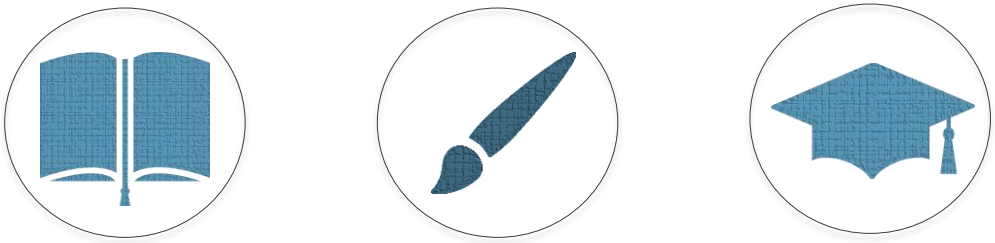


Data Science Concepts

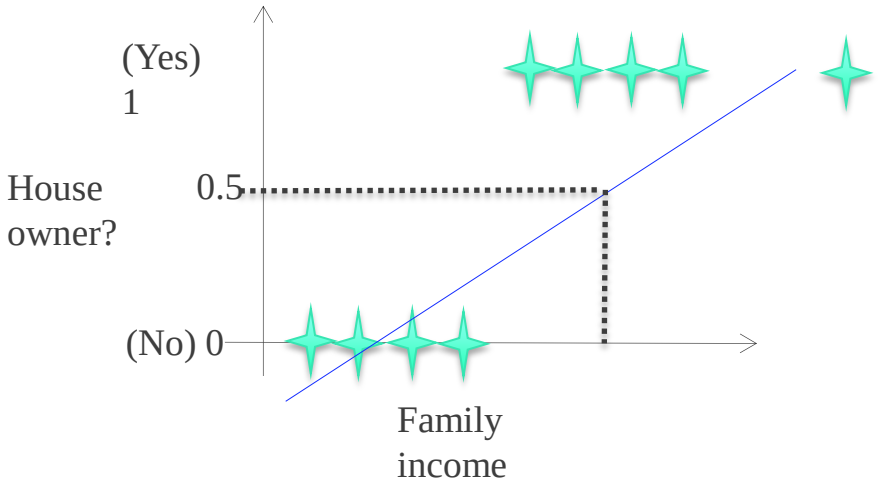
Lesson03–Logistic Regression Concepts

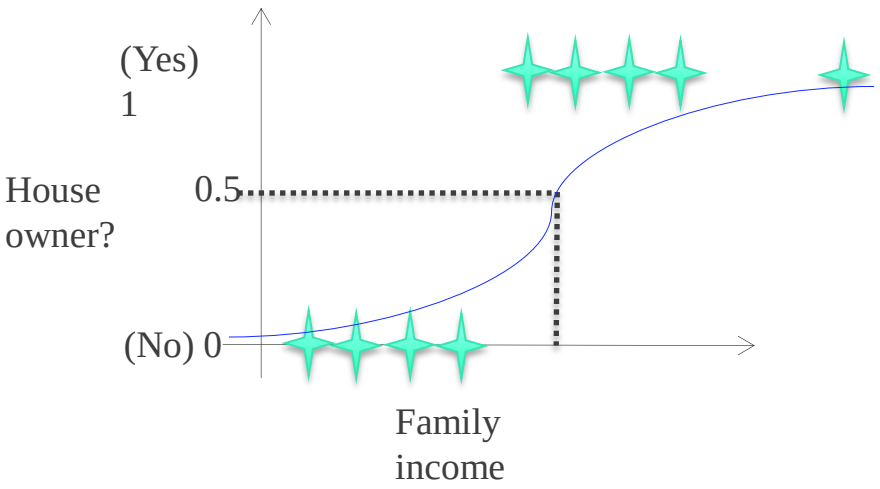
Objective

After completing this lesson you will be able to:

- Explain logistic regression analysis
- Describe the application areas of logistic regression
- Explain the various parameters derived to understand the validity of the model

Why Not Regression?





- Issues with regression to model qualitative response variable (owning a house at certain income level example):
  - Non-normality of error term
  - Heteroscedasticity in error term
  - Dependent variable beyond 0 and 1 values
  - Not logically attractive model.

Regression Vs. Logistic Regression

Regression models	Logistic regression models
Objective is to estimate the expected or mean value given the independent variables.	Objective is to find the probability of an event given the independent variables.

