

Multivariate Analysis Techniques

Multiple Regression Analysis

Logistic Regression Analysis

Discriminant Analysis

Multivariate Analysis of Variance (MANOVA)

Factor Analysis

Cluster Analysis

Multidimensional Scaling

Correspondence Analysis

Conjoint Analysis

Canonical Correlation

Structural Equation Modelling

First of all: What is it?

- **a choice model**

- variation of multiple regression

- allows for **prediction of an event**

- can utilize non-metric (typically **binary**) **dependent variables**, as the objective is to arrive at a **probabilistic assessment of a binary choice**

- independent variables can be either discrete or continuous

- **a contingency table** is produced, which shows the classification of observations as to whether **observed and predicted events match**

- **measure of the effectiveness of the model:**

- the sum of events that were predicted to occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events

Linear regression and the logistic regression model

linear regression analysis:

- ⇒ test whether 2 variables are linearly related
- &
- ⇒ calculate the strength of the relationship

Linear regression and the logistic regression model

linear regression analysis:

$$Y = \alpha + \beta X$$

Y = dependent variable

X = predictor

α = *intercept*, i.e. value of Y when X is zero

β = *slope*, i.e. change of Y associated with one-unit increase of X

Linear regression and the logistic regression model

multiple regression analysis:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

Y = dependent variable

X = predictor

α = *intercept*

β_i = *partial slope coefficients*

Linear regression and the logistic regression model

multiple regression analysis:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

Y = dependent variable

X = predictor

α = *intercept*

β_i = *partial slope coefficients*

ϵ = *error term, a random variable*

Linear regression and the logistic regression model

multiple regression analysis:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

*estimates of intercept and regression coefficients (α, β) are obtained mathematically using the method of ordinary least squares (OLS)
(cf. practically every intro to statistics text)*

Linear regression and the logistic regression model

multiple regression analysis:

$$\ddot{Y} = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k (+\epsilon)$$

\ddot{Y} value of Y predicted by the linear regression equation

a is the OLS estimate of the intercept, α

b is the OLS estimate for the slope, β

residuals (values for ϵ) are equal to $(Y_j - \ddot{Y}_j)$

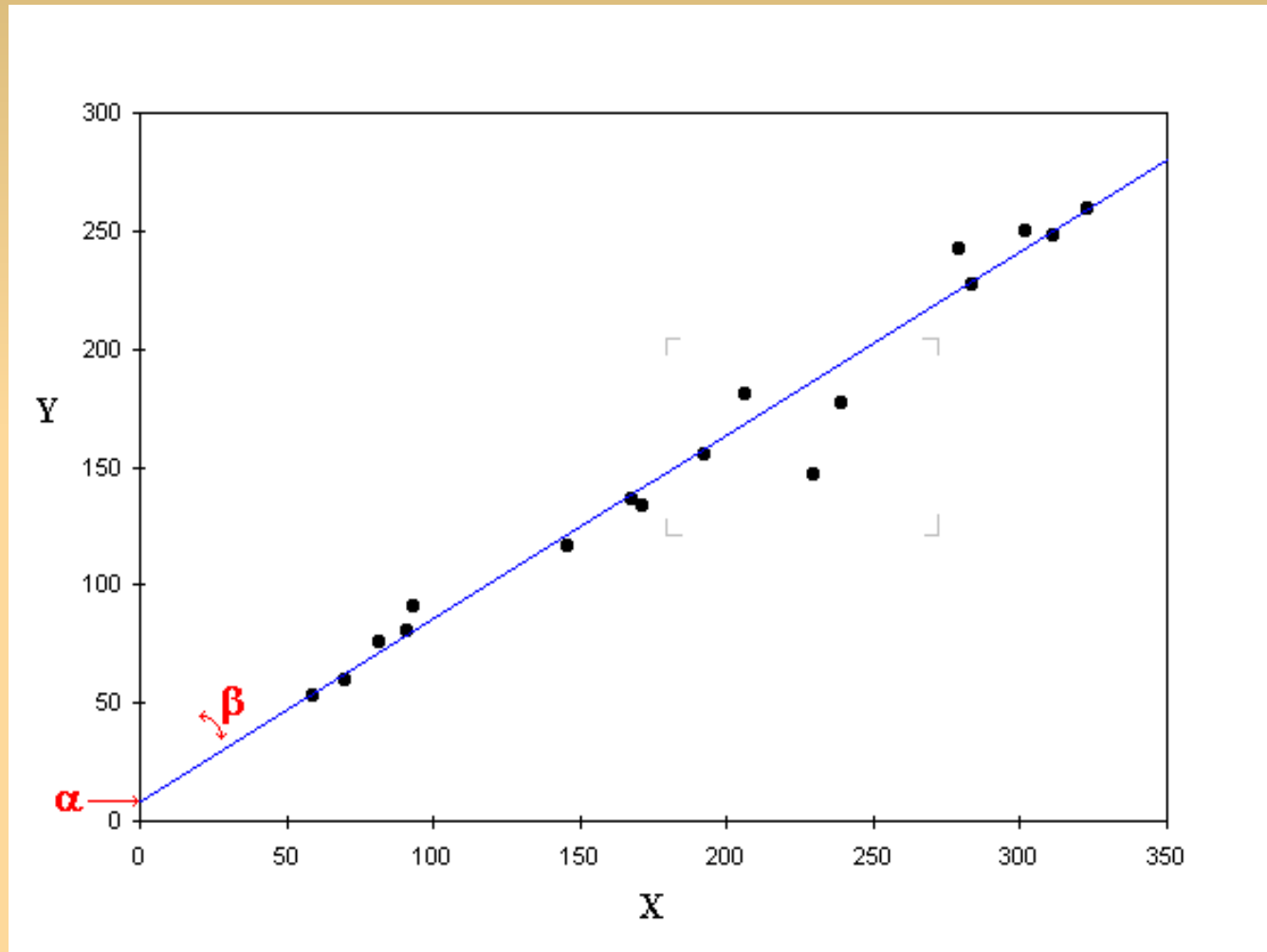
Linear regression and the logistic regression model

