Statistical Modelling

An Introduction to
Generalised Linear Models

Ioannis Ntzoufras

E-mail: ntzoufras@aegean.gr

Department of
Business Administration,
University of the Aegean

CONTENTS

WHAT IS STATISTICS?

Every moment in life we make **CHOICES**

Our Choices are based on **INCOMPLETE INFORMATION**

For example:

Shall I take an umbrella with me?

WHAT IS STATISTICS?

Example

WHAT IS STATISTICS?

Example

Incomplete Information:

Weather Forecast

WHAT IS STATISTICS?

So every choice or decision is made upon **UNCERTAINTY**

STATISTICS is the Science which **QUANTIFIES UNCERTAINTY** and hence helps to decide which decision is optimal

WHAT IS STATISTICS?

And it is not just a matter of taking an umbrella or getting wet.

Sometimes involves matters of

LIFE AND DEATH

WHAT IS STATISTICS?

REAL LIFE EXAMPLE [1]

- 1986: Challenger Space Shuttle exploded killing 7 astronauts.
- The accident would have been avoided if they have done a simple Statistical analysis which indicated: HIGH PROBABILITY OF FAILURE IN LOW TEMPRETURE
- (That day the temperature was 0 C)

WHAT IS STATISTICS?

REAL LIFE EXAMPLE [2]

- 1954: The POLIO VACCINE
- Vaccine Trials were performed in 400,000 children
- Good Statistical Analysis have indicated the effectiveness of the vaccine and today POLIO is almost unknown

WHAT IS STATISTICS?

AREAS OF STATISTICS

- MEDICINE
- ECONOMETRICS (ECONOMICS)
- MARKETING
- PSYCHOMETRICS (PSYCOLOGY)
- SPORTS (ATHLOMETRICS)
- SOCIAL SCIENCES
- ARCHAEOMETRICS (ARCHAELOGY)
- AUTHOR IDENTIFICATION

WHAT IS STATISTICS?

AREAS OF STATISTICS

- MEDICINE
- ECONOMETRICS (ECONOMICS)
- MARKETING
- PSYCHOMETRICS (PSYCOLOGY)
- SPORTS (ATHLOMETRICS)
- SOCIAL SCIENCES
- ARCHAEOMETRICS (ARCHAELOGY)
- AUTHOR IDENTIFICATION

WHAT IS STATISTICS?

AREAS OF STATISTICS

- QUALITY CONTROL
- ELLECTION POLLS
- ENVIROMENTAL MONITORING
- RACIAL BIAS
- LAW
- PATTERN, IMAGE AND VOICE RECOGNICION

WHAT IS STATISTICS?

WHAT IS STATISTICS?

STATISTICS IS THE SCIENCE OF SCIENCES

WHAT IS STATISTICS?

WHAT IS STATISTICS?

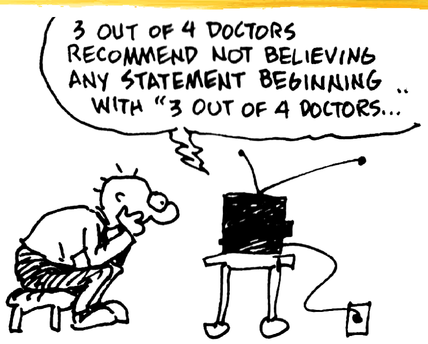
MAIN DIVISIONS OF STATISTICS

- DATA ANALYSIS
- PROBABILITY THEORY
- MATHEMATICAL STATISTICS

WHAT IS STATISTICS?

- GOOD STATISTICAL ANALYSIS IS ALMOST IMPOSSIBLE IN DAILY LIFE
- BE CAREFUL WITH STATISTICAL STATEMENTS
- BOOK:"HOW TO LIE WITH STATISTICS"
- DON'T TRUST A STATISTICAL FINDING UNLESS IT IS REPEATED CONSISTENTLY IN LITERATURE

WHAT IS STATISTICS?