The name, logistic regression, is derived from logistic function. Logistic regression or logit model is such that:

$$0 \leq f(x) \leq 1$$

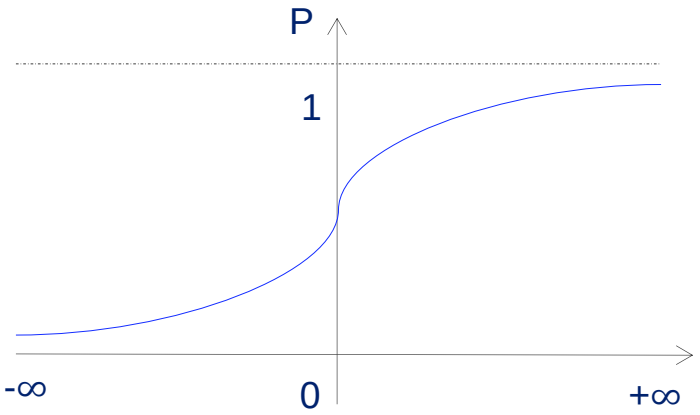
$$Y \in \{0,1\}$$

0: “Negative Class”
1: “Positive Class”

Applications of Logistic Regression

- Logistic regression is one of the most powerful technique to solve classification problem.
  - Email: Spam/Not Spam
  - Online Transaction: Fraudulent/Not Fraudulent (Yes/No)
  - HR Status: Joining/Not Joining
  - Credit Scoring: Defaulter/Non-defaulter

Deriving Logit Model



The logistic distribution function is given by:

$$P_i = P(Y = 1|X_i) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 * X_i)}}$$

Or

$$P_i = P(Y = 1|X_i) = \frac{1}{1 + e^{-Z_i}}$$

where $Z_i = \beta_1 + \beta_2 * X_i$

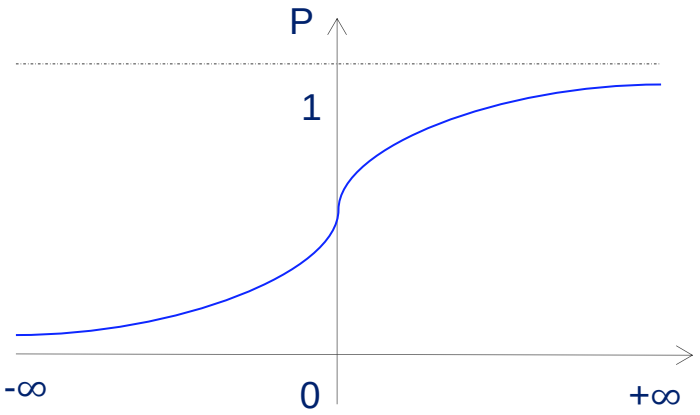
P_i is non-linear in βs. Linear transformation is required:

$$1 - P_i = P(Y = 0|X_i) = \frac{1}{1 + e^{Z_i}}$$

$$P_i/(1 - P_i) = e^{Z_i}$$

$P_i/(1 - P_i)$ is the odds ratio in favor of owning a house. The ratio of the probability that a family will own a house to the probability that it will not own a house.

Deriving Logit Model



The logistic distribution function is given by:

$$L_i = \ln(P_i/(1-P_i)) = Z_i$$

$$L_i = \ln(P_i/(1-P_i)) = \beta_1 + \beta_2 * X_i$$

L, the log of odds ratio is both linear in X and in parameters. This equation is called the logit model.

β_2, the slope, tells how the log odds in favor of say, owning a house change as income changes by a unit. If coefficient sign is positive, probability of owing a house increases. . If coefficient sign is negative, probability of owing a house decreases.

Odds Ratio Interpretation

- An example to illustrate the interpretation of coefficients.

The logistic regression equation is

$$\ln(P_i/(1-P_i)) = -1.59474 + 0.07862 * X_i$$

$$P_i/(1-P_i) = e^{-1.59474} * e^{0.07862*X_i}$$

where P is the probability of owning a house P(Y=1|X)

- $e^{0.07862} = 1.0817$ which means that every unit change in the income, the odds in favor of owning a house increases by 1.0817 or 8.17 %.

Estimation of the Logit Model

Estimation will need value of X and Y.

$$\ln(P_i/(1-P_i)) = \beta_1 + \beta_2 * X_i + u_i$$

Family	Y	X ('000 in $)
1	0	8
2	1	16
3	1	18
4	0	11
5	0	12
6	1	19
7	1	20
8	0	13
9	0	9
10	0	10

OLS will not work and maximum likelihood (ML) technique will be needed for estimating Logit. ML estimate holds good for large sample. Thumb rule >30 data points.

Logistic Regression–Model Validity

- Omnibus test of model coefficient:

Value less than 0.05 helps to reject the null hypothesis that that there is no difference between the model with only a constant and the model with independent variables

- Wald statistics: Equivalent of t – statistics in regression. Used to check the significance of individual explanatory variable.

