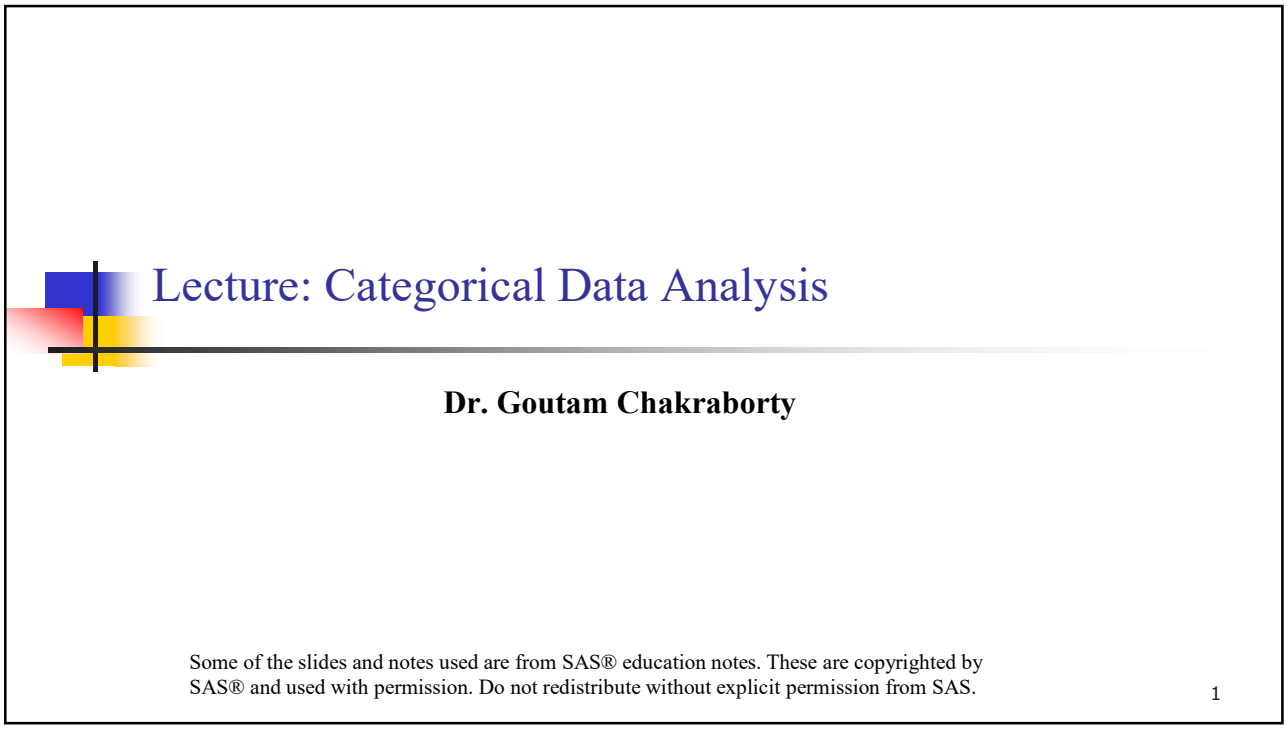
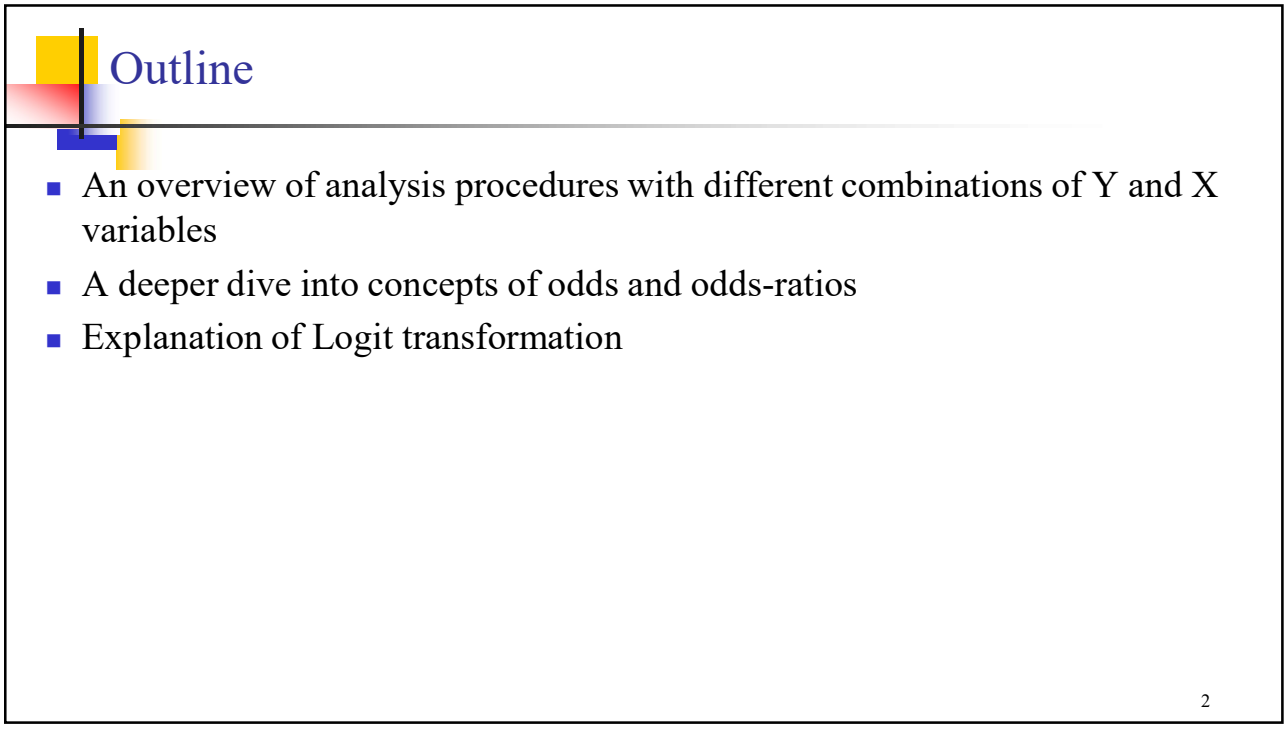

