ACIT4830 – Special Robotics and Control Subject

Topic5 – Regression methods

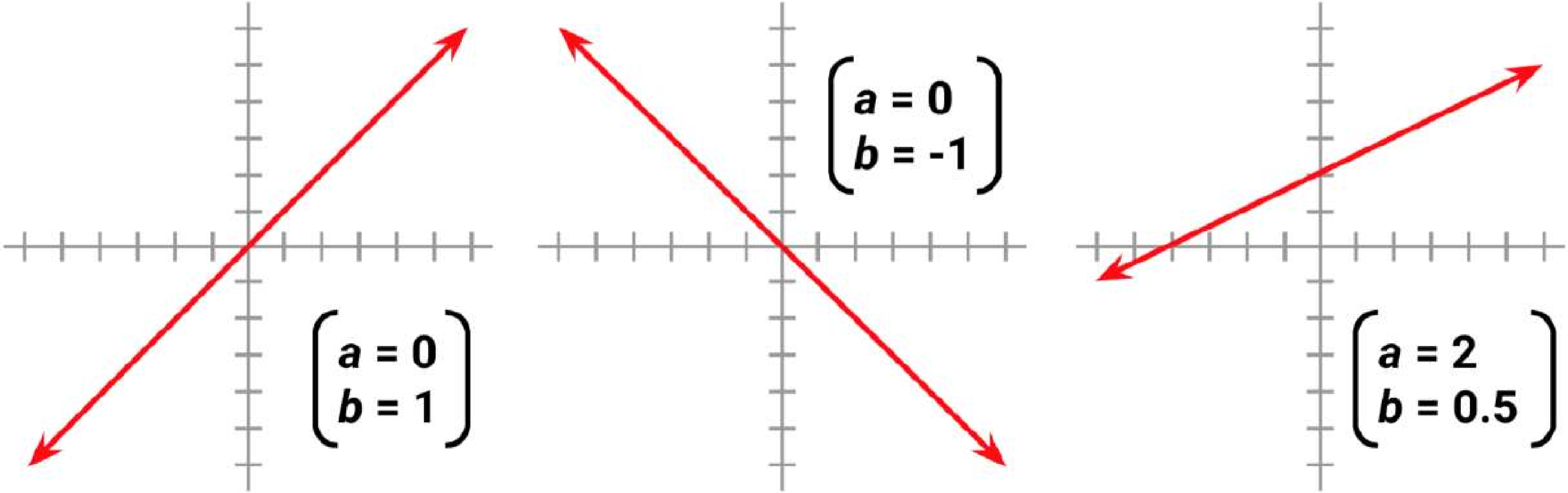
Evi Zouganeli
OsloMet – Oslo Metropolitan University
(evizou@oslomet.no)

Content

(Chapter 6 – Regression methods)

