SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

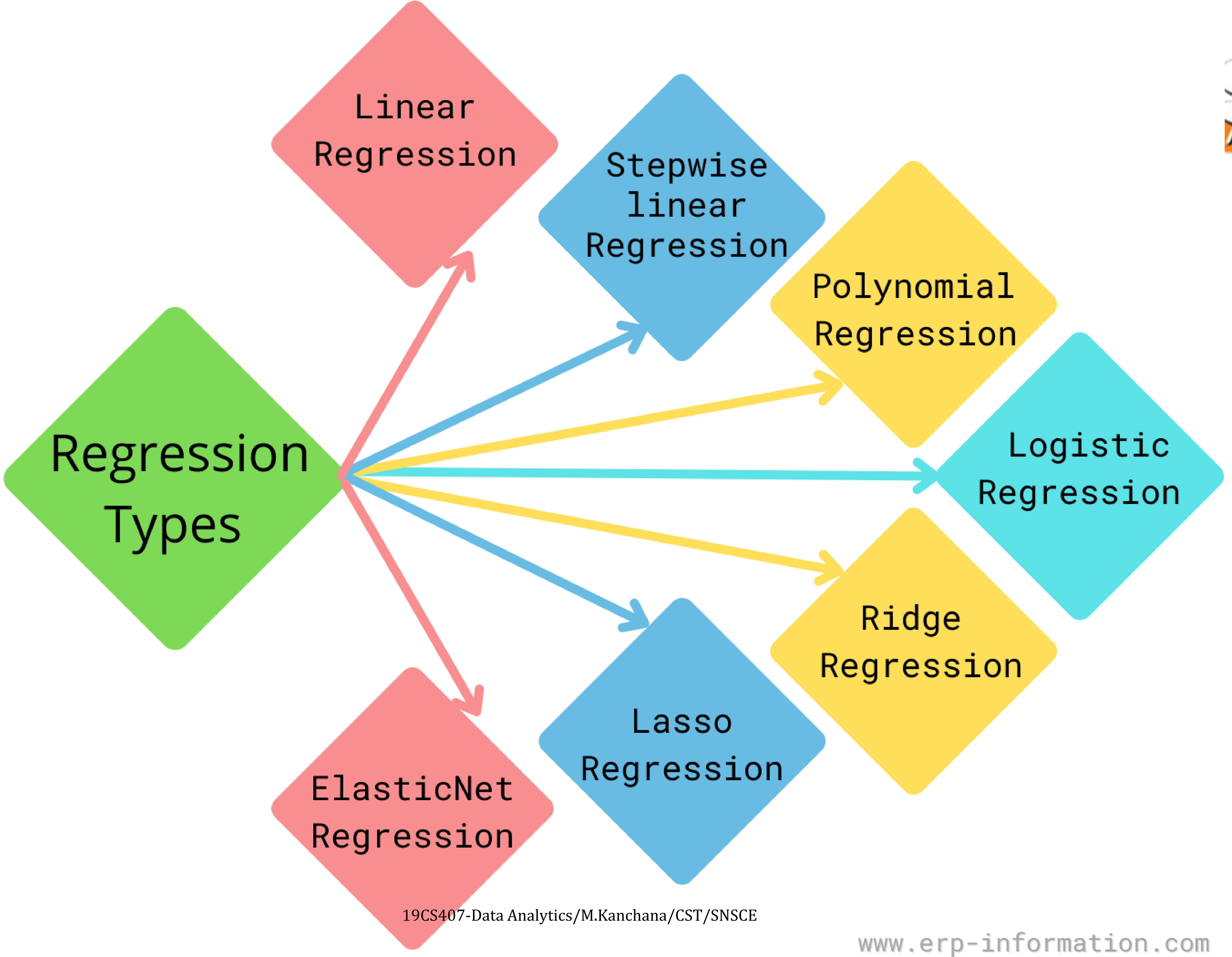
DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

COURSE NAME : 19CS407-DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit II – Statistics and Prescriptive Analytics

Topic : **Linear and Multiple Regression, Logistic Regression**


Regression Types
Linear Regression
Stepwise linear Regression
Polynomial Regression
Logistic Regression
Ridge Regression
Lasso Regression
ElasticNet Regression
www.erp-information.com

