Yesterday - Tomorrow Data

Aldo Solari

Yesterday

- From AS, Chapter 3
- Yesterday we observed $n = 30$ observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

- These data are artificially generated by the model

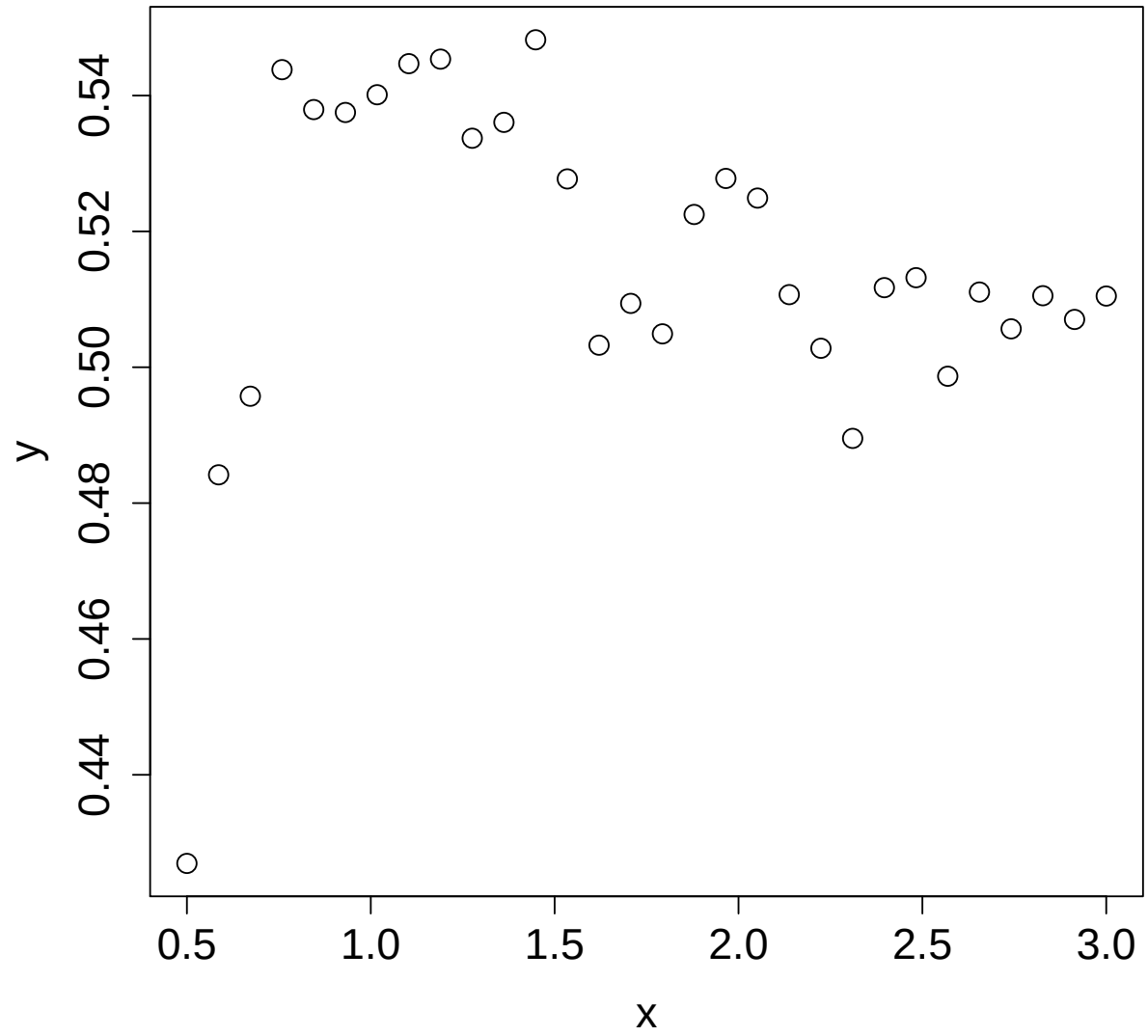
$$Y = f(x) + \text{error}$$

  where $f(x)$ is a unspecified smooth and regular function
- We wish to obtain a model

$$\hat{y} = \hat{f}(x)$$

  that allows us to predict y as new observations of x become available, say tomorrow

`yesterday` **data**



Mean squared error

- The MSE for the yesterday data (training data) is given by

$$\mathrm{MSE}_{\mathrm{Tr}} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2$$

- We would like to have a good performance

$$\mathrm{MSE}_{\mathrm{Te}} = \frac{1}{m}\sum_{i=1}^{m}[y_i^* - \hat{f}(x_i^*)]^2$$

on tomorrow data (test data)

$$(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_m^*, y_m^*)$$

but $y_1^*, \ldots, y_m^*$ are not available

Goal

- Training data: $n = 30$
- Test data: $m = n$ with $x_i^* = x_i$ for $i = 1, \ldots, 30$
- You are not allowed to use the test data $y_1^*, \ldots, y_{30}^*$
- Restrict attention to polynomial regression models

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon$$

  for $d \in \{0, 1, \ldots, n-1\}$
- Find the degree d that minimize $\mathbb{E}(\mathrm{MSE}_{\mathrm{Te}})$

```
# import data
load("poly.Rdata")

# plot training data
plot( y ~ x , train)

# 2nd-degree polynomial regression fit
fit <- lm( y ~ poly(x, degree=2), train)
yhat <- predict(fit, newdata=test)
lines( yhat ~ x, train)

# MSE.tr
MSE.tr <- mean( (train$y - yhat)^2 )
```

2nd-degree polynomial regression fit