Data Analytics Course

Logistic regression(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- Simple correlation and regression analysis
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

- Need of logistic regression?
- The logistic regression model
- Meaning of beta
- Goodness of fit
- Multicollinearity
- Prediction
- Stepwise regression

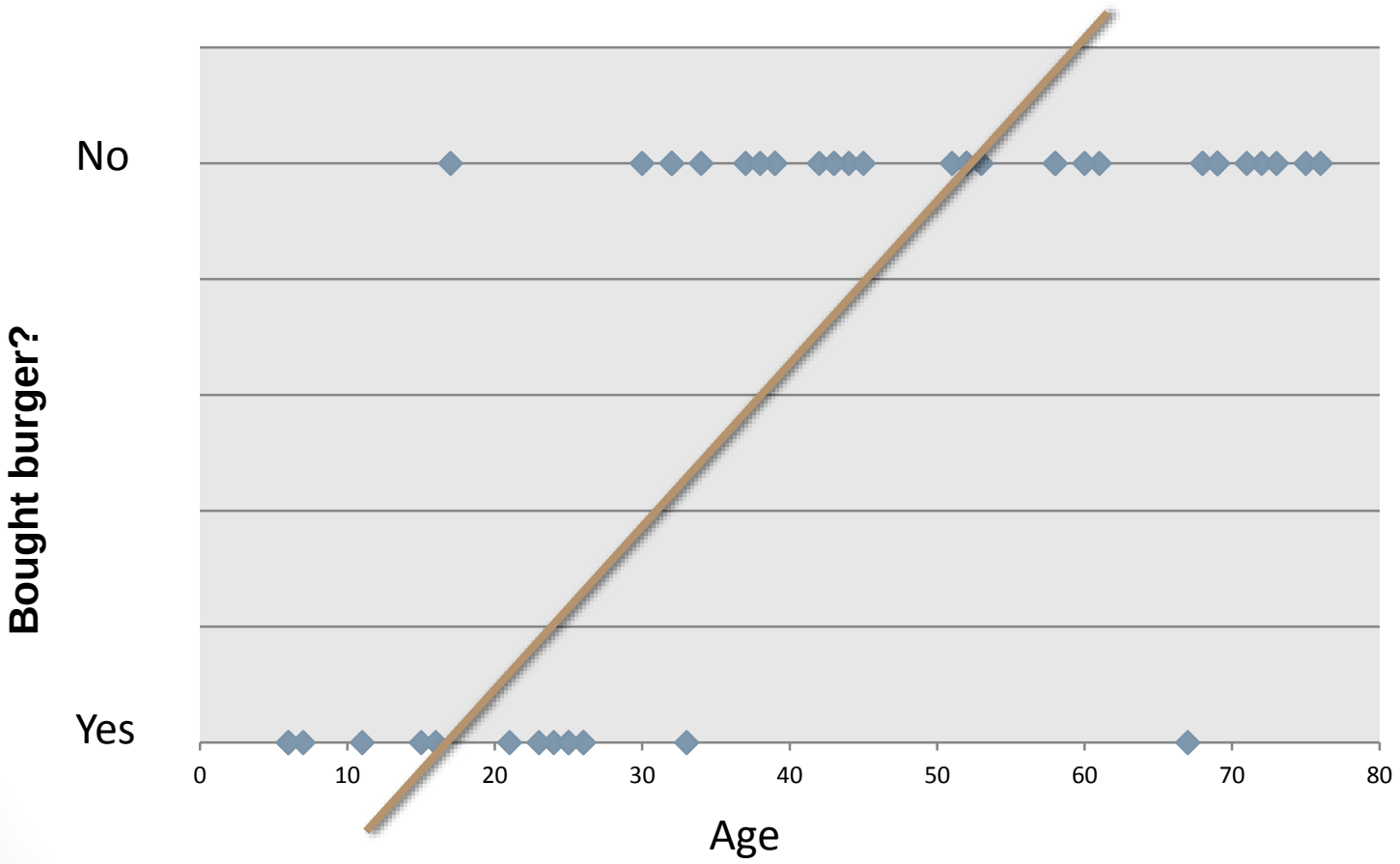
What is the need of logistic regression?

- Remember the burger example? Number of burgers sold vs number of visitors?
- What if we are trying to find whether a person is going to buy a burger or not, based on person's age.
- Download the data from here & fit a linear regression line.
- What is R squared
- If age increases, what happens to burger sales?
- 25 years old person, does he buy a burger?
- Can we fit a liner regression line to this data?

