LEARNING OBJECTIVES

Upon completing this session, you should be able to do the following:

- State the circumstances under which logistic regression should be used instead of multiple regression.
- Identify the types of dependent and independent variables used in the application of logistic regression.
- Interpret the results of a logistic regression analysis and assessing predictive accuracy
- Understand the strengths and weaknesses of logistic regression compared to multiple regression.

LOGISTIC REGRESSION DEFINED

- Logistic Regression . . . is a specialized form of regression that is designed to predict and explain a binary (two-group) categorical variable rather than a continuous dependent measure.
- Similar to regular regression the independent variables may be either continuous or categorical.

EXAMPLE
OBAMA V. ROMNEY 2012

- Question:
  - Who is going to vote for Obama in 2012 US presidential election?
  - Specifically, does party membership and race matter?
- Data come from a national survey study by ANES (American National Election Studies)

DATA SNAPSHOT

OBAMA V. ROMNEY 2012

ID	ft_dem	black	vote_obama
1	40	0	0
2	85	0	1
3	15	0	0
4	15	0	0
5	40	0	0
6	85	1	1
7	30	0	0
8	100	1	1
9	20	0	0
10	0	0	0
11	40	1	1
12	50	0	1
13	60	0	1
14	30	0	0
15	50	0	0
16	15	0	0
17	75	0	1
18	5	0	1
19	70	0	1
20	85	1	1

n = 3,858

OLS REGRESSION RESULTS

OBAMA V. ROMNEY 2012

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i$$

$$\text{Vote\_Obama} = b_0 + b_1 \times \text{ft\_dem} + b_2 \times \text{black} + e$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.741
R Square	0.549
Adjusted R Square	0.548
Standard Error	0.331
Observations	3858

