

# Logistic Regression





# Learning Objectives

- ◆ Rationale for Logistic Regression
- ◆ Identify the types of variables used for dependent and independent variables in the application of logistic regression
- ◆ Describe the method used to transform binary measures into the likelihood and probability measures used in logistic regression
- ◆ Interpret the results of a logistic regression analysis & assessing predictive accuracy
- ◆ Strengths & weakness of logistic regression



# Chapter Preview

- ◆ Logistic Regression (LR) is the appropriate statistical technique when the dependent variable is a categorical (nominal/ non-metric) variable and the independent variables are metric/ non-metric variables
- ◆ LR has the advantage of being less affected, when the basic assumptions, particularly normality of the variables are not met
- ◆ LR may be described as estimating the relationship between a single non-metric (binary) dependent variable & a set of metric/ non-metric independent variables, in this general form:

$$Y_1 = X_1 + X_2 + \dots + X_N$$

(Binary non-metric)

(Non-metric & metric)

- ◆ LR has widespread application in situations in which the primary objective is to identify the group to which an object (e.g., person, firm or product) belongs, where the outcome is binary (yes/ no)
- ◆ Situations include deciding whether a person should be granted credit/ predicting whether a firm will be successful/ success or failure of a new product
- ◆ Objective is to predict & explain the basis for each objects group membership through a set of independent variables selected by the researcher



# Decision Process for Logistic Regression

- ◆ Application of LR can be viewed from a six-stage model-building perspective.
  1. Setting the Objectives of LR
  2. Research design for LR
  3. Underlying assumptions of LR
  4. Estimation of the LR and assessing overall fit
  5. Interpretation of the results
  6. Validation of the results

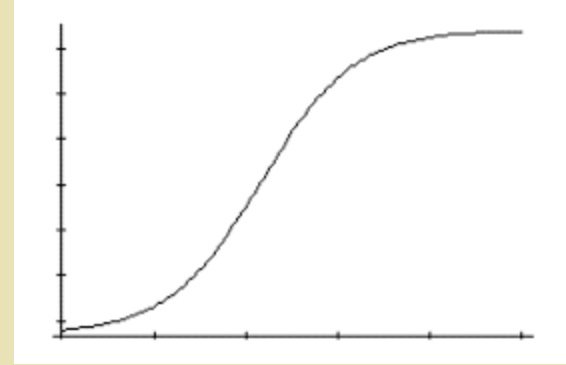
# Contd...

## ◆ Stage 1: Objectives of LR

- LR best suits to address 2 objectives:
  - Identifying the independent variables
  - Establishing a classification system
- In the classification process, LR provides a basis for classifying not only the sample, but also, any other observations that have values for all independent variables, into defined groups

## ◆ Stage 2: Research Design for LR

- Representation of binary dependent variable – LR represents binary variables with values 0 and 1
- Use of the logistic curve – LR uses logistic curve to represent the relationship between the independent & dependent variables
- Unique nature of dependent variable – First, the error term of a discrete variable follows binomial distribution and second, the variance of a dichotomous variable is not constant
- Overall sample size – LR uses maximum likelihood (MLE) as the estimation technique. MLE requires larger samples
- Sample size per category of the dependent variable – Recommended sample size for each group is at least 10 observations







# Contd...

## ◆ Stage 3: Assumptions of LR

- Lack of assumptions required
- Doesn't require linear relationships between the independent & dependent variables
- Can also address non-linear effects

## ◆ Stage 4: Estimation of the LR model & assessing overall fit

- **How do we predict group membership from the logistic curve?**
  - For each observation, LR technique predicts a probability value between 0 & 1
  - Predicted probability is based on the value(s) of the independent variable(s) & the estimated coefficients
  - If the predicted probability is  $>0.50$ , outcome is 1; otherwise, outcome is 0
- **How can we ensure that estimated values do not fall outside the range 0-1?**
  - In original form, probabilities are not restricted to values between 0 & 1
  - We restate the probability by expressing a probability as **odds** – the ratio of the probability of the 2 outcomes,  $p(i) / (1 - p(i))$
- **How do we keep the odds values from going below 0?**
  - Solution is to compute the **logit value**, which is calculated by taking the logarithm of the odds
  - Odds ratio  $< 1$ , will have negative logit value; odds ratio  $> 1$  will have positive logit values, and odds ratio of 1.0 has a logit value of 0



# Contd...

- **Estimating the coefficients** – independent variables are estimated using either the logit value or the odds value as the dependent measure

$$\text{Logit}_i = \ln \left( \frac{\text{prob}_{\text{event}}}{1 - \text{prob}_{\text{event}}} \right) = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n \text{ or}$$
$$\text{odds}_i = \left( \frac{\text{prob}_{\text{event}}}{1 - \text{prob}_{\text{event}}} \right) = e^{b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n}$$