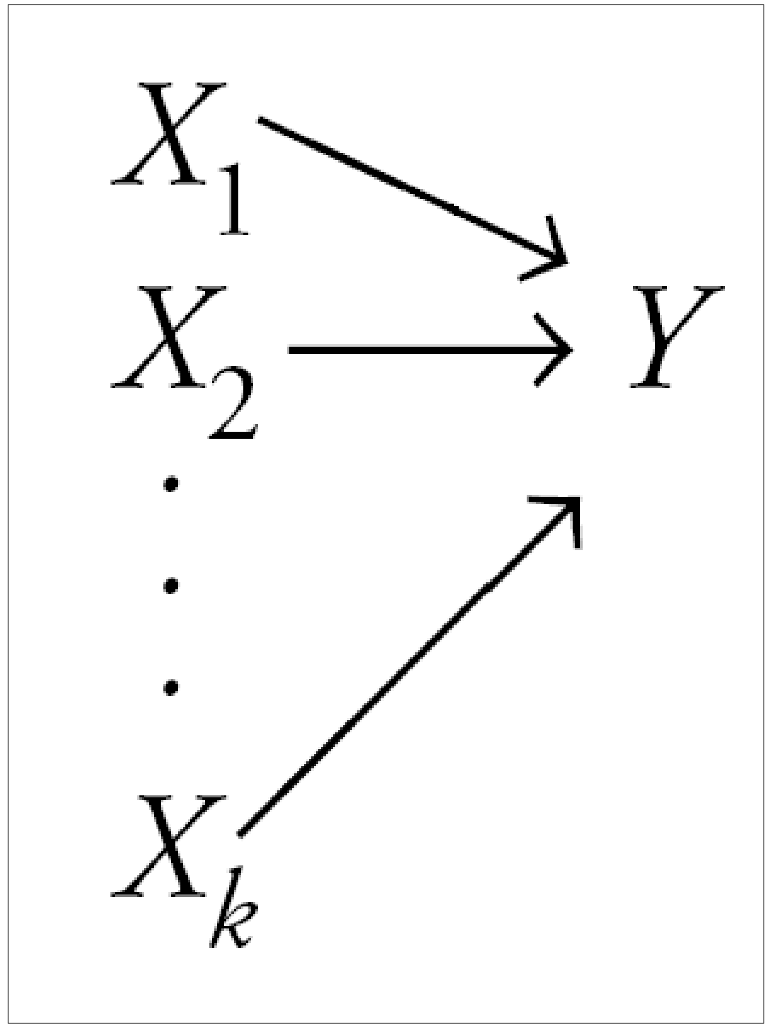
The General Idea

Simple regression considers the relation between a single explanatory variable and response variable

