MATH-6010: Your Project

Name
your Unid

August 20, 2022

I Data Retrieval

Our data sets (**Medical Cost Personal Datasets**) were obtained from the `Kaggle` website [1].

The data set has six independent variables:

1. `age`
2. `sex`
3. `bmi`
4. `children`
5. `smoker`
6. `region`

and one dependent variable: `charge`.

In Fig. 1 the charges ($) as function of bmi are displayed.

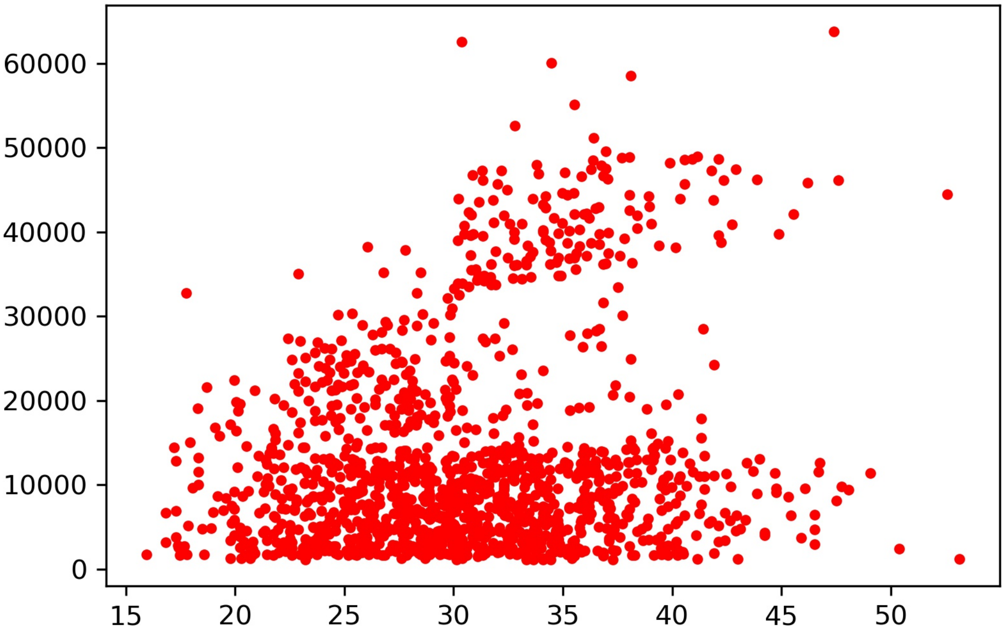

Figure 1: Charges ($) as function of bmi.

In Fig. 2 the age histogram is displayed.

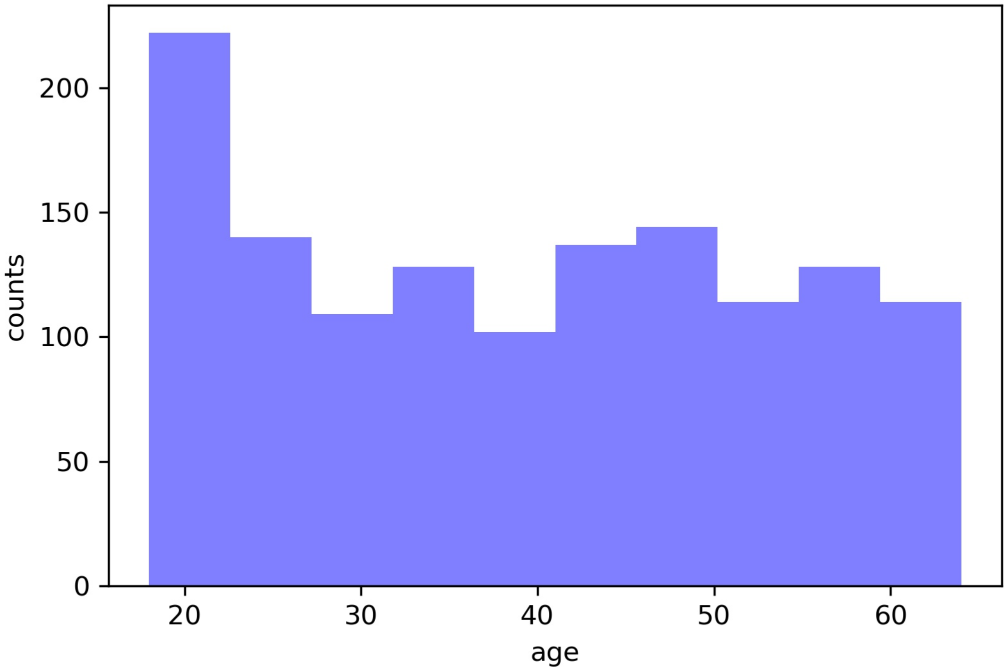

Figure 2: Age histogram.

II Statistical Analysis

II.1 Model

In what follows we will use the following linear model:

$$\begin{aligned} Y_i &= \beta_0 + x_{i,1}\,\beta_1 + x_{i,2}\,\beta_2 + x_{i,3}\,\beta_3 + x_{i,4}\,\beta_4 + \epsilon_i \\ &= \sum_{k=0}^{4} x_{i,k}\,\beta_k + \epsilon_i \qquad (1) \end{aligned}$$

where $x_{i,0} := 1$.

Eq. (1) can also be rewritten in matrix form[1]:

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \qquad (2)$$

The test of the null hypothesis can be achieved by calculating the value for the following F-statistic [2]:

$$f_{1,n-p} = \frac{(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})^{\boldsymbol{T}} \left[\boldsymbol{A}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{A}^{\boldsymbol{T}}\right]^{-\boldsymbol{1}} (\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})}{S^2} \qquad (3)$$

where the expression $(\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{c})$ imposes a constraint on $\widehat{\boldsymbol{\beta}}$.

[1] In what follows we will display vectors in bold.

References

[1] *Kaggle: Medical Cost Personal Datasets*, https://www.kaggle.com/datasets/mirichoi0218/insurance, Accessed: 2022-08-20.

[2] G.A.F. Seber and A.J. Lee, *Linear Regression Analysis*, Wiley Series in Probability and Statistics, ch. 4. Hypothesis Testing, Wiley, 2012.