Lecture: Categorical Data Analysis
Dr. Goutam Chakraborty
Some of the slides and notes used are from SAS® education notes. These are copyrighted by SAS® and used with permission. Do not redistribute without explicit permission from SAS.
1


Outline
An overview of analysis procedures with different combinations of Y and X variables
A deeper dive into concepts of odds and odds-ratios
Explanation of Logit transformation
2

Overview of Different Analysis

Type of Predictors (X) / Type of Response (Y)	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Multiple Regression (**MR**)	Analysis of Covariance (ANCOVA or MR)
Categorical	**Contingency Table (or, Cross-Tab Analysis)** or, Logistic Regression	Logistic Regression	Logistic Regression

Odds and Odds Ratios

- Odds of an event is defined as the probability of an event divided by the probability of a non-event
  - Odds of getting a head in toss of a fair coin is 1:1

$$\text{Odds} = \frac{p_{event}}{1 - p_{event}}$$

- An *odds ratio* between two groups indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Probability versus Odds of an Outcome

	Outcome		
	Yes	**No**	**Total**
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Total **Yes** outcomes in Group **A** ÷ Total outcomes in Group **A**

Probability of a **Yes** in Group A=**60÷80=0.75**

Probability of a **Yes** in Group B=**90÷100=0.90**

Probability versus Odds of an Outcome

	Outcome		
	Yes	**No**	**Total**
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Probability of **Yes** in Group A=0.75 ÷ Probability of **No** in Group A=0.25

Odds of **Yes** in Group A=**0.75÷0.25=3**

Probability versus Odds of an Outcome

	Outcome		
	Yes	**No**	**Total**
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

Probability of **Yes** in Group B=0.90 ÷ Probability of **No** in Group B=0.10

Odds of **Yes** in Group B=**0.90÷0.10=9**

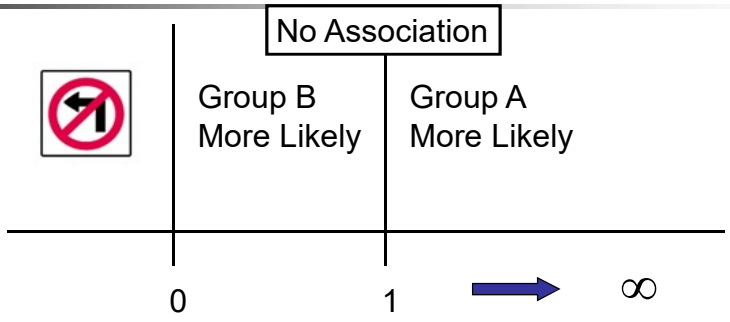
Odds Ratio

	Outcome		
	Yes	**No**	**Total**
Group A	60	20	80
Group B	90	10	100
Total	150	30	180

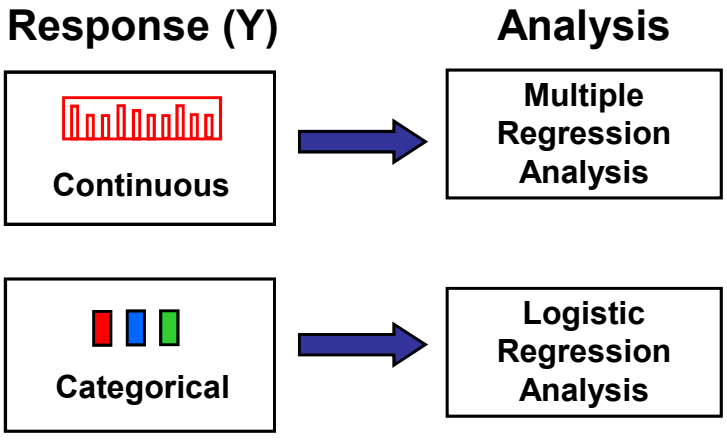
Odds of Yes in **Group A=3** ÷ Odds of Yes in **Group B=9**