WHAT IS STATISTICS?

In what it follows I will try to present elements of Statistical Modelling as simple as possible.

All you need is …
little Patience
some Thought
little bit of Maths

WHAT IS A STATISTICAL MODEL?

- STATISTICAL MODEL IS ANY GROUP OF MATHEMATICAL AND PROBABILISTIC EQUATIONS USED TO DESCRIBE, SUMMARIZE AND PREDICT REALITY
- USUALLY IT CONTAINS
  - STOCHASTIC RELATIONSHIPS [Y~NORMAL]
  - DETERMINISTIC RELATIONSHIPS [Y=Z+X]
- MOST POPULAR MODELS:
  GENERALISED LINEAR MODELS (GLM)

GENERALISED LINEAR MODELS

- **3.1. INTRODUCTION**
- **3.2. DATA**
- **3.3. THREE MAIN COMPONENTS**
- **3.4. TYPES OF GLM**
- **3.5. GENERAL PRINCIPLES OF MODELLING**

GENERALISED LINEAR MODELS

3.1. INTRODUCTION

- **IT IS A GENERALIZATION OF THE REGRESSION MODELS**
- **STARTED FROM LEGENDRE (1805) AND GAUSS (1809)**

GENERALISED LINEAR MODELS

3.2. DATA

- RESPONSE VARIABLE (Y): also called dependent or endogenous variable
  - Y is a random variable
- EXPLANATORY VARIABLES (X_j): Independent or Exogenous variables
  - X_j are usually assumed fixed by the experiment

GENERALISED LINEAR MODELS

3.3. THREE MAIN COMPONENTS

- (1) RANDOM COMPONENT
  - $Y_i \sim$ DISTRIBUTION ($\boldsymbol{\theta}$)
  - $\boldsymbol{\theta}$: VECTOR OF MODEL PARAMETERS
- (2) SYSTEMATIC COMPONENT
  - $\eta_i = \beta_0+\beta_1 X_{1i}+\ldots+ \beta_p X_{pi}$
  - η_i : LINEAR PREDICTOR OF THE MODEL

GENERALISED LINEAR MODELS

- (3) LINK FUNCTION
  - LINKS RANDOM COMPONENT AND LINEAR PREDICTOR
  - $g(\underline{\boldsymbol{\theta}}) = \eta_i = \beta_0+\beta_1 X_{1i}+\ldots+ \beta_p X_{pi}$
  - USUALLY $\underline{\boldsymbol{\theta}}$ IS THE MEAN OF Y

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: NORMAL MODEL

- RANDOM COMPONENT
  - Y QUANTITATIVE VARIABLE (WEIGHT)
  - $Y_i \sim$ NORMAL(μ_i , σ^2), $E(Y)=\mu$, $V(Y)= \sigma^2$
- SYSTEMATIC COMPONENT:
  - X_j QUANTITATIVE or QUALITATIVE
- LINK FUNCTION
  - $\mu_i = \eta_i = \beta_0+\beta_1 X_{1i}+\ldots+ \beta_p X_{pi}$

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: NORMAL MODEL

- QUANTITATIVE X's: REGRESSION MODEL
- QUALITATIVE X's:
  Analysis of Variance (ANOVA) Model
- BOTH TYPES OF X's:
  Analysis of Covariance (ANCOVA) Model
- All normal models are (sometimes) referred as Regression Models

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: BERNOULLI MODELS

- RANDOM COMPONENT
  - Y BINARY VARIABLE (0/1, e.g. die/survive)
  - $Y_i \sim$ Bernoulli (p_i), $E(Y)=p$
- SYSTEMATIC COMPONENT:
  - X_j QUANTITATIVE or QUALITATIVE
- LINK FUNCTION
  - $\log (p_i/(1- p_i)) = \eta_i = \beta_0+\beta_1 X_{1i}+\ldots+ \beta_p X_{pi}$
  - g(p)=logit(p)

