COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

ICAI – GITI/GITT

Linear regression
Estadística II

Eugenio Sánchez Úbeda

January 2024

comillas.edu

Bibliography

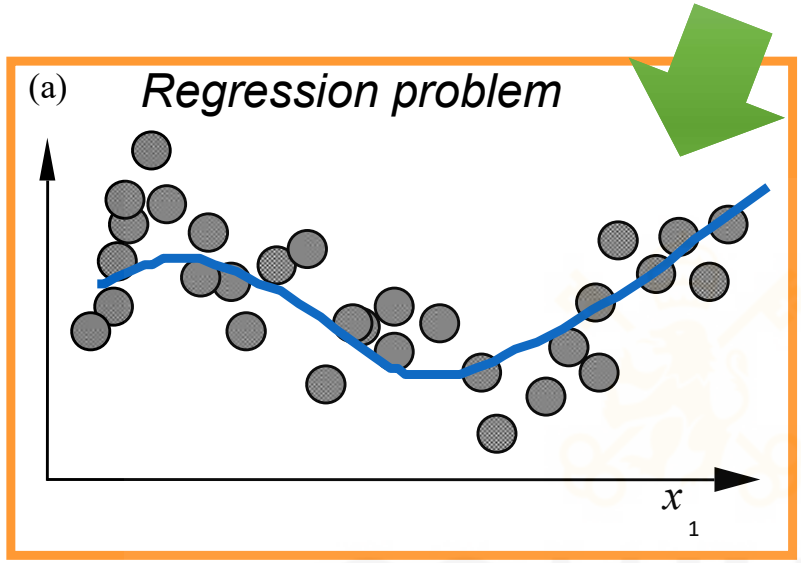
- G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, *An Introduction to Statistical Learning with Applications in Python*, Springer, 2023 (https://www.statlearning.com)
- G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013, (8th printing 2017). Chapter 3
- M.H. Beale, M.T. Hagan, and H.B. Demuth *Deep Learning Toolbox. Getting Started Guide* 2018 https://es.mathworks.com/help/pdf_doc/deeplearning/nnet_gs.pdf
- Tensorflow — Neural Network Playground

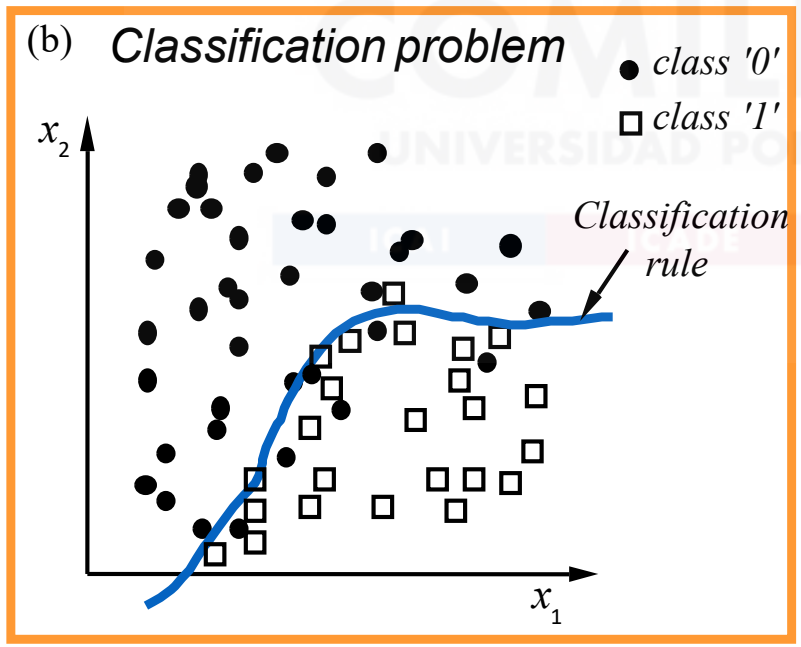
1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

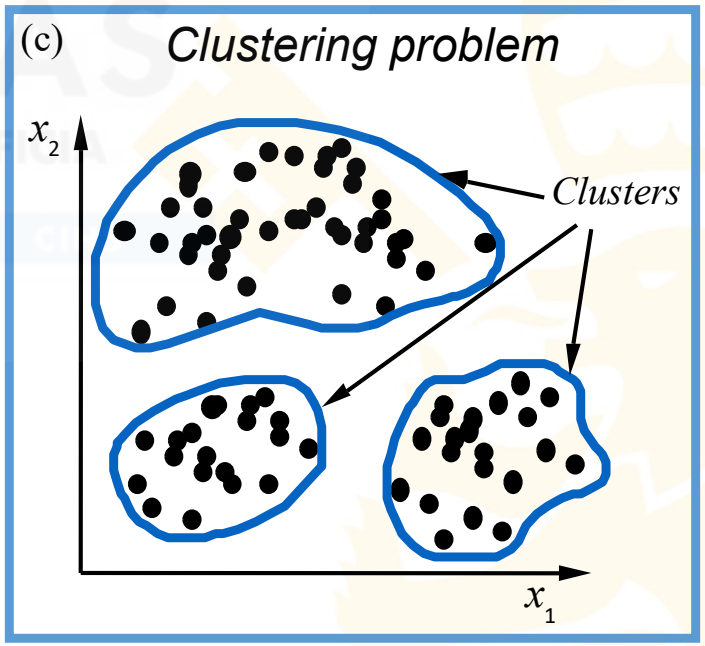
1

Introduction

Fundamental types of problems
Regression problem







Supervised learning

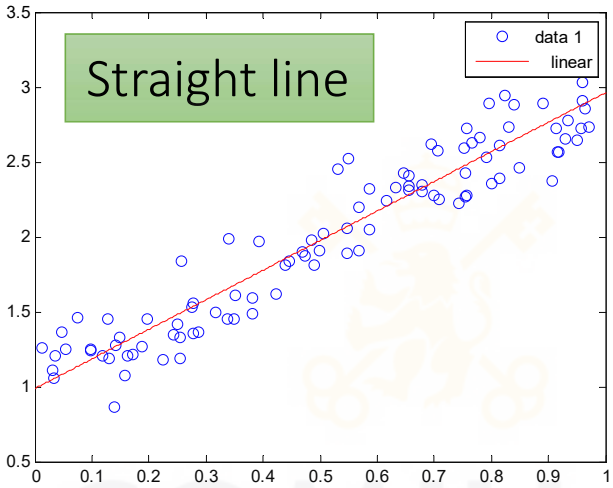
Unsupervised learning

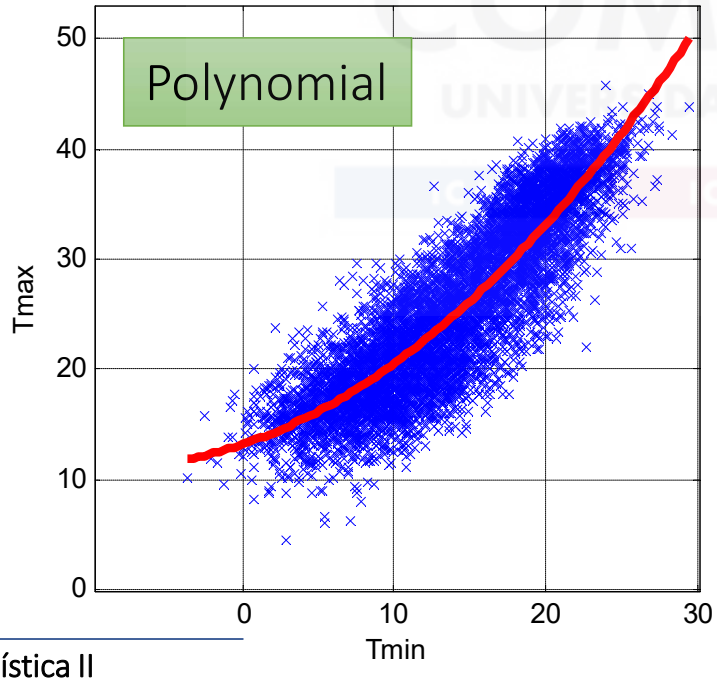
Regression problem
Types of models

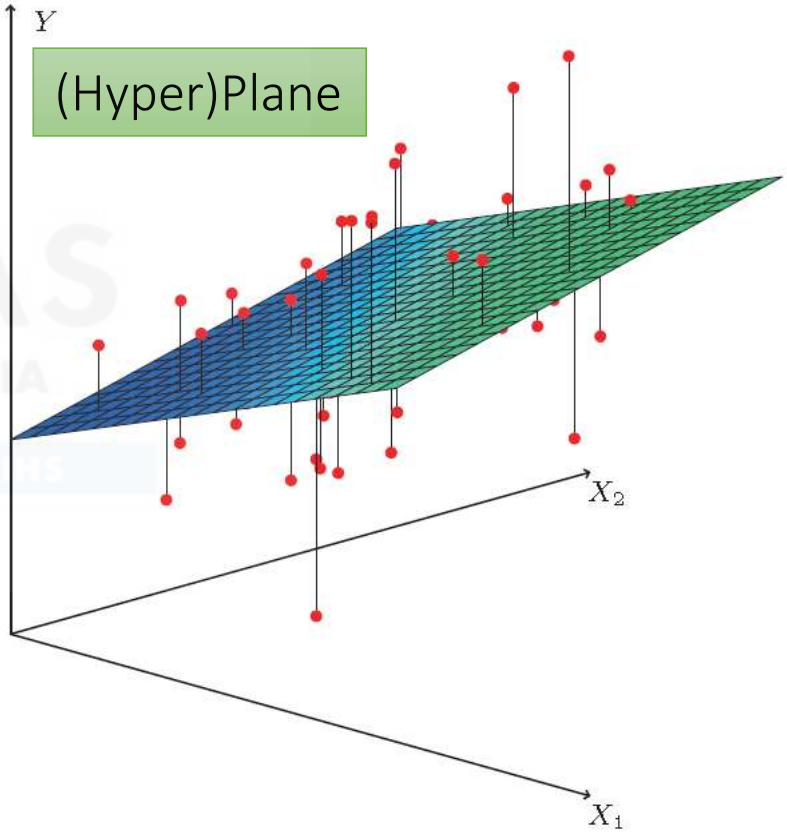
- There exist many types of regression models
- Can be classified according to different characteristics, for example:
  - Number of inputs and outputs
  - Complexity of the parameter fitting
  - Interpretability
- According to the number of input variables
  - Simple
  - Multiple
- According to the difficulty in parameter fitting
  - Simple
  - Complex