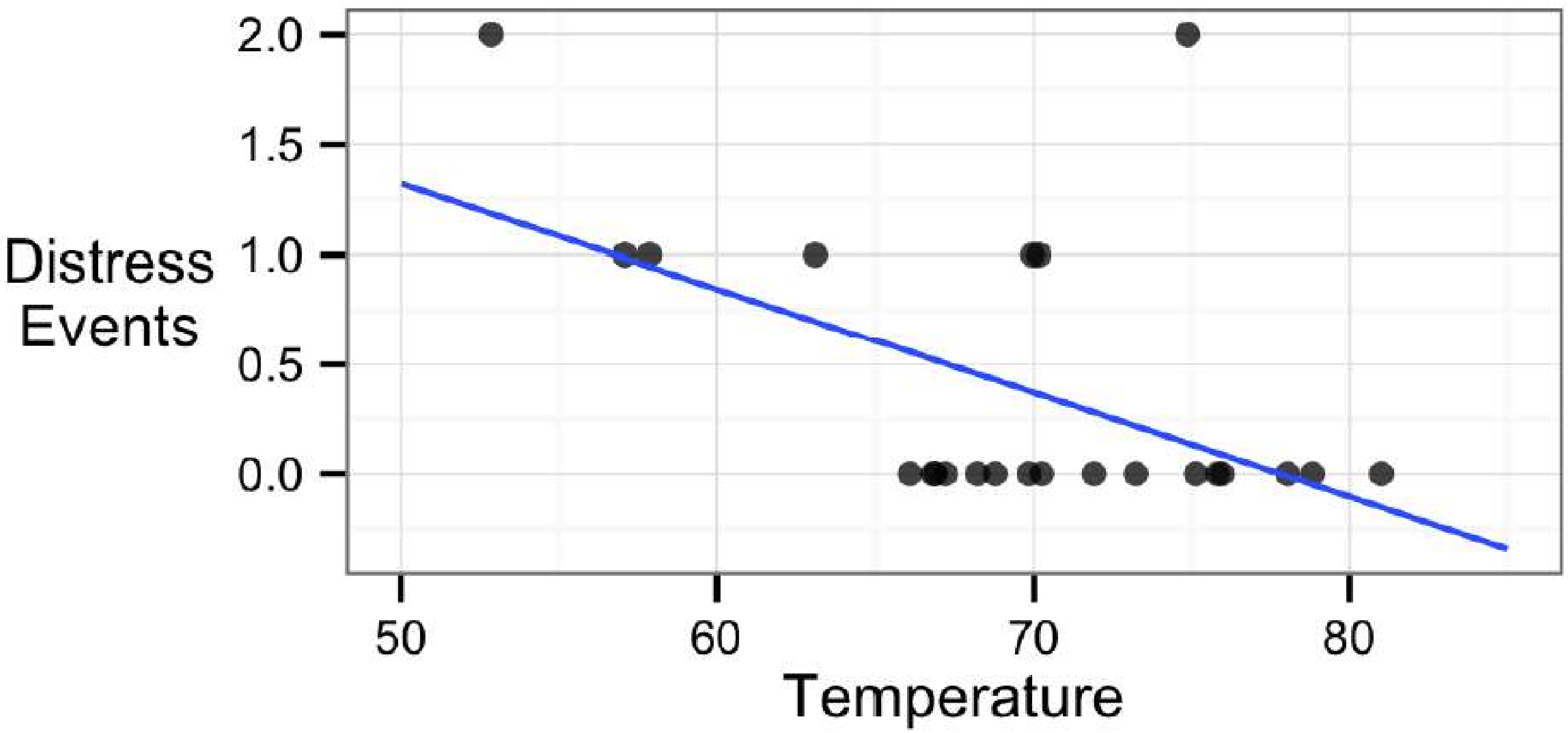
- Linear and Multiple regression
- Non-linear regression
- Regression trees
- Model trees

