Slide-1

Analyzing Attribute Data

Slide-2

Predicting Binary Outcomes

COMPANY CEO, start date	FIRST-DAY stock price	LONG-TERM stock price*
NEWELL RUBBERMAID Mark Ketchum, 10/17/05	↗ 9%	↘ 36%
OFFICE DEPOT Steve Odland, 3/14/05	↗ 8%	↘ 79%
RADIOSHACK Julian Day, 7/7/06	↗ 23%	↘ 5%
PROLOGIS Walter Rakowich, 11/12/08	↘ 35%	↗ 247%
VIACOM Philippe Dauman, 9/5/06	↘ 5%	↗ 47%
WALGREEN Alan McNally, 10/10/08	↘ 8%	↗ 16%

Predict the direction of a numerical series (up/down)

Will an event occur or not at time *t+k*?

Will a value cross a threshold of interest at time *t+k*?

Slide-3

Techniques Used

1) LOGISTIC REGRESSION 2) LOGIT ANALYSIS 3) PROBIT ANALYSIS

Slide-4

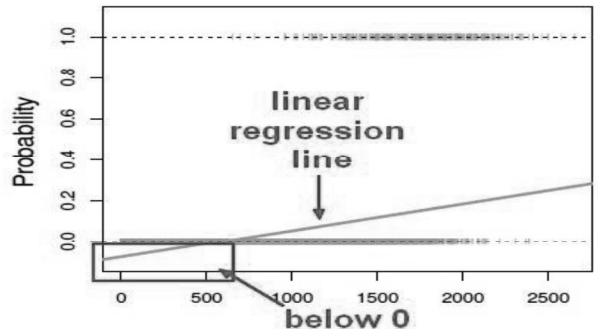
Logistic Regression:

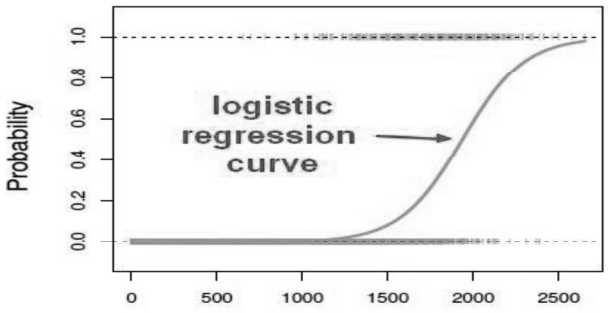
Logistic Regression

- Logistic Regression model predicts the probability associated with each dependent variable Category.

How does it do this?

- It finds linear relationship between independent variables and a link function of this probabilities. Then the link function that provides the best goodness-of-fit for the given data is chosen





Slide-5

Types of Logistic Regression:

- **Logistic Regression**

  Simple

- A Single Independent Variable Model is called as Simple Logistic Regression

Multiple

- A Multiple Independent Variable Model is called as Multiple Logistic Regression

Slide-6

Logistic Regression:

Multiple Logistic Regression Model:

Multiple Logistic Regression Model is quite similar to the Multiple Linear Regression Model, Only β coefficients vary.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

Where,

β_0 = the y axis intercept

β_i = the model coefficient for the linear effect of variable i on y

e = the random error

slide-7

Logistic Regression:

Probability:

The probability in a logistic regression curve:

$$p = \frac{e^y}{1 + e^y}$$

Where,

'e' is a real number constant, the basic of natural logarithm and equals 2.7183

'y' is the response value for an observation.

Slide-8

Methods:

There are three methods of Logistic Regression based on nature of the attribute data.

- Binary
- Nominal
- Ordinal

- Binary Logistic Regression
  - Binary logistic Regression is performed on the Binary response variables. It has only two categories, such as presence or absence of disease, pass or fail, defective or non-defective products.
- Nominal Logistic Regression
  - Nominal Logistic Regression is performed on the Nominal variables. These are categorical variables

that have three or more possible categories with no natural ordering

Example: Food is crunchy, mushy and crispy

Slide-9

Methods:

- Ordinal Logistic Regression
    - Ordinal Logistic Regression is performed on ordinal response variables. These are categorical variable that have three or more possible categories with a natural ordering.

Example: Survey on quality of a shirt material; strongly disagree, disagree, neutral, agree and strongly agree.

