Format: Mixture of conceptual questions (Multiple choice, short answer) and small analysis questions (short answer. Code and output will be given. The focus will be on communicating what the output is saying).

Key methods and concepts

Methods

<https://www.methodology.psu.edu/resources/AIC-vs-BIC/>

1. Multiple linear regression (by now with project 1 behind you, you should be pretty comfortable with this procedure)

- Make sure you know what the ASE is. How is it computed.

ASE – average square errors

A screenshot of a social media post

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Formula internet

Comfortable with model adequacy diagnostics

* 1. Comfortable with interpretation of regression coefficients – Unit 11 – Slides and Solution – Stats1
  2. Concept of multicollinearity (when is it a problem and when is it not)

definition – intenet

Talk inside

* 1. Concept of LASSO versus OLS.

Penelized....

* 1. Concept of k-fold cross validation

Talk

* 1. Importance of feature selection. When would feature selection / LASSO actually hurt you

Remove feature that necessary

* 1. Bias/Variance trade off

Overfit – underfit

Outliner can be for high variance

https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html

* 1. Benefits of CV and test sets

Help model selection - test sets: validation

**ross Validation** is used to assess the predictive performance of the models and and to judge how they perform outside the sample to a new data set also known as test data. The motivation to use **cross validation** techniques is that when we fit a model, we are fitting it to a training dataset.

1. Two Way Anova
   1. General workflow of analysis.

A screenshot of a social media post

Description automatically generated

(2 way anova independence is on the observation)

https://stats.stackexchange.com/questions/110519/assumptions-of-two-way-anova

It’s just a regression model so the assumption of normality is on residuals with constant variance. There is no assumption of normality on the response or the predictors. The only difference is repeated measures does not assume each observation is independent like MLR does.

* 1. What is the purpose of the F-tests? Contrasts?

A screenshot of a social media post

Description automatically generated

* 1. Definition of interaction.

 In **statistics**, an **interaction** is a special property of three or more variables, where two or more variables **interact** to affect a third variable in a non-additive manner. In other words, the two variables **interact** to have an effect that is more than the sum of their parts.

* 1. Do outliers matter in Two Way ANOVA? What about Leverage?

Outlier: Yes, can significant change result.

No - Leverage – continuous not categorical

Key methods and concepts

Methods

(repeated features – repeat on different predictors and

And times series on one predictor overtime

One on time - timeseries – have seasonal

repeated not on time

1. Time Series
   1. What is the major pitfall of not appropriately identifying a data set that is time series?

Wrong model for prediction

* 1. What is the major advantage of appropriately identifying a data set that is time series?

Good model and comprision of different model

* 1. Comfortable with interpretation of ACF, PACF plots. **Revisit the Durbin Watson test.**  We did not really cover it in class, but it may creep up in the midterm.

The **Durbin Watson Test** is a measure of autocorrelation (also called serial correlation) in residuals from regression analysis. Autocorrelation is the similarity of a time series over successive time intervals.

* 1. If a time series model include predictors, then is the original time series stationary or not?

“not” ???

* 1. General understanding of stationarity.

Average mean is constant

A screenshot of a social media post

Description automatically generated

* 1. There will be no ARIMA type concepts other than just suggesting models based on rules of thumb.

A screenshot of a cell phone

Description automatically generated

* 1. Repeated measures
  2. Same question #1 from time series.

Subject is one and its time

You can take multiple measurements and subjects

* 1. Do repeated measures have to look like two way ANOVA analysis?

No

Can be continuous not only categorical

* 1. Can you identify a data set is repeated measures through simply looking at the data set and model diagnostics or do you need to have a good understanding of the study design?

We need identify multiple subject multiple times

Study of design too

* 1. Know the basic idea of a covariance matrix. (don’t worry in exam)

<https://v8doc.sas.com/sashtml/stat/chap41/sect20.htm>

* 1. Know that the correlation is the standardized covariance (covariance divided by the standard deviations of each variable.)

(no in exam)

* 1. When we have repeated measures over time, we may need to account for a non zero covariance for observations from the same subject.

This non zero covariance may be modeled by using one of several variance covariance structures. Be familiar with:

* + - Compound Symmetry
    - Unstructured
    - Independent
    - AR(1)
* <https://v8doc.sas.com/sashtml/stat/chap41/sect20.htm>

7. Be able to determine an appropriate covariance structure to use on a given dataset. Model workflow using AIC to help make decisions.

It’s just a regression model so the assumption of normality is on residuals with constant variance. There is no assumption of normality on the response or the predictors. The only difference is repeated measures does not assume each observation is independent like MLR does.