- The non-linear nature of the logistic transformation requires the maximum likelihood procedure to be used in an iterative manner to find the most likely estimates for the coefficients
- LR maximizes the likelihood that an event will occur
- **Assessing Goodness-of-Fit (GoF) of the estimated model**
  - GoF for a LR model can be assessed in 2 ways:
    - Assess model estimation fit using pseudo R<sup>2</sup> values
    - Examine predictive accuracy
  - Likelihood Value – How well the maximum likelihood estimation procedure fits
  - LR measures model estimation fit with the -2 times the log of the likelihood value, referred to as -2LL or -2 Log Likelihood
  - Minimum value for -2LL is 0, which corresponds to perfect fit

# Contd...

- **Pseudo R<sup>2</sup>** value is interpreted in a manner similar to the coefficient of determination. The pseudo R<sup>2</sup> for a logit model (R<sup>2</sup><sub>LOGIT</sub>) can be calculated as

$$R^2_{LOGIT} = \frac{-2LL_{null} - (-2LL_{model})}{-2LL_{null}}$$

- Logit R<sup>2</sup> value ranges from 0.0 to 1.0
  - Perfect fit has a -2LL value of 0.0 & a R<sup>2</sup><sub>LOGIT</sub> of 1.0
  - Higher values of 2 other R<sup>2</sup> measures (Cos & Snell) indicate greater model fit. Nagelkerke R<sup>2</sup> measure ranges from 0 to 1
- **Predictive accuracy** has 2 most common approaches –  
Classification matrix and Chi-square based measures of fit
    - Classification Matrix – measures how well group membership is predicted by developing a **hit ratio**, which is the **percentage correctly classified**
    - Chi-square based measure – Hosmer & Lemeshow developed a classification test where cases are first divided into approximately 10 equal classes. No. of actual & predicted events is compared in each class with the chi-square statistic. Appropriate use of this test requires a sample of at least 50 cases, each class with at least 5 observations & predicted events should never fall below 1





# Contd...

## ◆ Stage 5: Interpretation of the results

- LR tests hypotheses about individual coefficients
- **Wald statistic** provides statistical significance for each estimated coefficient
- Logistic coefficients are difficult to interpret in their original form because they are expressed in logarithms
- Most computer programs provide an exponentiated logistic coefficient, a transformation (antilog) of original logistic coefficient
- The sign of the original coefficients (+ve/ -ve) indicate the direction of the relationship
- Exponentiated coefficients above 1.0 reflect a positive relationship & values less than 1.0 reflect negative relationship

## ◆ Stage 6: Validation of the results

- **How do we predict group membership from the logistic curve?**
  - For each observation, LR technique predicts a probability value between 0 & 1
  - Predicted probability is based on the value(s) of the independent variable(s) & the estimated coefficients
  - If the predicted probability is  $>0.50$ , outcome is 1; otherwise, outcome is 0



# Caselet – Stereotaxic Surgery

College students ( $N = 315$ ) were asked to pretend that they were serving on a university research committee hearing a complaint against animal research being conducted by a member of the university faculty. The complaint included a description of the research in simple but emotional language. Cats were being subjected to stereotaxic surgery in which a cannula was implanted into their brains. Chemicals were then introduced into the cats' brains via the cannula and the cats given various psychological tests. Following completion of testing, the cats' brains were subjected to histological analysis. The complaint asked that the researcher's authorization to conduct this research be withdrawn and the cats turned over to the animal rights group that was filing the complaint. It was suggested that the research could just as well be done with computer simulations.

In defence of his research, the researcher provided an explanation of how steps had been taken to assure that no animal felt much pain at any time, an explanation that computer simulation was not an adequate substitute for animal research, and an explanation of what the benefits of the research were.



# Contd...

Each participant read one of five different scenarios which described the goals and benefits of the research. They were:

- COSMETIC - testing the toxicity of chemicals to be used in new lines of hair care products
- THEORY - evaluating two competing theories about the function of a particular nucleus in the brain
- MEAT - testing a synthetic growth hormone said to have the potential of increasing meat production
- VETERINARY - attempting to find a cure for a brain disease that is killing both domestic cats and endangered species of wild cats
- MEDICAL - evaluating a potential cure for a debilitating disease that afflicts many young adult humans

After reading the case materials, each participant was asked to decide whether or not to withdraw Dr. Wissen's authorization to conduct the research and, among other things, to fill out D. R. Forysth's Ethics Position Questionnaire, which consists of 20 Likert-type items, each with a 9-point response scale from "completely disagree" to "completely agree."