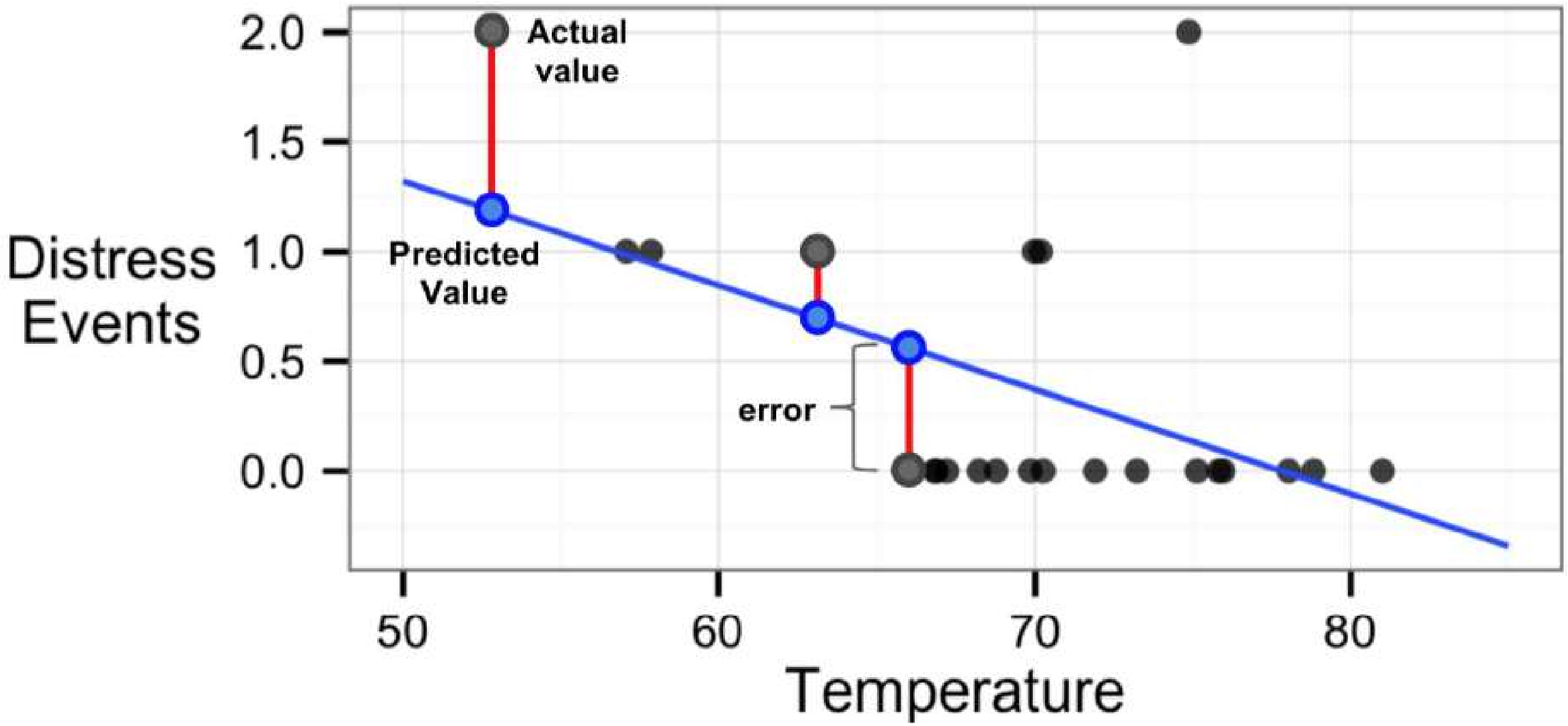
Linear Regression



$$y = \alpha + \beta x$$

Linear Regression





Minimize error: $$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

sum least squares

Linear Regression

$$y = \alpha + \beta x$$

$$a = \overline{y} - b\overline{x}$$

$$b = \frac{\sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum\left(x_i - \overline{x}\right)^2}$$

$$b = \frac{Cov(x, y)}{Var(x)}$$

$$\mathrm{Var}(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$ variance

$$\mathrm{Cov}\ (X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n}$$ covariance

Pearson's correlation
(-1 to +1)

$$\rho_{x,y} = Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Multiple Regression

