Introduction to Machine Learning Applications

Spring 2023

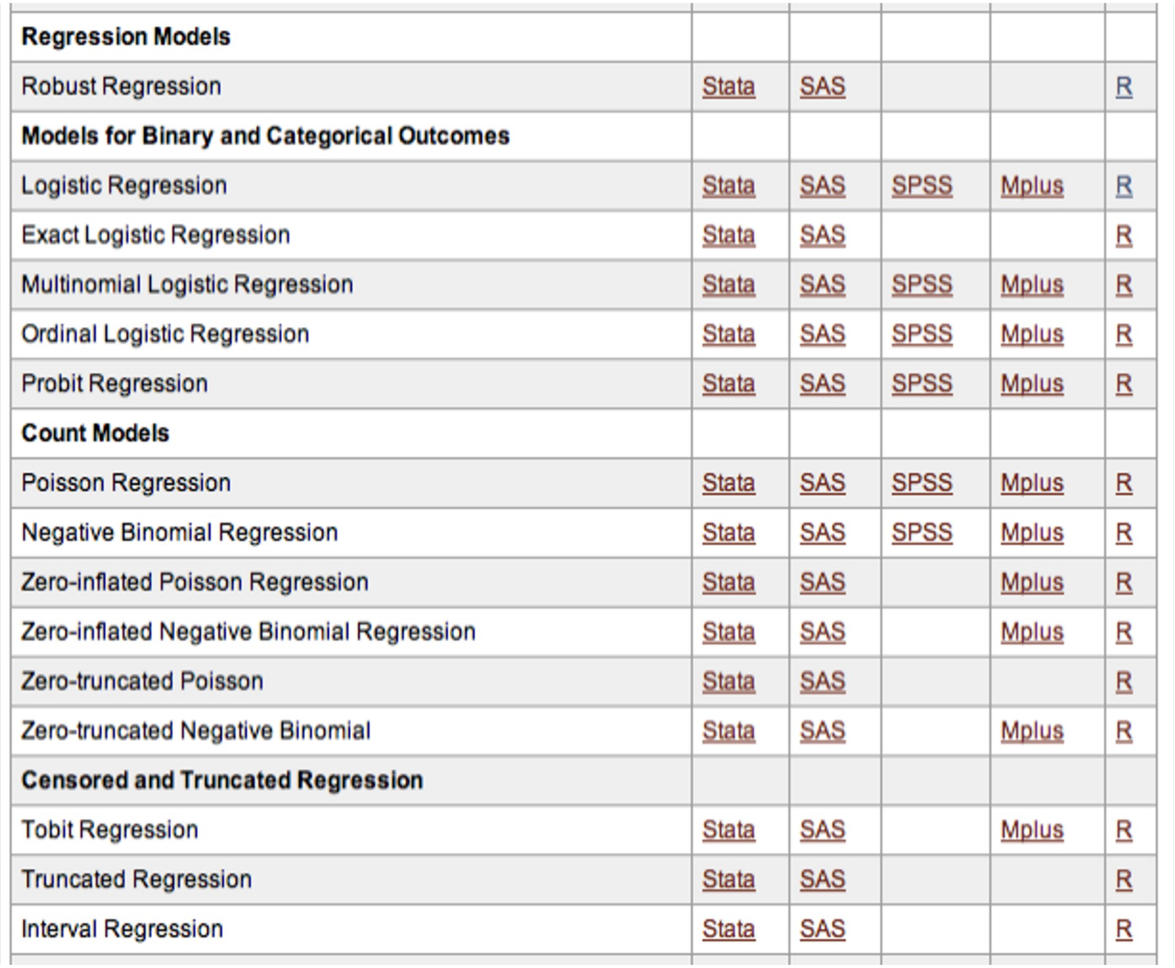
Regression

Minor Gordon

gordom6@rpi.edu

RENSSELAER POLYTECHNIC INSTITUTE
• 1824 •
Rensselaer

Distributions and Regression



Regression Models					
Robust Regression	Stata	SAS			R
Models for Binary and Categorical Outcomes					
Logistic Regression	Stata	SAS	SPSS	Mplus	R
Exact Logistic Regression	Stata	SAS			R
Multinomial Logistic Regression	Stata	SAS	SPSS	Mplus	R
Ordinal Logistic Regression	Stata	SAS	SPSS	Mplus	R
Probit Regression	Stata	SAS	SPSS	Mplus	R
Count Models					
Poisson Regression	Stata	SAS	SPSS	Mplus	R
Negative Binomial Regression	Stata	SAS	SPSS	Mplus	R
Zero-inflated Poisson Regression	Stata	SAS		Mplus	R
Zero-inflated Negative Binomial Regression	Stata	SAS		Mplus	R
Zero-truncated Poisson	Stata	SAS			R
Zero-truncated Negative Binomial	Stata	SAS		Mplus	R
Censored and Truncated Regression					
Tobit Regression	Stata	SAS		Mplus	R
Truncated Regression	Stata	SAS			R
Interval Regression	Stata	SAS			R

Different regression models for different dependent variable distributions

https://stats.idre.ucla.edu/other/dae/

Regression, Different DVs

- Normally distributed DV [Regression]
- Binary outcome – [Logistic Regression]
  - Like Titanic
- Count model – [Poisson Regression]
  - Number of likes on a Facebook post
- Count model, with lots of 0s [Zero-inflated Poisson Regression]
  - Number of shares on a Facebook post

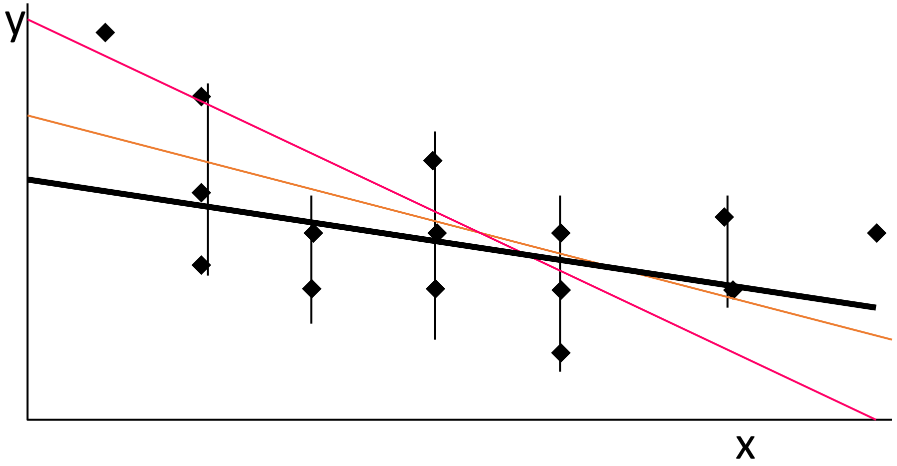
Linear Regression

The technique is used to **predict** the value of one variable (the dependent variable - y) **based on** the value of other variables (independent variables x_1, x_2,…x_k) where ε is the error.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Estimating the coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



The question is:
Which straight line fits best?

Logistic Regression

- Special case of linear regression where the target variable is categorical in nature