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: BINOMIAL MODELS

- RANDOM COMPONENT
  - Y # of successes in a total of n trials
  - $Y_i \sim$ Binomial (p_i, n_i), $E(Y)=np$
- SYSTEMATIC COMPONENT:
  - X_j QUANTITATIVE or QUALITATIVE
- LINK FUNCTION
  - $\log (p_i/(1- p_i)) = \eta_i = \beta_0+\beta_1 X_{1i}+\ldots+ \beta_p X_{pi}$
  - g(p)=logit(p)

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: BINOMIAL MODELS

- OTHER LINK FUNCTIONS:
  - g(p)=logit(p): LOGIT FUNCTION
  - $g(p)=\Phi^{-1}(p)$: PROBIT FUNCTION
  - g(p)=LOG(-LOG(1-p)) COMPLEMENTARY LOG-LOG

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: BINOMIAL MODELS

- OTHER LINK FUNCTIONS:
  - g(p)=logit(p): **LOGISTIC REGRESSION MODELS**
  - $g(p)=\Phi^{-1}(p)$: **PROBIT MODELS**
  - g(p)=LOG(-LOG(1-p)) COMPLEMENTARY LOG-LOG

GENERALISED LINEAR MODELS

3.4. TYPES OF GLM: POISSON MODELS

- RANDOM COMPONENT
  - Y # of successes in a fixed time period
  - $Y_i \sim$ Poisson (λ_i), $E(Y)=\lambda$
- SYSTEMATIC COMPONENT:
  - X_j QUANTITATIVE or QUALITATIVE
- LINK FUNCTION
  - $\log(\lambda_i) = \eta_i = \beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}$
  - $g(\lambda)=\log(\lambda)$

GENERALISED LINEAR MODELS

3.5. GENERAL PRINCIPLES OF MODELING

IT IS ART

- ALL MODELS ARE WRONG
  - SOME OF THEM ARE MORE USEFUL THAN OTHERS
  - WE SEEK FOR MODELS WHICH DESCRIBE REALITY
  - WE FIT AND CHECK MANY DIFFERENT MODELS
- ALWAYS USE SOME DIAGNOSTICS FOR CHECKING THE GOODNESS OF FIT

PRACTICAL EXAMPLES

4.1. EXAMPLE 1: Study of the Relationship Between Estriol and Birthweight

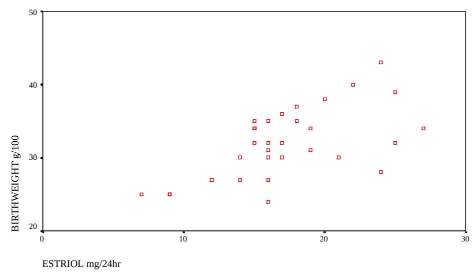
4.2. EXAMPLE 2: The case of Challenger Explosion

PRACTICAL EXAMPLES

4.1. EXAMPLE 1

- Green & Touchston (1963, *Am.Jour. Of Obsterics & Gynecology*)
- STUDY OF THE RELATIONSHIP
  - Y : Birthweight (gr/100)
  - X : Estriol level of women
  - Sample Size n=31
- The relationship can be examined in the following graph

PRACTICAL EXAMPLES



PRACTICAL EXAMPLES

- RESPONSE: Birthweight
- EXPLANATORY VARIABLE: Estriol Level
- RANDOM COMPONENT: $Birth_i \sim Normal(\mu_i, \sigma^2)$
- SYSTEMATIC COMPONENT: $\eta_i = \alpha + \beta \times Estriol_i$
- LINK FUNCTION: $\mu_i = \eta_i = \alpha + \beta \times Estriol_i$
- for i=1,...,31

PRACTICAL EXAMPLES

- ESTIMATE THE EFFECTS USING A STATISTICAL PACKAGE: FOR EXAMPLE SPSS

Birthweight ~ Normal(μ, 14.61)
Expected Birthweight= $\mu = 21.52 + 0.608 \times Estriol_i$