Regression problem
Types of models: Examples

- Simple



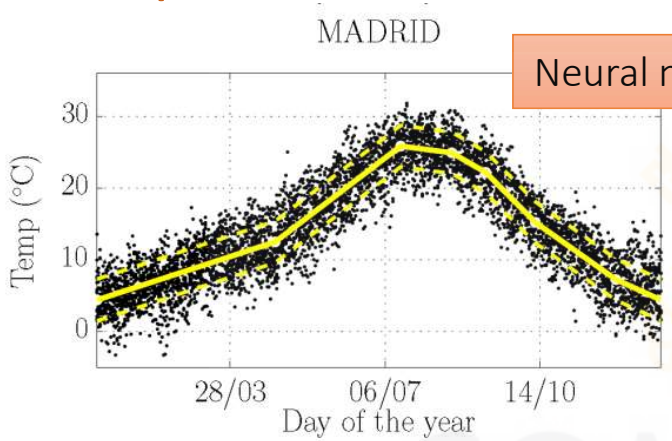




Regression problem
Types of models: Examples

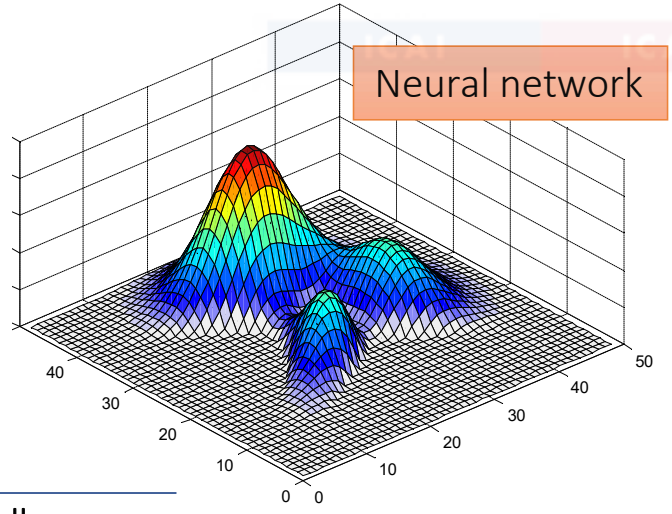
- Complex

Neural network

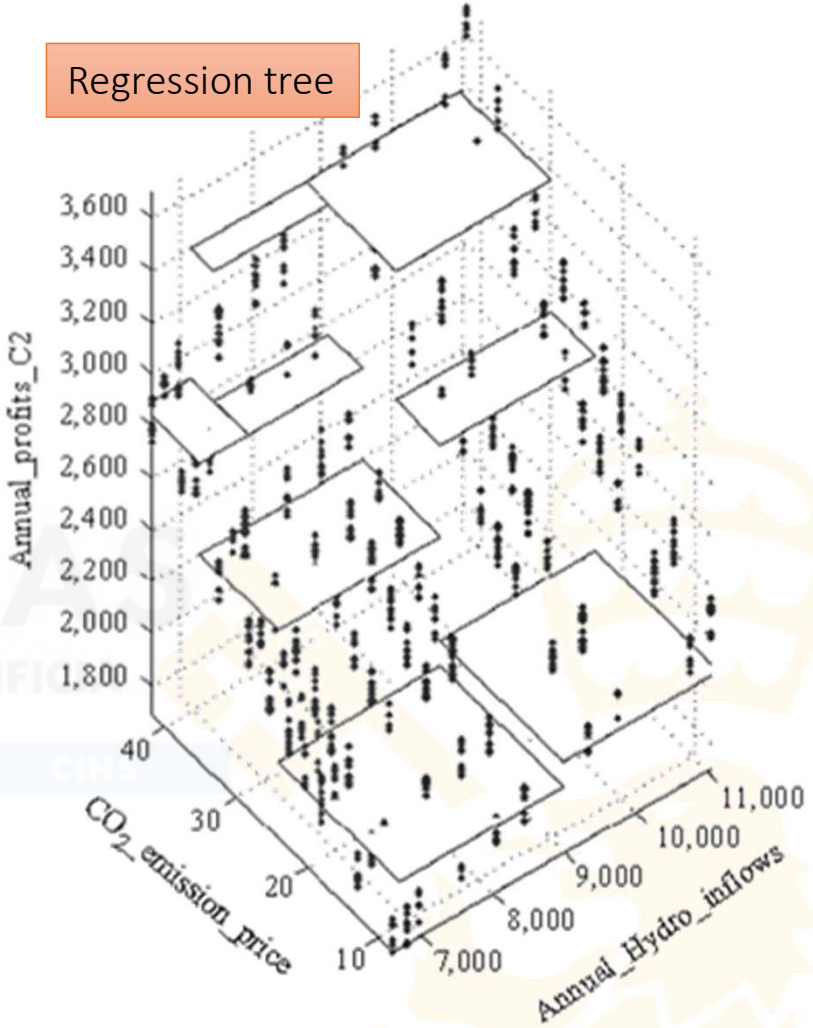


E.F. Sánchez-Úbeda, A. Berzosa, *New variables to improve electricity and natural gas consumption forecasting: dynamic degree-days*, 14th Conferencia de la Asociación Española para la Inteligencia Artificial - CAEPIA'11, Nov. 2011

Neural network



Regression tree



N. Mosquera, J. Reneses, E.F. Sánchez-Úbeda, "*Medium-term risk analysis in electricity markets: A decision-tree approach*", Int. Journal of Energy Sector Management. vol. 2, no. 3, pp. 318-339, Sept. 2008.

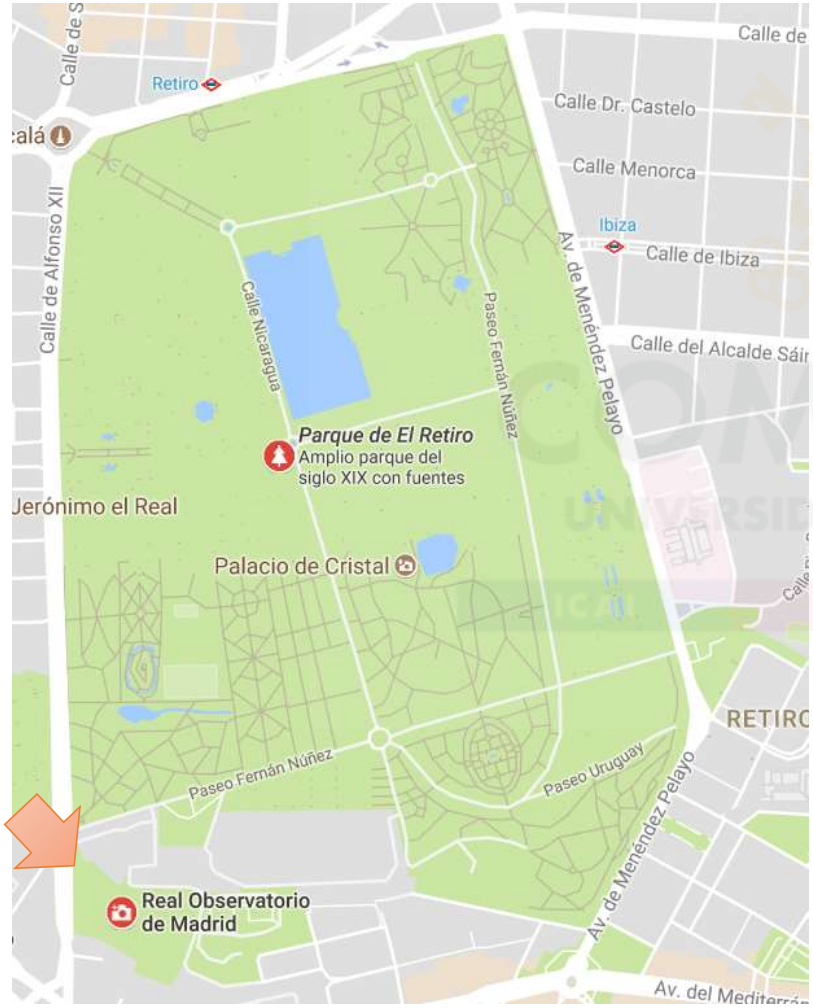
1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

2

Constant Model

Constant Model
Problem

- What is the maximum daily temperature in the Retiro Park of Madrid in December? How much change?



Fit a statistical model for estimating the maximum daily temperature in December

De Fanattiq - Trabajo propio, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=12745390

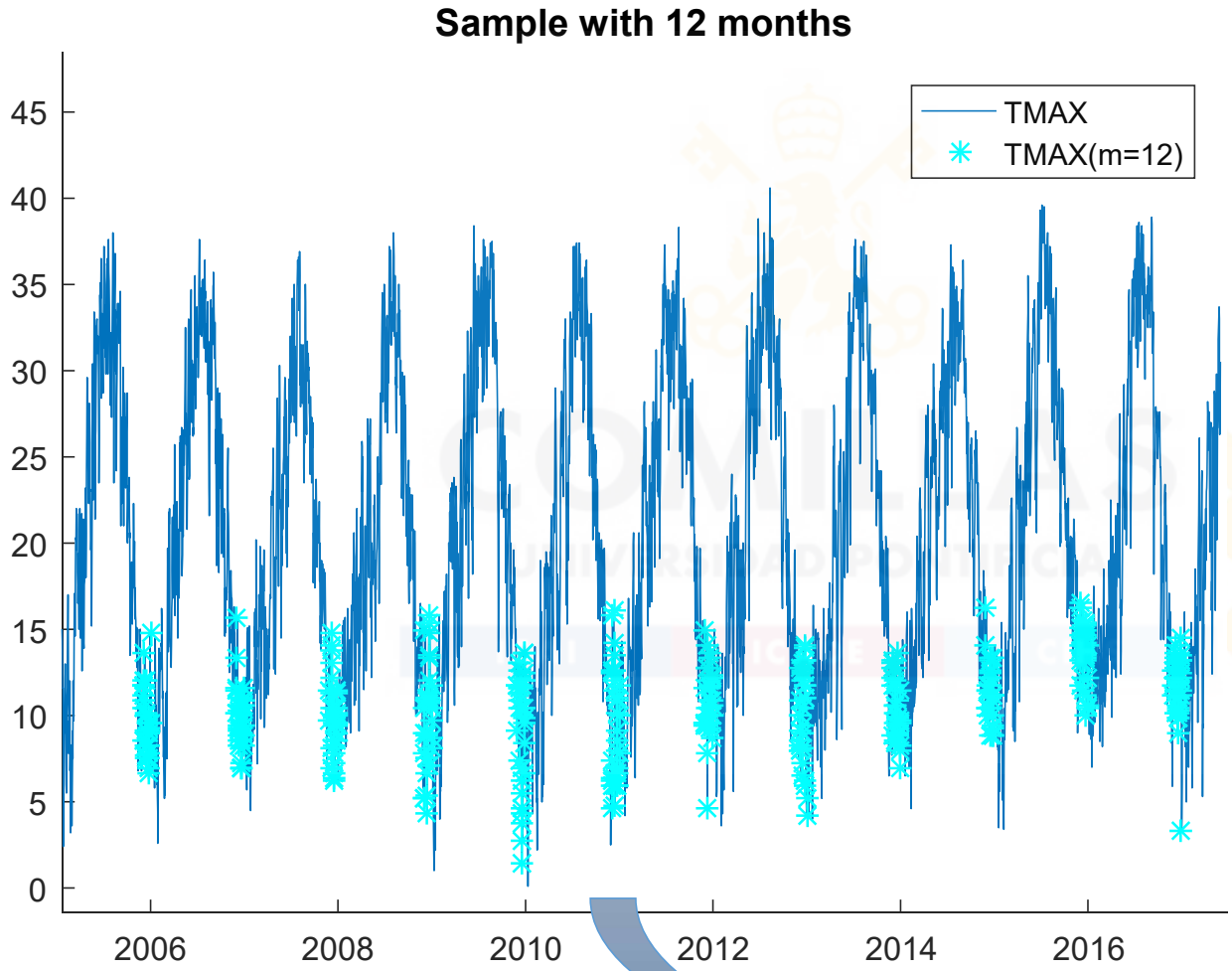
Available data: minimum and maximum daily temperatures since 1920

Constant Model Problem

Sample size n

$y_1, \cdots, y_n$

- We take a sample of several Decembers (12 years)