Odds Ratio, A to B=**3÷9=0.3333**

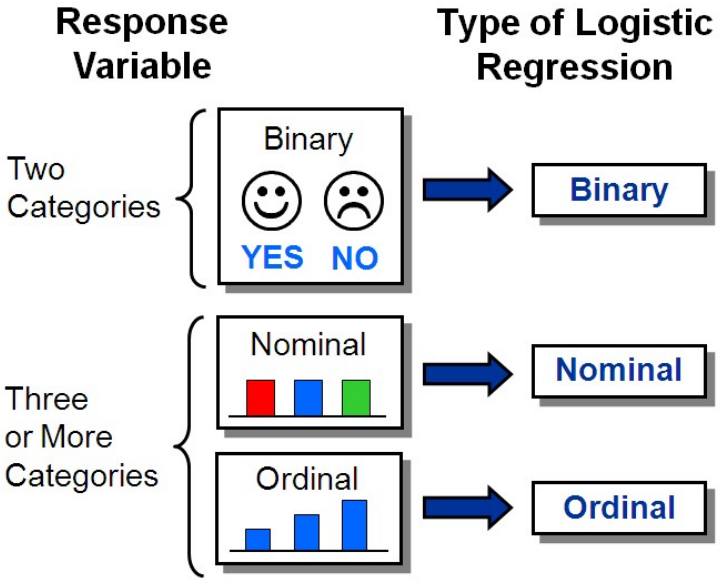
Properties of the Odds Ratio, A to B



Overview



Types of Logistic Regression



Why Not Ordinary Regression?

Regression: $Y_i=\beta_0+\beta_1X_{1i}+\varepsilon_i$

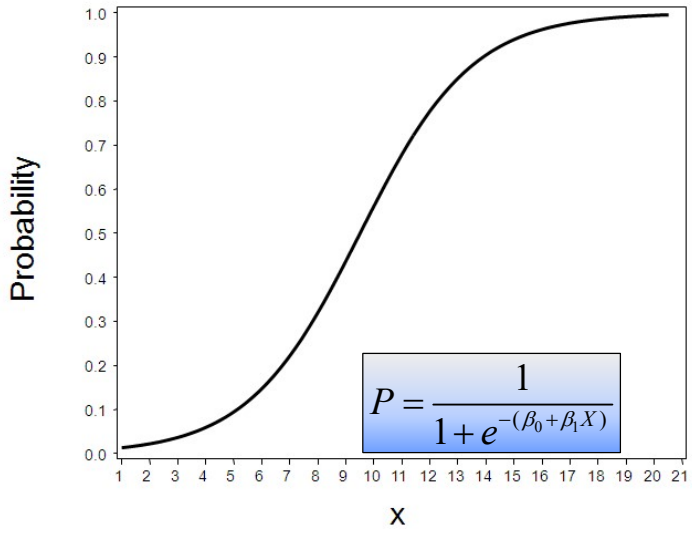
- If the response variable is categorical, then how do you code the response numerically?
  - If the response, Y is coded (1=Yes and 0=No) and your regression equation predicts 0.5 or 1.1 or -0.4, what does that mean practically?
- If there are only two (or a few) possible response levels, is it reasonable to assume constant variance and normality?

What about a Linear Probability Model?

Linear Probability Model: $p_i=\beta_0+\beta_1 X_{1i}$

- Probabilities are bounded, but linear functions can take on any value. (Once again, how do you interpret a predicted value of -0.4 or 1.1?)
- Given the bounded nature of probabilities, can you assume a linear relationship between X and *p* throughout the possible range of X?
- Can you assume a random error with constant variance?
- What is the *observed probability for an observation*?

Logistic Regression Model



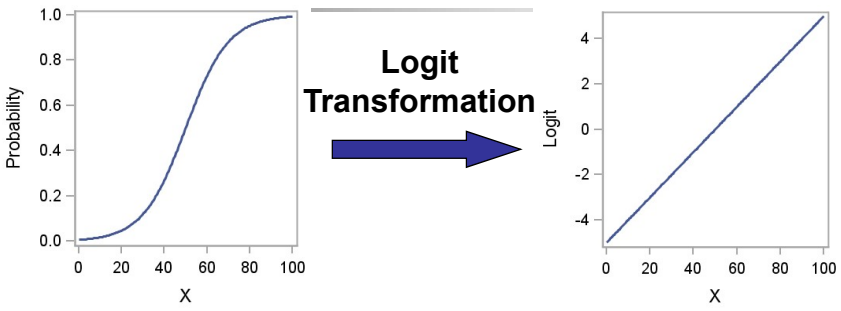
Logit Transformation

- Logistic regression models transformed probabilities, called *logits**,
- $$\text{logit}(p_i) = \ln\left(\frac{p_i}{(1-p_i)}\right)$$

where

- i indexes all cases (observations)
- p_i is the probability that the event (for example, a sale) occurs in the i^{th} case
- ln is the natural log (to the base e).
- The logit is the natural log of the odds.

