Interpreting the coefficients

Linear regression model

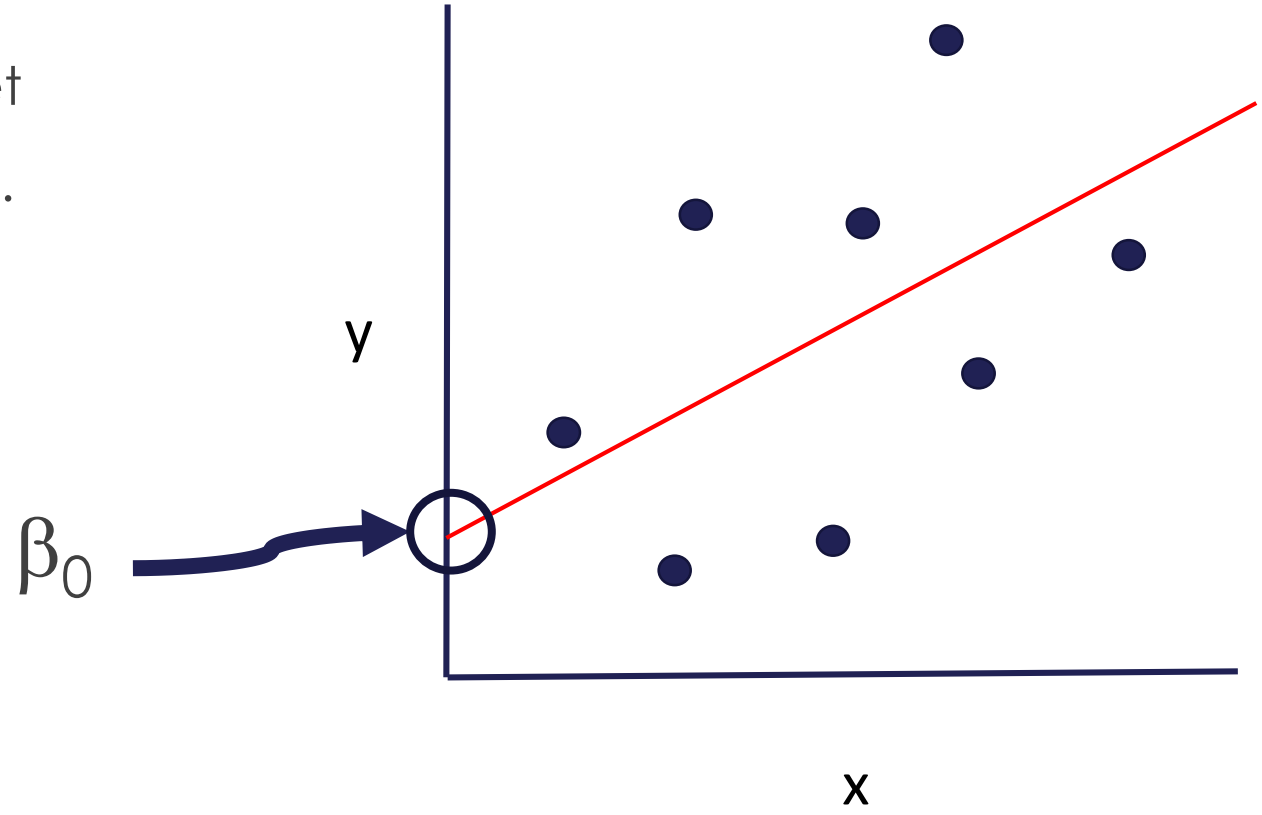
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i$$

- β are the coefficients.
- β_0 is the intercept.

The intercept

β_0 is the value of the target when all variables equal 0.

