STA130 Winter 2020

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Github.com/vivianngo97/STA130-Winter-2020

viv.ngo@mail.utoronto.ca

Agenda

- Reminders
- Vocabulary
- Group work & presentations
- Time to work on poster project
- Ticket out the door

Reminders

- Poster fair is in two weeks
  - Office hours
  - Piazza up until Friday, March 27th – TAs and profs will NOT be answering questions the weekend before the poster fair.
  - You should be meeting outside of tutorial/class to work on it. The weekend before the fair should be reserved for practicing your presentation.
  - You should already have the project completed and knitted by Friday, March 27th.
- SSU career panel

Extra: SSU Career Panel

Extra: SSU Career Panel

- Thinking about a career with a statistics degree? Debating between graduate school or finding a full-time job after graduation? Join us on our annual career panel, a networking event that is a great opportunity to network with 8 professionals from various industries and have a heads-up of what to expect in the workplace.
- Our panelists will share:
  - Their personal career paths
  - The realities of sectors they work in
  - How their university education is relevant to their particular position
- Attending Panelists: https://docs.google.com/document/d/1tdzjUVoJ8s9Aqc42TP7gwQRyDWeF0wN9xJ2A-uEihes/edit?usp=sharing
- Date: Saturday, March 14th, 202
- Time: 1:00PM - 5:00P
- Location: SS Room 1069 & SS Room 1071
- ***Registration required, space is limited. Sign up here: https://docs.google.com/forms/d/1VF5_Yd9WcyEFRRmCD-HGXAHZQHAmLs9G2GOvCCp0r-A/edit?usp=sharing

- Feel free to reach out to uoftssu@gmail.com if you have any questions!

Vocabulary

- Linear Relationship
- Approximately linear
- Non-linear
- Slope
- Intercept
- (Simple) Linear Regression
- Regression model
- Parameter
- Regression coefficients
- Fitted regression line
- Explanatory/Independent variable
- Dependent variable
- Measure of model fit
- Coefficient of determination
- Root mean square error
- Error
- Residual
- Least squares
- Least squares estimator

Vocabulary

- QUESTION: What is the correlation coefficient (r)?

Vocabulary

- **QUESTION:** What is the correlation coefficient (r)?
- ***Answer:***
  - *1= perfect correlation; 0 = no correlation*
  - *Positive values = positive correlation; e.g. as X increases, Y increases*
  - *Negative values = negative correlation (i.e., inverse association); e.g. as X increases, Y decreases or vice versa*

Vocabulary

- QUESTION: what is the standard linear regression equation? Define each part.

Vocabulary

- **QUESTION:** what is the standard linear regression equation? Define each part.
- ***Answer:***
  - $Y = B_o + B_1 X_i + e_i$
  - *Where y = linear outcome, B_o = intercept, B_1 = regression coefficient, X_i = explanatory variable, e_i = error, and I = number of individuals in the sample*
  - ***A hat symbol signifies an estimated value**.*
- *Example: Interested in brain weight (grams) as our dependent variable and head size (cm^3) as our independent variable.*
  - $\hat{Y} = 325.57342 + 0.26343 X_i$

Vocabulary

- **QUESTION:** What are some measures of correlation you've learned so far? What do they measure?

Vocabulary

- **QUESTION:** What are some measures of correlation you've learned so far? What do they measure?
- ***Answer:***
- ***R^2****: relative measure of fit, ranging from 0 to 1. E.g. $R^2 = 0.80$ means that 80% of the variation in your outcome can be explained by your model.*
- ***RMSE:*** *an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit.*

Vocabulary

- QUESTION: What are the assumptions of linear regression?

Vocabulary

- **QUESTION: What are the assumptions of linear regression?**
- ***Answer:***
- *There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.*
- *Multivariate Normality–Multiple regression assumes that the residuals are normally distributed. (More on this next week)*
- *No (or little) Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values. (more on this next week)*
- *Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.*

Vocabulary

- **QUESTION:** How do classification tress differ from linear regression? In which circumstances might you use one vs the other?
- ***Answer:***
- *Most notably, classification trees are limited to binary variables (or dichotomizing categorical or continuous variables). SLR requires a continuous outcome with some sort of linear relationship.*

Vocabulary

- **APPLIED QUESTION:**
- During the years 1790 to 1820, the correlation between the number of churches built in New England and the barrels of Rum imported into the region was a perfect 1.0. What does this tell you - that church building causes rum drinking, that rum drinking causes church building, or something else? If something else, what?
- ***Answer:***
- *Probably something else happening here since neither would make sense. Could just be that both are associated with some other third variable; e.g. increased population. We'll talk about confounders in the last week of class.*

Oral Presentations

- **In poster project groups**
- **Remember:**
- **THE 4 C'S:** Calm; Confident; Clear; Concise

- ***Tips for giving a great oral presentation: Content***
  - What is the main message you want to get across?
  - Create an (organized) outline of your presentation
  - Define terms early
  - Make clear transitions between parts of your presentation
  - Make your data/ figures meaningful
  - Summarize

- ***Tips for giving a great oral presentation: Delivery***
  - Be confident, make eye contact and avoid reading
  - Avoid filler words – "ummm", "like", "you know"
  - Speak slowly and it's ok to pause (and breathe!)
  - Remember to enunciate all the parts of each word
  - Practice! Practice! Practice!

Oral presentations

- **ACTIVITY 1: Based on questions 1a-b**
  - Describe your plot produced in question 1a. Make sure to note the x- and y-axis and to describe the association you observe, if any. E.g. the association linear, positive, negative, strong, weak, etc.?
  - Does there seem to be an association between log number of transistors and year for GPUs? For CPUs? Does this make sense based on your prior expectations?
  - Do there appear to be many outliers? Why might this matter?

- **ACTIVITY 2: Questions 1c-f**
  - Why did you plot the log number of transistors instead of the unmodified count?
  - Is the association between log transistor count and year stronger for CPUs or GPUs? Explain. Does this make sense based on your prior expectations?

- **ACTIVITY 3: Question 1g-i**
  - Provide a simple linear regression equation for the association between log transition count and year – separately for CPUs and GPUs. Explain what each part of the model means in lay terms.
  - What is the model equation and estimated values? What is the coefficient of determination? Explain what these values mean and an interpretation in lay terms.

- **ACTIVITY 4: Question 1j-k**
  - Briefly explain why or why not the interpretation of the intercept is helpful for understanding Moore's Law.
  - How well does your model perform as a predictive model?
  - Are there any other variables you think may be important factors influencing this association?

- **ACTIVITY 5**
  - Under what conditions would you use correlation and/or regression analysis?
  - Include comments on the type of data needed and a suggestion for their use.
- You can refer to this week's practice problems for examples.

Ticket out the door

- Write down two questions you may have regarding the poster project
- Hand these in anonymously before leaving