**Are idealism and relativism (and gender and purpose of the research) related to attitudes towards animal research in college students?**



# Contd...

The criterion variable is dichotomous & Predictor variables may be categorical or continuous

- Criterion variable - Decision
- Predictor Variables – Gender, Ethical Idealism (9-point Likert), Ethical Relativism (9-point Likert), Purpose of the Research
- gender - 0 = Female and 1 = Male
- decision - 0 =

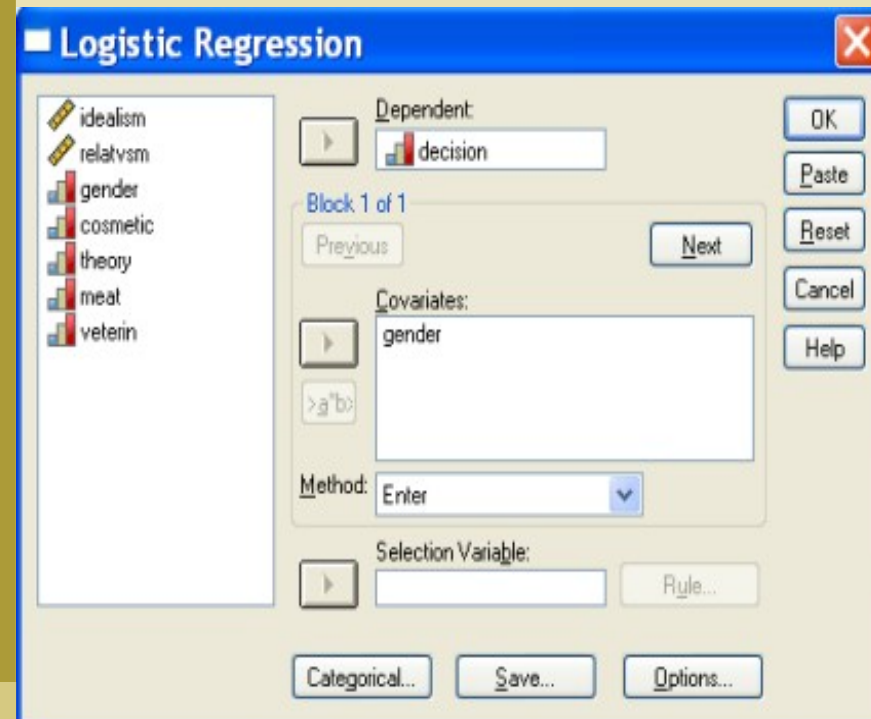
Model is ..... logit =

$$\ln(ODDS) = \ln\left(\frac{\hat{Y}}{1 - \hat{Y}}\right) = a + bX$$

Let's run Logistic Regression

Click Analyze, Regression, Binary Logistic

- Scoot the decision variable into the Dependent box and the gender variable into the Covariates box
- Click OK
- Looking at the statistical output, we see that there are 315 cases used in the analysis



Case Processing Summary

Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	315	100.0
	Missing Cases	0	.0
	Total	315	100.0
Unselected Cases		0	.0
Total		315	100.0

a. If weight is in effect, see classification table for the total number of cases.

◆ The Block 0 output is for a model that includes only the intercept / constant component

- Decision options:  $187/315 = 59\%$  decided to stop the research, 41% allow to continue
- You would be correct 59% of the time

◆ Under **Variables in the Equation** model the intercept is  $\ln(\text{odds}) = -.379$

- If we exponentiate both sides of this expression we find that predicted odds of deciding to continue the research is  $[\text{Exp}(B)] = .684$
- 128 voted to continue the research, 187 to stop it

◆ Block 1 output includes gender variable as a predictor. **Omnibus Tests of Model Coefficients** gives a Chi-Square of 25.653 on 1 df, significant beyond .001

- Tests null hypothesis. Adding gender variable has not significantly increased our ability to predict the decisions

Observed			Predicted		
			decision		Percentage Correct
			stop	continue	
Step 0	decision	stop	187	0	100.0
		continue	128	0	.0
Overall Percentage					59.4

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.379	.115	10.919	1	.001	.684

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	25.653	1	.000
	Block	25.653	1	.000
	Model	25.653	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	399.913 <sup>a</sup>	.078	.106

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than .001.

- ◆ Model Summary, -2 Log Likelihood = 399.913, measures how poorly the model fits the data. The smaller the statistic the better the model

For intercept only, -2LL = 425.566 . Add gender and -2LL = 399.913.

**Omnibus Tests:** Drop in -2LL = 25.653 ; Model  $\chi^2$   $df = 1, p < .001$

- ◆ Cox & Snell R<sup>2</sup> cannot reach a maximum value of 1. Nagelkerke R<sup>2</sup> can reach a maximum of 1

# Contd...

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step	gender	1.217	.245	24.757	1	.000	3.376
1	Constant	-.847	.154	30.152	1	.000	.429

a. Variable(s) entered on step 1: gender.