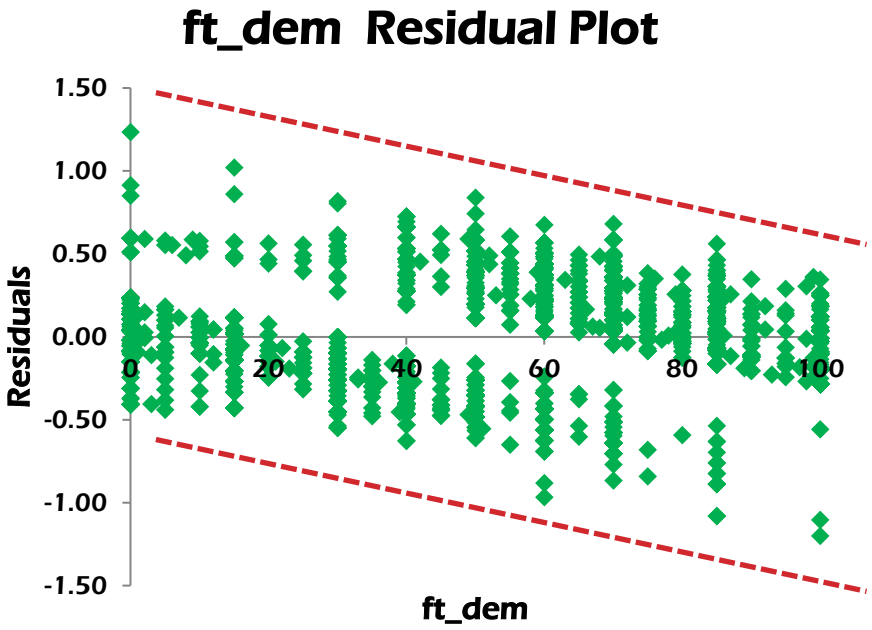
ANOVA

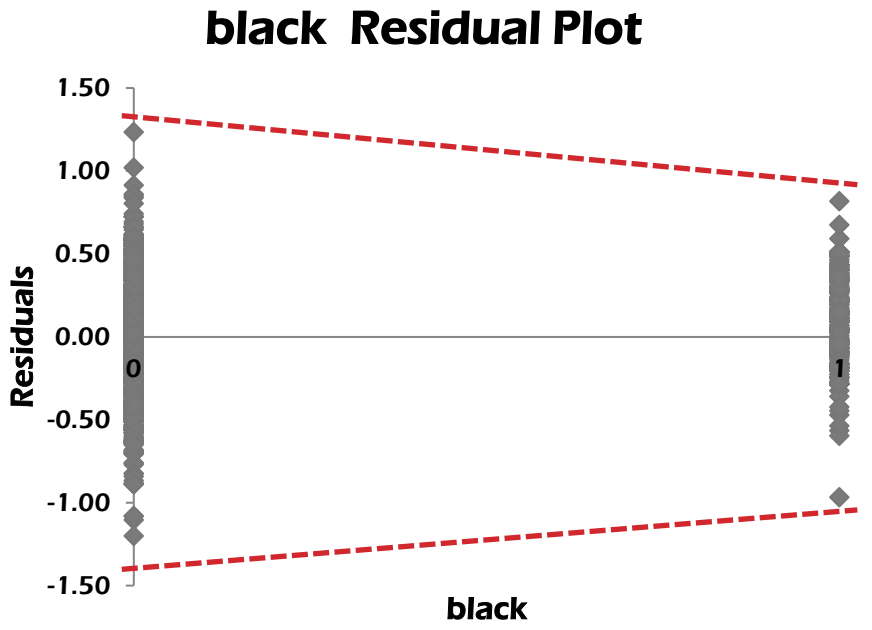
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	513.179	256.589	2343.637	0.000
Residual	3855	422.059	0.109		
Total	3857	935.237			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.065	0.011	-5.780	0.000	-0.087	-0.043
ft_dem	0.012	0.000	59.136	0.000	0.011	0.012
black	0.113	0.015	7.447	0.000	0.083	0.143

MODEL DIAGNOSTICS

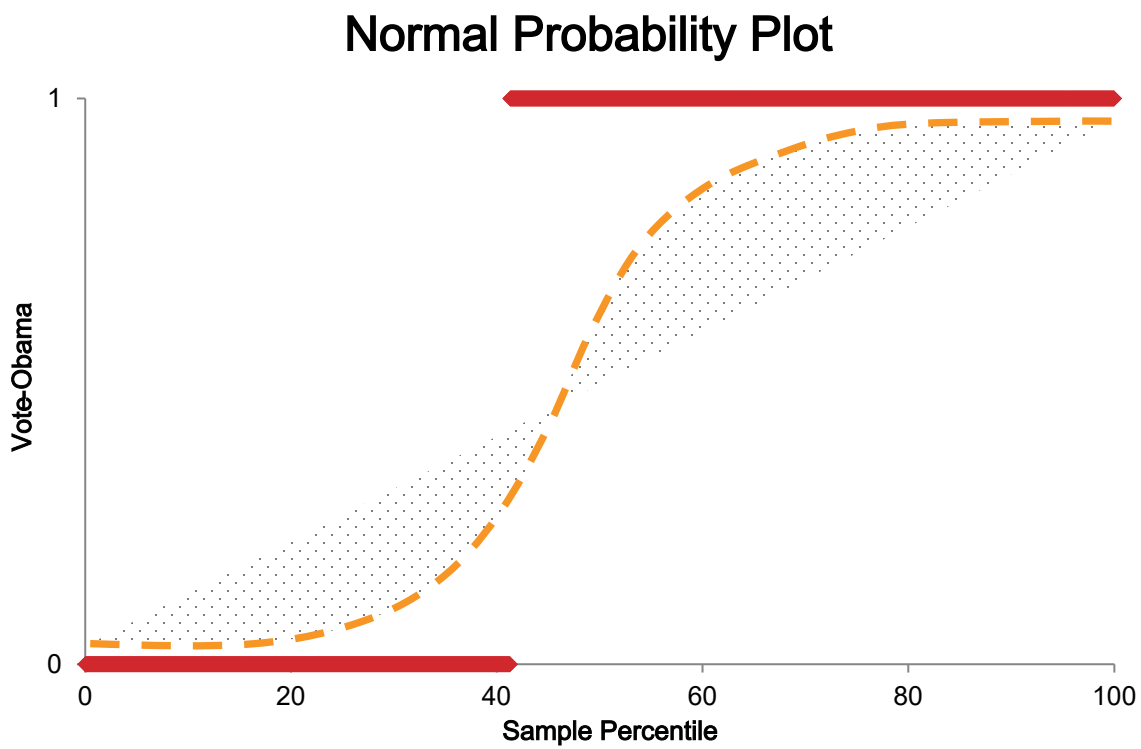
INDEPENDENCE OF ERRORS & EQUAL VARIANCES





MODEL DIAGNOSTICS

NORMALITY OF RESIDUALS



WHY NOT OLS REGRESSION?

