Day 3, statistical modeling

ESTP Use of R in Official Statistics

Contents

- Why statistical modeling?
- Inference
- Predict

Official Statistics

Statistics can be:

- *descriptive*: describing/summarizing the "state")
- *predictive*: estimate an unknown/incomplete parameter/indicator
- *forcasting*: predicting the future value of indicators.
- *inferential*: find out the mechanism, contributing parts.

Prediction aka Machine Learning

Official Statistics

- By origin Official Statistics is **descriptive**: it describe the "state" (of the state :-))
- But **observations are never complete**. A survey is a method to do a small (incomplete) measurement to calculate statistics on (groups in) the population: *estimation is a prediction method*.
- But **observations can have errors and missing values**. Fixing values is a form of *prediction*.
- But analysing time series, finding trend and seasonal patterns is very similar to *forcasting*: often it is *now-casting*.

Prediction

- survey estimation predicts for groups: totals (and or) means.
- most prediction works on observations.
- can be used together: *small area estimation* combines survey and prediction methods

Statistical models

Statistical models are:

"smart", sensible guessing machines, calibrated with known data:

If we measure the happyness of 0.1% of the population (smartly selected), this will (probably) hold for the whole population.

If we do not know the NACE code for this enterprise, but it has similar properties as whole-trade enterprises, it (probably) is a whole-trade enterprise.

Descriptive stats

- Can be done easily with `dplyr` and R:
- Counting / tabulating: `table`, `dplyr::count`, `dplyr::summarize`, `dplyr::n()`
- variables: `mean`, `median`, `sum`
- per group: `dplyr::group_by` (or `tapply`)

Prediction

- For surveys the `survey` package is of interest.
- For now we focus on predicting values for individual records:

Useful in official stats for

- classification (e.g. enterprises, economic activity)
- imputation for missing or erroneous values.
- estimating statistics for small groups (small area estimation)
- clustering (technically not prediction, but can reveal structure in data that can be used to predict which cluster an observation belongs to)

Statistical prediction

Model

$$Y = f(X) + \varepsilon$$

- Y: Predicted variable
- f: Relation between Y and X
- $X = (X_1, X_2, \ldots, X_p)$: predictive variables (predictors), aka independent variables
- ε: Noise, independent of X en Y.

Prediction

Estimated model

$$\hat{Y} = \hat{f}(X)$$

- $\hat{Y}$ the estimated value for Y.
- $\hat{f}$ the estimated model.