Assumption



Logistic Regression Model

$$\textbf{logit}(p_i)=\beta_0+\beta_1 X_{1i}+\ldots+\beta_k X_{ki}$$

where

- **logit (p_i)**= logit of the probability of the event
- β_0= intercept of the regression equation
- β_k= parameter estimate of the k^{th} predictor variable

Lecture: Mechanics and Metrics of LR

Dr. Goutam Chakraborty

18

Outline

- How does Logistics Regression (LR model) work?
- What are some of the important metrics/diagnostics from LR and how are these interpreted?

19

Logistic Regression (LR) Model

- The idea is instead of modeling the dependent (**response**) variable, Y (that takes only two values 1 and 0), we will model a function of this **response** variable
- Model: $G = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ..+ \beta_n X_n$
- Where, G= Logit of a yes response, β and X's are coefficients and predictor variables.
- G or Logit (yes) = log{p /(1-p)} where p = probability of yes response and log is natural log (i.e., logarithm base is e)

20

Mechanics of Logistic Regression (LR)

- The coefficients (β) of logistic regression will be estimated using ML (*maximum-likelihood*) method. ML is an iterative and numerical method
  - It starts by assuming *a process* that generates your data
    - For example, the process could be a Sigmoid curve of relationship between your X and probability of Y
    - Then, it generates data with different values for (β) and compare those *generated values* with *observed values (using log-likelihood)*
    - It selects values for (β) such that the likelihood of observed data pattern is maximized
  - ML method has assumptions that tend to work well with large data

21

Metrics in LR (Statistical Significance)

- Whole Model Test : Likelihood (Chi-square) based
  - Null hypothesis is that the model does not fit the data any better than random!
  - If the p-value from this test is less than 0.05, that means…
- Parameter Estimates : values for each coefficient and test if each coefficient estimate is different from 0
  - If p-value for any of the coefficient test is less than 0.05, then…
- Relative importance of each estimates is indicated by LogWorth (higher value of LogWorth means more important the variable is in predicting Y)

22

Metrics in LR (To Evaluate Model Fit)

- Misclassification rate (easiest to understand)
  - Between 0 and 1
- Multiple metrics related to R-square
  - Between 0 and 1
- Goodness-Of-Fit (GOF) statistics such as AICc and BIC
  - These are similar to MSEs in Multiple Regression – the lower the values the better the model is.
  - But absolute values of each of these metrics is hard to understand

23

LR metrics: How to Interpret Coefficients

- LR Model: $G = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ..+ \beta_n X_n$
  - Where, G= Logit of a yes response, β and X's are coefficients and independent variables.
- One way: As X_1 changes by 1 unit, the G (or logit of yes response) changes by β_1 units holding everything else constant
- Perhaps a better way is to use odds ratios for each variable

24

A Bit More on Odds

- Both probabilities and odds have lower bounds of 0.
- Unlike probabilities, Odds have no upper bounds!
  - This is an important issue for ratio comparisons!
    - If customer A has a probability of buying our brand as 0.6; we cannot find a customer whose probability of buying our brand will be twice as great!
    - What happens to such comparisons for odds?
- Odds and odds ratios will be used in interpreting coefficients in logistic regression

25

LR Metrics: Odds-Ratios

- Response (Yes=1, No =0)
- Independent variables : Gender (M=1, F=0) and Age (measured in years)
- If odds-ratio of gender is 1.25, it means..
  - Odds of **males** for response are 25% higher than odds for **females** for response, holding everything else constant
  - If we take two persons of *the same age*, the male is 25% more likely than the female to respond yes
- If odds ratio of Age is 1.15, it means..
  - 15% increase in odds for 1-unit increase in age, holding everything constant
  - If we take two persons of *the same gender* but with one aged 49 and another aged 50, then the 50 year old person's odds of response is 15% more than the 49-year old person's odds of response

26

Metrics in LR

- LR will create for each case,
  - Predicted probability that Y=1, p_1
  - Predicted probability that Y=0, p_0
  - Predicted response for Y
    - 1 if $p_1 > p_0$
    - 0 otherwise
- A classification table will show cross-tab of *observed values of Y versus predicted values of Y* (confusion matrix)