TMAXdec

Daily data 12 Decembers

8.6
11.1
11.8
11.7
...
10.4
10.8
11.7
13.6
10.1

Constant Model
Statement

- The maximum daily temperature in the Retiro Park in December can be considered as a random variable Y given by the theoretical model

Deterministic component

Stochastic component

$$Y = \beta_0 + \epsilon$$

CONSTANT POPULATION MEAN
(unknown theoretical value)

POPULATION VARIANCE
(unknown theoretical value)

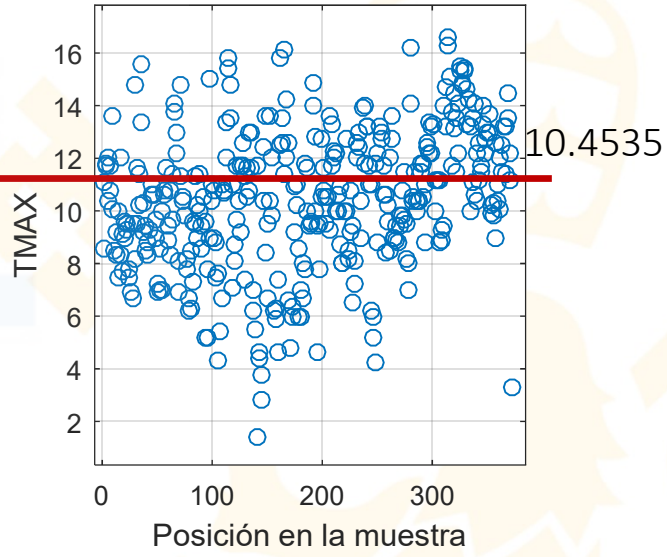
$$\epsilon \approx N(0, \sigma^2)$$

A random variable that considers the temperature variations with respect to the expected value (error, noise)

Constant Model
Coefficient estimation

- From the available sample, we can estimate the parameters of the deterministic and stochastic components
- Deterministic component represents the expected value of the random variable Y
  - It can be estimated by computing the sample mean

$$\hat{\beta}_0 = \bar{y} \longrightarrow \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$$



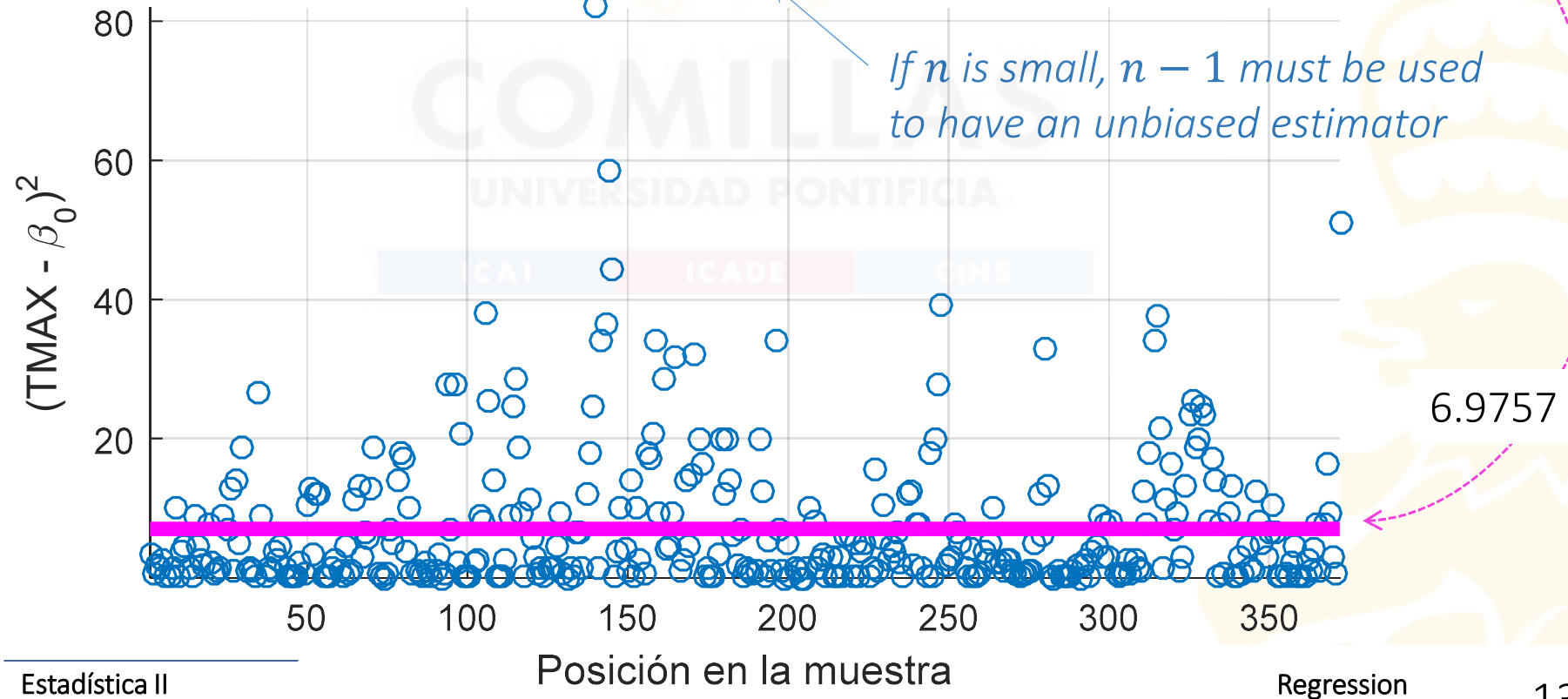
- Why is the simple mean a reasonable estimator?
- Are there other reasonable estimators?

Constant Model
Coefficient estimation

- Stochastic component represents the variability with respect to the expected value of the random variable Y
  - It can be estimated by computing the sample variance

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

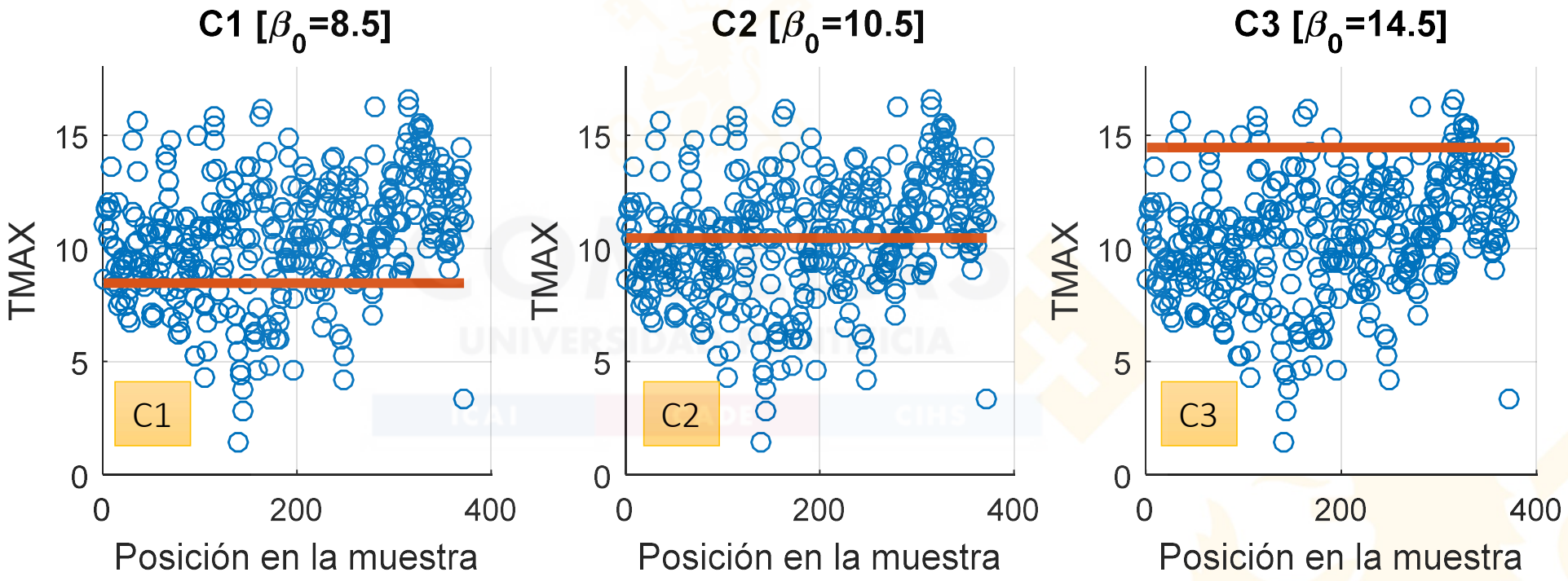
If n is small, $n-1$ must be used to have an unbiased estimator



Constant Model
Constant estimation

- In the temperature example

Which constant (C1, C2, or C3) is better?

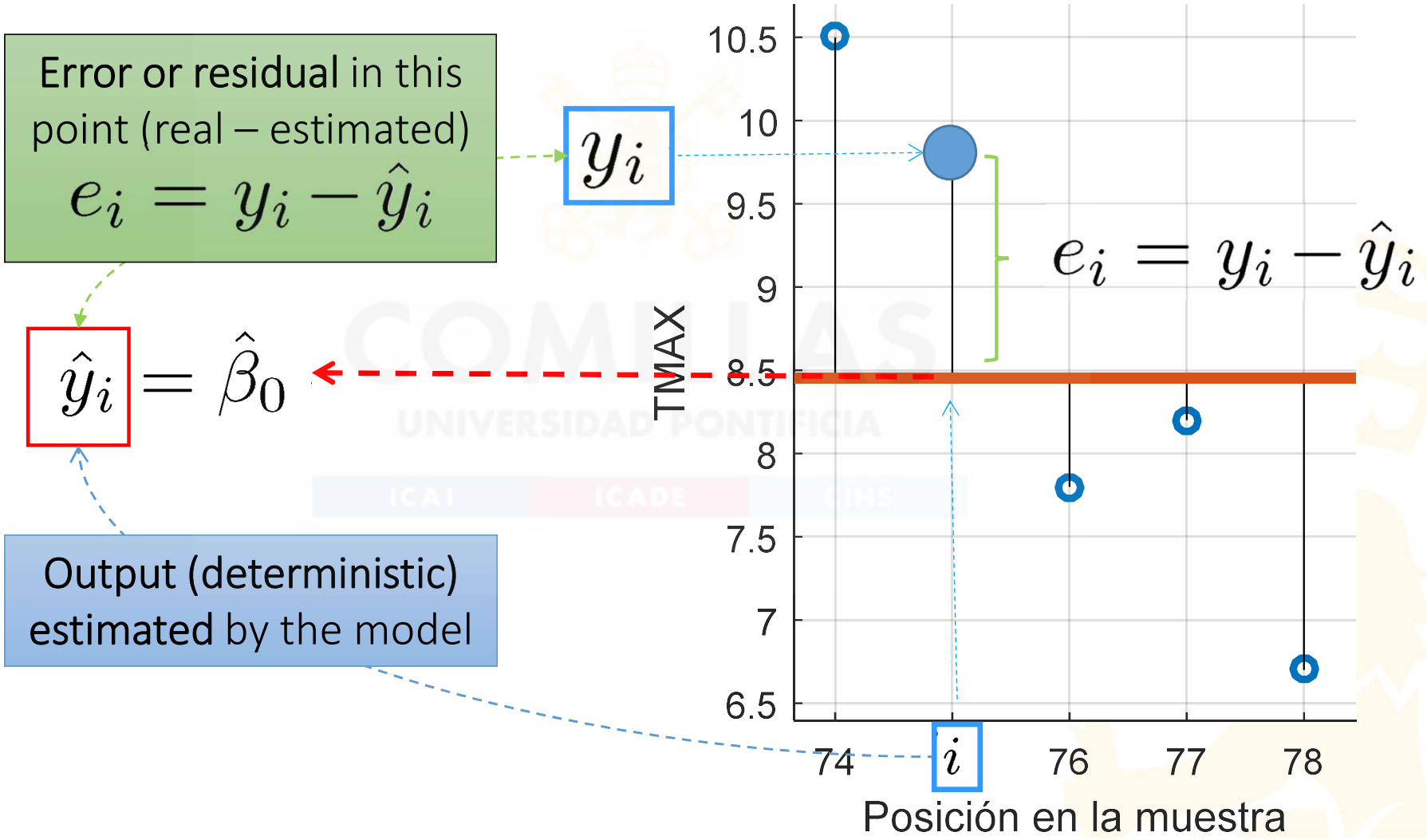


Why?