- The Variables in the Equation output shows us that the regression equation is  **$\ln(\text{ODDS}) = -0.847 + 1.217 * \text{Gender}$**  or  **$\text{ODDS} = e^{a+b*\text{Gender}}$**

$$\text{ODDS} = e^{-.847 + 1.217(0)} = e^{-.847} = 0.429 \quad \text{ODDS} = e^{-.847 + 1.217(1)} = e^{.37} = 1.448$$

- A woman (code=0) is only .429 as likely to decide to continue the research as she is to decide to stop the research; and a man (code=1) is 1.448 times more likely to decide to continue the research than to decide to stop the research
- We can easily convert odds to probabilities

$$\frac{\text{male\_odds}}{\text{female\_odds}} = \frac{1.448}{.429} = 3.376 = e^{1.217}$$

- 1.217 is the **B** (slope) for Gender, 3.376 is the **Exp(B)**, that is, the exponentiated slope, the **odds ratio**. Men are 3.376 times more likely to vote to continue the research than are women
- For women

$$\hat{Y} = \frac{\text{ODDS}}{1 + \text{ODDS}} = \frac{0.429}{1.429} = 0.30$$

men

$$\hat{Y} = \frac{\text{ODDS}}{1 + \text{ODDS}} = \frac{1.448}{2.448} = 0.59$$



- ◆ We need to have a decision rule to classify the subjects
- ◆ Our decision rule will take the following form:
  - If  $p(E) \geq \text{threshold}$ , we shall predict that event will take place. By default, SPSS sets this threshold to .5
  - our model leads to the prediction that the probability of deciding to continue the research is 30% for women and 59% for men
  - The Classification Table shows us that
    - overall success rate is  $(140+68) / 315 = 66\%$
    - Percentage of occurrences correctly predicted i.e.,  $P(\text{correct} \mid \text{event did occur}) = 68 / 128 = 53\%$ . This is known as the sensitivity of prediction
    - Percentage of non-occurrences correctly predicted i.e.,  $P(\text{correct} \mid \text{event did not occur}) = 140 / 187 = 75\%$ . This is known as the specificity of prediction
    - We could focus on error rates in classification.
    - False Positive Rate -  $P(\text{incorrect prediction} \mid \text{predicted occurrence})$ , Of all those for whom we predicted a vote to Continue the research, how often were we wrong  $= 47 / 115 = 41\%$
    - False Negative Rate -  $P(\text{incorrect prediction} \mid \text{predicted nonoccurrence})$ , Of all those for whom we predicted a vote to Stop the research, how often were we wrong is  $60 / 200 = 30\%$



Observed			Predicted		Percentage Correct
			decision		
			stop	continue	
Step 1	decision	stop	140	47	74.9
		continue	60	68	53.1
Overall Percentage					66.0

a. The cut value is .500

**Logistic Regression: Options**

Statistics and Plots

☐ Classification plots

☒ Hosmer-Lemeshow goodness-of-fit

☐ Casewise listing of residuals

☐ Correlations of estimates

☐ Iteration history

☒ CI for exp(B): 95 %

☐ Outliers outside 2 std. dev.

☐ All cases

Display

☒ At each step

☐ At last step

Probability for Stepwise

Entry: .05 Removal: .10

Classification cutoff: .5

Maximum Iterations: 20

☒ Include constant in model

Continue

Cancel

Help

# Multiple Predictors, both Categorical and Continuous

**Model Summary**

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	346.503 <sup>a</sup>	.222	.300

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

**Classification Table<sup>a</sup>**

Observed			Predicted		
			decision		Percentage Correct
			stop	continue	
Step 1	decision	stop	151	36	80.7
		continue	55	73	57.0
Overall Percentage					71.1

a. The cut value is .500

- ◆ Conduct Logistic Regression
- ◆ Click Analyze, Regression, Binary Logistic
  - decision variable - Dependent variable
  - gender, idealism, and relativism – Independent variables
- ◆ Click Options and check “Hosmer-Lemeshow goodness of fit” and “CI for exp(B) 95%.”
- ◆ In Block 1 output, the -2 Log Likelihood statistic has dropped to 346.503, indicating that our expanded model is doing a better job at predicting decisions than was our one-predictor model
- ◆ The R<sup>2</sup> statistics have also increased.
- ◆ Overall success rate in classification has improved from 66% to 71%

# Contd...

Hosmer and Lemeshow Test


Step	Chi-square	df	Sig.
1	8.810	8	.359

Contingency Table for Hosmer and Lemeshow Test

		decision = stop		decision = continue		Total
		Observed	Expected	Observed	Expected	
Step 1	1	29	29.331	3	2.669	32
	2	30	27.673	2	4.327	32
	3	28	25.669	4	6.331	32
	4	20	23.265	12	8.735	32
	5	22	20.693	10	11.307	32
	6	15	18.058	17	13.942	32
	7	15	15.830	17	16.170	32
	8	10	12.920	22	19.080	32
	9	12	9.319	20	22.681	32
	10	6	4.241	21	22.759	27

- ◆ Hosmer-Lemeshow tests the (null hypothesis) predictions made by the model fit against the observed group memberships
- ◆ Cases are arranged in order by their predicted probability on the criterion variable
- ◆ Ordered cases are then divided into ten (usually) groups
- ◆ For each of these groups we then obtain the predicted group memberships and the actual group memberships
- ◆ This results in a 2 x 10 contingency table
- ◆ A chi-square statistic is computed comparing the observed frequencies with those expected under the linear model.
- ◆ A nonsignificant chi-square indicates that the data fit the model well.
- ◆ This procedure suffers from several problems
  - With large sample sizes, the test may be significant, even when the fit is good.
  - With small sample sizes it may not be significant, even with poor fit. Even
  - Hosmer and Lemeshow is no longer recommended





