believe the time series is stationary or not? If not what assumption(s) is it violating?

A screenshot of a cell phone

Description automatically generated

It is not stationary because there is no constant mean and constant variance.

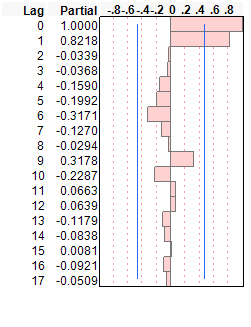
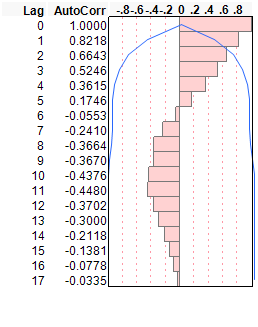
13. Consider the data set below of average hourly wages of textile and apparel workers for the 18 months from January 1986 through June 1987. A plot of the data over time as well as additional diagnostics are provided below.



A close up of a map

Description automatically generated

ACF PACF



**Question:** The time series only has 18 observations so assessment of stationarity will prove difficult. Assuming that the data set is stationary, provide an explanation as to if serial correlation is present in the data set or not. If you conclude that there is serial correlation present, provide a suggested time series model that could appropriately account for the correlation.

There is no way to tell if there is serial correlation because there is no pattern in the plot or data that supports that. The ACF and PACF plots do suggest that there is serial correlation in the data. There is a drop off from lag 0 for both plots.

AR2 – AR2 fits better because looking at the PACF plot, the correlation values where decreasing from lag 0 to lag 1 and then cuts off at lag 2.

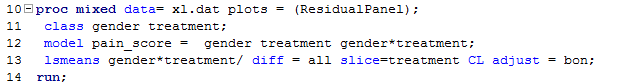
P.S If we consider moving Average, then ACF also suggests MA2 but AR2 is preferable.

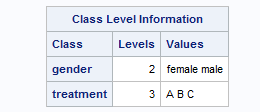
**Analysis Questions (SEE BELOW)**

Consider the data below from a study of 3 pain medicines.