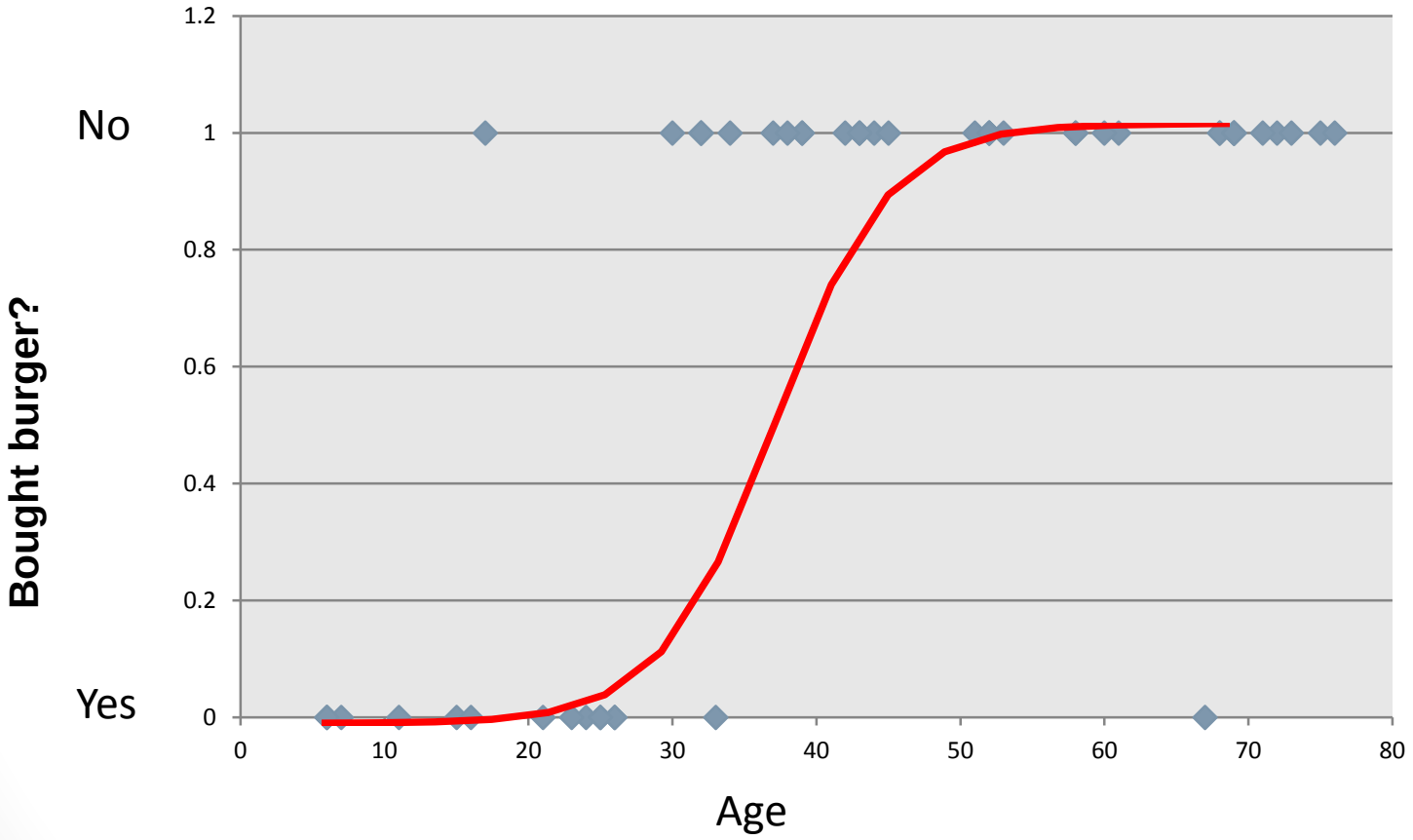
Real-life Challenges

- Gaming - Win vs Loss
- Sales - Buying vs Not buying
- Marketing – Response vs No Response
- Credit card & Loans – Default vs Non Default
- Operations – Attrition vs Retention
- Websites – Click vs No click
- Fraud identification –Fraud vs Non Frau
- Healthcare –Cure vs No Cure

Why not liner ?