27

Demo: Basics of LR

Dr. Goutam Chakraborty

28

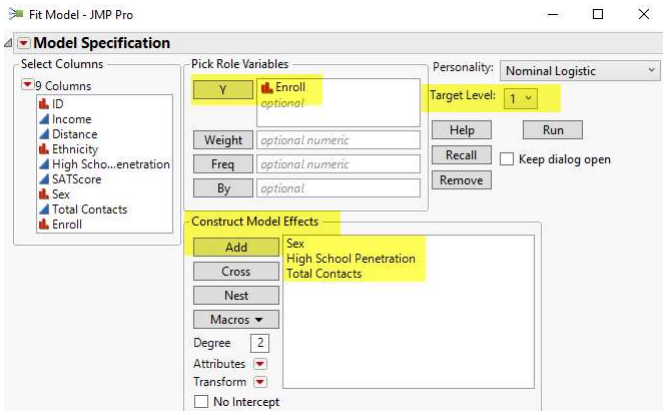
Data Set: Enrollment Sample Data

- A sample of 2,500 observations from a large university's admission **prospect** database
- Target (Y) variable is Enroll: Whether student enrolled (Yes=1, No =0)
- Potential predictor (X) variables are: Income (estimated average household income), Distance (in miles from university and prospect's address), Ethnicity (multiple categories such as C= Caucasian, H=Hispanic, B=Black, A=Asian, U=Unknown), High School Penetration (Percent of enquirers from the prospect's high school over last 5 years who enrolled at the university), SATScore, Sex (M= Male, F= Female) and Total Contacts (number of total contacts made by admission office and prospect)

29

Demo1

- Goal: Use only a few X variables (High School Penetration, Sex and Total Contacts) to predict and explain Y variable (Enroll)
- Open data in JMP
- Analyze > Fit Model > Select Enroll as Y > Change Target Level to 1 > Select High School Penetration, Sex and Total Contacts > Click Add button under Construct Model Effects > Run



30

Demo1 Output

Converged in Gradient, 7 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	997.1234	3	1994.247	<.0001*
Full	686.8613			
Reduced	1683.9846			

RSquare (U)	0.5921
AICc	1381.74
BIC	1404.91
Observations (or Sum Wgts)	2433

Is the Whole Model Test significant? What does that mean to you?

31

Demo1 Output (Continued)

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-4.3717443	0.1784928	599.88	<.0001*
Sex[F]	0.16994218	0.0708671	5.75	0.0165*
High School Penetration	21.304506	1.5256178	195.01	<.0001*
Total Contacts	0.78640594	0.0357452	484.01	<.0001*

For log odds of Yes/No

Are all of the coefficients statistically significant?

32

Demo1 Output (Continued)

Nominal Logistic Fit for Enroll

Effect Summary

Source	LogWorth	PValue
Total Contacts	238.052	0.00000
High School Penetration	83.190	0.00000
Sex	1.787	0.01632

Which is the most important variable?

33

Demo1 Output (Continued)

How well does the model fit?

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	997.1234	3	1994.247	<.0001*
Full	686.8613			
Reduced	1683.9846			

RSquare (U)	0.5921
AICc	1381.74
BIC	1404.91
Observations (or Sum Wgts)	2433

34

Demo1 Output (Continued)

Click red triangle next to Nominal Logistic and select "Save Probability Formula.

	ID	Income	Distance	Ethnicity	High School Penetration	SATScore	Sex	Total Contacts	Enroll	Lin[Yes]	Prob[Yes]	Prob[No]	Most Likely Enroll
43	2242	•	•	B	0.014084507	•	M	3	No	-1.882405157	0.1321128564	0.8678871436	No
44	2299	•	•	U	0.010869565	•	F	1	No	-3.183825427	0.0397789581	0.9602210419	No
45	2435	•	•	C	0.25	1270	M	6	Yes	5.5028756907	0.9959415025	0.0040584975	Yes
46	2533	•	•	H	1	940	M	4	Yes	19.908443284	0.9999999977	2.2587748e-9	Yes
47	2615	•	•	B	0.091954023	1190	F	2	No	-0.669955168	0.3385068795	0.6614931205	No
48	2629	•	•	U	0.091954023	1050	F	2	Yes	-0.669955168	0.3385068795	0.6614931205	No
49	2642	•	•	B	0.091954023	1150	M	2	Yes	-1.009839529	0.2670112564	0.7329887436	No
50	2684	•	•	A	0.091954023	•	F	1	No	-1.456361108	0.1890245153	0.8109754847	No
51	2686	•	•	A	1	1090	F	4	Yes	20.248327645	0.9999999984	1.6079148e-9	Yes
52	2687	•	•	H	1	1140	M	1	Yes	17.549225464	0.9999999761	2.3903869e-8	Yes
53	2738	129811	1619.28026	C	1	1300	M	6	Yes	21.481255164	0.9999999995	4.686077e-10	Yes
54	2768	71174	1606.402312	C	0.333333333	1120	M	10	Yes	10.423874941	0.9999702864	0.0000297136	Yes
55	2810	50879	1641.800269	U	0.5	1070	M	4	Yes	9.2561903018	0.9999044906	0.0000955094	Yes

35

Demo1 Output (Continued)

How well does the model fit?