- In Obama v. Romney case, the assumptions of linear regression are not valid:
  - Underlying relationships between Xs and Y are NOT linear.
  - Variance of residuals 'ε', is NOT the same for all values of the independent variable (i.e. heteroscedasticity).
  - Values of residuals 'ε' are NOT independent.
  - Residuals 'ε' are NOT normally distributed.
- If proceed to fit a linear model:
  - Magnitude of the effects of the independent variables may be greatly underestimated.

WHY LOGISTIC REGRESSION?

- The advantages of logistic regression are primarily the result of the general lack of assumptions.
  - Logistic regression does not require any specific distributional form for the independent variables.
  - Homoscedasticity of residuals is not required.
  - Linear relationships between the dependent and independent variables are not required.

WHAT IS LOGISTIC REGRESSION?

- Logistic regression is based on one principal:

  It expresses the multiple linear regression equation in logarithmic terms (called the *logit*) and thus overcomes the assumption of linearity.

Multiple Linear Regression Function ⇨ $$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i$$

Logistic Regression Function ⇨ $$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i})}}$$

HOW DOES LOGISTIC REGRESSION WORK?
ASSESSING THE MODEL

- Logistic regression calculates the probability of and event P(Y) occurring for a given person based on observation of whether or not the event (Y) did occur for than person.
- Probability of a voter to vote for Obama =

  P(Y) (Probability of the event)
- Whether or not the voter actually votes for Obama =

  Y (actual event)

P(Y) could be a value between 0 and 1.

Y could be either 0 (not Obama) or 1 (Obama).

LOG-LIKELIHOOD

$$\mathbf{log-likelihood} = \sum_{i=1}^{n} [Y_i \ln(P(Y_i)) + (1 - Y_i) \ln(1 - P(Y_i))]$$

- The log-likelihood is based on summing the probabilities associated with the predicted and actual outcomes.
- The log-likelihood statistic is analogous to the residual sum of squares (SSR) in multiple regression in the sense that it is an indicator of how much unexplained information there is after the model has been fitted.

DEVIANCE (-2LL)

$$\text{Deviance} = -2 \times \log - \text{likelihood}$$

- Deviance is often referred to as -2LL.
- It is convenient to use deviance rather than log-likelihood since it has a chi-square distribution, which makes it possible to calculate significance of the value.

-2LL(baseline) = a logistic regression model in which only the constant is included

-2LL(New) = a logistic regression model in which one dependent variable is added to the baseline model.

In logistic regression, we merely take the new model deviance and subtract from it the deviance for the baseline model:

Likelihood Ratio = [-2LL(baseline) – (-2LL(new))]

COMPARISON TO OLS REGRESSION...

Correspondence of Primary Elements of Model Fit

OLS Multiple Regression	Logistic Regression
Total Sum of Squares	-2LL of Baseline Model
Error Sum of Squares	-2LL of New Model
Regression Sum of Squares	Likelihood Ratio
F test of model fit	Chi-square test of Likelihood Ratio
Coefficient of determination (R^2)	"Pseudo" R^2 measures

"PSEUDO" R^2 MEASURES

- Hosmer and Lemeshow's R_L^2

$$\mathbf{R_l^2} = \frac{-\boldsymbol{\chi^2_{Model}}}{\mathbf{-2LL(baseline)}}$$

- Cox and Snell's R_{CS}^2

$$\mathbf{R_{CS}^2} = \frac{\mathbf{(-2LL(baseline))} - \mathbf{(-2LL(new))}}{\mathbf{-2LL(baseline)}}$$

This measure has limitations as it could never reach 1.00.

- Nagelkerke's R_N^2

$$\mathbf{R_{CS}^2} = \mathbf{1} - \mathbf{exp}\left(\frac{\mathbf{(-2LL(new))} - \mathbf{(-2LL(baseline))}}{\boldsymbol{n}}\right)$$

Both R_{CS}^2 and R_N^2 reflect the amount of variation accounted for by the logistic model, with 1.00 indicating a perfect fit.

