Dr Straw's D208 on One Page

Data

Download the data file from D208. Do not use your data file from D207 or D206.

Essential Learning Resources

- If you prefer videos: D208 videos (The links are in the Welcome lesson in the Python section)
    - Linear regression – lessons 1, 2, 3 and 7, plus Python (lessons 6 & 10) or R (lessons 5 & 9)
    - Logistic regression – lessons 1, 2, and 3, plus Python (lesson 6) or R (lesson 5)
- If you prefer reading: D208 lessons (Both tracks should read the Welcome lesson in the Python section)

Python lessons	R lessons	Description
1.2 through 1.3	1.6 through 1.9	Getting started
2.1 through 2.4	2.3 through 2.5	Foundations of linear regression
3.1 through 3.4	3.2 through 3.3, and 3.4 (last three sub-sections)	Multiple linear regression
4.1 (first sub-section), 4.2, and 4.5	4.1 (first sub-section), 4.2, and 4.5	Logistic regression
6.1 through 6.2	6.1 through 6.2	Model verification

- Additional help: D207 textbook, *Practical statistics for data scientist*, Includes Python and R.
    - Chapter 4 (*Introduction* through *Regression Diagnostics*) for Multiple linear regression
    - Chapter 5 (*Introduction*, *Logistic Regression* and *Evaluating Classification Models*).

Vocabulary

Response attribute: Attribute selected as Y in a regression equation [aka response variable, dependent variable, effect].

Predictor attributes: Attributes selected as $X_1, X_2, \ldots, X_j$ in a regression equation [aka predictor variables, independent variables, explanatory variables, cause].

Continuous: A type of data that has an infinite number of values between any two values (e.g. temperature) [aka numerical, quantitative, physical measurement]. Includes interval (different by distance with an arbitrary zero, e.g. Fahrenheit and Celsius) and ratio (different by distance with a true zero, e.g. Kelvin).

Categorical: A type of data that has a finite number of distinct groups (e.g. yes/no, BS/MS/PhD) [aka factor]. Includes nominal (different by name, e.g. yes/no) and ordinal (different by name and order, e.g. BS/MS/PhD). Categories can be represented by dummy attributes (e.g. yes = 1 and no = 0).

Multiple Linear Regression (Task 1) vs Logistic Regression (Task 2)

Multiple Linear Regression is prediction through the slope of a line

1. Can be visualized as a straight line.
2. Requires a continuous response attribute. (Review your D207 Task 1 submission)
3. Creates a linear regression model that can be used to calculate the value of the response attribute Y when you know the predictor attributes [$Y= b_0 + b_1X_1+b_2X_2 + \ldots + b_jX_j$]. A single unit increase in X_1, e.g. from 15 to 16 or from 32 to 33, causes the same increase in Y regardless of where the single unit increase occurs on the line. This holds true for all predictor attributes.

Logistic Regression is prediction through classification

1. Can be visualized as an S-curved line (like an S-curve in a road).
2. Requires a categorical response attribute. (Review your D207 Task 1 submission)
3. Creates a probability output that ranges from 0 to 1 and is the probability that the response attribute Y will fit into a specific category when you know the predictor attributes [$Logit(Y=PhD) = b_0 + b_1X_1+b_2X_2+\ldots+ b_jX_j$] A single unit increase in X_1, e.g. from 15 to 16 or from 32 to 33, does not necessarily cause the same increase in Y. Instead, the increase in Y is dependent on where the single unit increase is located on the S-curved line. This holds true for all predictor attributes.

Eric Straw, PhD (eric.straw@wgu.edu; 877-435-7948, extension 2685) 2021.5.26