multiple regression analysis:

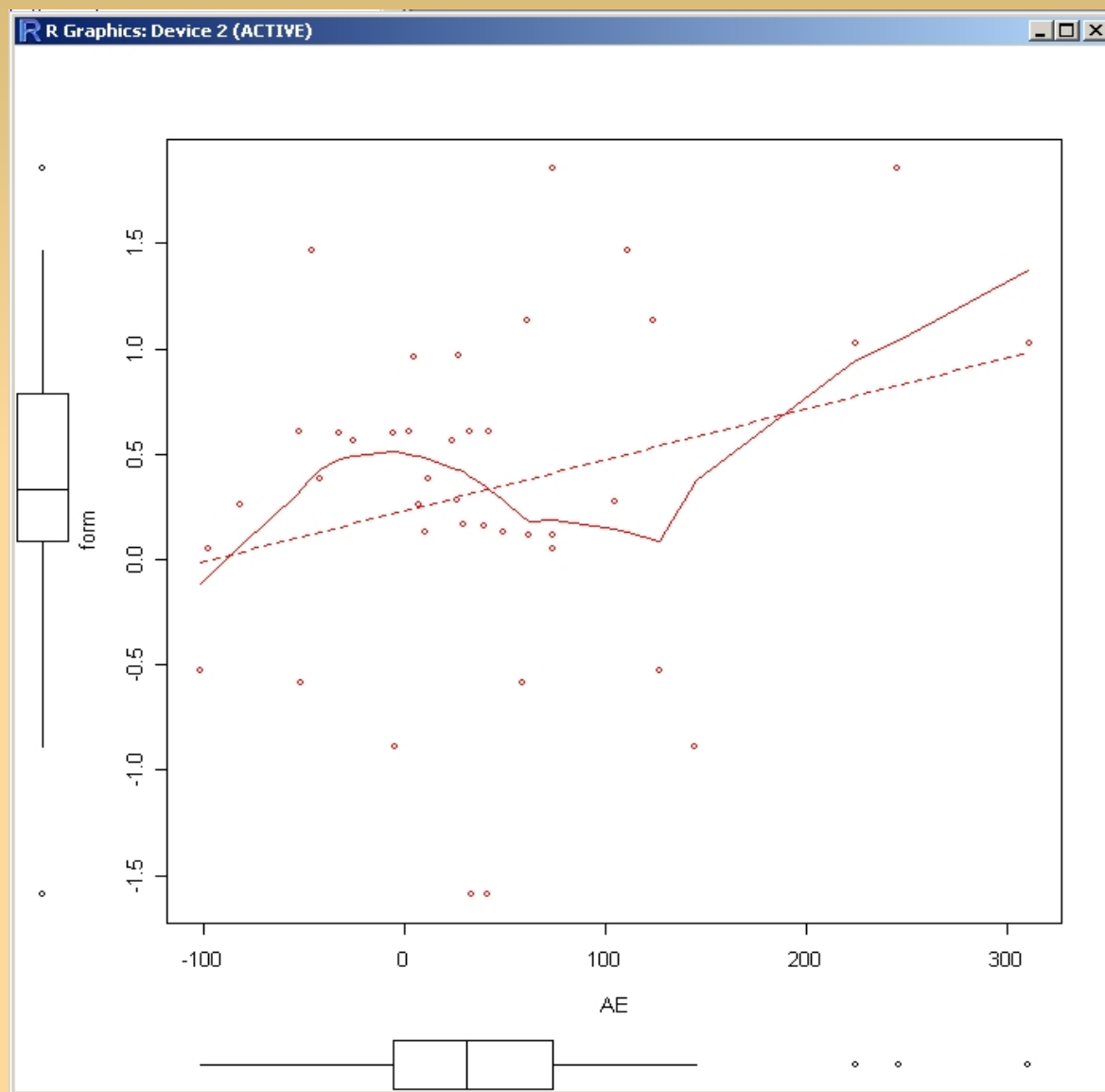
for bivariate regressions, we can represent these data graphically by the vertical distance between each point in a **bivariate scatterplot** and the regression line