LOGISTIC REGRESSION RESULTS

OBAMA V. ROMNEY 2012

CLASSIFICATION TABLE

	Suc-Obs	Fail-Obs	Total
Suc-Pred	1979	166	2145
Fail-Pred	286	1427	1713
Total	2265	1593	3858
Accuracy	0.874	0.896	0.883
Cut-off			*0.50*

Assessing Practical Significance of the Model

SUMMARY OUTPUT

R^2_L	0.553
R^2_{CS}	0.527
R^2_N	0.710

OVERALL MODEL FIT

Chi-Sq	2890.716
df	2
p-value	0.000
alpha	0.050
sig	yes

Assessing Overall Model Fit (Statistical Significance)

Assessing the Contribution of Predictors

Variables	B_i	SE	Wald	df	Significance	EXP(B_i)
Intercept	-4.868	0.181	724.618	1	0.000	0.008
ft_dem	0.096	0.003	837.098	1	0.000	1.101
black	3.214	0.357	80.950	1	0.000	24.876

CLASSIFICATION TABLE
(CONFUSION MATRIX)

CLASSIFICATION TABLE

	Suc-Obs	Fail-Obs	Total
Suc-Pred	1979	166	2145
Fail-Pred	286	1427	1713
Total	2265	1593	3858
Accuracy	**0.874**	**0.896**	**0.883**
Cut-off			*0.50*

- It is an indicator of classification accuracy.
- For equal group sizes, cut-off is **0.50**;

C_{EQUAL} = 1 ÷ Number of Groups

- For unequal group sizes, cut-off can be calculated as:
  - Maximum Chance Criterion: C_{MAX} = Proportion of individuals in the larger group
  - Proportional Chance Criterion: $C_{PRO} = p^2 + (1 - p)^2$; where p is proportion of individuals in group 1
  - How high does hit ratio have to be? Hit ratio should be at least one-fourth larger than that achieved by chance (e.g. 50%).

R^2 MEASURES

SUMMARY OUTPUT	
R^2_L	0.553
R^2_{CS}	0.527
R^2_N	0.710

- These measures are comparable to R^2 measure in multiple regression and are based on reduction in -2LL value.
- For Obama v. Romney model, Pseudo, Cox and Snell, and Nagelkerke R^2 values are 55%, 52% and 71%. These indicate that more than half (two third in the case of R^2_N) of variation between two groups of voters can be explained by the logistic regression model.

INTERPRETING LOGISTIC COEFFICIENTS

Variables	B_i	SE	Wald	df	Significance	EXP(B_i)
Intercept	-4.868	0.181	724.618	1	0.000	0.008
ft_dem	0.096	0.003	837.098	1	0.000	1.101
black	3.214	0.357	80.950	1	0.000	24.876

- Significance of the Coefficients

  All predictors (IVs) are significant at *p*-value = .001 and therefore can be used to identify the relationships affecting the predicted probabilities and subsequently group membership.

- Direction of the Relationships

  Original logistic coefficients (B_i) signs indicates the direction of relationship with negative sign indicating a negative relationship and vice versa.

  For instance, as positive feeling towards democratic party increases, the predicted probability will increase and thus the likelihood that an individual will vote for Obama.

INTERPRETING LOGISTIC COEFFICIENTS

Variables	B_i	SE	Wald	df	Significance	EXP(B_i)
Intercept	-4.868	0.181	724.618	1	0.000	0.008
ft_dem	0.096	0.003	837.098	1	0.000	1.101
black	3.214	0.357	80.950	1	0.000	24.876

- Magnitude of the Relationships

  The most direct way of assessing magnitude of relationships is by examining exponentiated coefficients (last column) and using this straight forward formula:

  Percentage Change in Odds = (Exponentiated Coefficients – 1.0) × 100

  One unit if increase in positive feeling towards democratic party increases the odds of voting for Obama by 10.1 percent.

USING LOGISTIC REGRESSION FOR PREDICTION

Variables	B_i	SE	Wald	df	Significance	EXP(B_i)
Intercept	-4.868	0.181	724.618	1	0.000	0.008
ft_dem	0.096	0.003	837.098	1	0.000	1.101
black	3.214	0.357	80.950	1	0.000	24.876

- Predict the probability of voting for Obama for a White-American voter who scored 50 Democratic Party Feeling thermometer?