# Caselet

## Sinking of the Titanic?

On April 14th, 1912, at 11.40 p.m., the Titanic, sailing from Southampton to New York, struck an iceberg and started to take on water. At 2.20 a.m. she sank; of the 2228 passengers and crew on board, only 705 survived. Data on Titanic passengers have been collected by many researchers, but here we shall examine part of a data set compiled by Thomas Carson. It is available on the Internet (<http://hesweb1.med.virginia.edu/biostat/s/data/index.html>).

For 1309 passengers, these data record whether or not a particular passenger survived, along with the age, gender, ticket class, and the number of family members accompanying each passenger.

We shall investigate the data to try to determine which, if any, of the explanatory variables are predictive of survival.



# Which of the explanatory variables are predictive of the response, survived or died?

- ◆ For a binary response – probability of died is  $p=0$  and survived is  $p=1$
- ◆ Logistic regression model given by

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q$$

- ◆ The log-odds of survival is modeled as a linear function of the explanatory variables
- ◆ Parameters in the logistic regression model can be estimated by maximum likelihood
- ◆ Estimated regression coefficients in a logistic regression model give the estimated change in the log-odds corresponding to a unit change in the corresponding explanatory variable conditional on the other explanatory variables remaining constant
- ◆ The parameters are usually exponentiated to give results in terms of odds
- ◆ In terms of  $p$ , the logistic regression model can be written as

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_q x_q)}$$

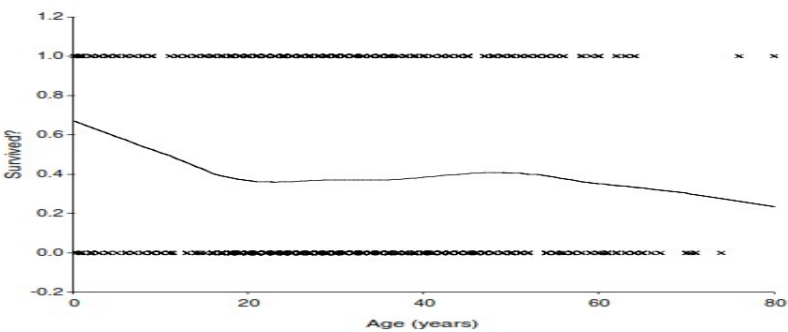


# Analysis using SPSS

- ◆ Analyses of Titanic data will focus on establishing relationships between the binary passenger outcome survival (survived=1, death=0) and five passenger characteristics that might have affected the chances of survival, namely:
  - Passenger class (variable pclass, with “1” indicating a first class ticket holder, “2” second class, and “3” third class)
  - Passenger age (age recorded in years)
  - Passenger gender (sex, with females coded “1” and males coded “2”)
  - Number of accompanying parents/children (parch)
  - Number of accompanying siblings/spouses (sibsp)
- ◆ Our investigation of the determinants of passenger survival will proceed in three steps
  - First, we assess (unadjusted) relationships between survival and each potential predictor variable singly
  - We adjust these relationships for potential confounding effects
  - Finally, we consider the possibility of interaction effects between some of the variables



- Begin with using simple descriptive tools to provide initial insights
- Crosstabs command – measures the associations between categorical explanatory variables and passenger survival
- The results show that in our sample of 1309 passengers the survival proportions were:
  - Clearly decreasing for lower ticket classes
  - Considerably higher for females than males
  - Highest for passengers with one sibling/spouse or three parents/children accompanying them
- Scatterplot - Examines the association between age & survival, by including a Lowess curve
  - Graph shows survival chances is highest for infants and generally decreases with age although the decrease is not monotonic, rather there appears to be a local minimum at 20 years of age & a local maximum at 50 years



Display 9.3 Scattergraph of binary survival against age enhanced by inclusion of a Lowess curve.

Survived?\* Passenger class Crosstabulation

		Passenger class			Total
		first class	second class	third class	
Survived? no	Count	123	158	528	809
	% within Passenger class	38.1%	57.0%	74.5%	61.8%
yes	Count	200	119	181	500
	% within Passenger class	61.9%	43.0%	25.5%	38.2%
Total		323	277	709	1309
		% within Passenger class	100.0%	100.0%	100.0%

b)

Survived?\* Gender Crosstabulation

		Gender		Total
		male	female	
Survived? no	Count	682	127	809
	% within Gender	80.9%	27.3%	61.8%
yes	Count	161	339	500
	% within Gender	19.1%	72.7%	38.2%
Total		843	466	1309
		% within Gender	100.0%	100.0%

c)