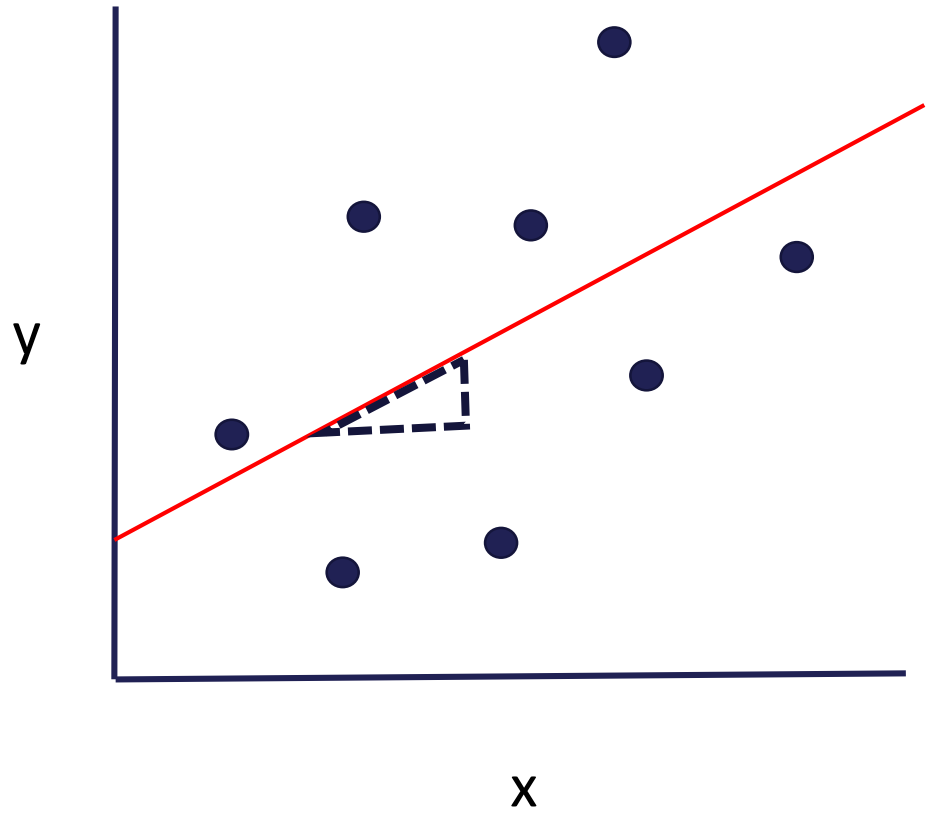
Also called the intercept.