Method	Description of categorical response variable	Example
Binary	Two Categories	Presense/ Absence of disease
Nominal	There are more categories with no natural ordering to the levels	Crunchy/Mushy/ Crispy
Ordinal	There are more categories with ordering of the levels	Strongly disagree/Disagree/Neutral/Agree/Strongly agree

Slide-10

Assumptions and Steps:

- *Logistic Regression Assumptions:*
    - *Only one outcome per event – Like pass or fail*
    - *The outcomes are statistically independent*

- *All relevant predictors are in the model*
- *One Category at a time- Means mutually exclusive and collectively exhaustive*
- *Sample sizes are larger than for linear regression*

Steps Involved in Logistic Regression

- *Step-1: Collect and organize sample data*
- *Step-2: Formulate Logistic Regression Model*
- *Step-3: Check the model's validity*
- *Step-4: Determine Probabilities using Probability equation*
- *Step-5: Compile the results*

Slide-11

Example:

Imagine that you are a Data Scientist at a very large scale integration circuit manufacturing company. You want to know whether the time spent inspecting each product impacts the quality assurance department's ability to detect a designing error in the circuit.

→ *Step-1: Collect and organize the sample data.*

→ Number of Observations

→ Error Identification

→ Inspection Time

Number of Observations: 55 Observations of circuits with errors, and determine whether those errors were detected by QA.

Slide-12

Observations	Inspection time (minutes)	Error detection Yes = 1 No = 0
1	38	1
2	29	0
3	34	0
4	35	1
5	32	1
6	41	1
7	35	0
8	41	1
9	24	0
10	28	1
11	32	0
12	36	1
13	20	0
14	21	0
15	39	1
16	42	1
17	27	0
18	20	1
19	37	1
20	39	1
21	30	0
22	37	1
23	41	1
24	23	0
25	18	0

Slide-13

Example:

- *Step-2: Formulate Logistic Regression Model*

- *The Logistic regression equation is derived from any statistical software*

y = -9.94645 + 0.328827x

- *Step-3: Check model's validity*
- *Check model's validity involves performing Hypothesis testing, or validations.*
- *Creating Logistic Regression Tables*

Logistic Regression Table:

- *Logistic Regression Table shows the estimated coefficients, standard error of the Coefficients, Z Statistics and p-values.*

Slide-14

Example:

→ ***Logistic Regression Table:***

The important thing about the table is p-value is 0.00 which is less than alpha value 0.05, this indicates that regression coefficients are not 0, and there is a significant relationship between the independent variable, -Inspection time and the dependent variable- error detection.

Predictor	Coef	SE coef	Z Statistic	P-Value
Constant	-9.947	2.498	-3.98	0.000
Inspection Time	0.329	0.081	4.05	0.000

Slide-15

Example:

Step-4: Determining Probabilities using Probability equation

Probability equation: $p = \frac{e^y}{1 + e^y}$ e = 2.7183

Logistics Regression equation:

y = -9.94645 + 0.328827x

y = -9.94645+(0.328827)(40)

y = 3.20663

$p = e^y / (1 + e^y)$

$p = 2.7183^{3.2066} / (1+ 2.7183^{3.2066})$

p = 24.69572 / 25.69572

p = 0.9611 or 96.11%

Interpretation:

Probability 96.11% indicates that the error will be detected if QA spent 40 minutes of inspection during an observation.

Slide-16

Example: ***Step-5: Compile the Results***

Slide-17

Example: ***Step-5: Compile the Results***

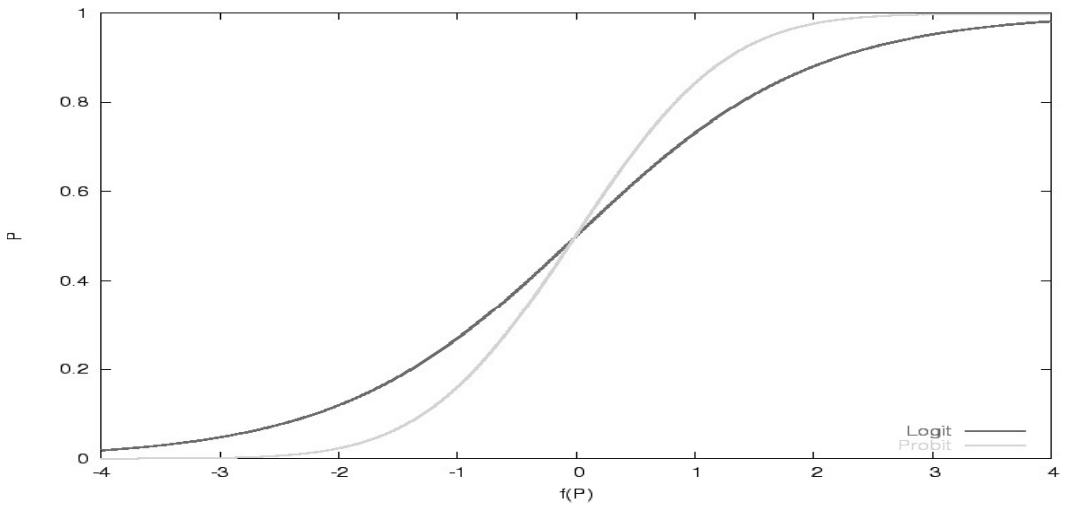
Inspection Time	Probabilities
38.000	0.928
29.000	0.399
34.000	0.774
35.000	0.827
32.000	0.640
41.000	0.972
35.000	0.827
41.000	0.972
24.000	0.114
28.000	0.323
32.000	0.640
36.000	0.869
20.000	0.033
21.000	0.046
39.000	0.947
42.000	0.979
27.000	0.256
20.000	0.033
37.000	0.902