Three step process for predicting the probability based on a set of given values of independent variables

1. Calculate Logit = $b_0 + b_1X_{1i} + b_2X_{2i}$
2. Calculate Odds = e^{Logit}
3. Calculate Predicted Probability = $\frac{\text{Odds}}{(1+\text{Odds})}$

$$-4.868 + 0.096 \times \text{ft\_dem} + 3.214 \times \text{black}$$

$$-4.868 + 0.096 \times (50) + 3.214 \times (0) = -0.068$$

$$e^{-0.068} = 0.934$$

$$\frac{0.934}{(1+0.934)} = 0.438$$

The probability of this particular type of voter (i.e. White American with a score of 50 on ft-Dem) to vote for Obama is 0.438 (or 43.8%).

USING LOGISTIC REGRESSION FOR PREDICTION

Variables	B_i	SE	Wald	df	Significance	EXP(B_i)
Intercept	-4.868	0.181	724.618	1	0.000	0.008
ft_dem	0.096	0.003	837.098	1	0.000	1.101
black	3.214	0.357	80.950	1	0.000	24.876

- Similarly, predict the probability of voting for Obama for an African-American voter who scored 50 Democratic Party Feeling thermometer?

1. Calculate Logit = $b_0 + b_1X_{1i} + b_2X_{2i}$
2. Calculate Odds = e^{Logit}
3. Calculate Predicted Probability = $\frac{\text{Odds}}{(1+\text{Odds})}$

-4.868 + 0.096× ft_dem + 3.214 × black

-4.868 + 0.096× (50) + 3.214 × (1.0) → **3.146**

$e^{-0.068}$ → **23.243**

$\frac{0.934}{(1+0.934)}$ → **0.958**

The probability of this particular type of voter (i.e. African-American with a score of 50 on ft-Dem) to vote for Obama is 0.958 (or 95.8%).

IMPLICATION...

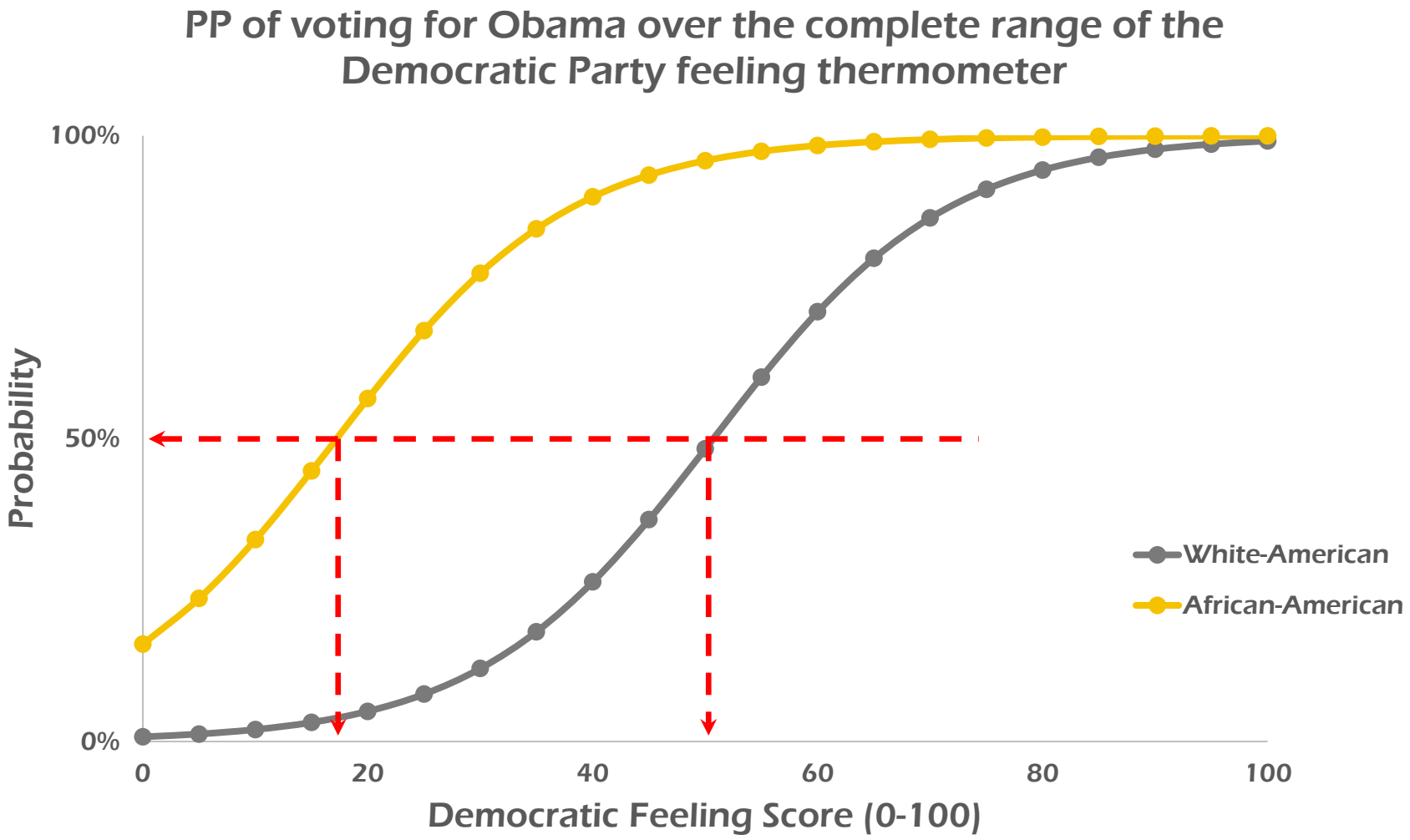
	Estimated Probability of Voting for Obama					
White-American	0.031	0.120	0.366	0.709	0.911	0.978
African-American	0.447	0.773	0.935	0.984	0.996	0.999
	15	30	45	60	75	90