$$X \rightarrow Y$$

The General Idea

Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y

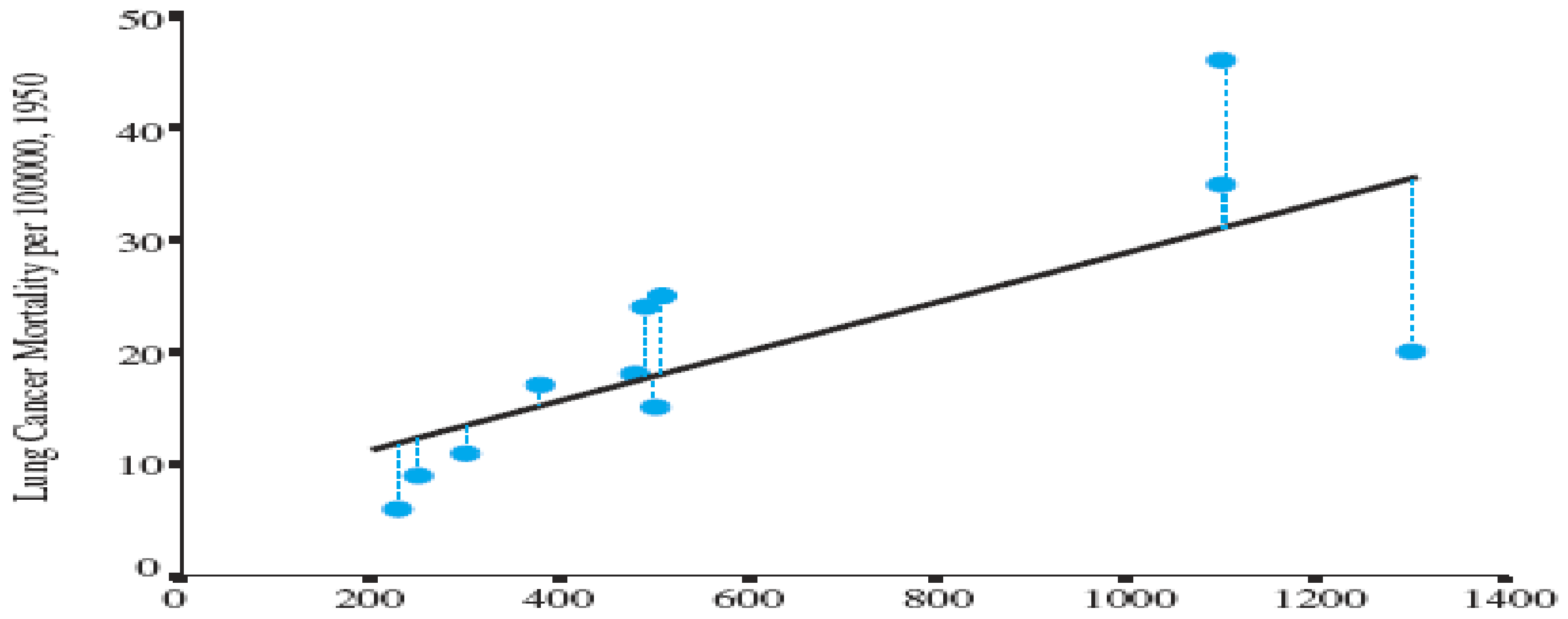


The intent is to look at the independent effect of each variable while “adjusting out” the influence of potential confounders

Regression Modeling

- A simple regression model (one independent variable) fits a regression *line* in 2-dimensional space

- A multiple regression model with two explanatory variables fits a regression plane in *3*-dimensional space



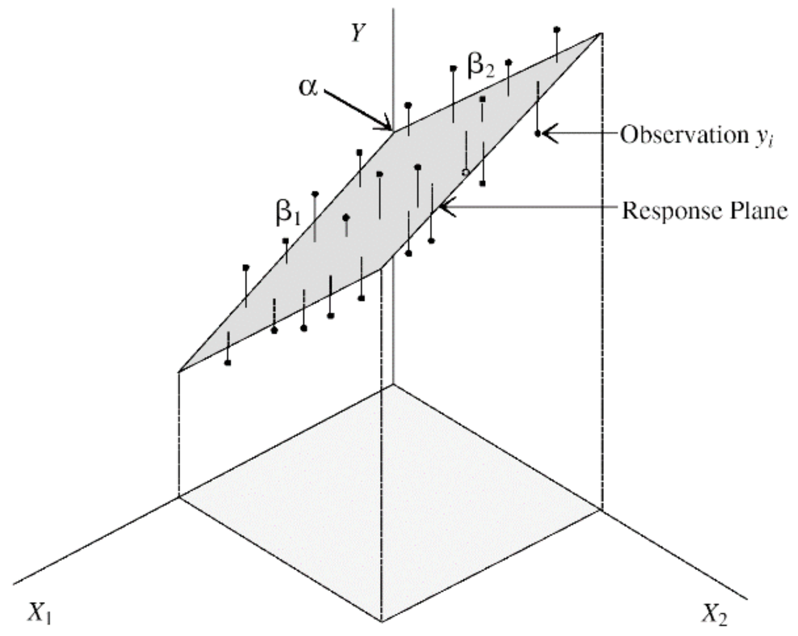


FIGURE 15.1 Three-dimensional response plane.

Simple Regression Model

Regression coefficients are estimated by minimizing ∑residuals² (i.e., sum of the squared residuals) to derive this model:

$$\hat{y} = a + bx$$

The standard error of the regression ($s_{Y|x}$) is based on the squared residuals:

$$S_{Y|x} = \sqrt{\sum \text{residuals}^2 / df_{\text{res}}}$$

Multiple Regression Model

Again, **estimates for the *multiple* slope coefficients** are derived by minimizing ∑residuals^2 to derive this multiple regression model:

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Again, the standard error of the regression is based on the ∑residuals^2:

$$S_{Y|x} = \sqrt{\sum \text{residuals}^2 / df_{\text{res}}}$$

Multiple Regression Model

- Intercept α predicts where the regression *plane* crosses the Y axis
- Slope for variable X_1 (β_1) predicts the change in Y per unit X_1 holding X_2 constant
- The slope for variable X_2 (β_2) predicts the change in Y per unit X_2 holding X_1 constant