Click red triangle next to Nominal Logistic and select Confusion Matrix

Confusion Matrix

Training

Actual Enroll	Predicted Count Yes	No
Yes	1120	151
No	132	1030

Total observations used in analysis = 1120+151+132+1030 = 2,433
Total correctly classified (predicted) = 1120+1030 = 2,150
Total Misclassified = 151+132 = 283
Misclassification Rate = 283/2433 = 0.12 or, 12%

36

Demo1 Output (Continued)

How do you explain the values of these coefficients?

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-4.3717443	0.1784928	599.88	<.0001*
Sex[F]	0.16994218	0.0708671	5.75	0.0165*
High School Penetration	21.304506	1.5256178	195.01	<.0001*
Total Contacts	0.78640594	0.0357452	484.01	<.0001*

For log odds of Yes/No

37

Demo1 Output (Continued)

How do you explain the values of these coefficients?

Click red triangle next to Nominal Logistic and select Odds Ratios

Odds Ratios

For Enroll odds of Yes versus No

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
High School Penetration	1.788e+9	89912727	3.56e+10	5.592e-10
Total Contacts	2.195492	2.046941	2.354822	0.4554789

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
High School Penetration	1.788e+9	89912727	3.56e+10	5.592e-10
Total Contacts	1.665e+9	2.511e+8	1.1e+10	6.007e-10

Odds Ratios for Sex

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
M	F	0.7118526	0.0165*	0.5391949	0.9397979
F	M	1.4047851	0.0165*	1.0640586	1.8546171

Normal approximations used for ratio confidence limits effects: Sex
Tests and confidence intervals on odds ratios are Wald based.

38

In Summary

- Overall model is statistically significant
- All of the coefficients are statistically significant
  - Most important variable is "Total Contacts"
- Model fit is reasonable (Misclassification rate of 12%)

39

Explanation of How Prediction is Handled in JMP

- I mentioned in the video that I will give you a handout – so, here it is. :
  - Model estimated is: Logit or, Ln(p/1-p) = -4.371+21.304*High School Penetration +0.17 * Sex (if Female then 1) +0.786*Total contacts
  - Where p=probability (enrollment), 1- p = probability (No enrollment)
  - **For ID=151**, the values of predictors (X) are: High School Penetration=0.0377, Sex = F, Total Contacts = 1
  - Substituting the X values, **for ID=151** we get predicted Ln(p/1-p) = -4.371+ 21.304*(0.0377) +0.17 * (1) +0.786*(1)
  - Or, predicted Ln(p/1-p) = -4.371+0.80+0.17+0.786 = -2.628, this is Lin(Yes) in JMP terms
  - By the way, if a candidate is male, then the coefficient of Sex is -0.17 (mirror opposite of number for Female)

40

Explanation of How Prediction is Handled in JMP (Continued)

- **For ID=151**, Ln(p/1-p) = -2.628
- Exponentiation of both sides, (p/1-p) = exp(-2.628)
- Or, (p/1-p)= 0.97, Or, p= 0.97 *(1-p)
- Or, p=0.97-p, Or, 2p=0.97
- Or, p= 0.48, this is Prob(Yes) in JPM terms (rounding error because I am only using a few places of decimal)
- 1-p = 1- 0.48 = 0.52. this is Prob(No) in JMP terms
- Rule used is if Prob(Yes) > Prob(No), then Predicted enroll=Yes, otherwise Predicted Enroll=No. (*This is same as saying if Prob(yes) GT 0.5 then Predicted Enroll=Yes, otherwise Predicted(Enroll)=No*)
- **For ID=151**, using rule above, Predicted(Enroll) = No, this is Most Likley to Enroll in JMP terms

41

Lecture: Sensitivity, Specificity, Variable Selection and Multicollinearity in LR

Dr. Goutam Chakraborty

42

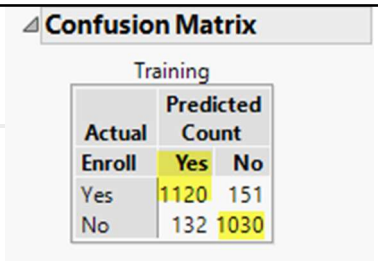
Outline

- More metrics based on confusion matrix:
  - Misclassification rate (already seen)
  - Sensitivity, Specificity, True Positive and True Negative
- How to select relevant variables for LR model when you have a large number of X variables and theory is not of much help!
- Assumptions in LR and how to check them
- Checking Multicollinearity in LR model

43

Metrics From Confusion Matrix

Actual \ Predicted	Yes (1)	No (0)	
Yes(1)	1120 (**TP**)	151 (**FN**)	**1271**
No(0)	132 (**FP**)	1030 (**TN**)	**1162**
	1252	**1181**	**2433**



Overall (Yes/no of Actual) % **correctly predicted** = (1120+1030)/2433 = 88.3%
Misclassification Rate= 11.7%