		Number of parents/children aboard									Total
		0	1	2	3	4	5	6	9		
Survived?	no	Count	666	70	56	3	5	5	2	2	809
		% within Number of parents/children aboard	66.5%	41.2%	49.6%	37.5%	83.3%	83.3%	100.0%	100.0%	61.8%
	yes	Count	336	100	57	5	1	1			500
		% within Number of parents/children aboard	33.5%	58.8%	50.4%	62.5%	16.7%	16.7%			38.2%
Total		Count	1002	170	113	8	6	6	2	2	1309
		% within Number of parents/children aboard	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

d)

Survived?\* Number of siblings/spouses aboard Crosstabulation

		Number of siblings/spouses aboard							Total
		0	1	2	3	4	5	8	
Survived? no	Count	582	156	23	14	19	6	9	809
	% within Number of siblings/spouses aboard	65.3%	48.9%	54.8%	70.0%	86.4%	100.0%	100.0%	61.8%
yes	Count	309	163	19	6	3			500
	% within Number of siblings/spouses aboard	34.7%	51.1%	45.2%	30.0%	13.6%			38.2%
Total		891	319	42	20	22	6	9	1309
		% within Number of siblings/spouses aboard	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Display 9.2 Cross-tabulation of passenger survival by (a) passenger class, (b) gender, (c) number of parents/children aboard, and (d) number of siblings/spouses aboard.

# Contd...

- ◆ Although the simple cross-tabulations and scatterplot are useful first steps, they may not tell the whole story about the data when confounding or interaction effects are present among the explanatory variables
- ◆ Cross-tabulations and grouped box plots (not presented) show that in our passenger sample:
  - Males were more likely to be holding a third-class ticket than females
  - Males had fewer parents/children or siblings/spouses with them than did females
  - The median age was decreasing with lower passenger classes
  - The median number of accompanying siblings/spouses generally decreased with age
  - The median number of accompanying children/parents generally increased with age
- ◆ To get a better picture of our data, a multiway-classification of passenger survival within strata defined by explanatory variable-level combinations might be helpful

- ◆ Before such a table can be constructed, the variables age, parch, and sibsp need to be categorized in some sensible way
  - Create two new variables - age\_cat and marital
  - Age\_cat categorizes passengers into 2 children - age < 21 yrs & adults - age >= 21 yrs
  - Marital categorizes into four
    - 1=no siblings/spouses and no parents/children
    - 2=siblings/spouses but no parents/children
    - 3=no siblings/spouses but parents/children
    - 4=siblings/spouses and parents/children
- ◆ The **Recode** command and the **Compute** command, in conjunction with the **If Cases sub-dialogue box** allows sequential assignment of codes according to conditions, which can be used to generate the new variables
- ◆ Crosstabs dialogue box is then employed to generate the required five-way table

Passenger class \* Survived? \* Gender \* AGE\_CAT \* MARITAL Crosstabulation

% within Passenger class

					Survived?
					yes
no sibs./spouse and no parents/childr.	Child	female	Passenger class	1	100.0%
				2	88.9%
				3	62.1%
		male	Passenger class	1	27.6%
				2	13.8%
				3	11.0%
	Adult	female	Passenger class	1	95.7%
				2	84.8%
				3	47.6%
		male	Passenger class	1	30.4%
				2	9.2%
				3	16.8%
sibs./spouse but no parents/childr.	Child	female	Passenger class	1	100.0%
				2	100.0%
				3	55.0%
		male	Passenger class	3	13.0%
	Adult	female	Passenger class	1	97.0%
				2	75.0%
				3	40.0%
		male	Passenger class	1	42.9%
				2	3.4%
				3	8.3%
no sibs./spouse but parents/childr.	Child	female	Passenger class	1	100.0%
				2	100.0%
				3	57.1%
		male	Passenger class	1	100.0%
				2	100.0%
				3	71.4%
	Adult	female	Passenger class	1	100.0%
				2	100.0%
				3	46.2%
		male	Passenger class	1	25.0%
				3	20.0%
sibs./spouses and parents/children	Child	female	Passenger class	1	66.7%
				2	90.9%
				3	34.2%
		male	Passenger class	1	75.0%
				2	88.9%
				3	19.2%
	Adult	female	Passenger class	1	95.2%
				2	93.8%
				3	37.5%
		male	Passenger class	1	33.3%
				2	9.1%
				3	10.0%

Crosstabs

name

age

sibsp

parch

ticket

fare

cabin

embarked

boat

body

home.des

sibsp1

parch1

Row(s):

pclass

Column(s):

survived

Previous

Layer 3 of 3

Next

marital

OK

Paste

Reset

Cancel

Help

Display clustered bar charts

Suppress tables

Exact...