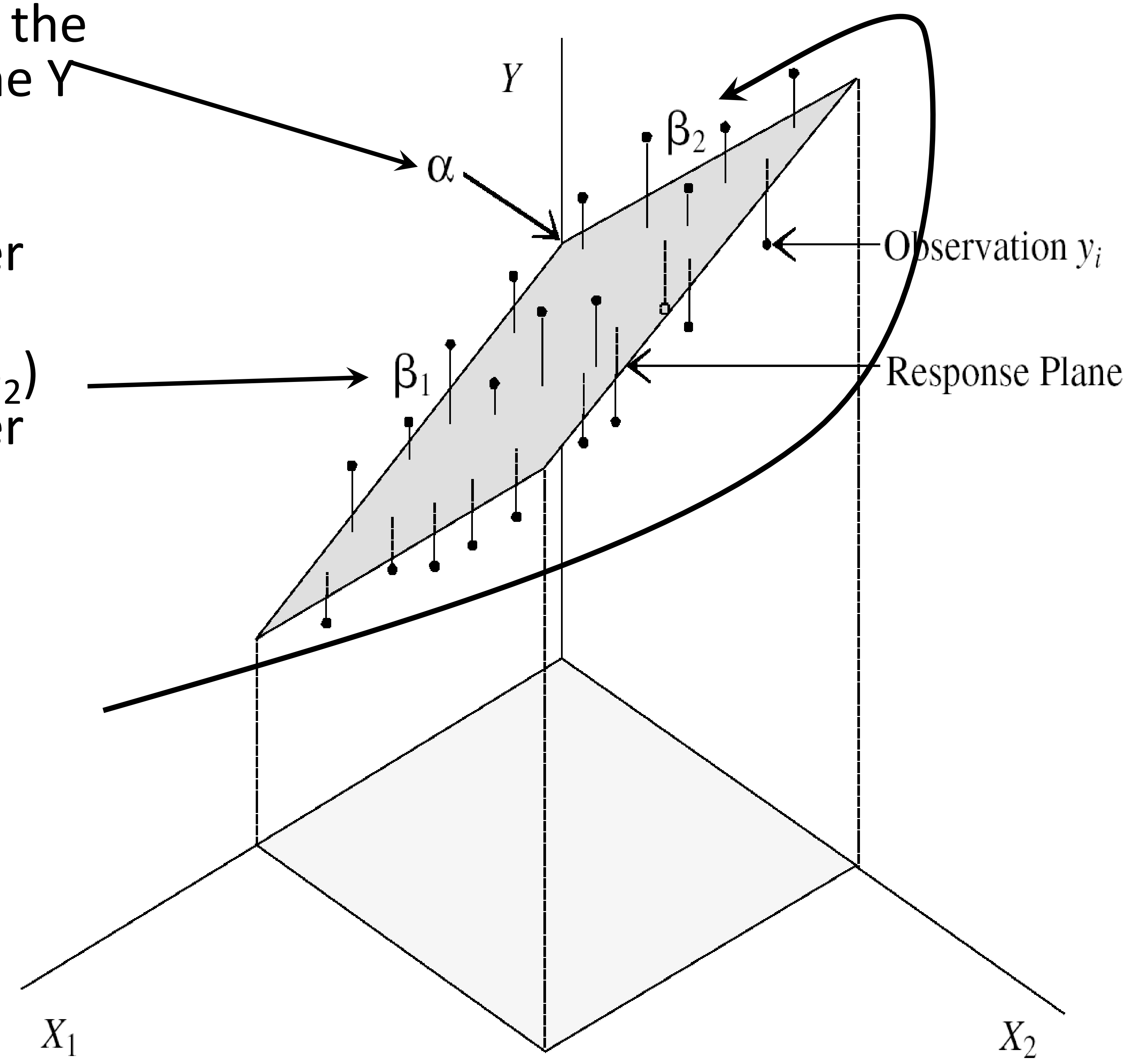