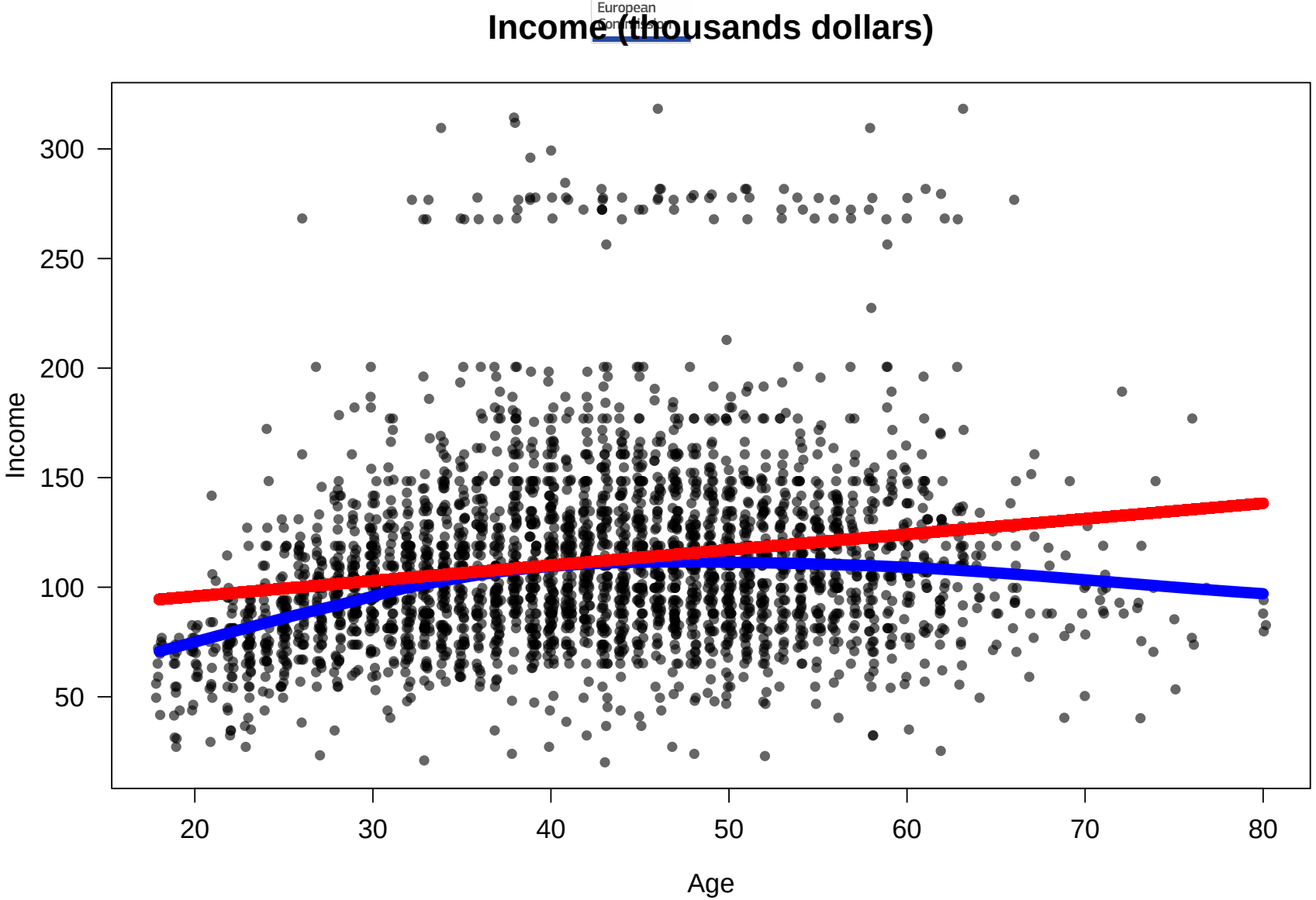
Remark

For individual observations

$$y_i = \hat{f}(x_i),$$

the relation f works as a *black box* / data generating machine: input x generates value y.

Example



X_1: age X_2: income f_1: lin. reg. f_2: Loess

Prediction error (classical)

Estimated prediction error

$$\begin{aligned} \mathrm{E}[(Y - \hat{Y})^2] &= \mathrm{E}[(f(X) + \varepsilon - \hat{f}(X))^2] \\ &= \underbrace{\mathrm{E}[(f(X) - \hat{f}(X))^2]}_{\text{Reducable}} + \underbrace{\mathrm{Var}(\varepsilon)}_{\text{Fundamental}} \end{aligned}$$

Estimated prediction error

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- y_i: Observed values
- $\hat{y}_i$: Predicted values $\hat{y}_i = \hat{f}(x_i)$
- MSE Mean Square Error for estimated model $\hat{f}$.

Prediction

Numerical Y

- e.g. `turnover`
- prediction is called **regression**

Categorical Y

- e.g. `nace`
- prediction is called **classification**

Prediction methods

- linear models: `lm`
- generalized linear models: `glm`
- decision trees: `rpart` (from R package `rpart`)
- randomForests: `randomForest` (from R package `randomForest`)
- etc.

Model in R

Formula :

```
wage ~  age + education # wage depends on age and educatio
```

```
model <- lm(wage ~ age, data = Wage) # create a linear mod
```

Linear model in R


```
model <- lm(wage ~ age, data = Wage) # create a linear mod
model
```

```
Call:
lm(formula = wage ~ age, data = Wage)

Coefficients:
(Intercept)          age
    81.7047       0.7073
```

Linear model in R


```
summary(model)
```

```
Call:
lm(formula = wage ~ age, data = Wage)

Residuals:
     Min       1Q   Median       3Q      Max
-100.265  -25.115   -6.063   16.601  205.748

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.70474    2.84624   28.71   <2e-16 ***
age          0.70728    0.06475   10.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.93 on 2998 degrees of freedom
Multiple R-squared:  0.03827,   Adjusted R-squared:  0.03795
F-statistic: 119.3 on 1 and 2998 DF,  p-value: < 2.2e-16
```