If the P value corresponding to Wald statistics is < 0.05, the coefficient of the explanatory variable is not zero.

- Hosmer Lemeshow test: Test for overall fitness for binary logistic regression.

P value < 0.05 signifies bad fit for the model. P value > 0.05 the model is accepted

Logistic Regression–Model Validity

- Likelihood ratio: Equivalent of F statistics in regression. Used to test the null hypothesis that all the slope coefficients are simultaneously equal to zero.

  Deviance D = - 2 * (LL). LL implies log likelihood.

Measures the deviance from the perfect model. The larger the value of D, the worse the fit. If the P value corresponding to D is < 0.05, the overall model is accepted.

- Conventional measure of R^2 is not meaningful. Different R^2 statistics prevalent:
  - McFadden R^2 value of .20 and above is considered good.
  - Cox and Snell R^2
  - Nagelkerke R^2: Modified Cox and Snell R^2 to maximum value of 1.
  - Count R^2 which is $\frac{no\ of\ correct\ predictions}{total\ number\ of\ observations}$ (If predicted probability is > 0.5 it is classified as 1 else as 0.)

Logistic Regression–Model Validity

$$Sensitivity = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Specificity = \left(\frac{TN}{TN + FP}\right) = \frac{17}{17} = 100\%$$

Classification matrix		
	Predicted	
Observed	Class=1 (Positive)	Class=0 (Negative)
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

$$Model\ accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) = \frac{21}{24} = 87.5\%$$

Sensitivity is the probability that predicted class is 1 when observed class is 1.
Specificity is the probability that the predicted class is 0 when the observed class is 0.

Logistic Regression–Model Validity

$$Recall = \left(\frac{TP}{TP + FN}\right) = \frac{4}{7} = 57.1\%$$

$$Precision = \left(\frac{TP}{TP + FP}\right) = \frac{4}{4} = 100\%$$

Classification matrix		
	Predicted	
Observed	Class=1 (Positive)	Class=0 (Negative)
Class =1 (Positive)	f_{11}= 4 [TP]	f_{10}= 3 [FN]
Class =0 (Negative)	f_{01}= 0 [FP]	f_{00}= 17 [TN]

Recall is same as sensitivity.

Higher precision means lower recall/sensitivity.

Higher precision means higher specificity.

Precision: Of all the cases where model predicted Y = 1, what fraction actually are positive?

Recall: Of all the cases which are actually Y = 1, what fraction actually are predicted positive?

Concordant and Discordant Pair

German Credit Rating Data

SL No	CHK_ACCT	Duration	Class	Class Code	Predicted Probability	Predicted class at 0.50
235	**no-account**	**4**	**good.**	**1**	**0.93653**	**1**
431	no-account	5	good.	1	0.93427	1
27	no-account	6	good.	1	0.93194	1
489	no-account	10	good.	1	0.92181	1
506	**no-account**	**10**	**bad.**	**0**	**0.92181**	**1**
746	**0DM**	**13**	**good.**	**1**	**0.58332**	**1**
52	less-200DM	27	good.	1	0.57500	1
797	0DM	18	bad.	0	0.53731	1
4	0DM	42	good.	1	0.32131	0
77	0DM	42	bad.	0	0.32131	0
651	0DM	48	good.	1	0.27447	0
702	0DM	48	bad.	0	0.27447	0
814	0DM	48	bad.	0	0.27447	0
815	0DM	48	bad.	0	0.27447	0
928	0DM	48	bad.	0	0.27447	0
678	less-200DM	72	bad.	0	0.20099	0
30	0DM	60	bad.	0	0.19455	0
974	**0DM**	**60**	**bad.**	**0**	**0.19455**	**0**

$$Greater\ than\ a\ Cut - off\ Value \rightarrow\ Predicted\ class\ is\ Good$$

Case	235	974
Actual	Good (1)	Bad (0)
Predicted Probability	0.9365	0.1945

There exist cut –off values which can correctly classify good credit rating as good and bad credit rating as bad. So case 235 and 974 are concordant pairs