(this is, of course, a little trickier for multiple regressions because it requires multiple dimensions)

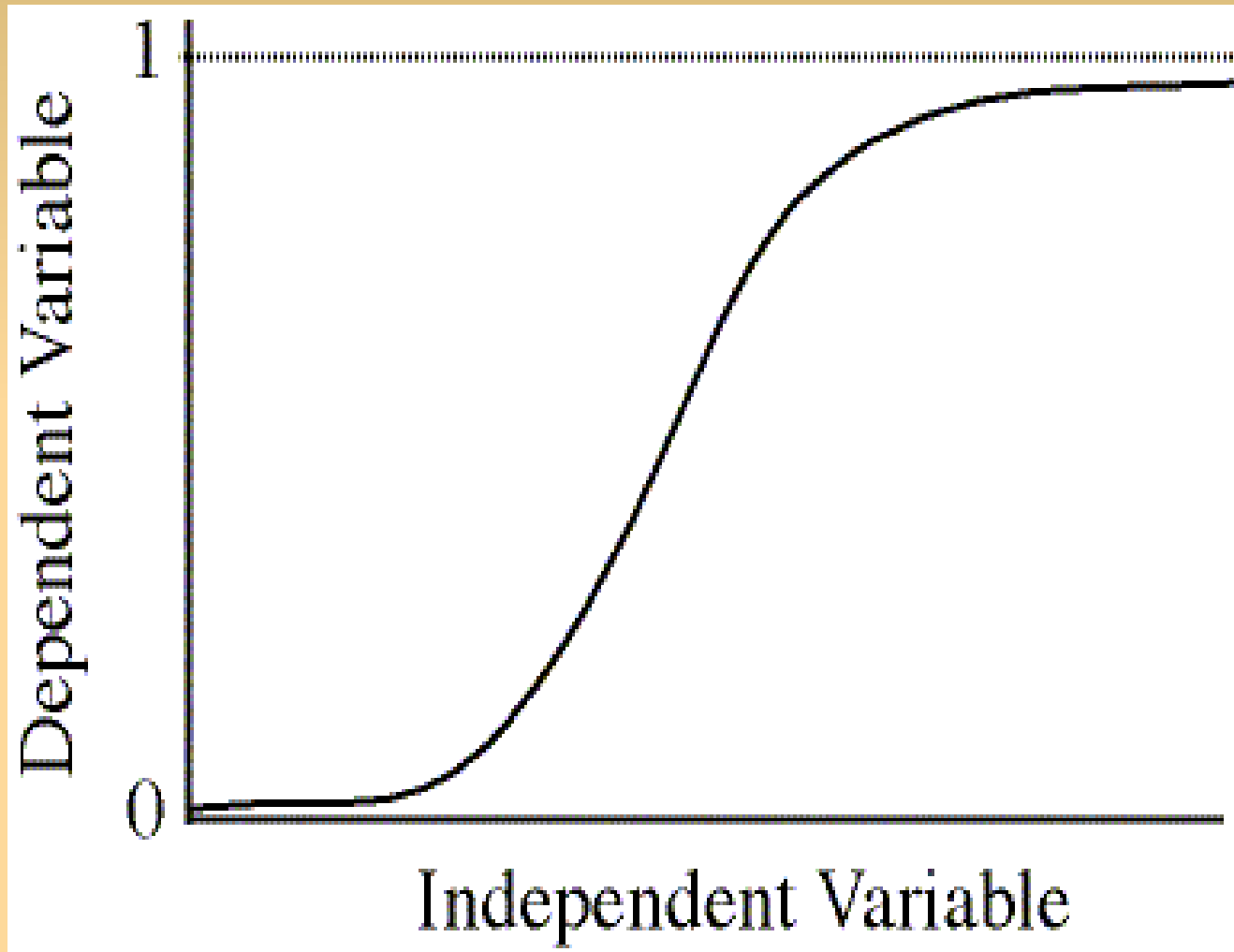
Linear regression



Linear regression - an example: Wiechmann (to appear)



Logistic curve model for a dichotomous dependent variable



Nonlinear Relationships and Variable Transformations

Q: So, what if a relationship appears to be nonlinear?

A: *Transform* either dependent or (one or more) predictor variables so that

substantive relationship remains *non-linear*,
but the form *form* or the relationship is *linear*

“nonlinear in terms of its variable but linear
in terms of its parameters” B&F 1985

One possible transformation that can be used to model
non-linearity is

logarithmic transformation

“nonlinear in terms of its variable but linear
in terms of its parameters” B&F 1985

logarithmic transformation

$$(Y+1) = e^{\alpha + \beta X}$$

or

$$Y = e^{\alpha + \beta X} - 1$$

note: Adding 1 avoids taking the nat. log. of zero (undefined)

$$e = 2.72$$

probabilities, odds and odds ratios

...our problem will always be to **predict the probability** that a case will be classified **one** as opposed to the other of the two **categories** of the **dependent variable**...

probabilities, odds and odds ratios