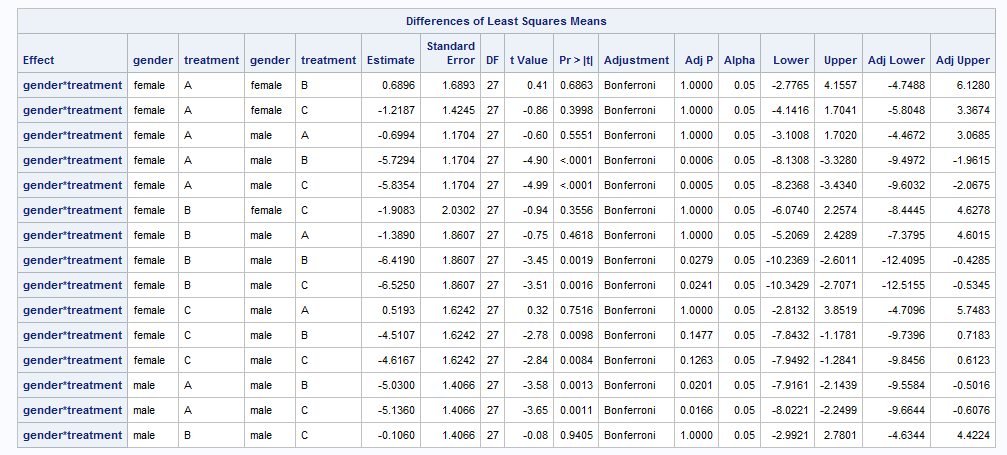


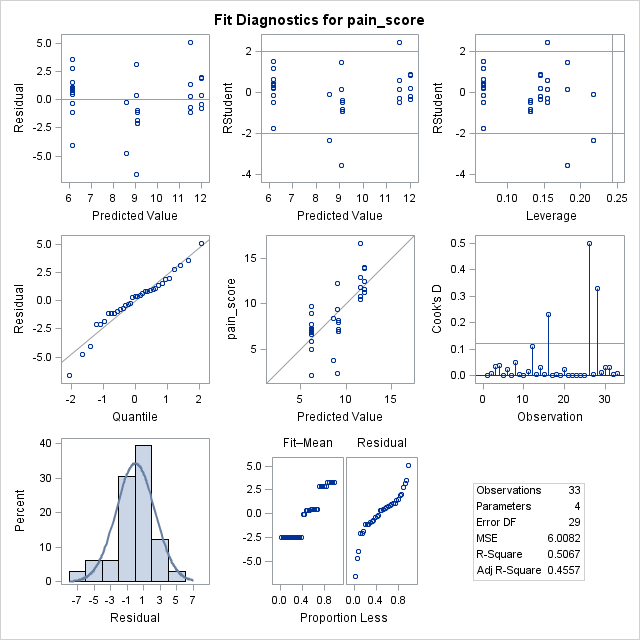
A two way analysis of variance was performed on this data with the results below:

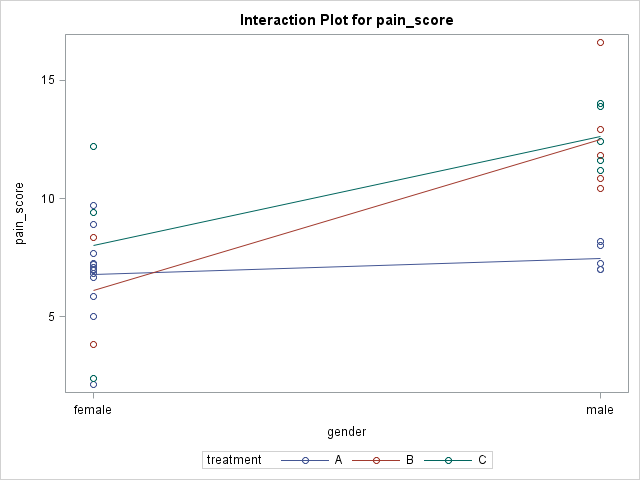










**QA (3pts):** Briefly state the assumptions in order for the analysis results to be valid.

Errors should be independent, normally distributed and have a constant variance. Data must be independent.

**QB (3pts):** Is there any evidence that changes in mean pain score between males and females depend on what treatment they got? Explain what output you used to answer the question and provide a simple explanation in words what the hypothesis test is asking?

If we examine the estimates from least square means table, we can see that females responded better generally than men. Females responded much better to treatment B than men compared with the A or C treatment. The interaction terms all have a low P-value so we reject the null hypothesis.

**QC (3pts):** Dr. Adams is a key opinion leader in the field of pain management and has written that men and women respond differently to pain medicine. Do the results support Dr. Adams statement? Explain.

Dr Adams is correct. Men and women certainly responded differently to the treatments. The interaction term in the model shows a P-value less than 0.05 so we would keep that interaction in the model which means that it’s statistically significant. We can confirm that B and C intersected in the interaction plot and also if you hold line A constant it would eventually intersect because none of the lines are parallel.

**QD (3pts)**: Provide a brief summary on what treatments potentially work better for men. Provide statistical results to back up your summary. Assume that a lower pain score means that the treatment is working better.

Treatment A worked better for men because the estimates from the least means table suggests a lower value even though the t-value is lower, the P-value is approximately the same with B and C.

***The Steve Bramhall Bonus Question (Pts based on quality of response)***

You are about to head into a meeting to discuss a new project in which you will be building a model for prediction. Before you head in, you jot down a series of questions (or general topics of discussion) that you feel have to absolutely be answered before you dive in and get started. Provide a list of questions that you would need so that you could effectively do your job. You may assume that the amount of data you will have to work with will be large. Your questions/topics can be both technical and logistical.

Questions

1. who are the stakeholders and their expectations.
2. Is the data readily accessible? Do we need to install special program or given access to retrieve data? Is the data sensitive?
3. What are the important variables are important? What is the nature of the data? Is it cleaned?
4. What problems are we to solve for you.
5. What information they are hoping to get out of the data set. (How to explore EDA).
6. What the risks if model doesn’t perform well. (Expections – Accuracy, Sensitivity, Specificity)
7. Timeline on when the project is due. (Resources involved).
8. Project goals