This example deals with simple logistic regression model with one dependent variable and one independent variable. The same rule applies for Multiple Logistic Regression.

Slide-18

Logit and Probit Analysis:

In dealing with attribute data, when you want to estimate the likelihood that an event will take place, you can use tools like Logit and Probit analysis, which helps to explain the past and predict the future.



Slide-19

Logit and Probit Analysis:

- *Dependent Variable*
- *Independent Variable*

  Logit and Probit analyses help to estimate the maximum likelihood of an event's occurrence or non-occurrence, by predicting an attribute dependent variable from one or more independent variables.

Slide-20

Example:

Vehicle Purchase

You could use Logit and Probit analysis if you want to model the effect consumer characteristics have on the type of vehicle purchased. For instance, what type of consumer buys a sport utility vehicle, van, car, light pickup truck, or motorcycle?

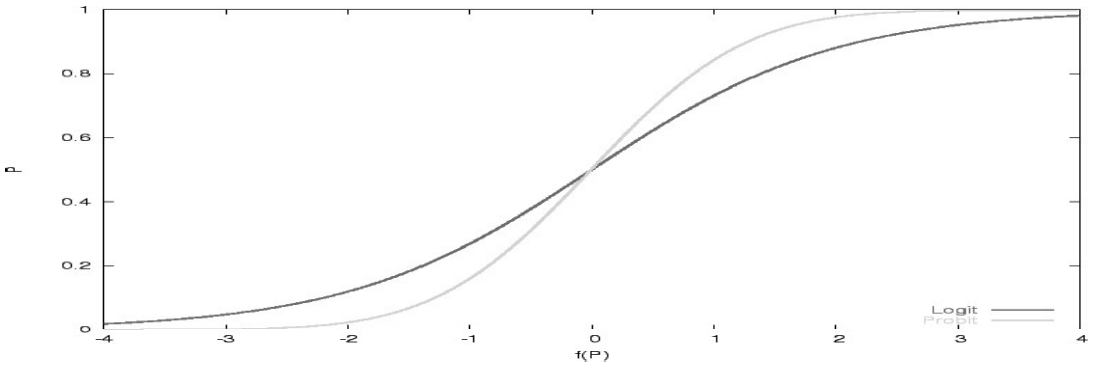
- Logit and Probit Assumptions:
  - Observations on y have been randomly sampled from the population.
  - Dependent variable y is caused by independent variable x's.
  - Relationship between y and x's is uncertain.
  - Assess the distribution of error terms

Slide-21

Differences:

Differences between Logit and Probit Analysis:

- The main difference is that the Logit model has slightly flatter tails; the normal or probit curve approaches the axes more quickly than the Logit curve.
- Although both deals with population parameters, the estimates might differ due to the difference in the purpose and methodologies of the two methods.

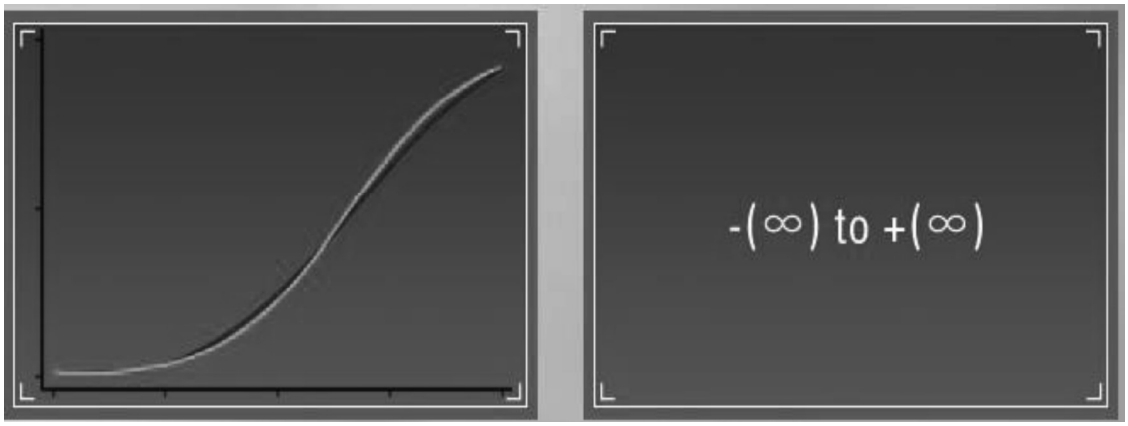


Slide-22

Logit Analysis:

Characteristics of Logit Analysis:

- Logit analysis produces results that are statistically sound.
    - It helps you avoid out of range estimates by transforming a dichotomous dependent variable into a continuous variable ranging from negative infinity to positive infinity.

