Data Science

Linear Regression

CS202 11.06.22

Linear Regression

Preparing Variables

There are a lot of conditions that need to be met to optimize linear regression. Remember that a liner regression model best fits a ‘line’ given many data points. This means that all our variables must contain data as integers or floats.

Unfortunately, not all data is numerical.

For example we have a lot of data/variables that are non-numerical. This is what we call: categorical data.

Linear Regression

Categorical data

Categorical variables represent types of data which may be divided into groups.

Examples of categorical variables are:

race, sex, age group, educational level.

While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.

Linear Regression

Categorical data

There are two types of categorical data, namely; the nominal and ordinal data:

1. **Nominal Data**

Nominal data is sometimes called “labelled” or “named” data. Examples of nominal data include name, hair colour, sex etc.

Mostly collected using surveys or questionnaires, this data type is descriptive, as it sometimes allows respondents the freedom to type in responses. Although this characteristic helps in arriving at better conclusions, it sometimes poses problems for researchers as they have to deal with so much irrelevant data.

Linear Regression

Categorical data

Second type is called:

2. **Ordinal Data**

This is a data type with a set order or scale to it. However, this order does not have a standard scale on which the difference in variables in each scale is measured.

Although mostly classified as categorical data, it is said to exhibit both categorical and numerical data characteristics making it in between. Its classification under categorical data has to do with the fact that it exhibits more categorical data character.

Some ordinal data examples include; Likert scale, interval scale, bug severity, customer satisfaction survey data etc. Each of these examples may have different collection and analysis techniques, but they are all ordinal data.

Categorical Data

General characteristics and features

- **Categories**: These consist of two categories of categorical data, namely; nominal data and ordinal data. Nominal data, also known as named data is the type of data used to name variables, while ordinal data is a type of data with a scale or order to it.

- **Qualitativeness**: Categorical data is qualitative. That is, it describes an event using a string of words rather than numbers.

- **Analysis**: Categorical data is analysed using mode and median distributions, where nominal data is analysed with mode while ordinal data uses both. In some cases, ordinal data may also be analysed using univariate statistics, bivariate statistics, regression applications, linear trends and classification methods.

Categorical Data

General characteristics and features

- **Graphical Analysis**: It can also be analysed graphically using a bar chart and pie chart. A bar chart is mostly used to analyse frequency while a pie chart analysis percentage. This is done after grouping it into a table.

- **Interval Scale**: In the case of ordinal data, which has a given order or scale, the scale does not have a standardised interval. This is not applicable for nominal data.

- **Numeric Values**: Although categorical data is qualitative, it may sometimes take numerical values. However, these values do not exhibit quantitative characteristics. Arithmetic operations can not be performed on them.

- **Nature**: Categorical data may also be classified into binary and non-binary depending on its nature. A given question with options “Yes” or “No” is classified as binary because it has two options while adding “Maybe” to the given options will make it non-binary.

Linear Regression

Cateogorical data

Analysis of categorical data generally involves the use of data tables. A ***two-way table*** presents categorical data by counting the number of observations that fall into each group for two variables, one divided into rows and the other divided into columns.

For example, suppose a survey was conducted of a group of 20 individuals, who were asked to identify their hair and eye color. A two-way table presenting the results might appear as follows:

	Eye Color				
Hair Color	Blue	Green	Brown	Black	Total
Blonde	2	1	2	1	6
Red	1	1	2	0	4
Brown	1	0	4	2	7
Black	1	0	2	0	3
Total	5	2	10	3	20

Linear Regression

Categorical data

The totals for each category, also known as ***marginal distributions***, provide the number of individuals in each row or column without accounting for the effect of the other variable (in the example above, the total number of individuals with blue eyes, regardless of hair color, is 5).

Since simple counts are often difficult to analyze, two-way tables are often converted into percentages. In the above example, there are 4 individuals with red hair. Since there were a total of 20 observations, this means that 20% of the individuals survered are redheads. One also might want to investigate the percentages within a given category -- of the 4 redheads, 2 (50%) have brown eyes, 1 (25%) has blue eyes, and 1 (25%) has green eyes.

Linear Regression

Categorical data

Here is a very simple example:

Suppose we have a variable which is “day of week”: Sun, Mon, Tue, Wed, Thur, Fri, Sat

To feed this into our machine learning model we can present numerical values for the day of week. i.e. Sun = 0, Mon = 1, Tues = 2, Wed = 3, Thurs = 4, Fri = 5, Sat = 6. Then our linear regression model will take in values from 0-6 to that it can represent that as a point in your space of dimension n (where n is the number of variables)

Linear Regression

Colinearity

Collinearity, in statistics, correlation between predictor variables (or independent variables), such that they express a linear relationship in a regression model. When predictor variables in the same regression model are correlated, they cannot independently predict the value of the dependent variable. In other words, they explain some of the same variance in the dependent variable, which in turn reduces their statistical significance.

Linear Regression

Colinearity

Collinearity becomes a concern in regression analysis when there is a high correlation or an association between two potential predictor variables, when there is a dramatic increase in the *p* value (i.e., reduction in the significance level) of one predictor variable when another predictor is included in the regression model, or when a high variance inflation factor is determined. The variance inflation factor provides a measure of the degree of collinearity, such that a variance inflation factor of 1 or 2 shows essentially no collinearity and a measure of 20 or higher shows extreme collinearity.

Linear Regression

Multicolinearity

Multicollinearity describes a situation in which more than two predictor variables are associated so that, when all are included in the model, a decrease in statistical significance is observed. Similar to the diagnosis for collinearity, multicollinearity can be assessed using variance inflation factors with the same guide that values greater than 10 suggest a high degree of multicollinearity. Unlike the diagnosis for collinearity, however, it may not be possible to predict multicollinearity before observing its effects on the multiple regression model, because any two of the predictor variables may have only a low degree of correlation or association.

Linear Regression

Colinearity

We have many variables that go into our linear regression.

We want to minimize those number of variables especially if one is highly correlated with the other.

For example: going back to predicting men's weight using other variables, suppose we have two variables: "#calories intake per day" and another variable "Junk food consumption" these two variables may determine one another. Someone who has a large calorie intake has a large junk food intake. Then you may want to discard one of the variables.