Note: The model holds only for range of values of ESTRIOL observed in sample

R^2 = % of variance explained by our model = 61%

PRACTICAL EXAMPLES

Birthweight ~ Normal(μ, 14.61)
Expected Birthweight= $\mu = 21.52 + 0.608 \times Estriol_i$

INTERPRETATION OF PARAMETERS

- If ESTRIOL= 0 =>

  expected Birthweight = 21.52 × 100= 2152 grams
  [Range(estriol)=(7 , 27) so Estriol=0 is out of range]

PRACTICAL EXAMPLES

Birthweight ~ Normal(μ, 14.61)
Expected Birthweight= $\mu = 21.52 + 0.608 \times Estriol_i$

INTERPRETATION OF PARAMETERS

- If ESTRIOL= Mean(estriol)=17.2 =>

  Expected Birthweight = (21.52+0.608 × 17.2) × 100= 3198 grams

PRACTICAL EXAMPLES

Birthweight ~ Normal(μ, 14.61)
Expected Birthweight= $\mu = 21.52 + 0.608 \times Estriol_i$

INTERPRETATION OF PARAMETERS

- If two women differ by one unit of ESTRIOL=>

  Expected difference of Birthweight = 0.608 × 100= 60.8 grams

PRACTICAL EXAMPLES

Hypothesis Test:
Is the effect of ESTRIOL important for the Determination of BIRTHWEIGHT?

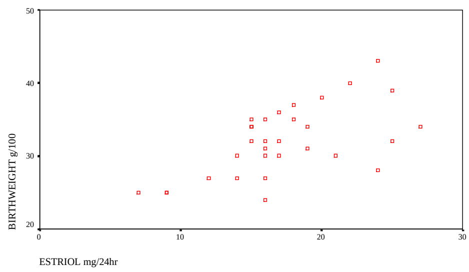
$H_0: \beta=0$ vs. $H_1: \beta \neq 0$

USE Statistical Functions and p.values
If p.value<0.05 the reject H_0

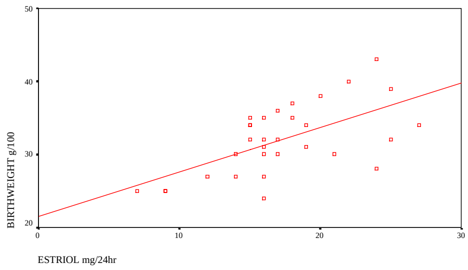
PRACTICAL EXAMPLES

Here p.value=0.000<0.05
so reject H_0
What is H_0?
H_0: $\beta=0$
What does this mean?
The effect of ESTRIOL level is significant for the Determination of BIRTHWEIGHT!

PRACTICAL EXAMPLES



PRACTICAL EXAMPLES



PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **January, 1986**: 25th flight in National Aeronautics and Space Administration's (NASA) space shuttle program.

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.40, January 28, 1986**
- **Temperature: 31 ^{o}F**
  -0.6 ^{o}C
- 7 Member Crew ready for take off

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.40, January 28, 1986**
- **Temperature: 31 ^{o}F**
  -0.6 ^{o}C
- 7 Member Crew ready for take off

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.40, January 28, 1986**
- **Temperature: 31 ^{0}F -0.6 ^{0}C**
- 7 Member Crew ready for take off

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.40, January 28, 1986**
- **Temperature: 31 ^{0}F -0.6 ^{0}C**
- 7 Member Crew ready for take off

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.40, January 28, 1986**
- **Temperature: 31 ^{0}F -0.6 ^{0}C**
- 7 Member Crew ready for take off

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.41, January 28, 1986**
- 73 seconds after the lift off …
- something seemed wrong

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.41, January 28, 1986**
- 73 seconds after the lift off …
- a large explosion destroyed the Challenger Space Shuttle

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- **11.41, January 28, 1986**
- 73 seconds after the lift off …
- a large explosion destroyed the Challenger Space Shuttle

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- All crew members were killed
- Billions of Dollars were lost
- The whole NASA program delayed

- WHAT HAPPENED IN THE MOST AMBITIOUS AND EXPENSIVE RESEARCH PROGRAM?