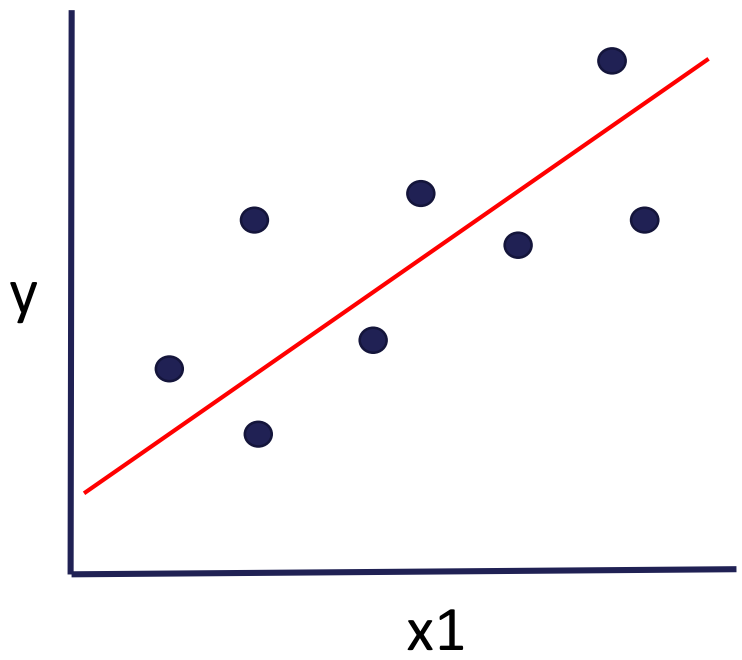
Coefficients

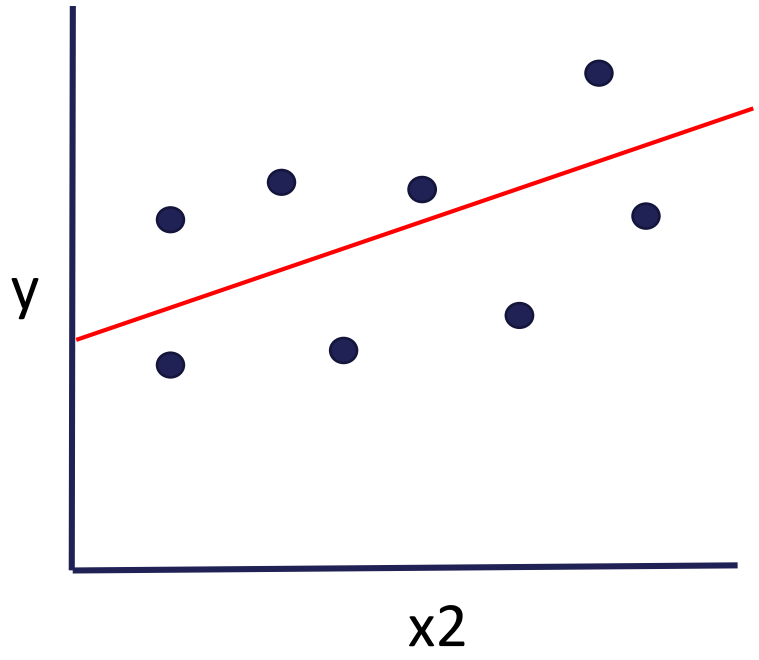
- β represents the gradient (slope) of the regression.
- β is the change in y, per unit change in x.

(provided all other variables stay the same.)