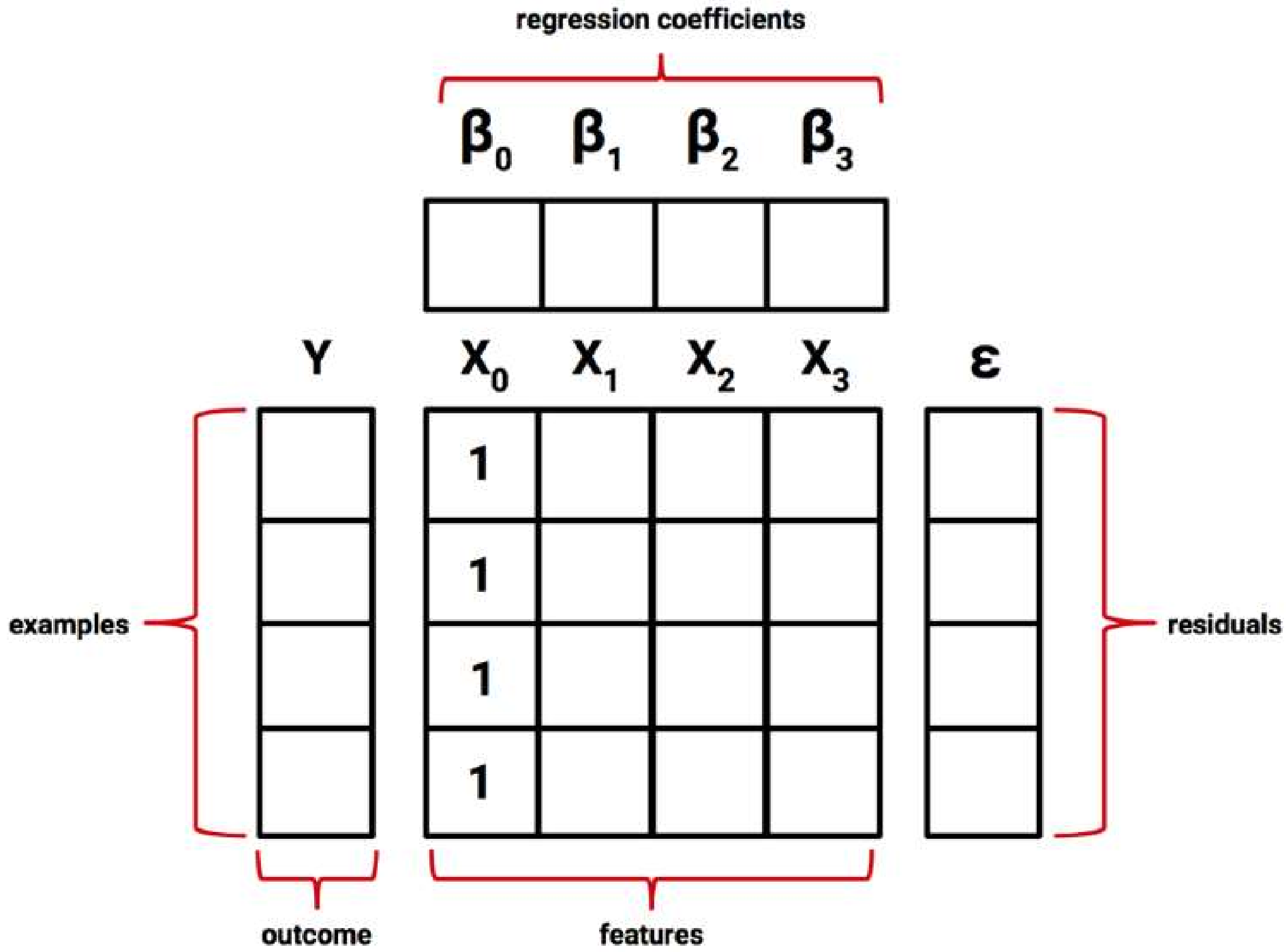
Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any data • Provides estimates of the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not do well with missing data • Only works with numeric features, so categorical data require extra processing • Requires some knowledge of statistics to understand the model

