STA130 Fall 2019 – T0107

Week 9: Simple Linear Regression

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

This presentation was prepared by Vivian Ngo.)

Github.com/viviannngo97/STA130-Fall-2019

viv.ngo@mail.utoronto.ca

Agenda

- **Material & Vocabulary Review**
- **Discussion Questions**
- **Group work and presentations**
- **Time to work on poster project**

Material and Vocabulary Review

- 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
- Prediction error
- Least squares
- Least squares estimator

Discussion questions

- **1**
- **What is the correlation coefficient (r)?**
  - 1= ???
  - 0 = ???
  - Positive values = ???
  - Negative values = ???

Discussion questions

- **1**
- **What is the correlation coefficient (r)?**
  - 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

Discussion questions

- **2**
- **What is the standard linear regression equation? Define each part.**
  - **$Y= B_o + B_1 X_i + e_i$**

Discussion questions

- **2**
- **What is the standard linear regression equation? Define each part.**
  - **$Y= B_o + B_1 X_i + e_i$**
- 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
- Hat symbol signifies an **estimated** value.

Discussion questions

- **2**
- **What is the standard linear regression equation? Define each part.**
  - **$Y= B_o + B_1 X_i + e_i$**
- 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
- Hat symbol signifies an **estimated** value.
- E.g. Interested in brain weight (grams) as our dependent variable and head size (cm^3) as our independent variable.
- $\hat{Y} = 325.57342 + 0.26343 X_i$

Discussion questions

- **2**
- **What is the standard linear regression equation? Define each part.**
  - **$Y= B_o + B_1 X_i + e_i$**
- 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
- Hat symbol signifies an **estimated** value.
- E.g. Interested in brain weight (grams) as our dependent variable and head size (cm^3) as our independent variable.
- $\hat{Y} = 325.57342 + 0.26343 X_i$
- This means that when head size is 0 cubic cm, the average brain weight is 325.57 grams. The intercept in this case is not very informative since the interpretation does not make sense. Humans can't have a head size of 0 cubic cm! The estimated slope is 0.26. This means that if an individual's head is one cubic centimeter larger, the average brain weight will be 0.26 grams heavier, on average.

Discussion questions

- **3**
- **QUESTION: What are some measures of correlation you’ve learned so far? What do they measure?**

Discussion questions

- **3**
- **What are some measures of correlation you’ve learned so far? What do they measure?**
- **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.

Discussion questions

- **4**
- **What are the assumptions of linear regression?**

Discussion questions

- **4**
- **What are the assumptions of linear regression?**
- 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.

Discussion questions

- **5**
- **How do classification tress differ from linear regression? In which circumstances might you use one vs the other?**

Discussion questions

- **5**
- **How do classification tress differ from linear regression? In which circumstances might you use one vs the other?**
- Most notably, classification trees are limited to binary variables (or dichotomizing categorical or continuous variables). SLR requires you have a continuous outcome with some sort of linear relationship.

Oral presentations

- **ACTIVITY 1: Based on questions 1a-c**
  - 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.?
  - What is the correlation between carbohydrates and calories weight? Make sure to explain how you calculated this value and what it means; i.e., provide an interpretation of the value.
  - Does this make sense based on your prior expectations? Are there any other variables you think may be important factors influencing the calories in a Starbucks food item?
  - Do there appear to be many outliers? Why might this matter?
- **ACTIVITY 2: Questions 1d-f**
  - Provide a simple linear regression equation for the association between calories and carbohydrates. Explain what each part of the model means in lay terms.
  - Based on your answer to part e, report the estimated values of your model and provide an interpretation of these values.
  - How well does your model fit the data? Explain what the coefficient of determination means and provide an interpretation.
- **ACTIVITY 3: Question 2e**
  - Present your regression model of miles per gallon (mpg) on acceleration based on the training set.
  - 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.
  - How well does your model perform as a predictive model?
- **ACTIVITY 4: Question 2f**
  - What is the predicted fuel efficiency (in mpg) for a hybrid vehicle that can accelerate from 0 to 60 m/h in 10 seconds.? Make sure to present your regression equation, including all coefficients.
  - From the plot produced earlier in this question we can see that there was one vehicle with an acceleration time of 10 seconds in the sample. The actual fuel efficiency of this vehicle was 21 mpg. What is the residual?
  - Provide an interpretation in lay terms. Is this a large difference? Based on previous work done in this question, why do you think this may be the case? Hint: Think about how well the model fits the data, if there may be other important factors, etc.