$$y_i \approx \hat{\beta}_0$$

Constant Model
Constant estimation

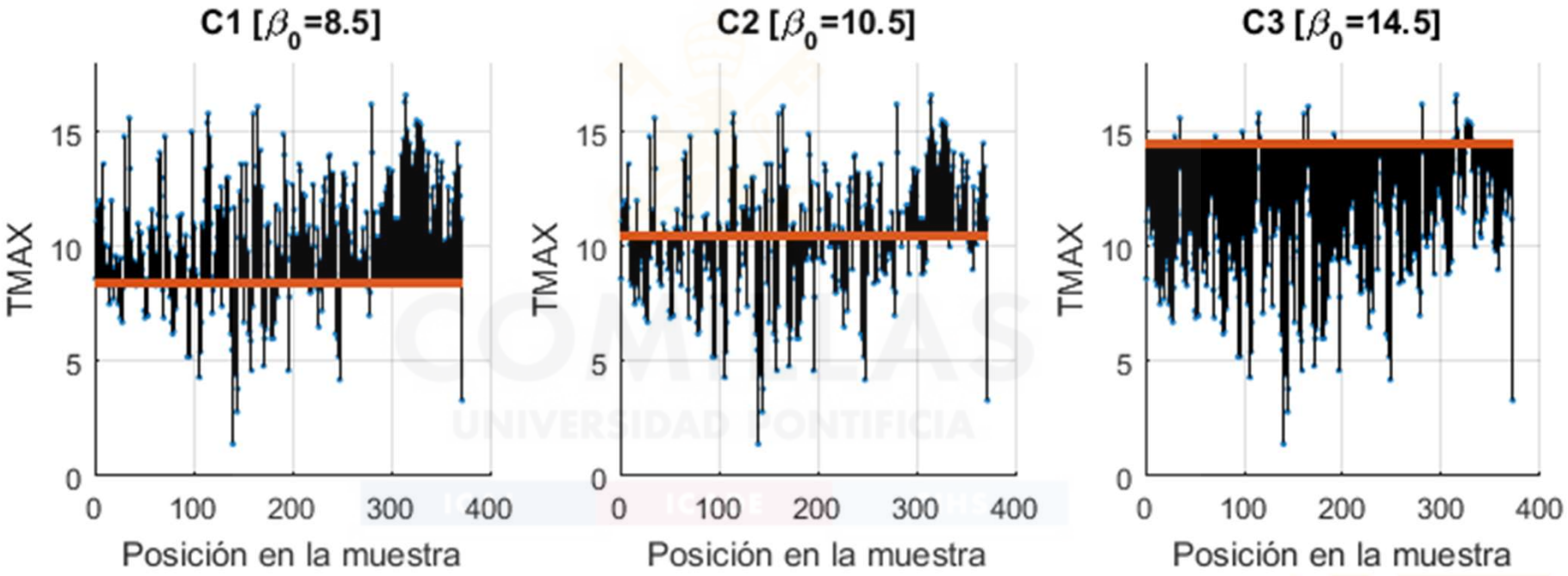
- We want to find the coefficient, such as the resulting constant is the closest to the data



Constant Model
Constant estimation

- Residuals can be positive or negative
- The smaller, the better

$$e_i = y_i - \hat{y}_i$$

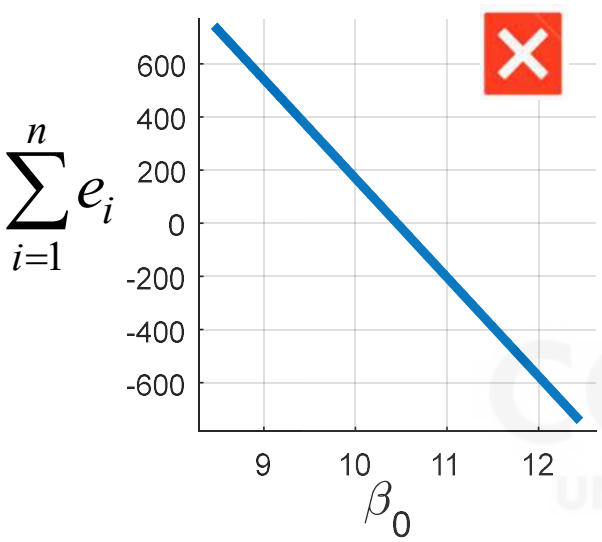


Which constant (C1, C2, or C3) produces smaller residuals?

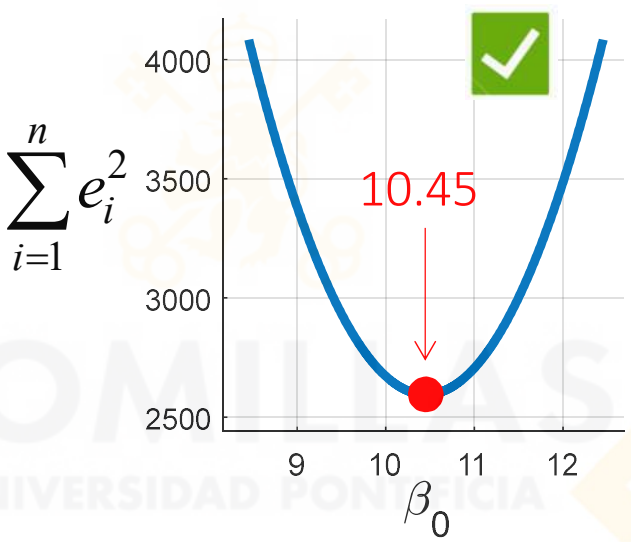
Consider these global error measures: $\sum_{i=1}^{n} e_i$ $\sum_{i=1}^{n} e_i^2$ $\sum_{i=1}^{n} |e_i|$

Constant Model
Constant estimation

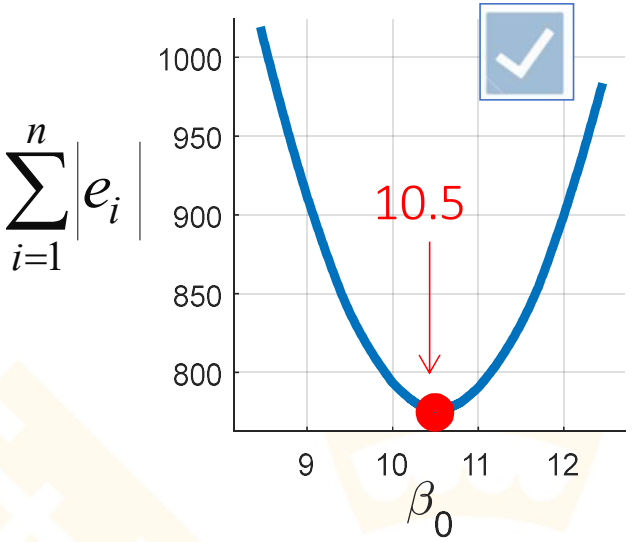
- Variation of the global error measures with the constant (temperature problem)



Not useful



Minimizing the sum of squared residuals we obtain the MEAN



Minimizing the sum of absolute residuals we obtain the MEDIAN

Constant Model
Point estimation

- We use exclusively the deterministic component

$$Y = \beta_0 + \epsilon$$

$$\hat{y} = \hat{\beta}_0$$

- Temperature problem:

What is the maximum daily temperature in Retiro Park in December?

$$\hat{y} = 10.45\,°\text{C}$$

Maximum daily temperature expected on any day in December in the Retiro Park

Accuracy of the estimation?

Constant Model
Accuracy in the constant estimation

- Sensitivity of the estimated coefficient to the sample
  - If the sample used changes, estimation can change. How much?
- The theoretical knowledge of the sample mean estimator can be used

$y_1, \ldots, y_n$

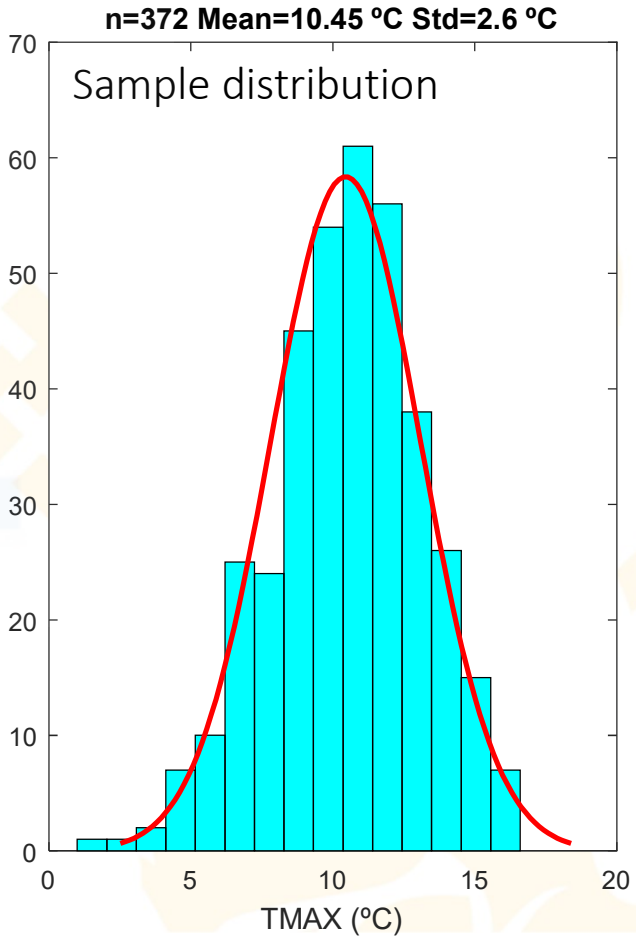
$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$

$\hat{\mu} = \bar{y}$

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \frac{\sigma^2}{n}$$

- Standard error (SE) of the estimator is the standard deviation of the distribution of this estimator and represents the expected variation if the sample changes
- SE decreases with the sample size (greater accuracy)



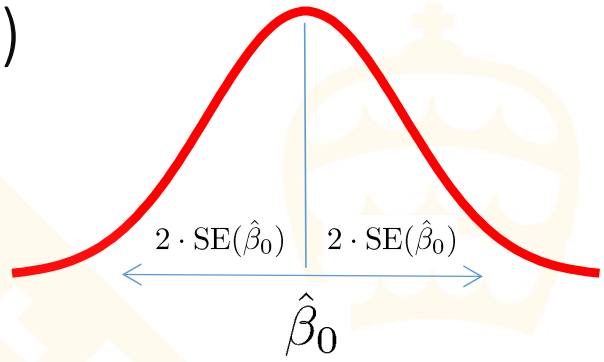
Constant Model
Interval estimation

- We can build a confidence interval (CI) for the constant using the information of both model components

$$Y = \beta_0 + \epsilon \qquad \epsilon \approx N(0, \sigma^2)$$

- Confidence interval of 95% (approximately)

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$



- CI of 95% in the temperature problem
  - Computed correctly estimating the population variance

$$\mathrm{SE} = 0.1371°\mathrm{C}$$

$$\beta_0 \in [10.18, 10.72]$$

If computation is repeated with 100 different samples (of the same size), we can expect the estimated value to be out of the CI 5 % of the cases.