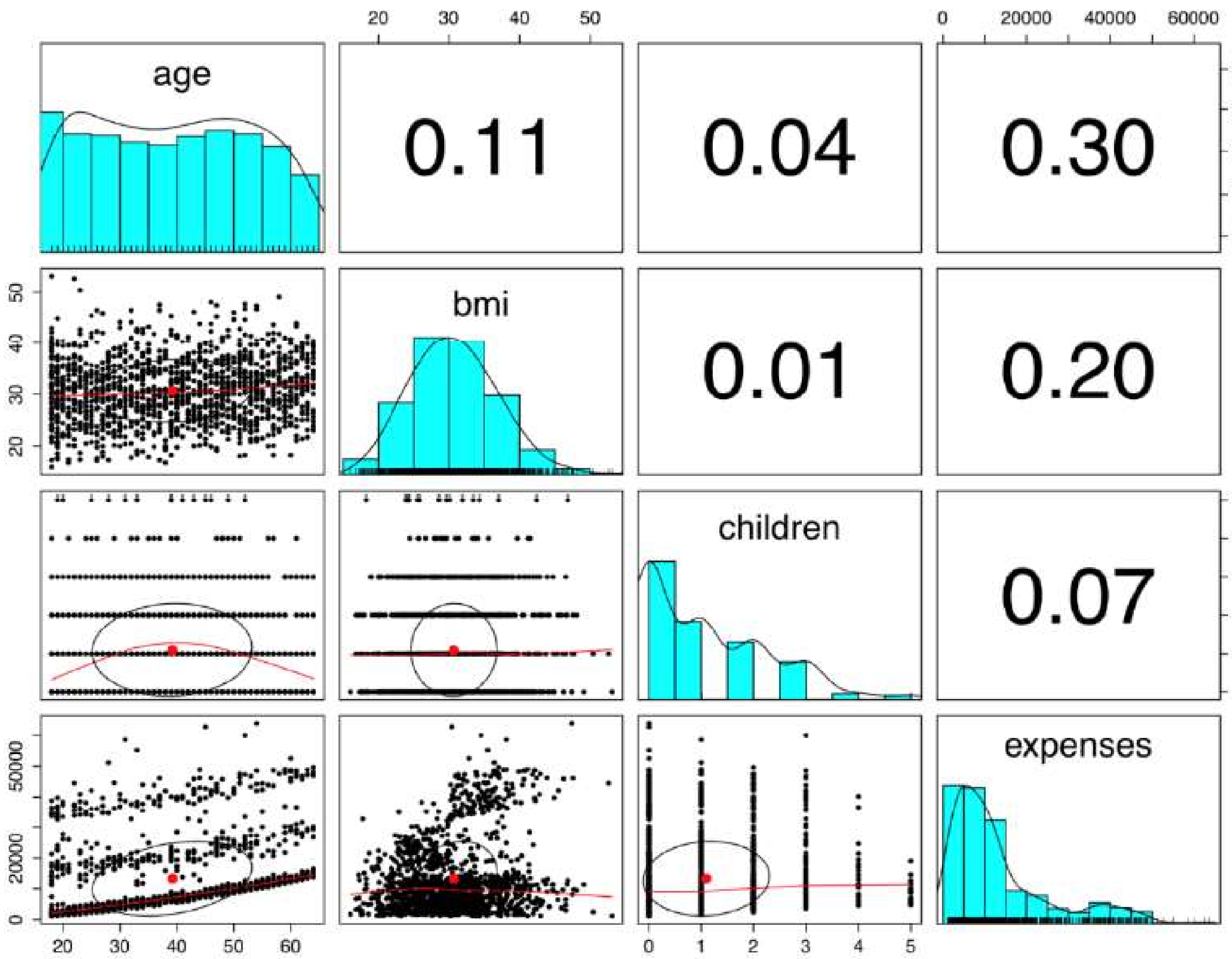
Multiple Regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Multiple Regression