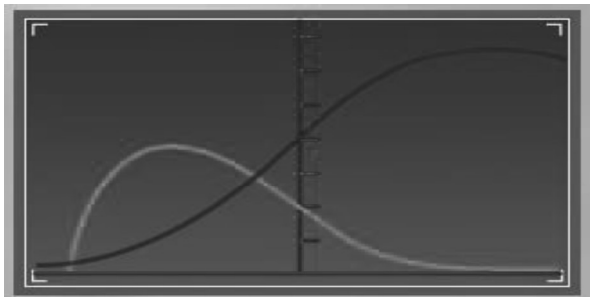


Slide-23

Logit Analysis:

Characteristics of Logit Analysis:

- Logit analysis results are easily interpreted and the method is easy to analyze.

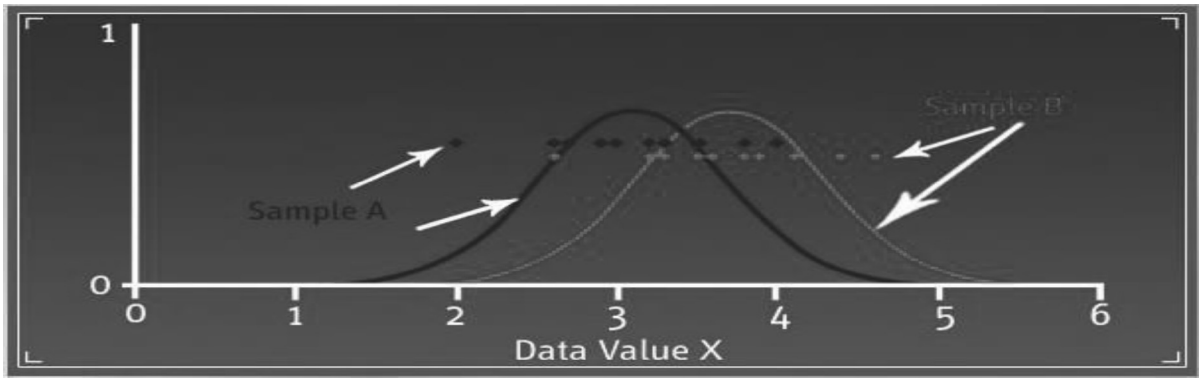


Slide-24

Logit Analysis:

Characteristics of Logit Analysis:

- Logit analysis allows you to apply the analog of the regression t-test by giving parameter estimates that are efficient and asymptotically consistent



Example:

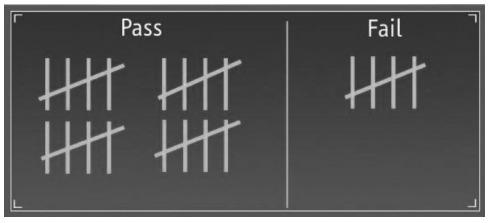
A marketing company might use logit analysis to test brand preference and loyalty for a certain product.

Slide-25

Logit Analysis:

Characteristics of Logit Analysis:

- It uses the odds to determine how much more likely it is that an observation will be a member of one category instead of another.



What exactly does this mean? A probability of p=.80 of being in-group A can be expressed in odds terms as four to one. In other words, there are four chances to pass versus one chance to fail.

Slide-26

Logit Analysis:

Logit Formula:

Logit formula

$$L = \text{logit} = \ln\left(\frac{p}{1 - p}\right)$$

The logit ranges from 0 to 1: $0 < L < 1$

The probability for a given L value is provided by the equation:

$$p = \frac{e^L}{1 + e^L}$$

Where,
p is the probability for a given L value
ln is the natural logarithm
e is a real number constant, the base of natural logarithm and equals 2.7183

Slide-27

Logit Analysis:

Logit Formula:

After knowing the p value we need to calculate the odds using the below formula.

$$\text{Odds} = \frac{p}{1-p}$$

Slide-28

Example:

Logit Analysis Example:

- Imagine that you are a Data Scientist at a very large scale integration circuit manufacturing company. You want to know whether or not the time spent inspecting each product impacts the quality assurance department's ability to detect a designing error in the circuit.