Coefficients slope

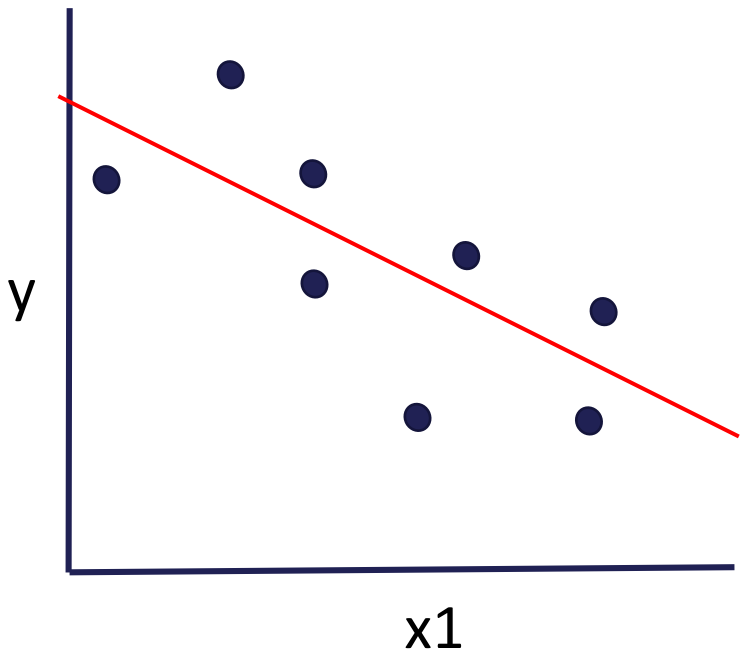


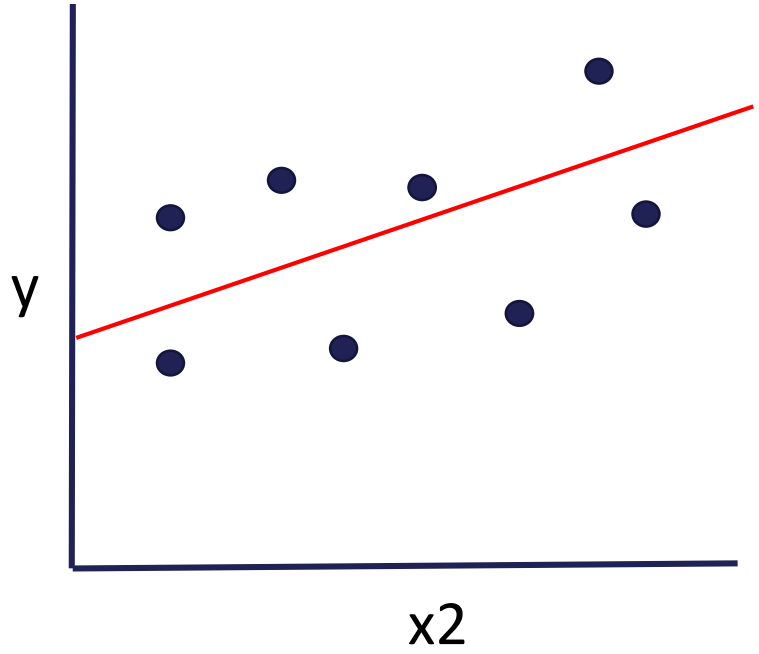


$\beta 1 > \beta 2$ ➔ x1 has a greater contribution than x2 to the target value.

(For a meaningful comparison features should be in a similar scale).

Coefficients sign



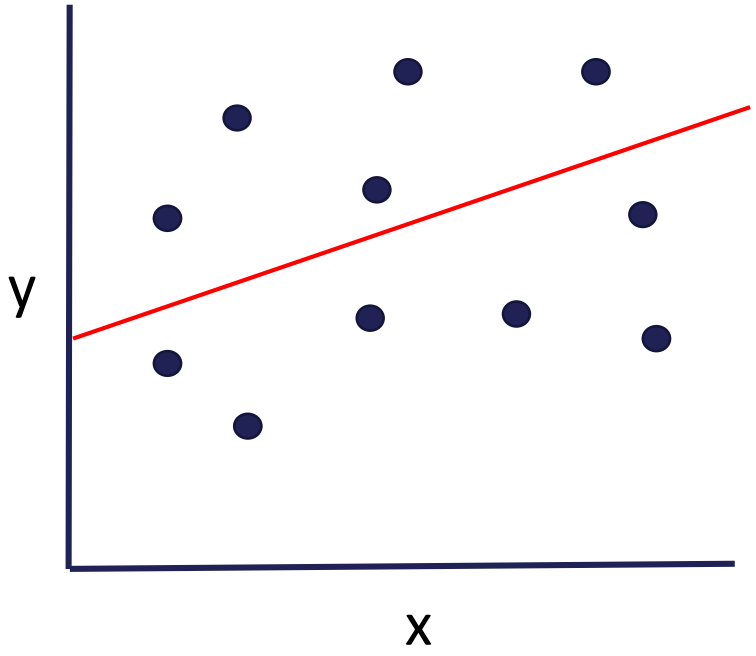


+ ➔ as x increases, so does y.

- ➔ as x increases, y decreases.

Coefficients significance

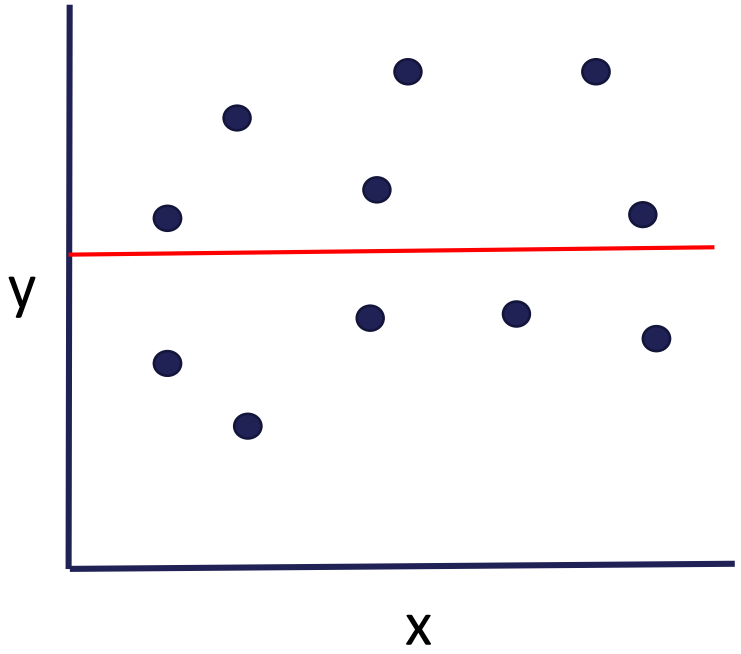
Does the red line show a good linear fit?



t-test

Bad model: we'd expect the change in y to be zero per unit change of x.