age
bmi
children
expenses
0.11
0.04
0.30
0.01
0.20
0.07

Non-linear Regression

$$y = \alpha + \beta_1 x + \beta_2 x^2$$

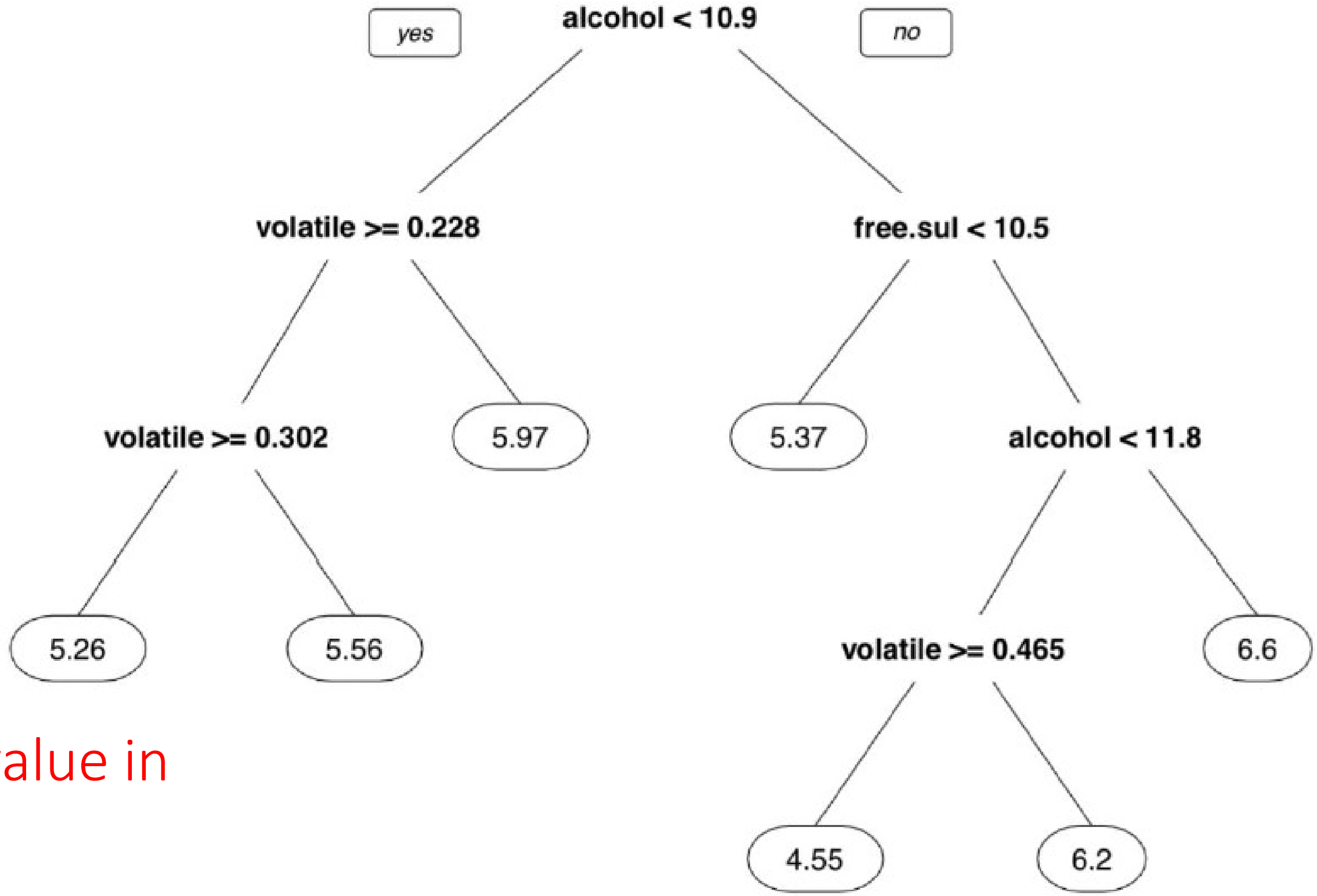
Model adjustment
– medical expences example in hands-on

- Added a non-linear term for age
- Created an indicator for obesity
- Specified an interaction between obesity and smoking

Adding Regression to Trees

Strengths	Weaknesses
• Combines the strengths of decision trees with the ability to model numeric data • Does automatic feature selection, which allows the approach to be used with a very large number of features • Does not require the user to specify the model in advance • May fit some types of data much better than linear regression • Does not require knowledge of statistics to interpret the model	• Not as commonly-used as linear regression • Requires a large amount of training data • Difficult to determine the overall net effect of individual features on the outcome • May be more difficult to interpret than a regression model