Observations	Inspection Time (minutes)	Error Detection (1 - Yes, 0 - No)
1	38	1
2	29	0
3	34	0
4	35	1
5	32	1
6	41	1
7	35	0
8	41	1
9	24	0
10	28	1
11	32	0
12	36	1
13	20	0
14	21	0
15	39	1
16	42	1
17	27	0
18	20	1
19	37	1
20	39	1
21	30	0

Slide-29

Example:

Logit Analysis Example:

- From 55 observation, only 27 of the defects are detected.

$$p = 27/55$$
$$p = 0.491$$

- Calculate the odds

$$\text{Odds} = \frac{p}{1-p} \qquad \text{Odds} = \frac{0.491}{0.509} = 0.965 \text{ or } 0.965{:}1$$

Slide-30

Example:

Logit Analysis Example:

Now recall the VLSI manufacturing example. Imagine that you know there is a 96.11% chance the error will be detected if QA spends 40minutes during inspection. Again, you could calculate the odds given the probability 0.9611 by dividing it by 1 minus 0.9611. The result is odds of 24.71 to 1

$$\text{Odds} = \frac{0.9611}{1-0.9611} = 24.71$$

Slide-31

Example:

Logit Analysis Example:

Calculate the Logit:

$$\text{Logit} = \ln(p/1-p)$$
$$\text{Logit} = \ln(24.71)$$
$$\text{Logit(L)} = 3.207$$

Calculate the Probability

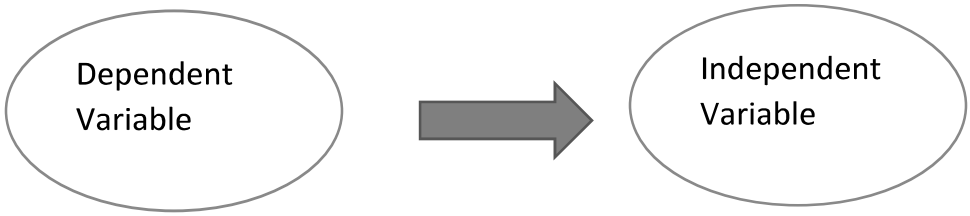
$$p = \frac{e^{3.207}}{1 + e^{3.207}} \qquad p = \frac{24.71}{25.71} = 0.9611$$

Slide-32

Probit Analysis:

Characteristics of Probit Analysis:

Probit analysis is a specialized regression model used with binomial dependent variables. It employs a Probit link function.



Slide-33

Probit Analysis

Characteristics of Probit Analysis:

- It is very similar to Logit analysis, but uses cumulative normal probability distribution and is preferred when data are normally distributed. It transforms the sigmoid regression curve to a straight line that can be analyzed by

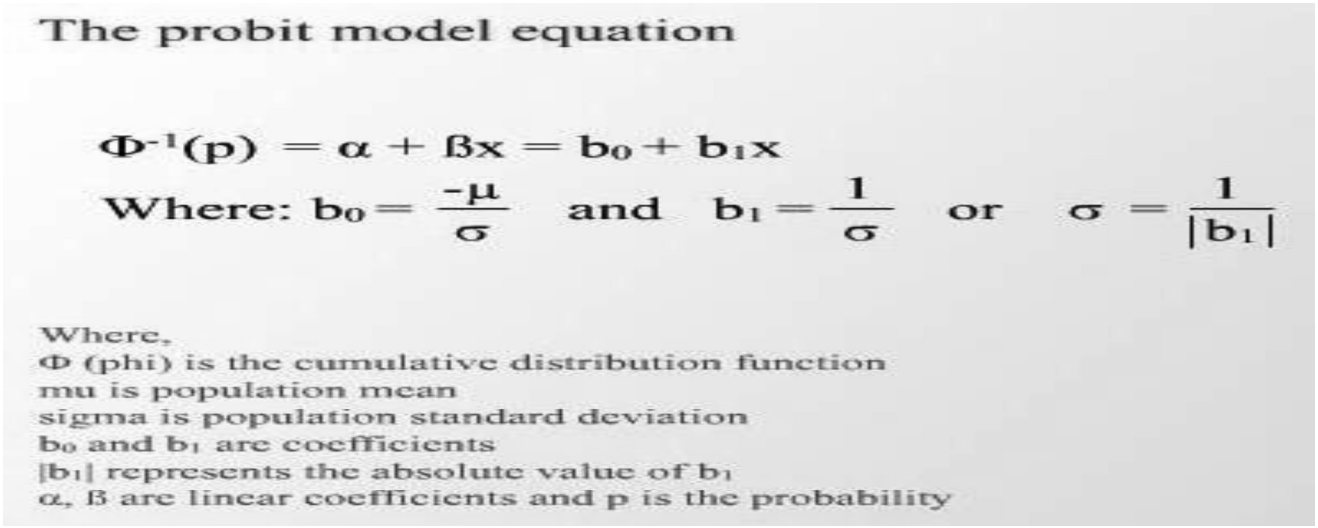
regression using either least squares or maximum likelihood method.