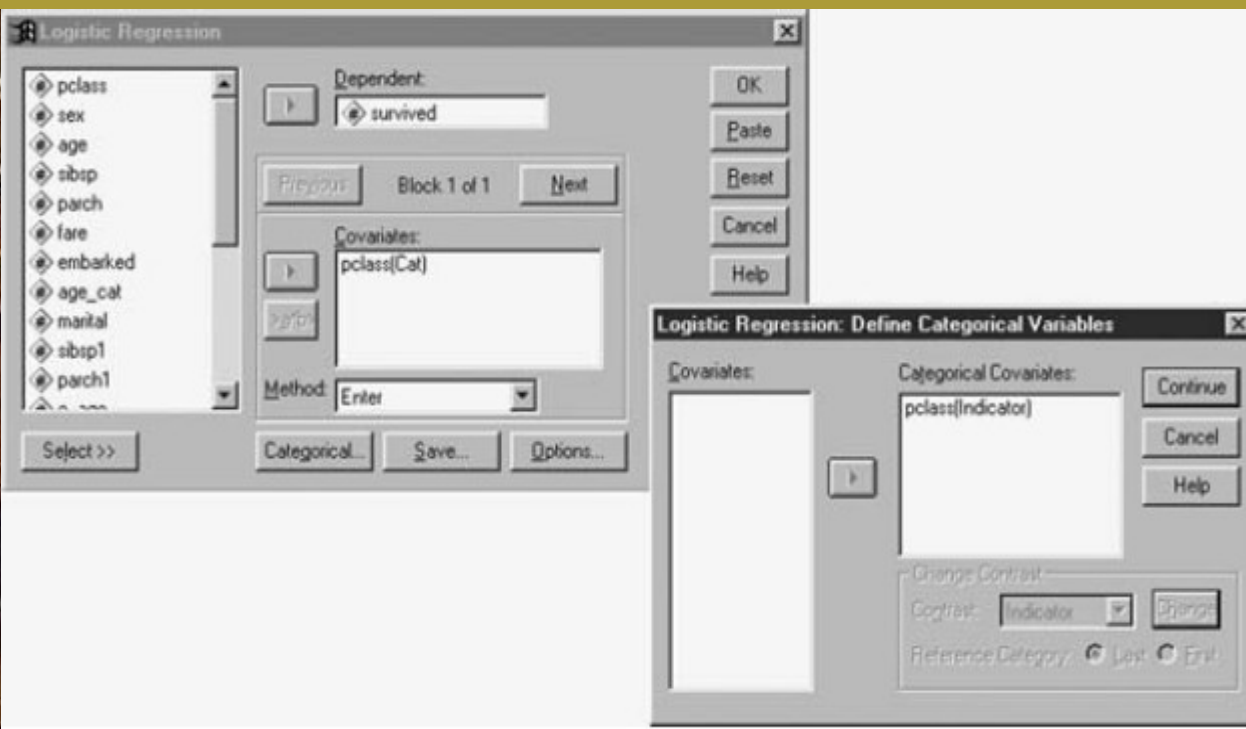
Statistics...

Cells...

Format...



- ◆ We can now proceed to investigate the associations between survival and the five potential predictors using logistic regression
- ◆ The SPSS logistic regression dialogue box is obtained by using the commands **Analyze – Regression – Binary Logistic...**
- ◆ Include single explanatory variable in the model at a time
- ◆ We start with the categorical explanatory variable pclass
- ◆ Binary dependent variable is declared under the Dependent list and the single explanatory variable under the Covariates list
- ◆ By default, SPSS assumes explanatory variables are measured on an interval scale
- ◆ To inform SPSS about the categorical nature of variable pclass, the **Categorical...** button is checked and pclass included in the Categorical Covariates list on the resulting **Define Categorical Variables** sub-dialogue box



Contd...

We also check CI for  $\exp(B)$  on the **Options sub-dialogue box** so as to include confidence intervals for the odds ratios in the output

# Contd...

- ◆ There are basically three parts to the output
- ◆ The first three tables inform the user about the sample size, the coding of the dependent variable, and the dummy variable coding of the categorical predictor variables
  - Here with only one categorical explanatory variable, (1) corresponds to first class, variable, (2) to second class, and third class represents the reference category

**Case Processing Summary**

Unweighted Cases <sup>a</sup>		N	Percent
Selected Cases	Included in Analysis	1309	100.0
	Missing Cases	0	.0
	Total	1309	100.0
Unselected Cases		0	.0
Total		1309	100.0

a. If weight is in effect, see classification table for the total number of cases.

**Dependent Variable Encoding**

Original Value	Internal Value
no	0
yes	1

**Categorical Variables Codings**

		Frequency	Parameter coding	
			(1)	(2)
Passenger class	first class	323	1.000	.000
	second class	277	.000	1.000
	third class	709	.000	.000

# Contd...

- ◆ SPSS automatically begins by fitting a null model
- ◆ Classification Table - compares survival predictions made on the basis of the fitted model with the true survival status of the passengers
  - On the basis of the fitted model, passengers are predicted to be in the survived category if their predicted survival probabilities are above 0.5
  - Here the overall survival proportion (0.382) is below the threshold and all passengers are classified as non-survivors by the null model leading to 61.8% (the non-survivors) being correctly classified
- ◆ Variables in the Equation table provides the Wald test for the null hypothesis of equal survival and non-survival proportions
- ◆ Variables not in the Equation table lists score tests for the variables not yet included in the model, here pclass
  - It is clear that survival is significantly related to passenger class (Score test:  $X(2) = 127.9$ ,  $p < 0.001$ )
  - Score tests for specific passenger classes are also compared with the reference category (third class)