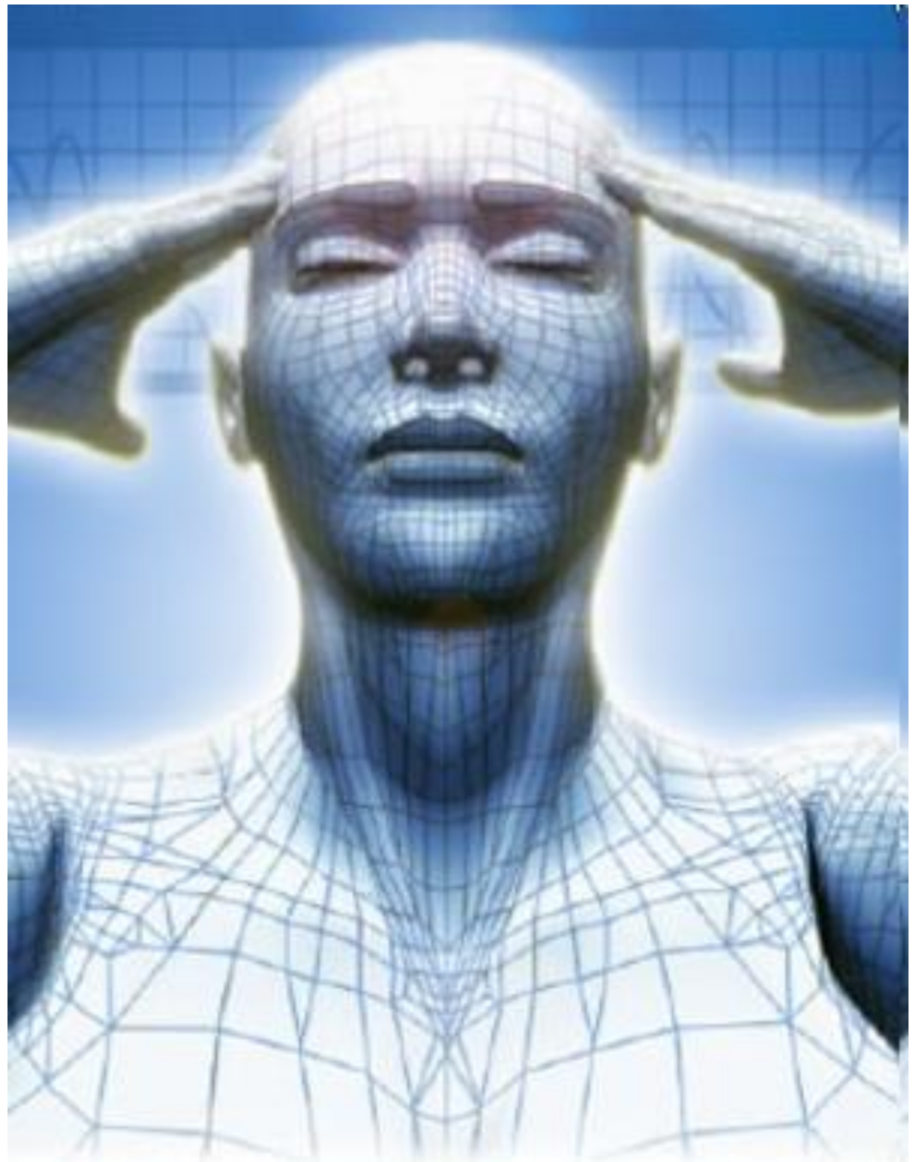