- Probit analysis is often used in reliability testing study involving destructive testing. Automotive and other product manufacturing companies use probit analysis

Slide-34

Probit Analysis:

Probit Formula:



Slide-35

Probit Analysis:

Probit Analysis Outputs:

- Goodness - of – fit
  - The goodness of fit test assesses whether an assumed distribution fits the data adequately

- Probability Plot
- Probability plot is used to determine if an assumed distribution fits the data. The closer the points fall to the fitted line in the plot, the better the fit.

Slide-36

Probit Analysis:

Probit Analysis Outputs:

- Table of Percentiles
  - The table of percentiles shows you what percentage of an event or non-event takes place at what level

Percent	Percentile	Std.Error	Lower	Upper
1	967.358	475.656	447.839	3514.07
2	547.785	229.358	284.102	1640.99
3	392.084	147.408	217.316	1049.06
4	308.884	106.749	179.465	762.640
5	256.463	82.5839	154.559	594.861
6	220.130	66.6357	136.689	485.113
7	193.322	55.3612	123.118	407.950
8	172.647	46.9929	112.388	350.853
9	156.168	40.5515	103.645	306.967
10	142.693	35.4515	96.3549	272.224
20	77.4379	13.4975	58.5786	121.115
30	52.8508	6.93546	42.6406	73.4859
40	39.4449	4.03922	33.1534	50.5454
50	30.7629	2.58183	26.4388	37.2397
60	24.5078	1.85505	21.1042	28.6869

Slide-37

Example:

Metal Manufacturer:

- Imagine that you are a Data Scientist at a Metal Manufacturer. You are validating the product's capability of sustaining stress during transportation.
    - You conduct number of destructive tests and collect samples.
    - You set the pressure machine at various pressure levels and record the breaking response- binary data.
    - If doesn't break-success recorded as one, if breaks-failure 0

Slide-38

Example:

Metal Manufacturer:

Step-1: Tested 50 samples with ranging pressure form 10 to 50n/m2

Pressure (N/m^2)	Result 1 = survives, 0 = fails	Frequency
10	1	45
10	0	5
20	1	38
20	0	12
30	1	30
30	0	20
40	1	20
40	0	30
50	1	10
50	0	40

Slide-39

Example:

Metal Manufacturer:

Step-2: Stress test data, use your statistical software to calculate that the coefficient beta0 equals 3.84018 and beta1 equals -1.22776

Step-3: *Goodness-of-fit test*

Method	Chi-square	DF	p
Pearson	6.99	3	0.07

- p value is 0.07 which is greater than alpha value 0.05. There is no difference between observed and predicted, so we conclude that Probit model fits the data.

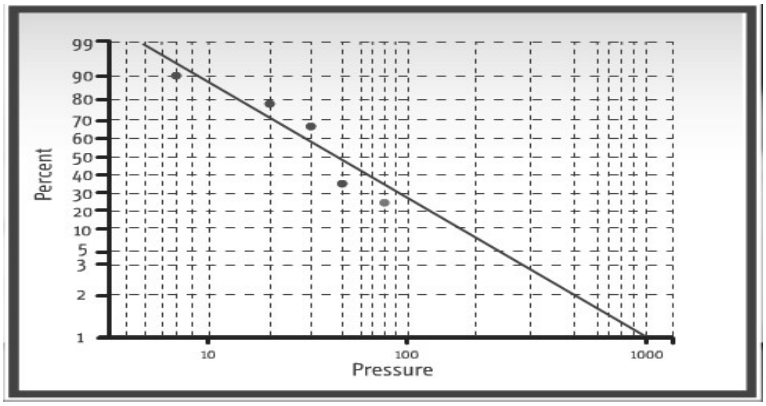
Slide-40

Example:

Metal Manufacturer:

Step-4: *Probability Plot*

- *Probability Plot clears says that model fits the data adequately. Data points fall close to the fitted line.*



Slide-41

Example:

Metal Manufacturer*:*

Step-4: *Table of Percentiles*

- Table of percentiles tells you what percentage of an event or non-event takes place at what pressure level.
- Interpretation: At 77.43 newton per square meter pressure, only 20 cans survive.