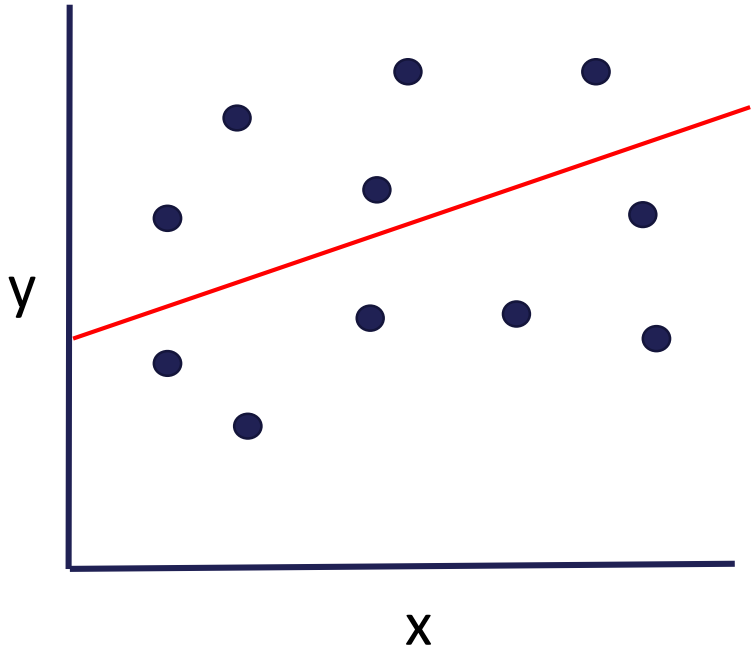
➔ $\beta = 0$



t-test

If we can predict y from x linearly

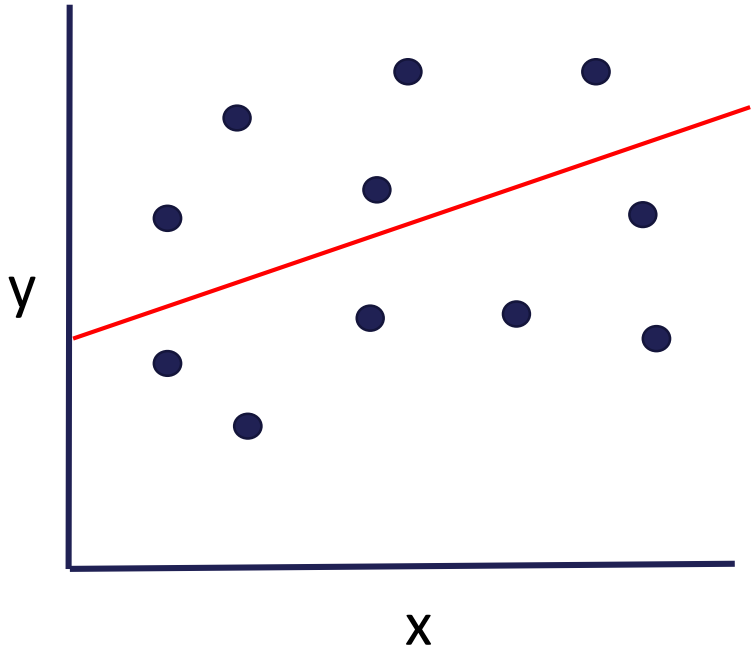
➔ then β is different from 0.



t-test

- t-test tests the null hypothesis: **β=0**.
- Tests how big β is, compared to its variability.