- Uses a log of odds as a dependent variable
- Predicts the probability of occurrence of an event using a sigmoid function (inverse of logit function)

$$p = 1/(1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)})$$

Linear vs Logistic Regression

- Output for linear regression is continuous
    - For example, stock prices
    - Or real estate price estimation
- Output for logistic regression is estimated as a constant
    - For example, predicting if a sample is tested +ve or –ve
    - Output >0.5 is +ve or 1 or yes; output <=0.5 is –ve or 0 or no

Linear vs Logistic Regression

- Linear regression is estimated using ordinary least squares
    - Distance minimizing approximation approach
    - Fits a regression line on a given set of data points that has the minimum sum of squared deviations (least squared error)

- Logistic regression is estimated using maximum likelihood estimation
    - “Likelihood” maximization method
    - Determines parameters (such as mean/variance) that are most likely to produce the set of data points.

Avoiding False Discoveries

- An algorithm applied to a set of data will usually produce some result(s)
    - There have been claims that the results reported in more than 50% of published papers are false. (Ioannidis)
- Results may be a result of random variation
    - Any particular data set is a finite sample from a larger population
    - Often significant variation among instances in a data set or heterogeneity in the population
    - Unusual events or coincidences do happen, especially when looking at lots of events
    - For this and other reasons, results may not replicate, i.e., generalize to other samples of data
- Results may not have domain significance
    - Finding a difference that makes no difference
- Data scientists need to help ensure that results of data analysis are not false discoveries, i.e., not meaningful or reproducible

Spurious correlations