FIGURE 15.1 Three-dimensional response plane.

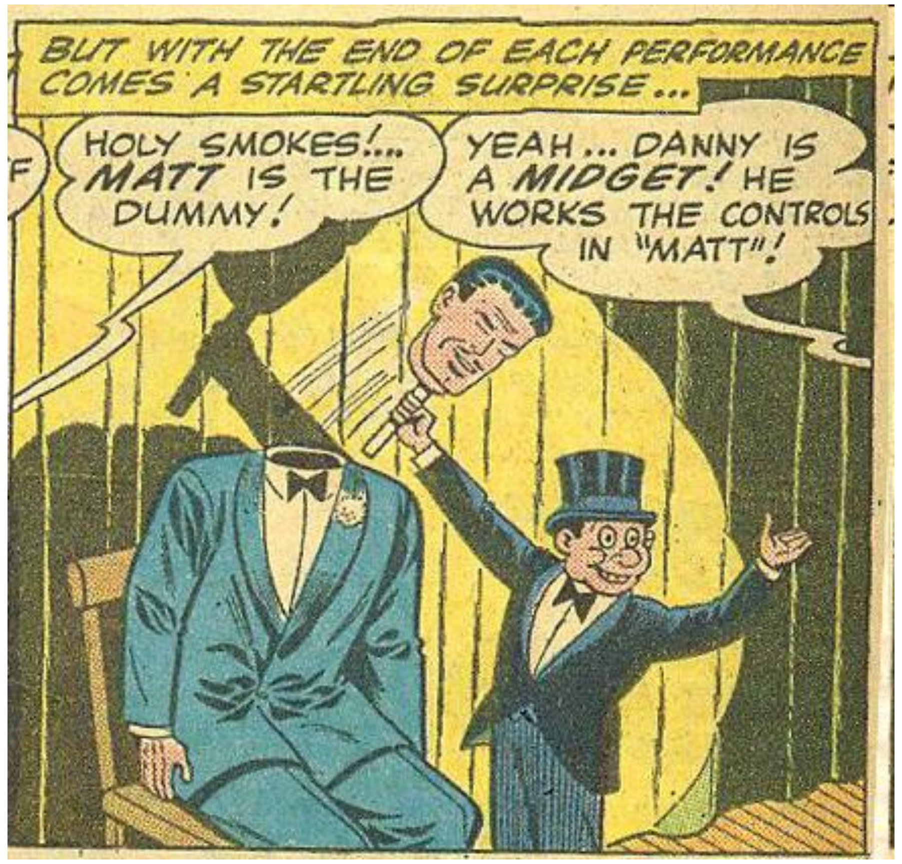
Multiple Regression Model

A multiple regression model with *k* independent variables fits a regression “surface” in k + 1 dimensional space (cannot be visualized)



Categorical Explanatory Variables in Regression Models

- Categorical independent variables can be incorporated into a regression model by converting them into 0/1 ("dummy") variables
- For binary variables, code dummies "0" for "no" and 1 for "yes"



Dummy Variables, More than two levels

For categorical variables with *k* categories, use *k*–1 dummy variables

SMOKE2 has three levels, initially coded
0 = non-smoker
1 = former smoker
2 = current smoker

Use *k* – 1 = 3 – 1 = 2 dummy variables to code this information like this:

SMOKE2	DUMMY1	DUMMY2
0	0	0
1	1	0
2	0	1

Illustrative Example

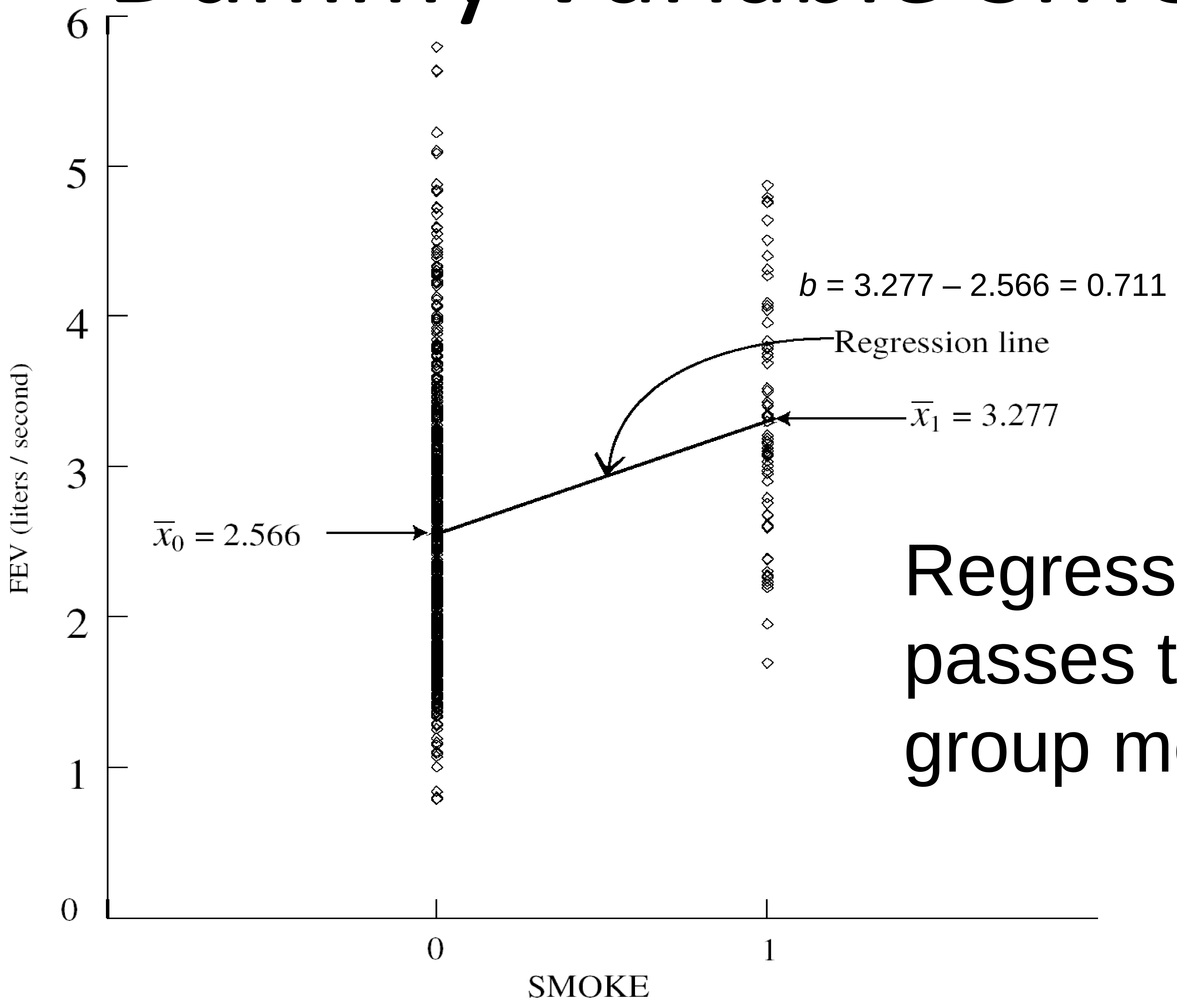
Childhood respiratory health survey.

