INFO 2950 Lecture Topics, Fall 2023 Midterm

Below is a list of broad topics covered thus far in 2950 (details of these topics, even if not listed, may still be on the midterm even if not explicitly listed on this handout). Topic length does not necessarily correspond to weight in the midterm.

Data types

- String, bool, int, float
- Datetime
- Conversion among data types

Python

- Libraries
- Virtual environments
- Github
- Importing data
    - File system
    - Working directories
- Numpy
    - Lists vs Arrays
    - Shapes of arrays
    - 1D arrays (vectors) have only axis 0 representing rows
    - 2D arrays (matrices) have axis=0 and axis=1
    - More than 2D arrays (tensors) are the same, with axis=2, etc...

SQL

- Choosing columns to return (SELECT)
- Specifying source dataframes (FROM)
- Filtering (WHERE)
- Joins ([INNER, FULL, LEFT, RIGHT] JOIN… ON)
- Making new columns (SELECT…AS)
- Sorting (ORDER BY)
- Aggregating (GROUP BY) with functions (AVG, SUM, COUNT)
- Restricting output (LIMIT)

Dataframes

- What are requirements of dataframes?
- Dimensions
- Reshaping wide <-> long

    - Wide dataframes often work by selecting a row and a column, and finding the number at the intersection
    - Tidy/Tall/Long dataframes have one row per data value
    - Functions melt and pivot to convert between wide and long dataframes

Time Series

- What makes a time series meaningful
    - Is regularly spaced and chronological
    - Has corresponding data per time step
    - Is unique
    - Dealing with missing values
- How to deal with missing data
    - Plotting with NaN vs. missing rows
    - Explain missingness in words

Visualization

- Histograms
- Line plot
    - Time series
- Facets

Statistics

- Statistics on single variables
    - Mean, variance, median
        - Mathematical definitions: write down these equations
- Statistics on multiple variables
    - Covariance, correlation
        - Mathematical definitions: write down these equations
        - Correlation != Causation
    - Normalization
        - Reasons to normalize: your variables have scales that are different orders of magnitude
        - Z-score definition: write down this equation

Regression on single variable

- 3 modes of interpretation
    - Predict
    - Summarize
    - Oddities / outliers
- Intuition
    - Springs physically want to minimize tension
    - Minimize sum of squared errors: write down this equation

- Notation
    - What do i’s mean
    - Residuals (epsilon ε)
    - Prediction (ŷ)
- Math
    - Slope can be expressed in terms of covariance, correlation, std dev (write down these equations)
- Input types
    - x’s can be numeric or binary (dummy) variables; how to interpret?
- Binary output
    - Sigmoid / logit functions: write down these equations
    - Converting among frequencies, log odds, probabilities, and odds
        - Probabilities must be between 0 and 1
        - Frequencies (ie counts of events) can be any non-negative number
        - To get probability from frequency, divide the frequency of the event by the sum of the frequencies of *all* events (including the one you want the probability of)
        - To get odds ratio from probability, divide the probability of the event by the probability of all other events (NOT including the one you want the probability of)
        - To get odds ratio from frequency, do the same thing
        - Odds = $e^{(\text{log odds})}$
        - Probability = σ(log odds)
    - Logistic regression
        - When do we use it? Binary outputs
        - Interpreting logistic regression output
            - Predict, summarize, oddities/outliers

Transformations

- If transformation is not linear (e.g. arithmetic, quadratic) we can still run linear regressions. (Why?)
- You can often use transforms to get better linear fit (e.g. use logarithm to smoosh a variable)
- Residual plots
    - X- and Y-axis definitions
        - X axis represents prediction ŷ
        - Y axis represents *difference* between real y and prediction ŷ
        - We can do this for multiple inputs, which would be hard to visualize otherwise
    - When is a residual plot good or bad with regard to randomness?
- What is heteroskedasticity?
    - How to identify heteroskedasticity in data
    - When does it occur in the wild?
    - How do you resolve heteroskedasticity?

- Logarithms
    - Practice log facts: log(ab) = log(a) + log(b), log(a^b) = b log(a), log(1/a) = -log(a) and related exponentiation facts
    - Why might you do log transforms? To smoosh values, for heteroskedasticity, etc.
- How to interpret regression output when x and/or y are logged (summarize, interpret, oddities/outliers)
    - How to rederive the summary interpretations (see table below)

Regression on multiple variables

- Notation & intuition
    - Regression with multiple inputs is hard because we need to account for the fact that inputs can be correlated
- Choosing what x's to include/exclude
    - Exclude x's with:
        - Collinearity
            - Find out using correlation matrix
        - Overfitting
            - Does your model generalize?
    - Update x's using transformations, or include more x's:
        - If there's lack of randomness in residual plots
- Dummy variables
    - What kind of variable data type are they generated from
    - Why do we drop one dummy variable in regressions
    - How to interpret dummy variables
- Summarize, predict, oddities/outliers

Model	**Regression Interpretation**
Linear $y = \alpha + \beta x$	If $x=0$, $y = \alpha$ 1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \ln(x)$	If $x=1$, $y = \alpha$ If x is multiplied by *e*, we expect a β unit change in y 1% change in x is associated with a $0.01*\beta$ unit change in y
Log-linear $\ln(y) = \alpha + \beta x$	If $x=0$, $y = e^{\alpha}$ For a 1 unit change in x, we expect y to be multiplied by e^{β} 1 unit change in x is associated with a $100*(\exp(\beta)-1)\%$ change in y
Log-log $\ln(y) = \alpha + \beta \ln(x)$	If $x=1$, $y = e^{\alpha}$ If x is multiplied by *e*, we expect y to be multiplied by e^{β} 1% change in x is associated with a $\beta\%$ change in y (*elasticity*)
Logistic $y \sim \sigma(\alpha + \beta x)$ (y must be binary)	The probability that $x=0$ yields output $y=1$ is $e^{\alpha}/(e^{\alpha}+1)$ For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β} 1 unit change in x is associated with a $100*(e^{\beta} – 1)\%$ change in y
In the above models where **x is binary**	You wouldn’t need to take the ln() of a binary variable. For the remaining models, interpret “a 1 unit change” as “going from x=0 to x=1”