Democratic Feeling Thermometer

- Table above shows the estimated probabilities of voting for Obama in 2012 presidential election for African-American and White-American voters with different levels of positive feelings towards Democratic party.

- How can Obama team use this information to better target voters in their election campaigns?

IMPLICATION...
VISUALISING PREDICTED PROBABILITIES



AN ILLUSTRATIVE EXAMPLE
BLITZ COUPON

- BLITZ promotional campaign includes a catalogue with a $50 discount coupon on purchases of $200 or more. The catalogues are expensive and BLITZ would like to send them only to those customers who have the highest probability of using the coupon. The management thinks that annual spending at BLITZ and whether the customer has BLITZ credit card are the two variables that might be helpful in predicting whether a customer who receives the catalogue will use the coupon.

- BLITZ conducted a pilot study using a random sample of 50 BLITZ credit card customers and 50 other customers who do not have a BLITZ credit card. BLITZ sent the catalogue to the selected 100 customers and at the end of the trial period noted whether the customers used the coupon.

DATA SNAPSHOT
BLITZ COUPON

ID	Spending ($,000)	BLITZ card	Coupon
1	$2.29	1	0
2	$3.22	1	0
3	$2.14	1	0
4	$3.92	0	0
5	$2.53	1	0
6	$2.47	0	1
7	$2.38	0	0
8	$7.08	0	0
9	$1.18	1	1
10	$3.35	0	0
11	$2.14	1	0
12	$3.26	0	1
13	$1.51	0	0
14	$2.15	0	1
15	$6.74	0	0
16	$6.49	0	0
17	$1.31	0	0
18	$3.47	1	0
19	$2.94	0	0
20	$6.40	0	1

n = 100

Which type of customer is most likely to take advantage of BLITZ promotional campaign?

AN ILLUSTRATIVE EXAMPLE
BLITZ COUPON

- Estimate the probability of using a coupon for customers who spend $2000 annually and do not have a BLITZ credit card?

- Similarly estimate the probability of using a the coupon for customers who spend $2000 last year and have a BLITZ credit card?

MANAGERIAL USE

	Estimated Probability of Using the Coupon						
With BLITZ card	.330	.410	.494	.579	.659	.731	.793
No BLITZ card	.141	.188	.245	.314	.392	.475	.561
	$1,000	$2,000	$3,000	$4,000	$5,000	$6,000	$7,000
	Annual Spending						

- Table above shows the estimated probabilities for values of annual spending ranging from $2000 to $7000 for both customers who have a BLITS credit card and customers who do not have a BLITZ credit card.
- How can we use this information to better target customers for the promotional catalogue?

IMPLICATION...
VISUALISING PREDICTED PROBABILITIES