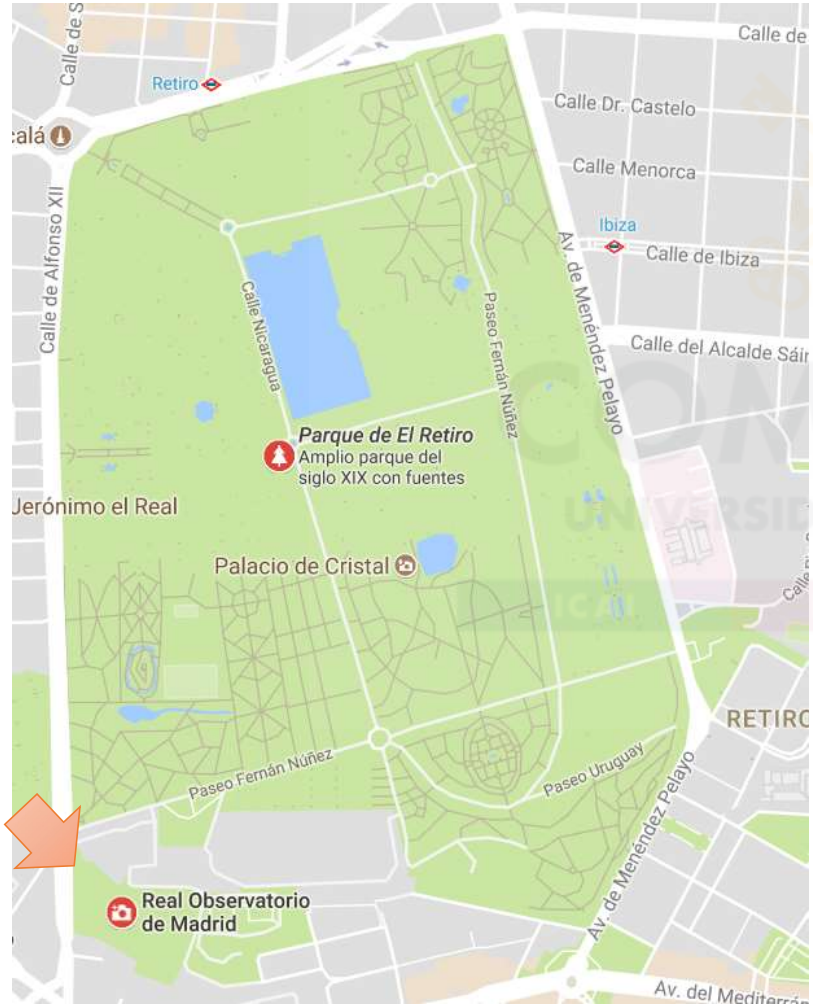
1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

3

Simple Linear Regression

Simple Linear Regression
Problem

- Estimate the maximum daily temperature in the Retiro Park of Madrid from the minimum daily temperature



Fit a statistical model for estimating the maximum daily temperature from the minimum daily temperature.

De Fanattiq - Trabajo propio, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=12745390

Available data: minimum and maximum daily temperatures since 1920

Simple Linear Regression
Problem

- Estimate the maximum daily temperature in the Retiro Park of Madrid from the minimum daily temperature
- We take a sample from a year (2016)

Daily data from 2016

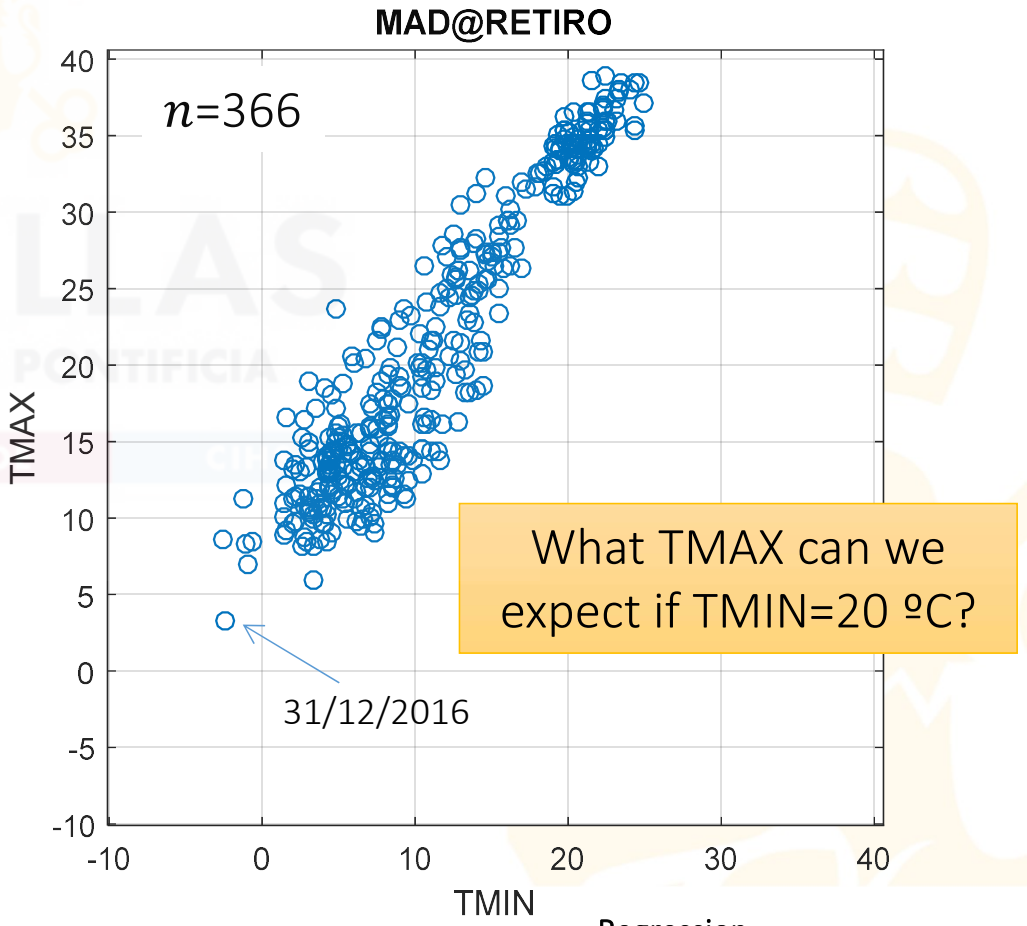
TMIN	TMAX
8.2	14.4
5	11.4
4.7	11.9
6.8	13.5
4.1	9.6
1.4	8.9
8.9	14.4
10.4	12.9
8.2	11.5
8.5	12
5.2	11.3
3.1	10.6
2	9.6
3.8	8.6
...	
4.5	13.2
3	11.4
2	13.2
5	14.5
4.4	13.5
1.6	12.2
-1.2	11.2
-2.4	3.3

(x_1, y_1)

(x_2, y_2)

$\cdots$

(x_n, y_n)



Simple Linear Regression
Statement

- Simple linear regression model allows estimating the continuous r.v. Y from the single input variable X
  - In the current problem, Y is variable TMAX and X is variable TMIN

Deterministic component

Stochastic component

$$Y = \beta_0 + \beta_1 X + \epsilon$$

POPULATION REGRESSION STRAIGHT LINE
(theoretical unknown relation)

$$\epsilon \approx N(0, \sigma^2)$$

A random variable that takes the variations of the output with respect to the expected value given by the deterministic component (error, noise)

Sample

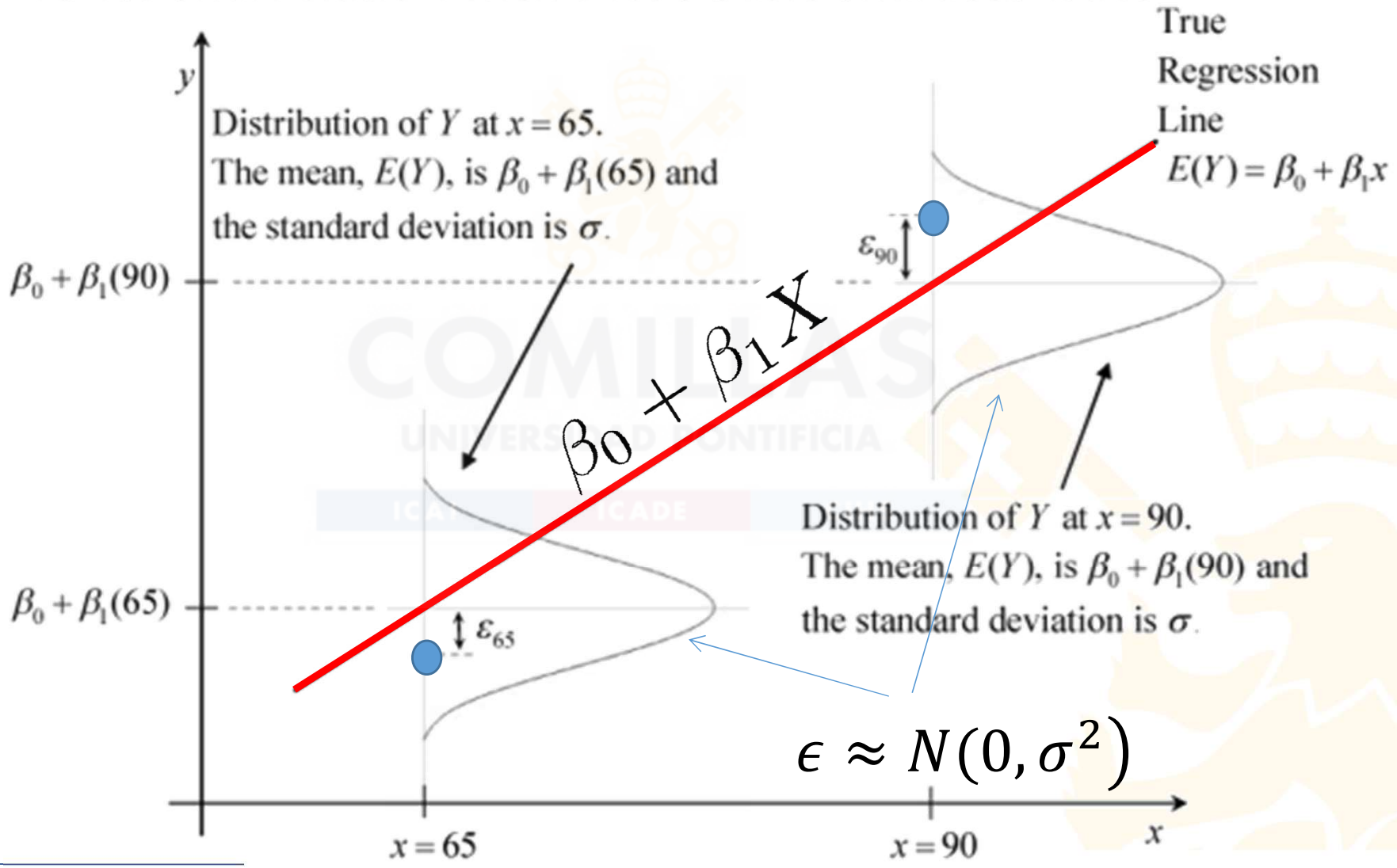
$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