- Binary explanatory variable (SMOKE) is coded 0 for non-smoker and 1 for smoker
- Response variable Forced Expiratory Volume (FEV) is measured in liters/second
- The mean FEV in nonsmokers is 2.566
- The mean FEV in smokers is 3.277

Example, cont.

- Regress FEV on SMOKE least squares regression line:
  $\hat{y} = 2.566 + 0.711X$
- Intercept (2.566) = the mean FEV of group 0
- Slope = the mean difference in FEV
  = 3.277 − 2.566 = 0.711
- $t_{stat} = 6.464$ with 652 *df*, $P \approx 0.000$ (same as equal variance *t* test)
- The 95% CI for slope β is 0.495 to 0.927 (same as the 95% CI for $\mu_1 - \mu_0$)

Dummy Variable SMOKE



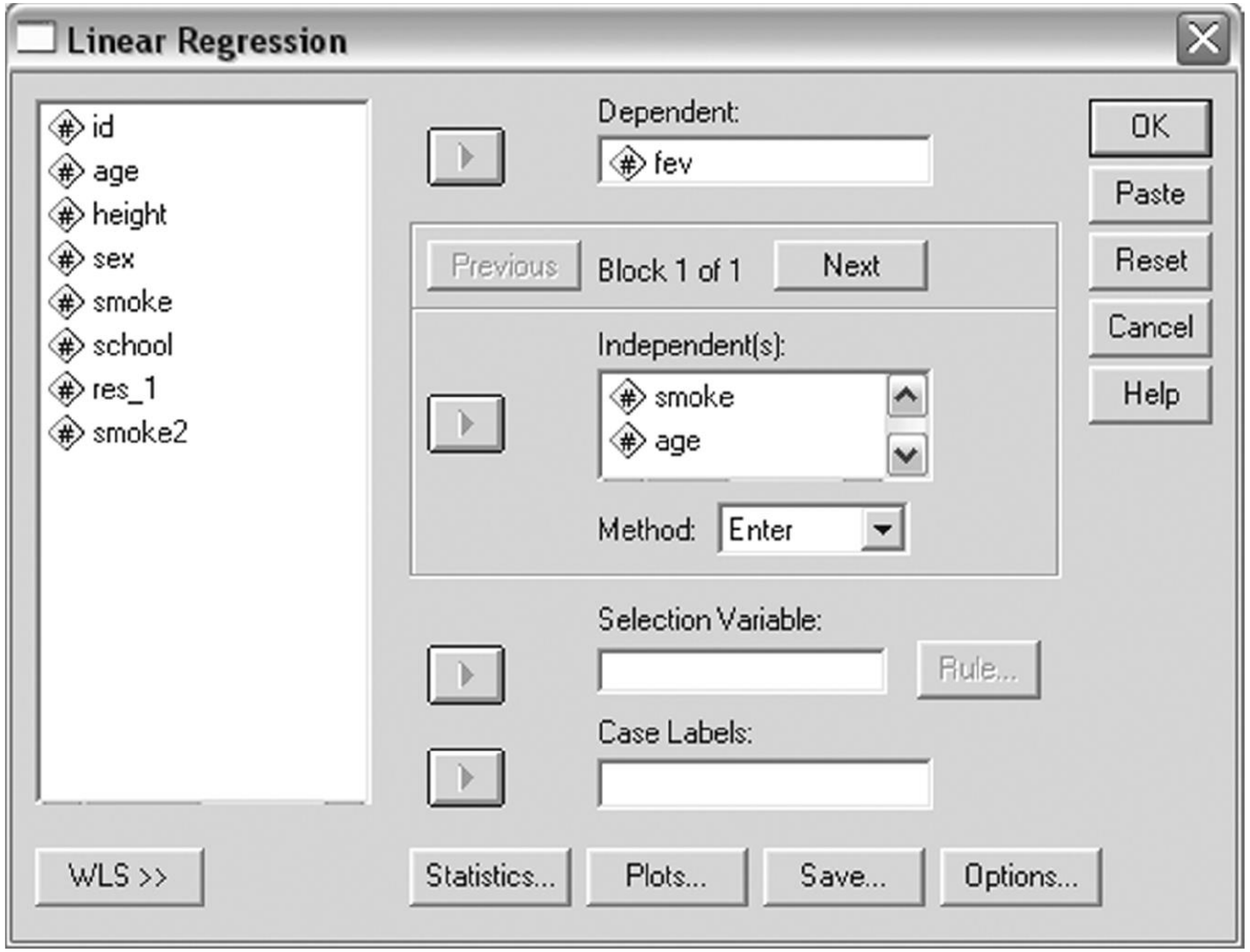
Regression line passes through group means