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

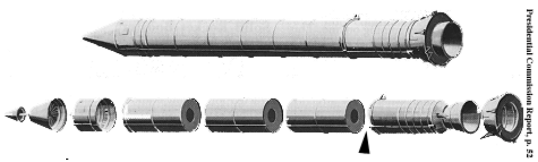
- A Presidential Commission was appointed to determine the cause of accident
- Head of the Commission: William Rogers
- Commission included: scientists+members of space exploration community.
- KEY PERSON: Richard Feynman (physicist)

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- The commission examined the accident and the events leading to the accident
- Two Volume Report: *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (1986)

PRACTICAL EXAMPLES



- Each Booster Rocket consists of several pieces whose joints are sealed with rubber O-rings

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

BACKGROUND INFORMATION:

O-rings:

- 37-foot (11.27 m) circles made of rubber
- designed to seal the booster sections of the rocket
- Prevent release of hot gases produced during combustion.
- Each joint between the segments contains two O-rings positioned concentric with the Solid Rocket Boosters (1 primary and 1 secondary).

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

BACKGROUND INFORMATION:

- Each Booster contains three Primary O-rings
- In the previous 23 flights they examined the hardware for O-ring damage (one was lost in the sea)
- The Forecasted temperature was 31 ^{0}F (-0.6 ^{0}C) while the coldest previous launch was on 53 ^{0}F (11.7 ^{0}C - ONE MAJOR MISTAKE)

PRACTICAL EXAMPLES

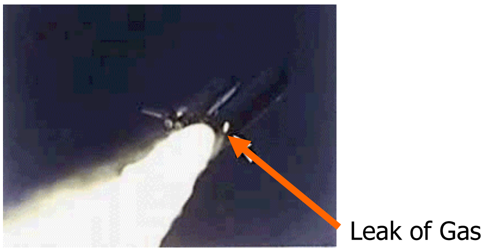
4.2. EXAMPLE 2

BACKGROUND INFORMATION:

- THE SENSITIVITY OF O-RINGS TO TEMPRETURE WAS WELL-KNOWN!!!
- WARM O-RING => Quickly Recover its shape after removal of compression
- COLD O-RING => Does not Recover its shape which may lead to gas leak and explosion!

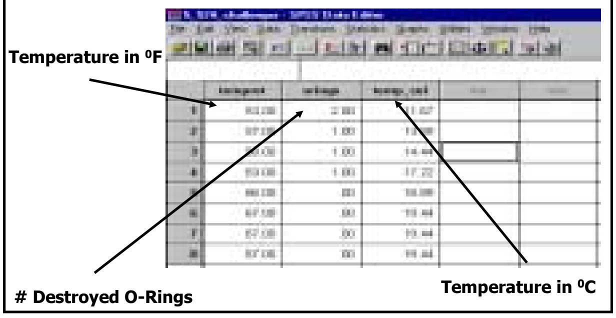
PRACTICAL EXAMPLES

4.2. EXAMPLE 2



PRACTICAL EXAMPLES

THE DATA

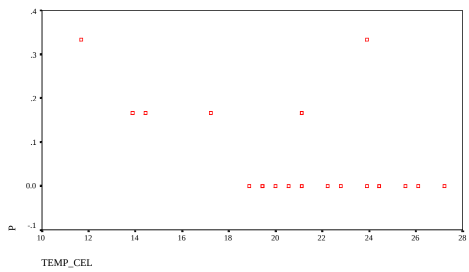


PRACTICAL EXAMPLES

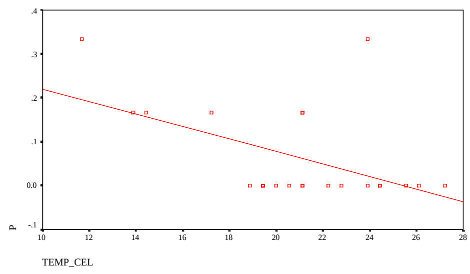
4.2. EXAMPLE 2

- 1st Step: Plot of O-Rings/6 per temperature
- 2nd Step: Fit a Bernoulli GLM to estimate the probability of at least one destroyed O-ring
- 3nd Step: Fit a Binomial GLM to predict number of destroyed O-rings