$$t = \frac{\beta}{SE\beta}$$



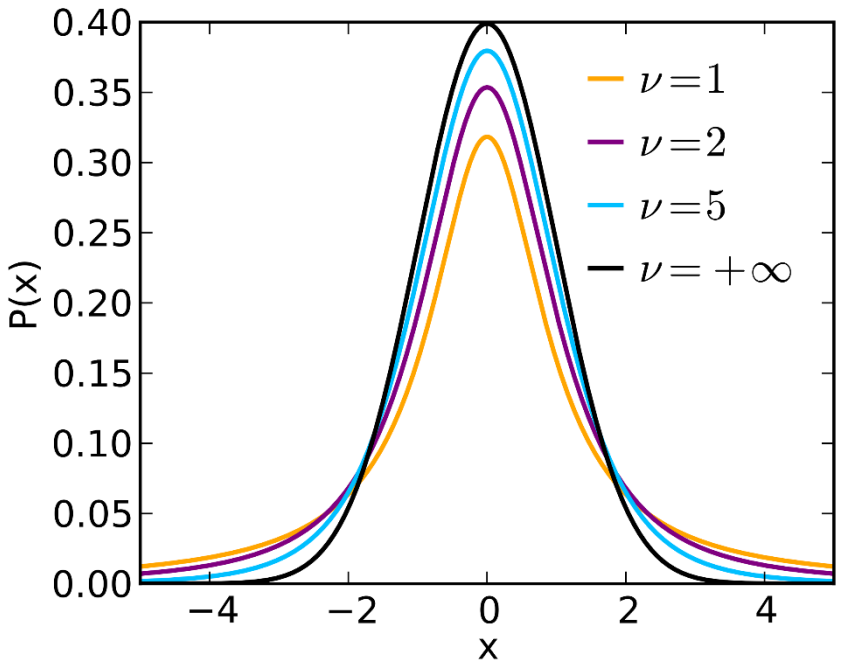
t-distribution

$$t = \frac{\beta}{SE\beta}$$

If *t* is too big or too small ➔ the probability that β=0 is small, then, the regression coefficient is statistically significant.

The t-distribution

https://commons.wikimedia.org/wiki/File:Student_t_pdf.svg



Coefficients variability

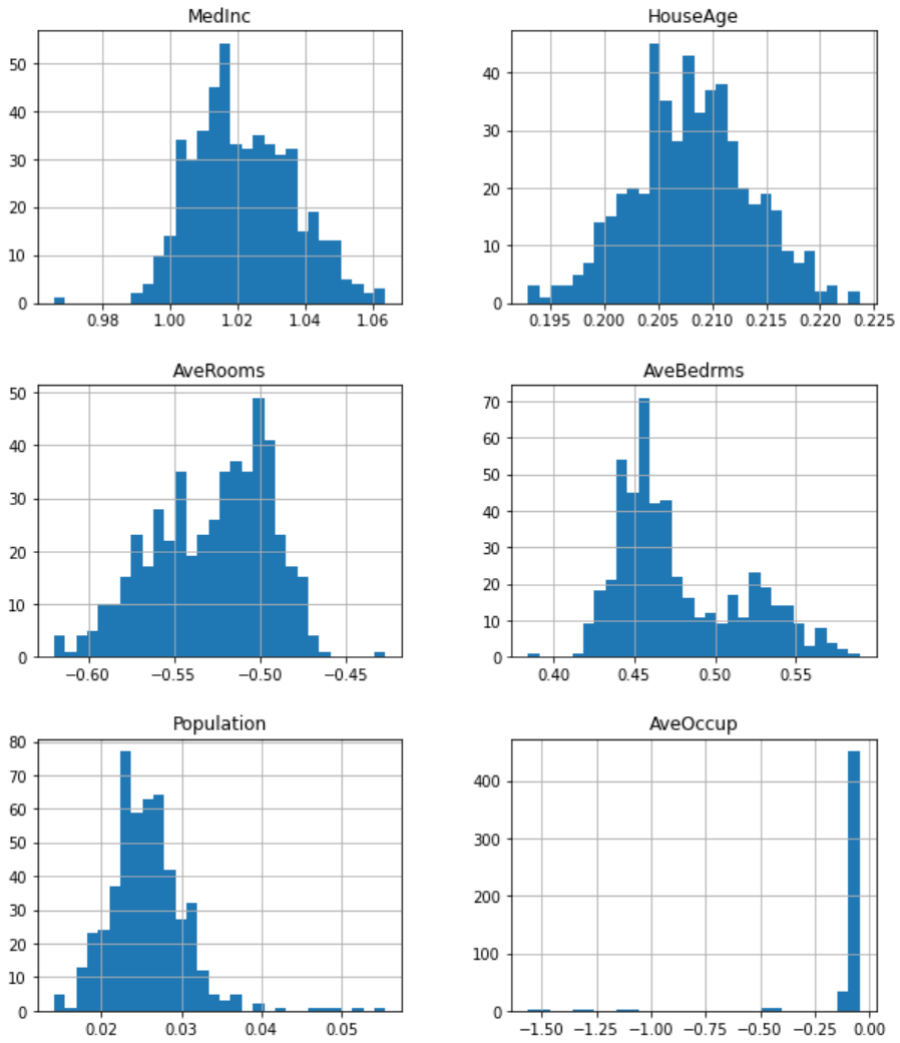
What is it?