ESTIMATED REGRESSION STRAIGHT LINE

Simple Linear Regression
Statement

- The simple linear regression model assumes a theoretical functional relation with these basic characteristics

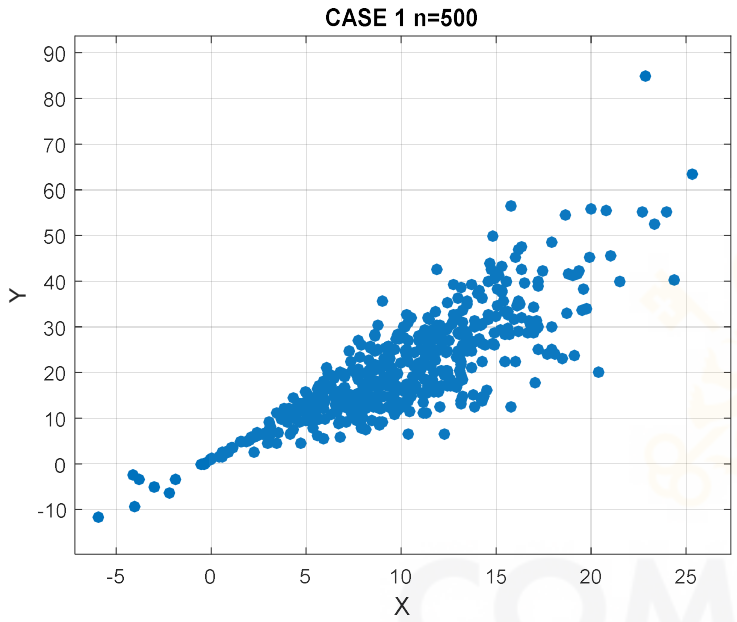


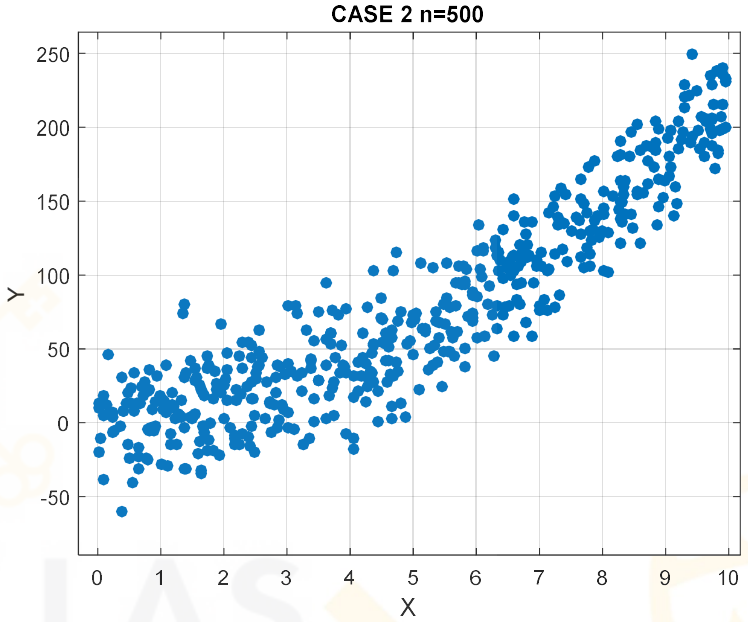
Source: ReliaWiki

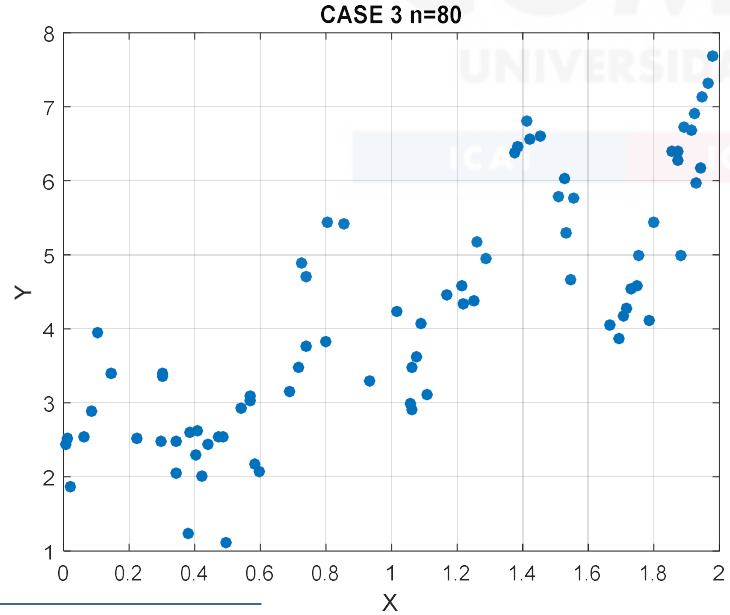
Simple Linear Regression
Statement

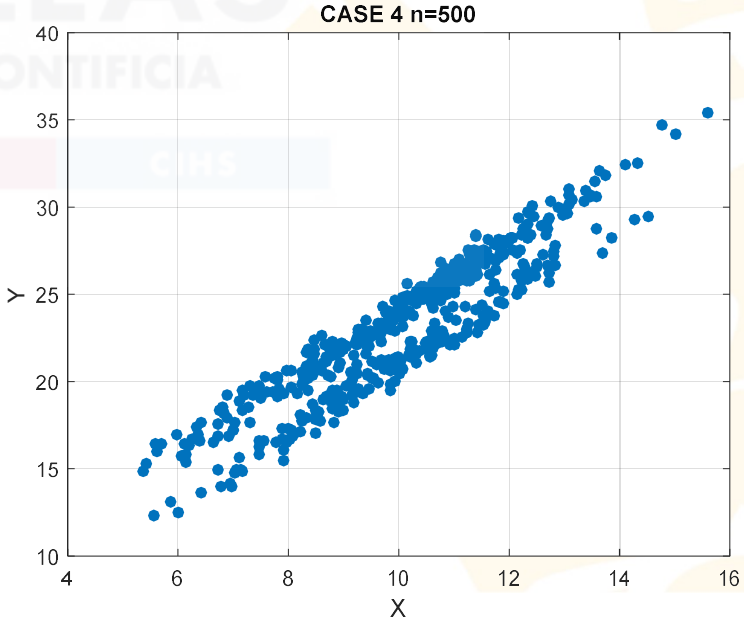
- The simple linear regression model assumes a hypothesis that must be satisfied to be used without any problem
  - It is common to use it, although some hypotheses may be violated, always with caution and knowing what that implies
- Basic assumptions
  - There is a linear relation between the variables
  - Possible noise has zero mean and constant variance
  - Noise follows a normal distribution
  - Noise has no structure; there is independence between errors

Simple Linear Regression
Problematic example cases





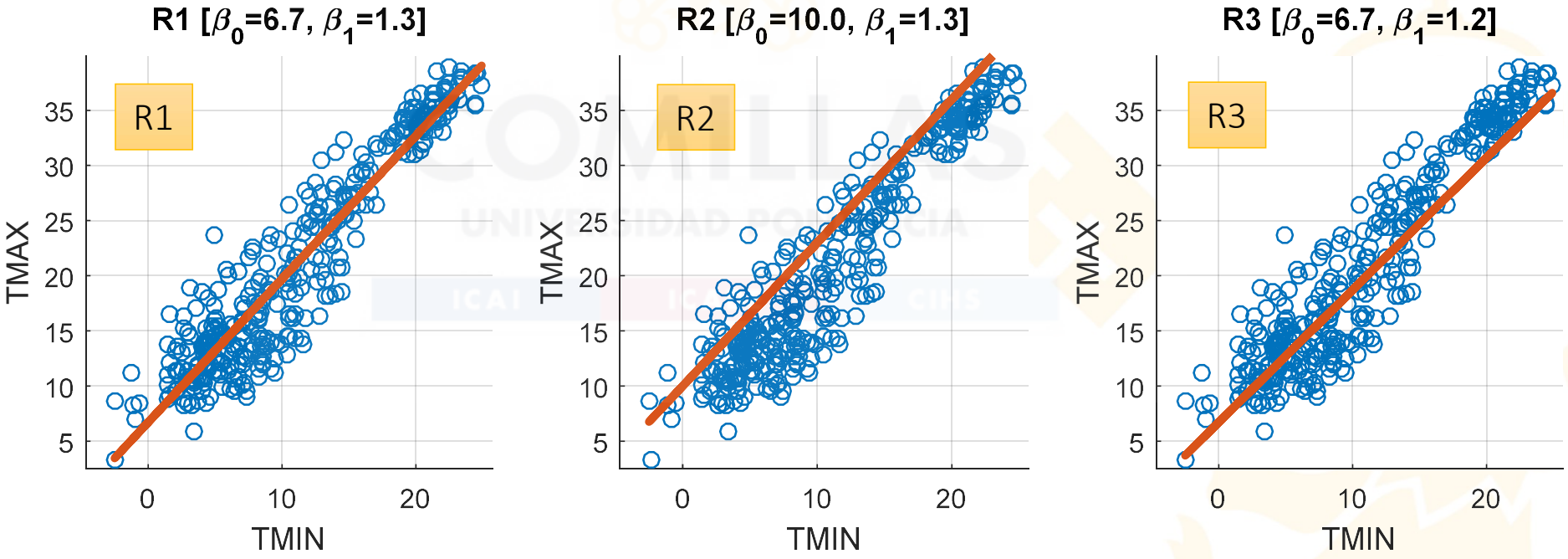




Simple Linear Regression
Coefficient estimation

- Objective: from the available data, obtain a (good) estimation of the model coefficients
- Temperature example

Which straight line (R1, R2, or R3) is better?