Classification Table<sup>a,b</sup>

			Predicted		
			Survived?		Percentage Correct
			no	yes	
Step 0	Survived?	no	809	0	100.0
		yes	500	0	.0
Overall Percentage					61.8

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.481	.057	71.548	1	.000	.618

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	PCLASS	127.856	2	.000
		PCLASS(1)	102.220	1	.000
		PCLASS(2)	3.376	1	.066
	Overall Statistics		127.856	2	.000

- ◆ Latest **Classification** Table shows that inclusion of the pclass factor increases the percentage of correct classification by 67.7%
- ◆ **Omnibus Tests of Model Coefficients** table contains the likelihood ratio (LR) test, test for assessing the effect of pclass
  - We detect a significant effect of passenger class LR test:  $X^2(2) = 127.8, p < 0.001$
- ◆ Finally, the latest **Variables in the Equation** table provides Wald's tests for all the variables included in the model
  - Consistent with the LR and score tests, the effect of pclass tests significant  $X(2) = 120.5, p < 0.001$
  - Parameter estimates (log-odds) are given in the column labeled "B," with the column "S.E." providing the standard errors of these estimates
  - Comparing each ticket class with the third class, we estimate that the odds of survival were 4.7 times higher for first class passengers (CI form 3.6 to 6.3) and 2.2 times higher for second class passengers (1.6 to 2.9)

Classification Table<sup>a</sup>

Observed			Predicted		
			Survived?		Percentage Correct
			no	yes	
Step 1	Survived?	no	686	123	84.8
		yes	300	200	40.0
Overall Percentage					67.7

<sup>a</sup>. The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	127.765	2	.000
	Block	127.765	2	.000
	Model	127.765	2	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1 <sup>a</sup>	PCLASS			120.536	2	.000			
	PCLASS(1)	1.557	.143	117.934	1	.000	4.743	3.581	6.282
	PCLASS(2)	.787	.149	27.970	1	.000	2.197	1.641	2.941
	Constant	-1.071	.086	154.497	1	.000	.343		

<sup>a</sup>. Variable(s) entered on step 1: PCLASS.

**Clearly, the chances of survival are significantly increased in the two higher ticket classes**



- ◆ The results for the remaining categorical explanatory variables considered individually are summarized in the adjacent table
- ◆ Table shows that the largest increase in odds is found when comparing the two gender groups — the chance of survival among female passengers is estimated to be 8.4 times that of males
- ◆ The shape of the Lowess curve plotted in earlier suggests that the survival probabilities might not be monotonically decreasing with age
- ◆ Such a possibility can be modeled by using a third order polynomial for the age effect
- ◆ To avoid multicollinearities, we center age by its mean (30 years) before calculating the linear (c\_age), quadratic (c\_age2), and cubic (c\_age3) age terms
- ◆ The three new age variables are then divided by their respective standard deviations (14.41, 302.87, and 11565.19) simply to avoid very small regression coefficients due to very large variable values
- ◆ Inclusion of all three age terms under the Covariates list in the Logistic Regression dialogue box gives the results shown in the adjacent display
- ◆ We find that the combined age terms have a significant effect on survival (LR:  $X^2(3) = 16.2$ ,  $p = 0.001$ ). The single parameter Wald tests show that the quadratic and
- ◆ cubic age terms contribute significantly to explaining variability in survival
- ◆ probabilities. These results confirm that a linear effect of age on the logodds
- ◆ scale would have been too simplistic a model for these data.

**Table 9.1 Unadjusted Effects of Categorical Predictor Variables on Survival Obtained from Logistic Regressions**

<i>Categorical Predictor</i>	<i>LR Test</i>	<i>OR for Survival</i>	<i>95% CI for OR</i>
Passenger class (pclass)	$X^2(2) = 127.8$ ,		
First vs. third class	$p < 0.001$	4.743	3.581–6.282
Second vs. third class		2.197	1.641–2.941
Gender (sex)	$X^2(1) = 231.2$ ,		
Female vs. male	$p < 0.001$	8.396	6.278–11.229
Number of siblings/spouses aboard (sibsp1)	$X^2(3) = 14.2$ , $p < 0.001$		
0 vs. 3 or more		2.831	1.371–5.846
1 vs. 3 or more		5.571	2.645–11.735
2 vs. 3 or more		4.405	1.728–11.23
Number of parents/children aboard (parch1)	$X^2(3) = 46.7$ , $p < 0.001$		
0 vs. 3 or more		1.225	0.503–2.982
1 vs. 3 or more		3.467	1.366–8.802
2 vs. 3 or more		2.471	0.951–6.415