Percent (Binary response)	Percentile (Continuous input)
1	967.358
2	547.785
3	392.084
4	308.884
5	256.463
6	220.130
7	193.322
8	172.647
9	156.168
10	142.693
20	77.4379
30	52.8508
40	39.4449
50	30.7629
60	24.5078
70	19.6209
80	15.4898
90	11.5707
91	11.1566
92	10.7310
93	10.2901
94	9.82845
95	9.33843
96	8.80767
97	8.21449
98	7.51408
99	6.57921

Slide-42

Logistic Function:

- $P_i=E(Y=1| X)= \beta_0+ \beta_1D_1+ \beta_2D_2+ \beta_3D_3+ \beta_4D_4+ \beta_6X$ - LPM representation
- Instead we represent:

  $P_i=E(Y_i=1| X)= 1/ (1+e^{-(\beta0+ \beta1D1+ \beta2D2+ \beta3D3+ \beta4D4+ \beta6Xi)})$
- For exposition we write:

  $P_i=E(Y_i=1| X)= 1/ (1+e^{-(z)})= e^z/(1+e^z)$

  Where $z_i= \beta_0+ \beta_1D_1+ \beta_2D_2+ \beta_3D_3+ \beta_4D_4+ \beta_6X_i$
- Above equation represents the Cumulative Logistic Distribution Function
- As z_i ranges from $-\infty$ to $+\infty$, P_i ranges between 0 and 1
- P_i non-linearly related to X

– The logistic curve resembles an S-shape.

Slide-43

Logistic Regression:

- Looks like we have hit the jackpot!- The logistic curve addresses all the above problems
- But how to estimate the model? Non-linear in parameters
- Any linearizing transformation?
  - Taking logarithm on both sides doesn't linearize it
  - Take ratio: $P_i/(1-P_i) = e^{zi}$
  - Then take logarithm on both sides:

$$L_i = \ln(P_i/(1-P_i)) = Z_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_6 X_i$$

This is called the Logit model/ Logistic Regression

Slide-44

Logistic Regression:

– This P_i is the probability that a claimant is not represented by an attorney

– Thus $(P_i/1-P_i)$ is simply the ratio of the probability that a claimant is not represented by an attorney to the probability that he is

- This ratio is called the odds ratio. Eg: Odds ratio=2 means that the odds are 2 to 1 in favor of not being represented by an Attorney

Useful features of the Logit Model:

- As P_i goes from 0 to 1, L_i goes from -∞ to +∞

Slide-45

Logistic Regression:

Refer to model output in R

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}}$$

$$Odds_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}$$

$$\log odds_i = ^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}}$$

Slide-46

Logistic Regression:

- Model output interpretation based on

- Consider the “claimant” dataset
- Dummy Variable ATTORNEY: Attorney = 0, if yes
  - = 1, if not
- Predict the outcome whether claimant is represented by an attorney or not on the following-
- Claimant’s age - CLMAGE (D_1)(in years)

- Claimant's sex - CLMSEX(D_2)(0 if Male, 1 if Female)
- Whether the claimant was wearing seatbelt - SEATBELT (D_3) (0 if yes, 1 if no)
- Whether the driver of the claimant's vehicle was uninsured - CLMINSUR (D_4) (0 if yes, 1 if no)
- The claimant's total economic loss (in thousands) – LOSS (X)

Specify the regression:

$$Y= \beta_0+ \beta_1D_1+ \beta_2D_2+ \beta_3D_3+ \beta_4D_4+ \beta_5X+ \varepsilon$$

Slide-47

Logistic Regression – Model Fit:

Residual deviance and AIC

Residual deviance = -2 log likelihood

AIC = Residual deviance + 2 number of parameters

Goodness of Fit: Hosmer – Lemeshow Test

$$\sum_{k=0}^{1} \sum_{l=1}^{g} \frac{(o_{kl}-e_{kl})^2}{e_{kl}}$$

Efron's / McFadden's / Cox & Snell

Slide-48

Confusion Matrix

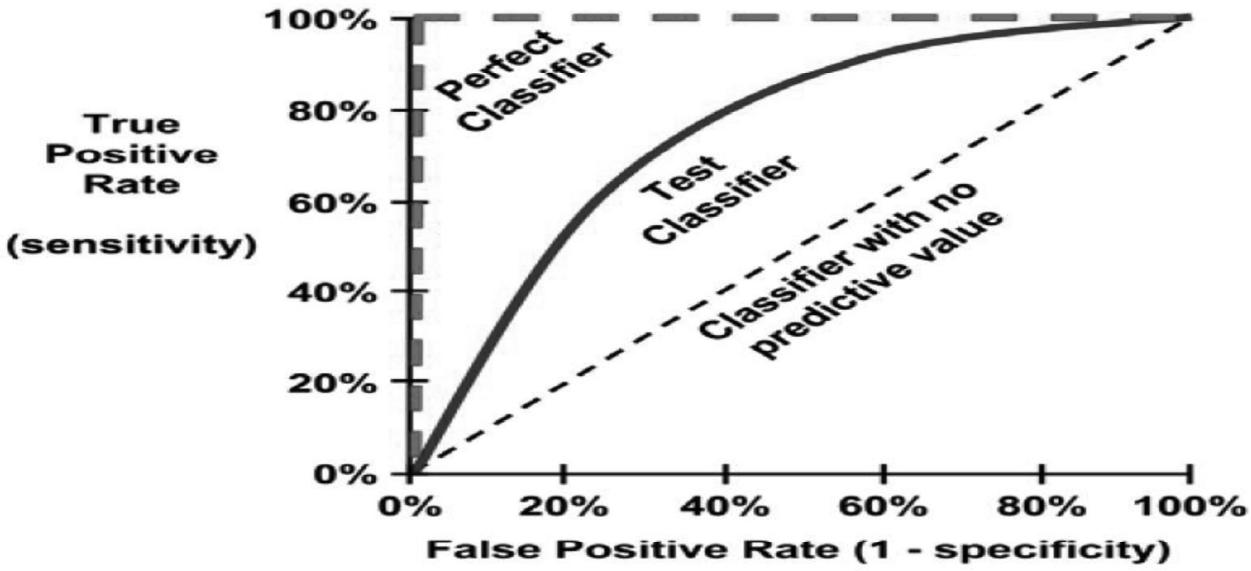
	Disease Present	Disease Absent

Test positive	True positive	False positive
Test negative	False negative	True negative

Slide -49

ROC – Receiver Operating Characteristic

- History is WW II to differentiate between true signals & false alarms



	Disease present	Disease absent
Test positive	True positives	False positives
Test negative	False negative	True negatives

0.9 - 1.0 = A (outstanding)

0.8 - 0.9 = B (excellent/good)

0.7 - 0.8 = C (acceptable/fair)

0.6 - 0.7 = D (poor)

0.5 - 0.6 = F (no discrimination)

Slide-50

Forecasting – Logistic Regression

Melbourne Rainfall Dataset

Potential Predictors:

Annual seasonality (sine, cosine)

Previous day(s) Rain indicator or rainfall amount

How about linear regression?

$Rain_t$ =

$b_0 + b_1 Rain_{t-1} + b_2 \sin(2pt/365.25) + b_3 \cos(2pt/365.25) + e$

Capturing Annual Seasonality

Slide-51

Forecasting – Logistic Regression

$Rain_t$ =

$b_0 + b_1\ Rain_{t-1} + b_2\ sin\ (2pt/365.25) + b_3\ cos\ (2pt/365.25) + e$

Replace with a function of "Rain" that guarantees forecasts in range [0, 1] and give probability of rain.

Slide-52

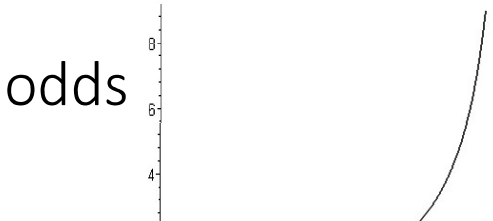
Forecasting – Logistic Regression

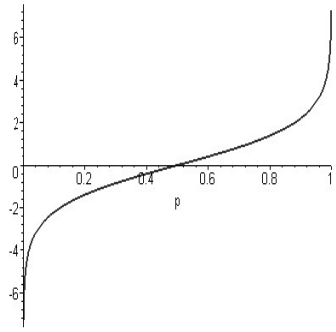
Output variable:

$Rain_t$ (binary variable)

p =Prob($Rain_t$ =1)

$$odds(Rain_t = 1) = \frac{p}{1-p}$$



logit 

Slide-53

Forecasting – Logistic Regression

Logit ($Rain_t$=1)

=

$b_0 + b_1 Rain_{t-1} + b_2 sin(2pt/365.25) + b_3 cos(2pt/365.25)$

Odds ($Rain_t$=1)

=

$$e^{b_0 + b_1\, Raint_{-1} + b_2\, sin\ (2pt/365.25) + b_3\, cos\ (2pt/365.25)}$$

Prob($Rain_t$=1)

=

$$\frac{\mathbf{1}}{\mathbf{1+}\ e^{-}\ \{b_0 + b_1\ Raint_{-1+}b_2 sin(2pt/365.25) + b_3 cos(2pt/365.25)\}}$$

$$Prob(Rain_t = 1) = \frac{1}{1 + e^{-logit}}$$