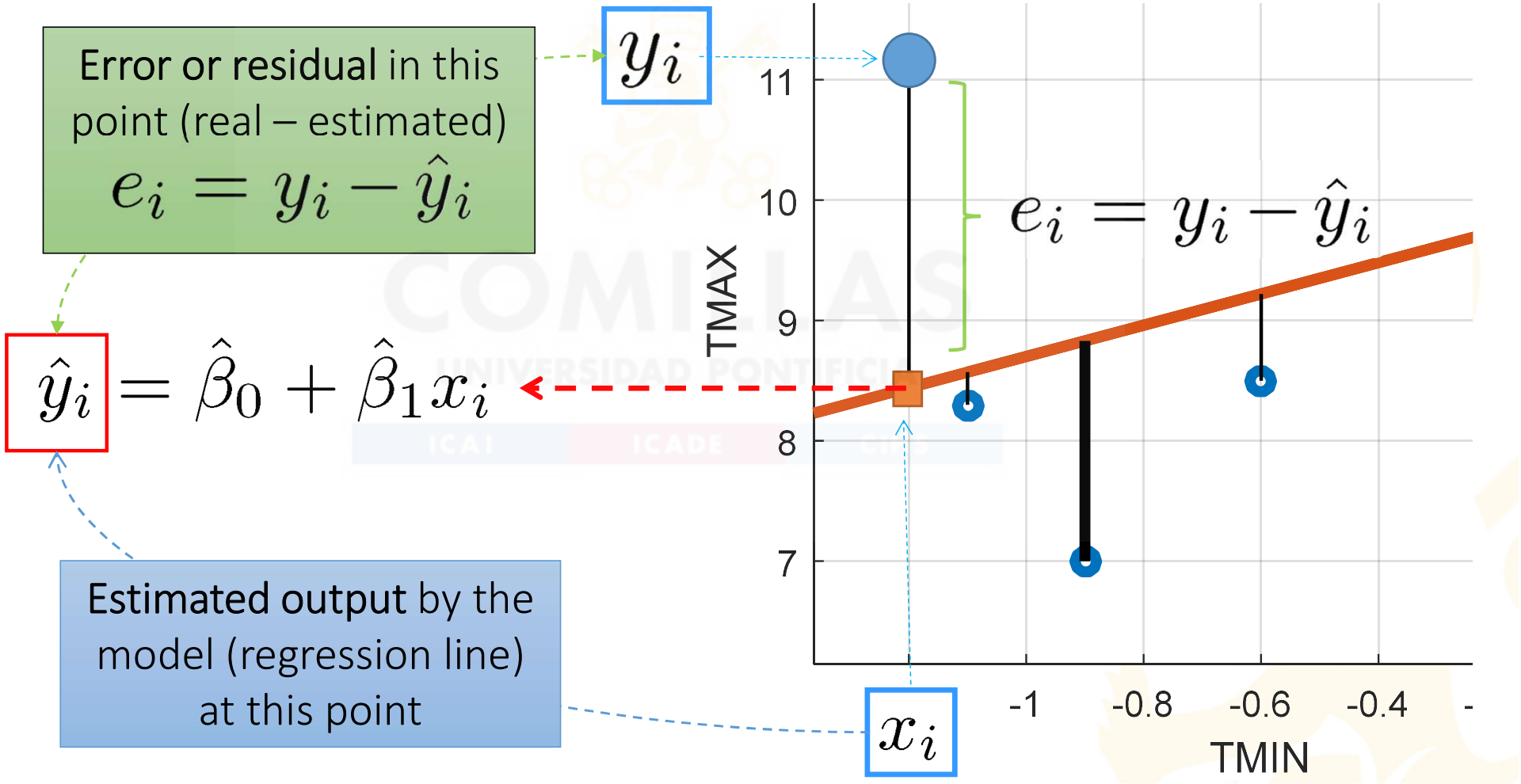


Why?

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Simple Linear Regression
Coefficient estimation

- We want to find the coefficients such that the resulting straight line is as closest as possible to the data

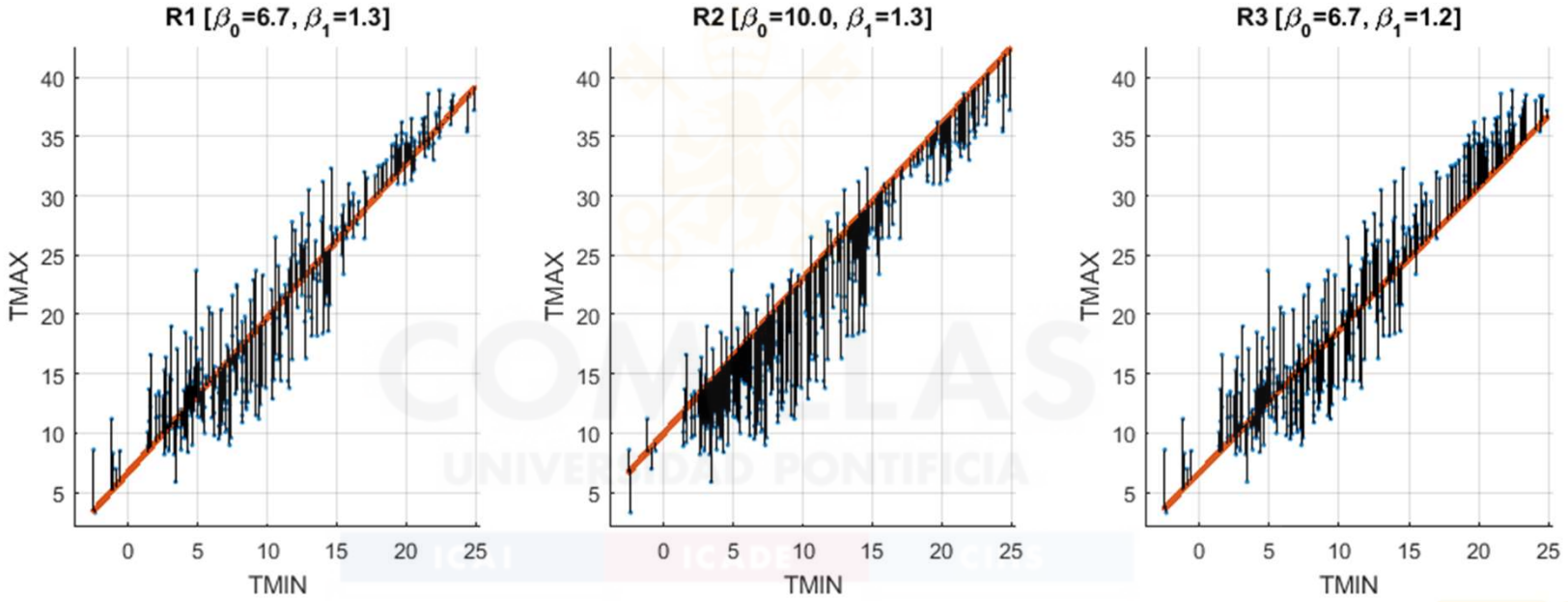


Simple Linear Regression
Coefficient estimation

- Residuals can be positive or negative

$$e_i = y_i - \hat{y}_i$$

- The smaller, the better



Which straight line (R1, R2, or R3) produces lower residuals?

Consider these global error measures: ✗ $\sum_{i=1}^{n} e_i$ ✓ $\sum_{i=1}^{n} e_i^2$ ✓ $\sum_{i=1}^{n} |e_i|$

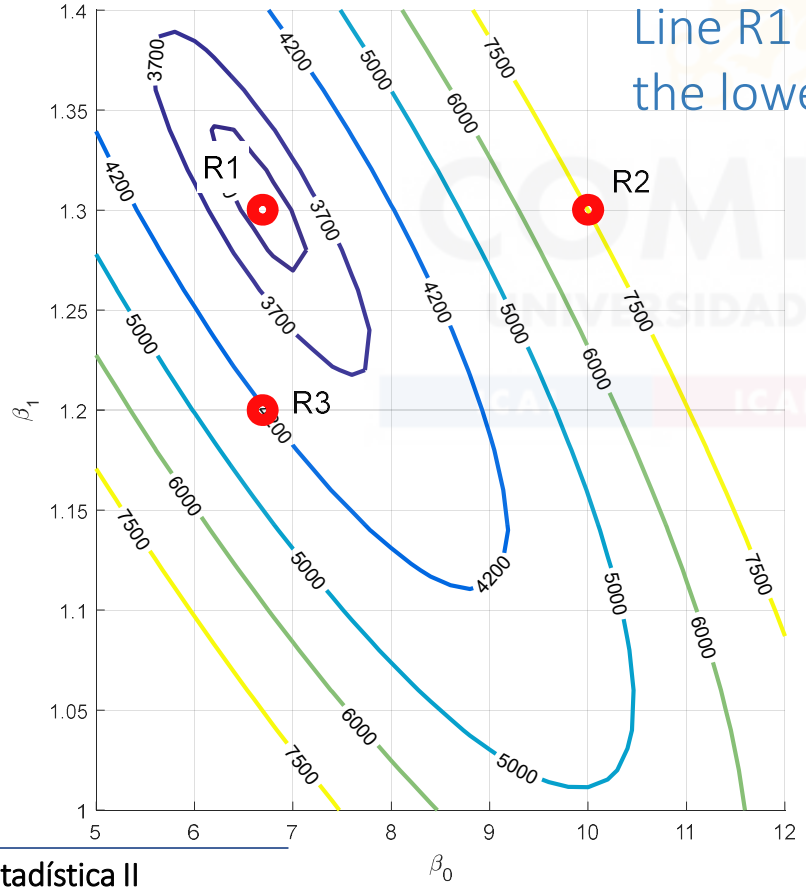
Simple Linear Regression
Coefficient estimation

- Common criterion: minimize the sum of the squared residuals

The residual sum of squares (RSS) is also known as the sum of squared residuals (SSR)

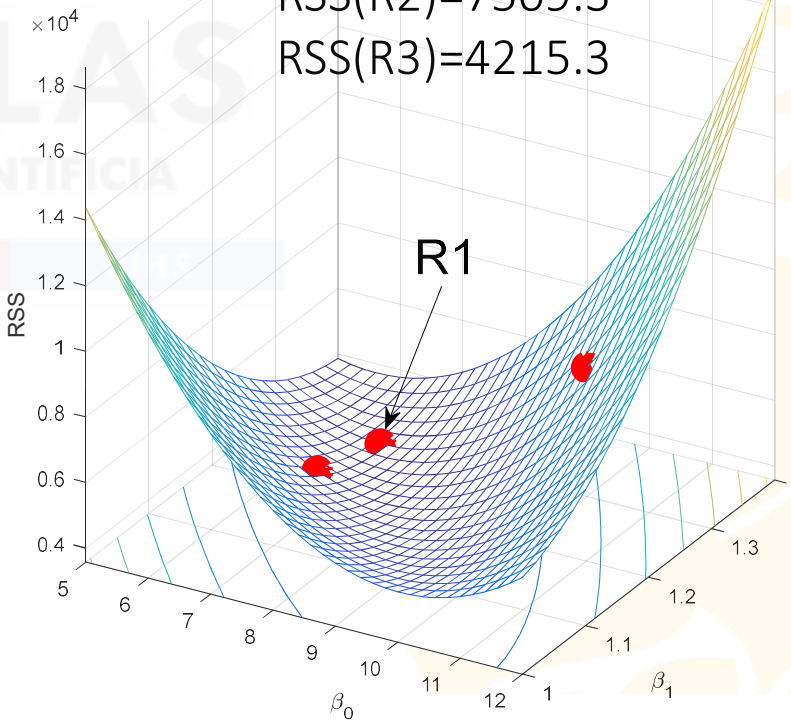
$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

- How does RSS change when the model parameters are changed?