Smoking increases FEV?

- Children who smoked had higher mean FEV
- How can this be true given what we know about the deleterious respiratory effects of smoking?
- ANS: Smokers were older than the nonsmokers
- AGE confounded the relationship between SMOKE and FEV
- A multiple regression model can be used to adjust for AGE in this situation

Multiple Regression Coefficients

Rely on software to calculate multiple regression statistics



Example

SPSS output for our example:

Intercept *a*   Slope b_1 Slope b_2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.367	.081		4.511	.000
	SMOKE	−.209	.081	−.072	−2.588	.010
	AGE	.231	.008	.786	28.176	.000

a. Dependent Variable: FEV

The multiple regression model is:

FEV = 0.367 + −.209(SMOKE) + .231(AGE)

Multiple Regression Coefficients, cont.

- The slope coefficient associated for SMOKE is −.206, suggesting that smokers have .206 *less* FEV on average compared to non-smokers (after adjusting for age)

- The slope coefficient for AGE is .231, suggesting that each year of age in associated with an increase of .231 FEV units on average (after adjusting for SMOKE)

Inference About the Coefficients

Inferential statistics are calculated for each regression coefficient. For example, in testing

H_0: $\beta_1 = 0$ (SMOKE coefficient controlling for AGE)

$t_{stat} = -2.588$ and $P = 0.010$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.367	.081		4.511	.000
	smoke	-.209	.081	-.072	-2.588	.010
	age	.231	.008	.786	28.176	.000

a. Dependent Variable: fev

$df = n - k - 1 = 654 - 2 - 1 = 651$

Inference About the Coefficients

The 95% confidence interval for this slope of SMOKE controlling for AGE is −0.368 to − 0.050.

Coefficients[a]

Model		95% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	.207	.527
	smoke	-.368	-.050
	age	.215	.247

a. Dependent Variable: fev

References

TEXT BOOKS

1.João Moreira, Andre Carvalho, Tomás Horvath – "A General Introduction to Data Analytics" – Wiley - 2018

2.An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics. W. N. Venables, D.M. Smith and the R Development Core Team. Version 3.0.1 (2013-05-16). URL: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

References:

1**. Dean J, —***Big Data, Data Mining and Machine learning***,** Wiley publications, 2014.

2. Provost F and Fawcett T, —*Data Science for Business*, O'Reilly Media Inc, 2013.

3. Janert PK, —*Data Analysis with Open Source Tools*, O'Reilly Media Inc, 2011.

4. Weiss SM, Indurkhya N and Zhang T, —*Fundamentals of Predictive Text Mining,* Springer-Verlag London Limited, 2010.

5.Marz N and Warren J,- Big Data, Manning Publications,2015

Thank You