Coefficients variability

- β is an estimate.
- As such, it comes with an estimation error.

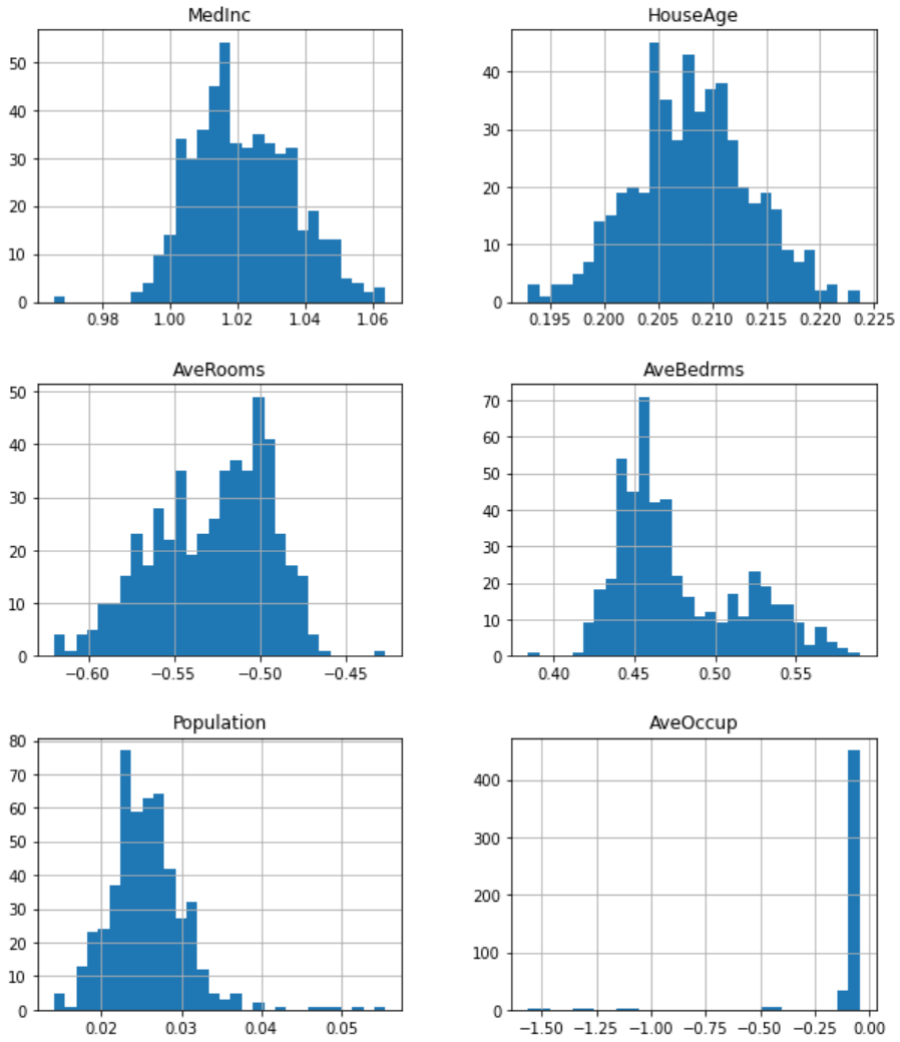
Coefficients variability

If we take 500 different samples of the data, we'll obtain 500 different coefficients.