% of **Yes** of actual correctly predicted (**Sensitivity**) = 1120/1271= 88.1%

% of **No** of actual correctly predicted (**Specificity**) = 1030/1162= 88.6%

% of **FP** (False Positive) of prediction= 132/1252= 10.5%

% of **FN** (False Positive) of prediction= 151/1181= 12.7%

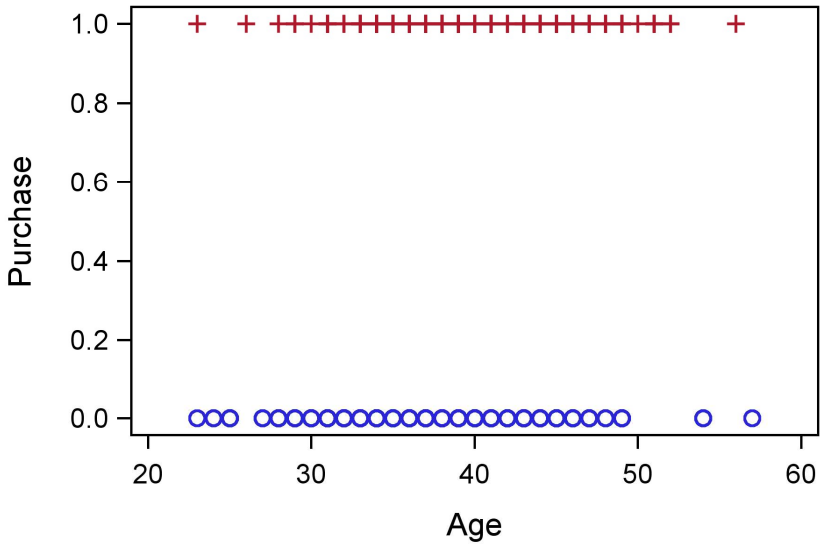
44

Other Issues in LR (contd.)

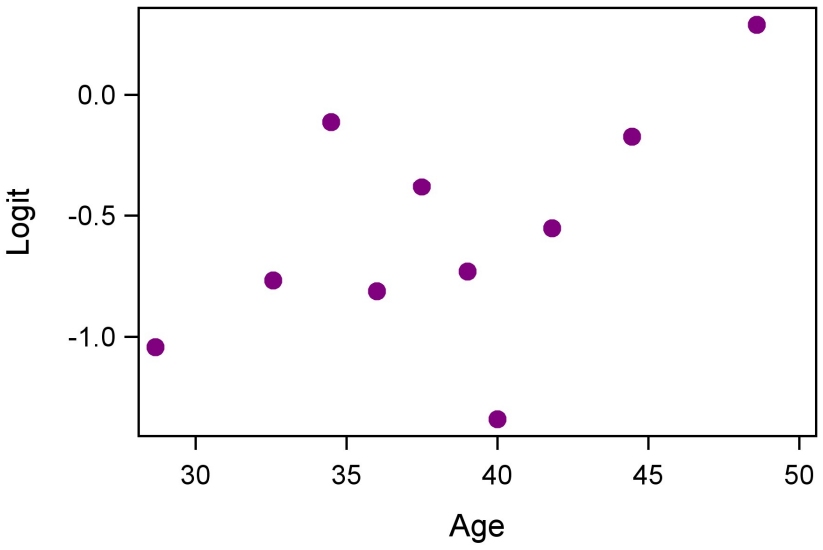
- What to do when you have a large number of independent variables to choose from
  - Same principle applies as in MR
    - Retain theoretically/managerially important variables
  - Forward, backward and mixed (stepwise) selection is available
  - Multiple criteria may be used such as: p-value threshold or, data mining metrics such as Minimum AICc or, BIC
- Multicollinearity
  - Problems remain same as in MR.
  - No diagnostics in LR – best is to fit an MR model (with 1/0 dependent variable) and use the collinearity diagnostics from the MR model
- LR Assumption: Logits are linearly related to X-variables