Case	746	506
Actual	Good (1)	Bad (0)
Predicted Probability	0.58332	0.9218

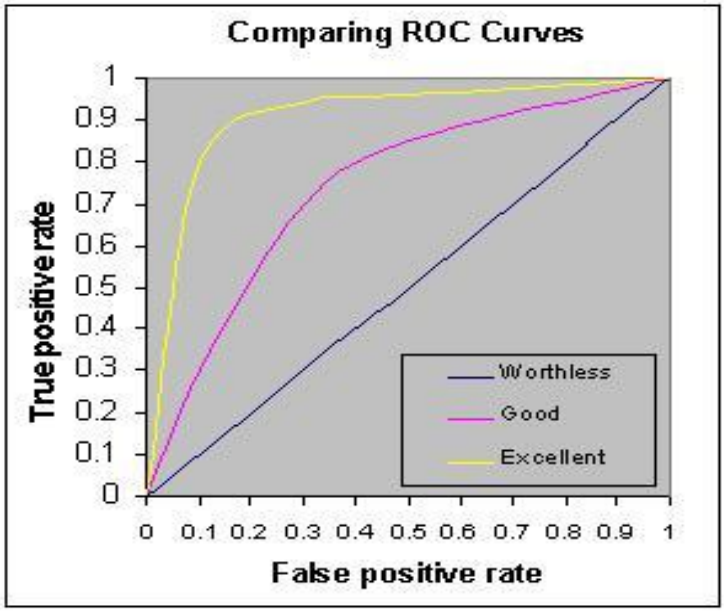
There exist no cut –off values which can correctly classify good credit rating as good and bad credit rating as bad. So case 506 and 746 are discordant pairs

ROC Curve

- Receiver operating characteristics (ROC) Curve is a useful way to determine cut-off point which maximizes sensitivity and specificity.
- Sensitivity and specificity measures are computed based on a sequence of cut-off points to be applied to the model for predicting observations into Positive or Negative.

An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC). AUC values between:

- 0.9-1 indicate perfect sensitivity and specificity,
- 0.8-0.9 indicate good sensitivity and specificity,
- 0.7-0.8 indicate fair sensitivity and specificity,
- 0.6-0.7 is poor
- 0.6 and below indicate by chance outcome



Given a pair of data (one positive class and one negative class), the AUC represents the probability that the pair will be correctly classified.

F 1 Score

Another measure to compare different models.

$$F\ Score = \frac{2 * P * R}{\mathrm{P} + \mathrm{R}}$$

Model	Precision	Recall	Average	F Score
Logistic	0.5	0.4	0.45	0.44
Decision Tree	0.7	0.1	0.4	0.17
Neural Network	.02	1.0	0.51	0.03

Simple average of precision and recall may not be a good metric to select the best model

Logistic Regression–Influential Cases and Outliers

- An **outlier** is an observation whose dependent variable value is unusual given its values on the predictor variables.
  - Residual is the difference between the actual probability and predicted probability.

If a case has a standardized residual larger than 3.0 or smaller than -3.0, it is considered an outlier and a candidate for exclusion from the analysis

- An observation is said to be **influential** if removing the observation substantially changes the estimate of coefficients.
  - Cook's distance: is a measure of the influence which a case has on the solution

A case is identified as influential if its Cook's distance is greater than 1.0

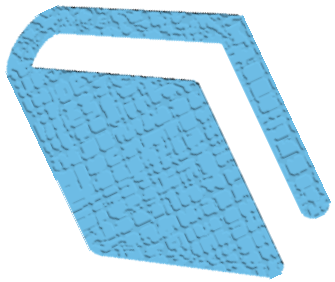
- An observation with an extreme value on a predictor variable is called a point with high **leverage**. Leverage is a measure of how far an observation deviates from the mean of that variable

If after removing the outliers and influential cases the model accuracy does not change by more than 2%, then retain the cases. More at: http://www.ats.ucla.edu/stat/r/dae/rreg.htm

Logistic Regression in R Using an Example

Summary

Summary of the topics covered in this lesson:



- The intent of performing logistic regression analysis is to predict the probability of an event outside the sample dataset.
- Validity of logistics regression is understood through various statistics.
- Validity is important to bring in better generalizability of the model.
- Usefulness of the model is understood by interpreting the signs of the coefficients and whether it is inline with the natural belief.

QUIZ TIME

Quiz Question 1

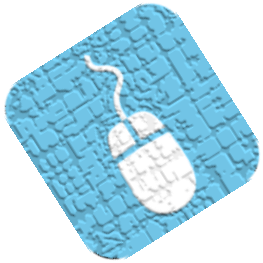
Quiz 1	What is the significance of Wald Statistics in logistic regression?

a. Used to check influential cases in the dataset.

b. Used to check the overall fit of the model.

c. Used to check the significance of individual explanatory variable.

d. None of the above.



Quiz Question 1

Quiz 1	What is the significance of Wald Statistics in logistic regression?

a. Used to check influential cases in the dataset.

b. Used to check the overall fit of the model.

c. Used to check the significance of individual explanatory variable.

d. None of the above.

Correct answer is: *c*

Equivalent of t – statistics in regression. Used to check the significance of individual explanatory variable.

End of Lesson03–Logistic Regression Concepts