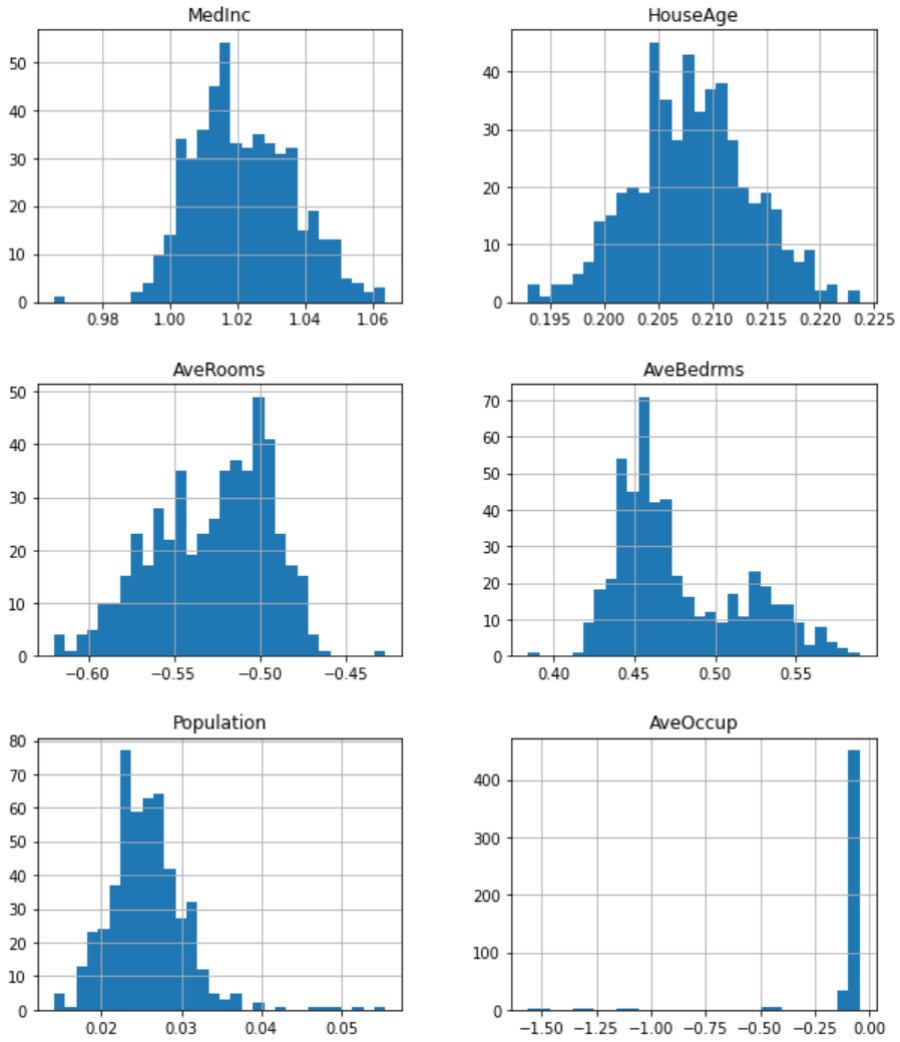
Coefficients variability

- Calculate the mean and standard deviation of β.
- Calculate *t*.
- Find the p-values.



Coefficients variability

- Calculate the mean and standard deviation of β.
- Calculate *t*.
- Find the p-values.
- **Cross-validation**



Statsmodels

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$se(\hat{\beta}_i) = \sqrt{(X^T X)^{-1}_{ii}}\hat{\sigma}.$$

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - p}$$

https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Summary

Intercept

- Value of y when all variables' values equal 0.
- Realistic scenario: variables are centred at mean =0.

Coefficients

- β indicates the change in y given a unit change in x.
- If β is positive, then y increases as x increases.
- If β is negative, then y decreases as x increases.
- To compare features by using β, they need to be in a similar scale.

Significance

- We can test if the β are significantly different from 0 by using a t-test.
- t tells us if β is statistically bigger than its estimation error (i.e., variability).

Coefficient significance in Python

- Sklearn
  - Cross-validation
  - Manual calculations
- Statsmodels
  - Coefficients
  - Coefficients error
  - t and the p-values
  - Confidence intervals

Train In Data

THANK YOU

www.trainindata.com