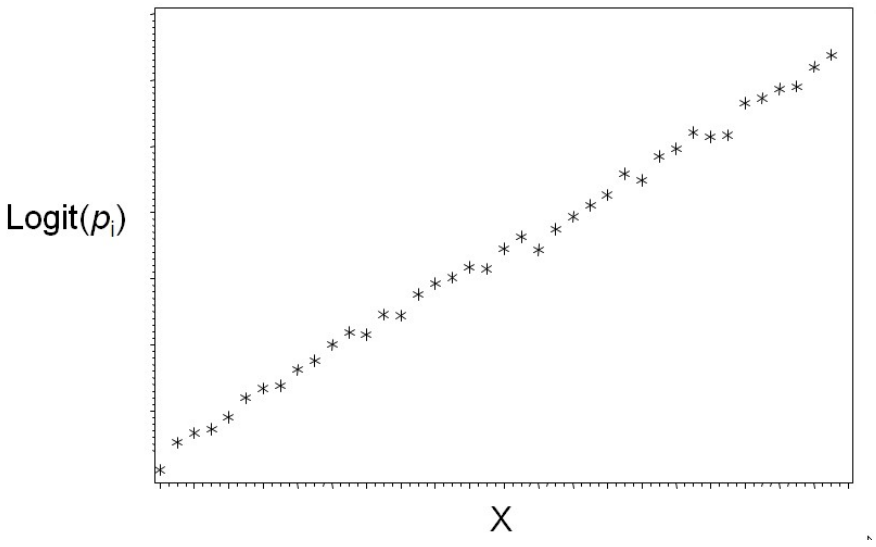
45

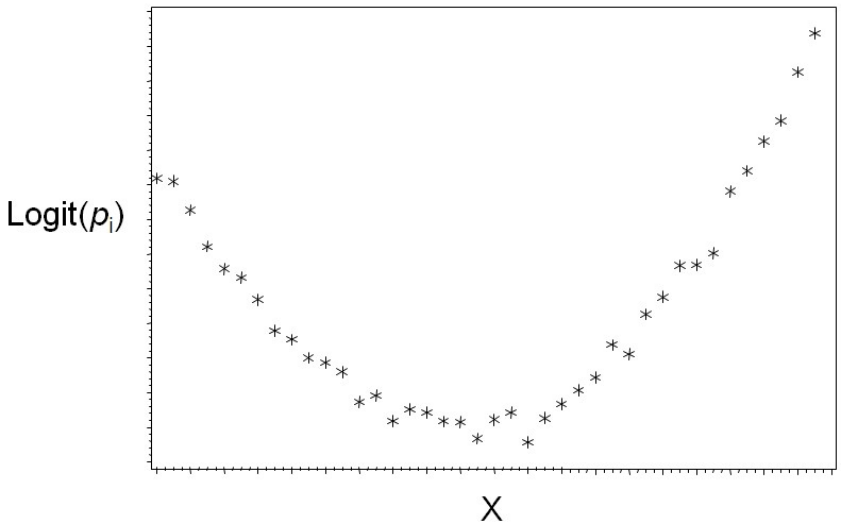
Scatter Plot of Binary Response Data



Estimated Logit Plot







Estimated Logits

$$\ln\left(\frac{E_i+1}{C_i-E_i+1}\right)$$

where

E_i= number of events in bin

C_i= number of cases in bin

Demo: Variable Selection and Multicollinearity in LR

Dr. Goutam Chakraborty

51

Demo Procedure

- Use Enrollment Sample data
- For variable selection
  - Use all of the available X variables as potential predictors and use Mixed (stepwise) selection with p-value threshold
  - Explain Rule for combining categories of nominal variables during this selection process
- For checking Multicollinearity in LR
  - First, change the target to a "continuous" variable
  - Then use MR and check Multicollinearity

52

Demo: Prediction of New Data

Dr. Goutam Chakraborty

53

Prediction (or, Scoring) of New Data

- Suppose after some trial-and-error, you built a model (MR or LR or others) and happy with its diagnostics. How do you use it to predict new data?
- Note: new data must have values for all of the X-variables that were in the final model built, but has no Y values
- Basic idea: take the estimated coefficients of each X-variable from the model built, multiply those with the corresponding X-values in the new data and add those up to get predicted value for the new data
  - For MR, those are the predicted Y
  - For LR, those are predicted logits (G).
  - We have to convert those logits to Probability (Y=1), and then apply a rule such as "if Probability (Y=1) GT 0.5, then Predicted(Y)=1, else Predicted(Y)=0"

54

Scoring Procedure

- If scoring data set is small, then:
  - In JMP, Save Probability Formula (under red triangle) from your final model.
    - This will create predicted values for all data used in building model
  - Copy and paste scoring data below the original data table and JMP will automatically create predictions for the newly pasted data
- If scoring data set is large, then:
  - Save scoring code (this contains the formula and estimates from the model built) and then apply this scoring code on new data
  - This process depends on the software used such as JMP, SAS, R, Python and others

55

Enrollment Scoring Data

- Use model built from stepwise to score this data using method described in previous slide (scoring data size is small)

	ID	Income	Distance	Ethnicity	High School Penetration	SATScore	Sex	Total Contacts	Enroll
1	1	•	•	C	0	•	M	2	
2	2	•	•	U	0.037651631	•	•	1	
3	3	•	•	N	0	1210	M	2	
4	4	68618	733.195905	U	0.111111111	1110	M	7	
5	5	34207	730.3963136	U	0	•	F	2	
6	6	55973	101.9126495	C	0.103448276	1130	F	7	
7	7	41521	99.25082955	C	0.103448276	1230	F	4	
8	8	41539	166.462944	A	0.035714286	1260	M	16	
9	9	49649	81.69713156	N	0.1875	1400	F	9	
10	10	83078	150.0847011	C	0.096385542	1210	M	8	

56