Regression Tree



mean value in subset

Regression Trees

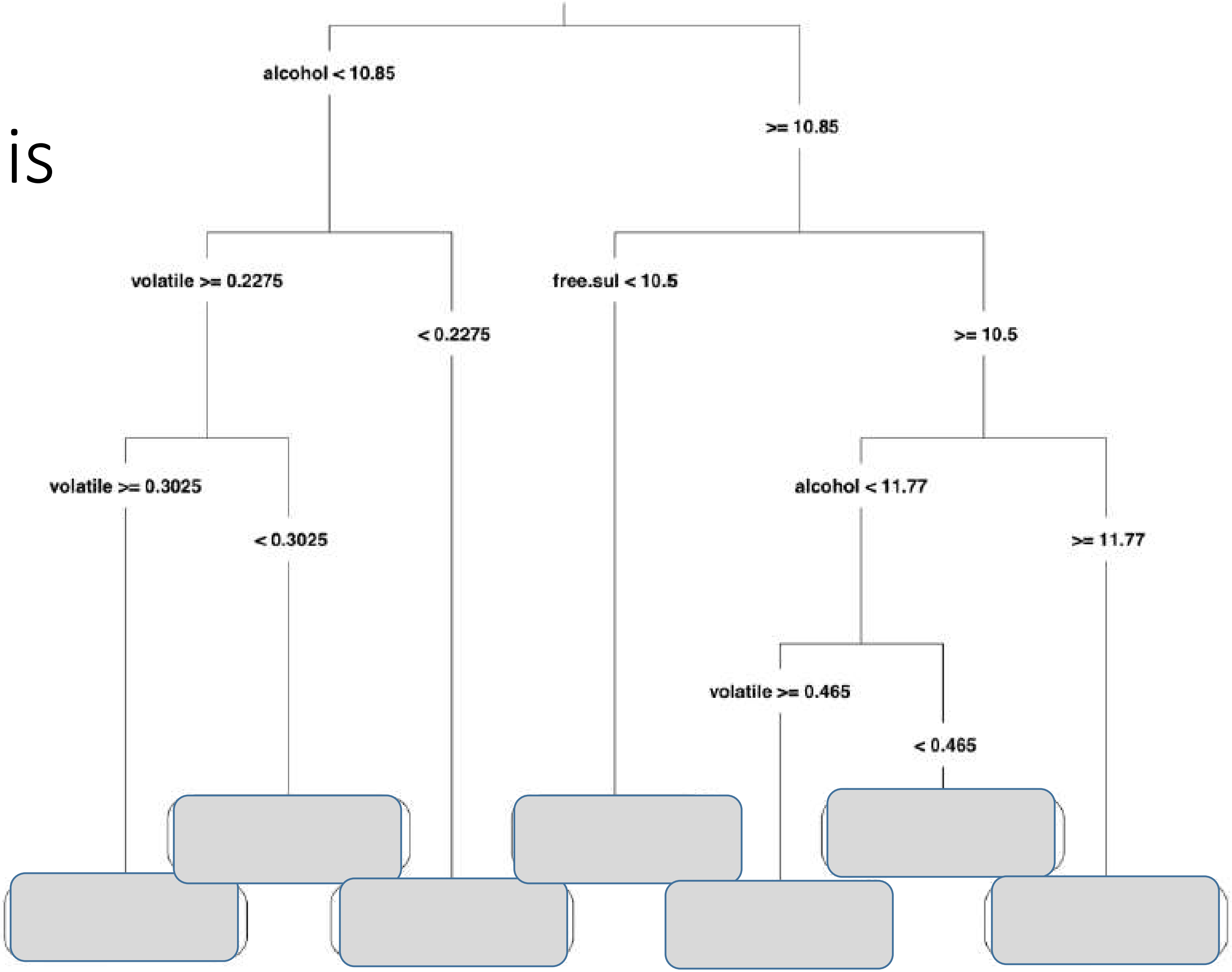
original data	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature A	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature B	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7

$\mathbf{T_1}$ $\mathbf{T_2}$

Standard Deviation Reduction $$\mathrm{SDR} = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

Choose the feature and split that **maximizes SDR**

Model Tree: every branch is a separate regression model



Performance evaluation

Mean Absolute Error $$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$