The probability of being classified into the first (or lower-valued) category, $P(Y=0)$, is equal to 1 minus the probability of being classified into the second (or higher-valued) category, $P(Y=1)$.

probabilities, odds and odds ratios

Hence, we know the one probability, if we know the other

$$P(Y=0) = 1 - P(Y=1)$$

We could try and model the probability that $Y=1$ as

$$P(Y=1) = \alpha + \beta X$$

probabilities, odds and odds ratios

But there is a catch...

observer values for $P(Y=1)$ must lie between 0 and 1,
predicted values may be less than 0 or greater than 1

⇒ replace the *probability* that $Y=1$
with the *odds* that $P=1$

Odds($Y=0$) is the ratio of the probability that $Y=1$ to the probability that $Y \neq 1$ or

$$\text{Odds}(Y=1) = P(Y=1) / [1 - P(Y=1)]$$

probabilities, odds and odds ratios

Logarithmic transformation of odds produces a variable that can vary from negative to positive infinity:

natural logarithm of the odds:

$$\ln \{P(Y=1)/[1-P(Y=1)]\}$$

is called logit (Y)

probabilities, odds and odds ratios

$$\text{logit}(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$$

$$\text{or: Odds}(Y=1) = e^{\ln[\text{Odds}(Y=1)]}$$

...and converting back to probabilities, we get:

$$P(Y=1) = \frac{e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{(1 + e^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k})}$$