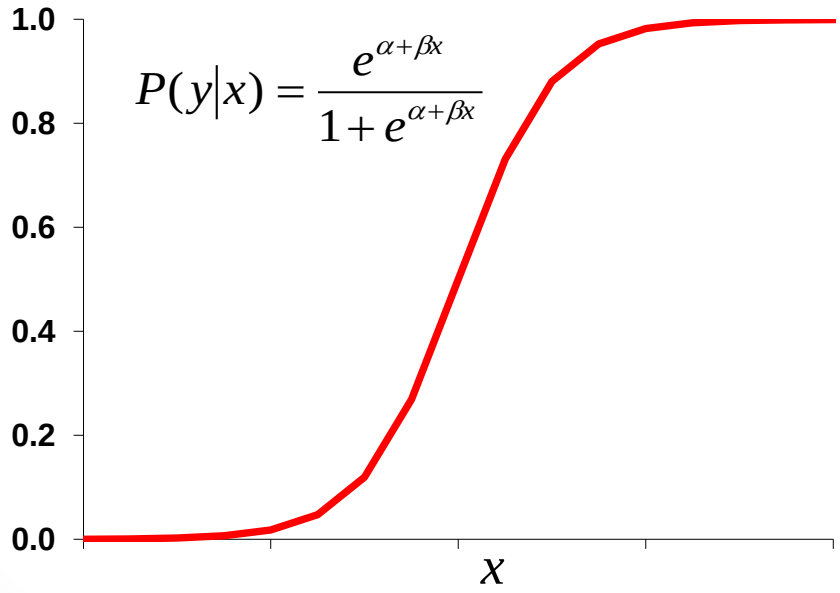


Any better line?



Logistic function

- We want a model that predicts probabilities between 0 and 1, that is, S-shaped.
- There are lots of S-shaped curves. We use the logistic model:
- Probability = $1/[1 + \exp(\beta_0 + \beta_1 X)]$ or $\log_e[P/(1-P)] = \beta_0 + \beta_1 X$
- The function on left, $\log_e[P/(1-P)]$, is called the logistic function.



Logistic regression function

Logistic regression models the logit of the outcome

=Natural logarithm of the odds of the outcome

=ln(Probability of the outcome (p)/Probability of not having the outcome (1-p))

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

β = log odds ratio associated with predictors

e^{β} = odds ratio

Curve fitting using MLE

- Remember OLS for linear models?
- Maximum Likelihood Estimator:
  - starts with arbitrary values of the regression coefficients and constructs an initial model for predicting the observed data.
  - Then evaluates errors in such prediction and changes the regression coefficients so as make the likelihood of the observed data greater under the new model.
  - Repeats until the model converges, meaning the differences between the newest model and the previous model are trivial.
- The idea is that you “find and report as statistics” the parameters that are most likely to have produced your data.

Meaning of beta

- The betas themselves are **log-odds ratios**. Negative values indicate a negative relationship between the probability of "success" and the independent variable; positive values indicate a positive relationship.
- Increase in log-odds for a one unit increase in x_i with all the other x_i s constant
- Measures association between x_i and log-odds adjusted for all other x_i

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

Goodness of fit for a logistic regression

Chi-Square

- The Chi-Square statistic and associated p-value (Sig.) tests whether the model coefficients as a group equal zero.
- Larger Chi-squares and smaller p-values indicate greater confidence in rejected the null hypothesis of no impact

Percent Correct Predictions

- The "Percent Correct Predictions" statistic assumes that if the estimated p is greater than or equal to .5 then the event is expected to occur and not occur otherwise.
- By assigning these probabilities 0s and 1s and comparing these to the actual 0s and 1s, the % correct Yes, % correct No, and overall % correct scores are calculated.
- **Note:** subgroups for the % correctly predicted is also important, especially if most of the data are 0s or 1s

Goodness of fit for a logistic regression

- **Hosmer and Lemeshow Goodness-of-Fit Test**
  - Chisquare test for Observed and expected bad by diving the variable into groups
  - The test assesses whether or not the observed event rates match expected event rates in subgroups of the model population.
  - The Hosmer–Lemeshow test specifically identifies subgroups as the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar are called well calibrated.
- **Other methods**
  - ROC curves – How to interpret ROC curve?
  - Somers' D
  - Gamma
  - Tau-a
  - C
  - More than a dozen “R2”-type summaries

Lab-Logistic Regression

- Fit a logistic regression line for burger example
- If age increases, what happens to burger sales?
- 25 years old person, does he buy a burger? What is the probability (or what are the odds of him buying the burger)
- How good is the regression line for burger example
- If age difference is 20 years, what is the difference in odds
- Write the below code to see

```
Proc logistic data=burger2
PLOTS(ONLY)=(ROC(ID=prob) EFFECT) descending;
model buy=Age / lackfit;
run;
```

Lab: Logistic regression

- Build a logistic regression line on credit card data(cred.training)
- Problem with monthly income, number of open cred limits, number of dependents? Drop them & build for the rest of variables
- How good is the fit?
- What is the impact of each variable
- Is there any interdependency between variables
- Remove 60 plus delinquency & 90 plus delinquency variables & build the model again
- How to identify & use the variables with data related issues? Data validation & cleaning

Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

www.TrendwiseAnalytics.com/venkat

+91 9886 768879