PRACTICAL EXAMPLES



PRACTICAL EXAMPLES



PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- 2nd Step: Logistic Regression (Bernoulli)
- Y (response):
  - 1 if at least one O-ring was damaged
  - 0 otherwise
- X (explanatory): Temperature in ^{0}C

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- p=probability of at least one damaged O-ring
- log(p/(1-p)) = 7.61 - 0.418 × ^{0}C

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- ODDS= p/(1-p) [Odds of at least one damaged O-Ring]
- ODDS = exp(- 0.418) = 0.658
- Increase one 1 ^{0}C decreases the odds of at least one damaged O-ring by 34.2% [=(1-0.658) × 100]

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- For -0.56 ^{0}C the model predicts
- $p = e^{7.61 - 0.418 \times (-0.56)}/[1+e^{7.61 - 0.418 \times (-0.56)}]$
- p = 0.99961 !!!!
- The probability to have at least one damaged O-Ring is almost 1.
- => P(# O Rings ≥ 3) = 0.957

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

- 3rd Step: Logistic Regression (Binomial)
- Y (response):
  - % of damaged O-rings (out of total 6)
- X (explanatory): Temperature in ^{0}C

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- p=proportion of damaged O-rings
- BEWARE THIS P IS DIFFERENT THAN THE P IN 2nd STEP MODEL
- E(Y)=np = Expected damaged O-Rings
- log(p/(1-p)) = 1.386 - 0.208 × ^{0}C

PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- ODDS= p/(1-p) [odds of damaged O-rings]
- ODDS = exp(- 0.208) = 0.812
- Increase one 1 ^{0}C decreases the odds of damaged O-rings by 18.8% [=(1-0.812) × 100]

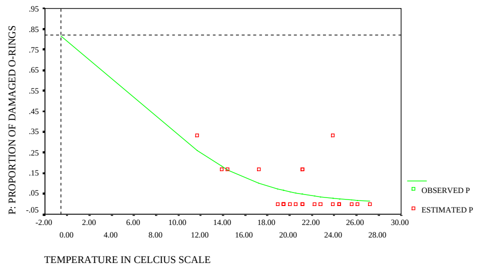
PRACTICAL EXAMPLES

4.2. EXAMPLE 2

RESULTS

- For -0.56 ^{0}C the model predicts
- $p = e^{1.386 - 0.208 \times (-0.56)}/[1+e^{1.386 - 0.208 \times (-0.56)}]$
- p = 0.818 ! =>
- Expected Damaged O-Rings = 6×0.818=4.91

PRACTICAL EXAMPLES



PRACTICAL EXAMPLES

4.2. EXAMPLE 2

CONCLUSIONS:

- ALL SERIOUS ANALYSIS LED TO THE CONCLUSION OF HIGH RISK OF EXPLOSION
- 7 LIVES AND BILLIONS OF DOLLARS WERE LOST BECAUSE NONE HAS DONE A SERIOUS ANALYSIS OF THE DATA

CONCLUSIONS

CLOSING REMARKS

- STATISTICS ARE USEFUL TOOLS
- WITH STATISTICAL MODELS:
  - WE CAN SEE RELATIOSHIPS
  - DESCRIBE REALITY
  - MAKE PREDICIONS
- A GOOD STATISTICAL ANALYSIS IS A GOOD ADVISOR

Statistical Modelling

End of Lecture

Ioannis Ntzoufras

E-mail: ntzoufras@aegean.gr

Department of
Business Administration,
University of the Aegean