Line R1 is the one with the lowest RSS

RSS(R1)=3573.6
RSS(R2)=7509.3
RSS(R3)=4215.3



Simple Linear Regression
Coefficient estimation

- Least squares approach (the common one):
  - The coefficients to determine will be those that minimize the sum of the squared errors

    The regression straight line crosses the center of the point cloud $(\bar{x}, \bar{y})$
  - RSS is a quadratic function for all the coefficients, with a clear minimum (there is a linear relation)
  - Deriving with respect to each parameter and setting equal to zero, we obtain the analytical expression of the estimated coefficients minimizing RSS

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$\boldsymbol{cov(x,y)}$ over the numerator; $\boldsymbol{var(x)}$ over the denominator

$$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$$

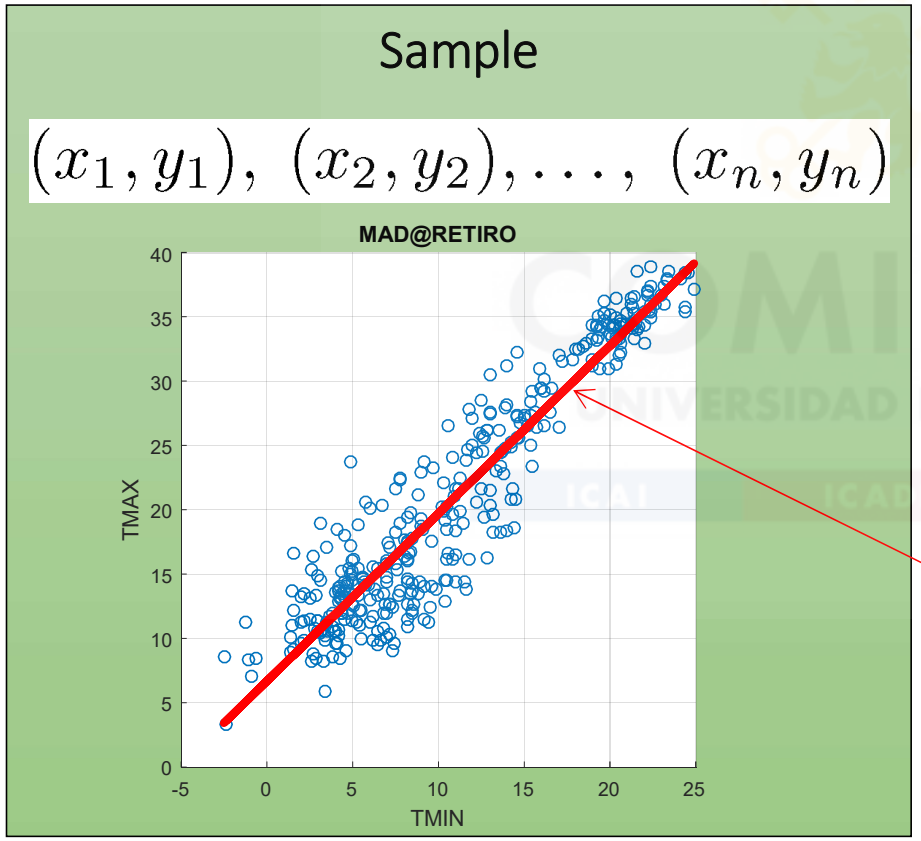
The regression straight line crosses the center of the point cloud $(\bar{x}, \bar{y})$

Simple Linear Regression
Coefficient estimation

- Estimation of the temperature problem
  - Estimate the maximum daily temperature in the Retiro Park from the minimum daily temperature

Sample

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



$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

6.6735 1.3043

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

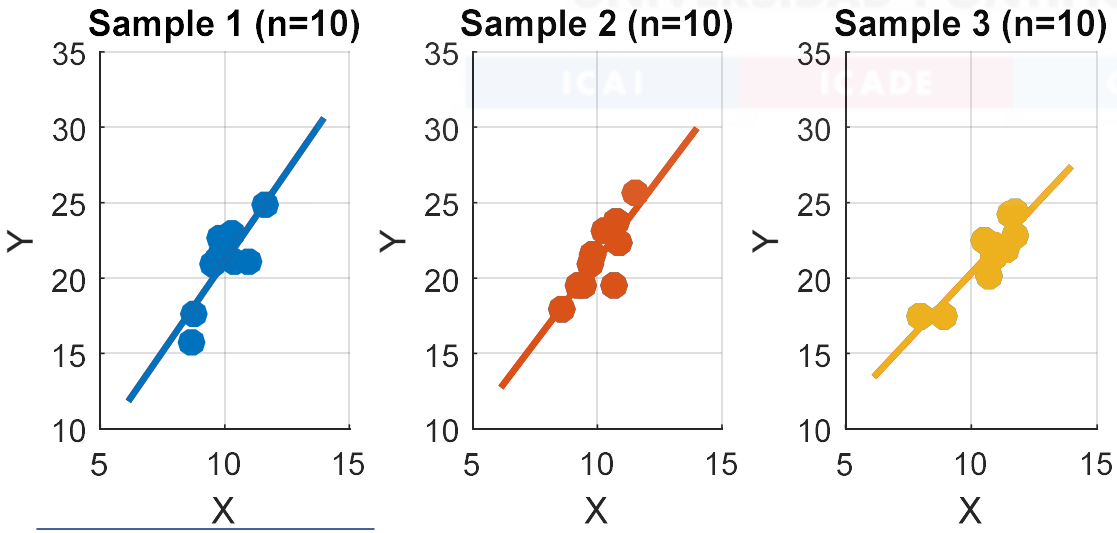
Estimated TMAX if TMIN = 20 ºC

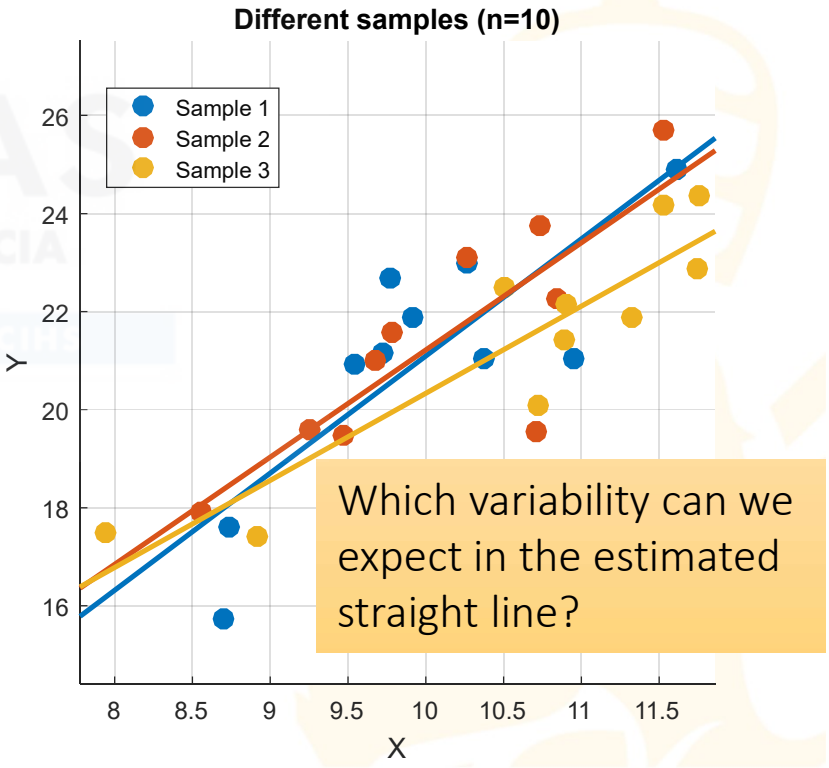
32.7595 = 6.6735 + 1.3043 * 20

Simple Linear Regression
Accuracy in the coefficient estimation

- The estimated coefficients can be different depending on the samples used
  - Variability in sample data is moved to the coefficients
  - The more precise the existing linear relation, the lower the variability among samples
- Example

```
n = 10; % sample size
m = 3; % number of samples
x = normrnd(10,1,n,m);
noise = normrnd(0,1,size(x));
b0 = 1; b1 = 2;
y = b0 + b1 * x + noise;
```





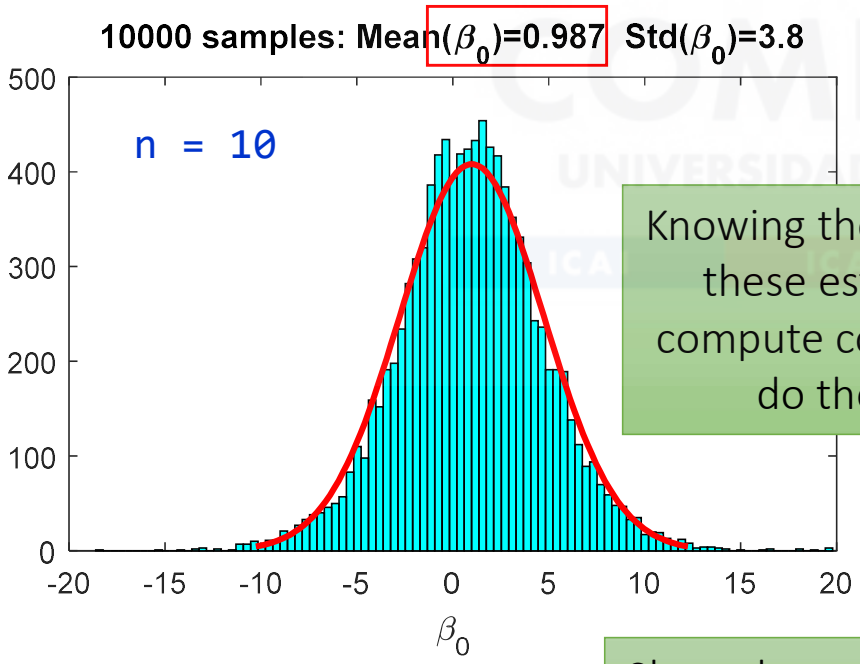
Which variability can we expect in the estimated straight line?

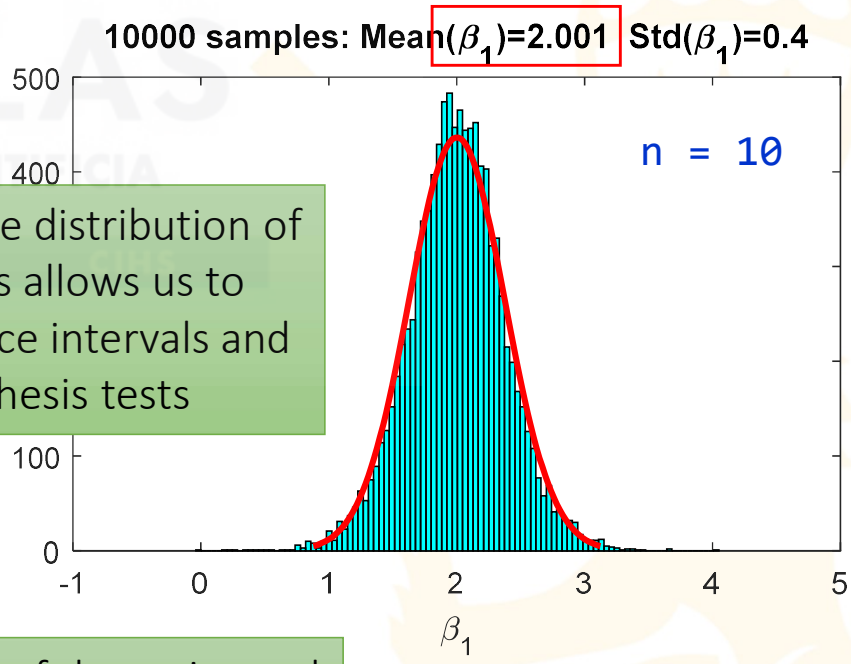
Simple Linear Regression
Accuracy in the coefficient estimation

- Sample distribution of the estimators (marginal ones)
  - Illustrative example

```
n = 10; % sample size
m = 10000; % number of samples
x = normrnd(10,1,n,m);
noise = normrnd(0,1,size(x));
b0 = 1; b1 = 2;
y = b0 + b1 * x + noise;
```

- Generate many samples (10000), with 10 data each
- Estimate the model parameters in each sample by least squares
- Observe the empirical distribution of the estimated coefficients drawing the histogram





Knowing the sample distribution of these estimators allows us to compute confidence intervals and do the hypothesis tests

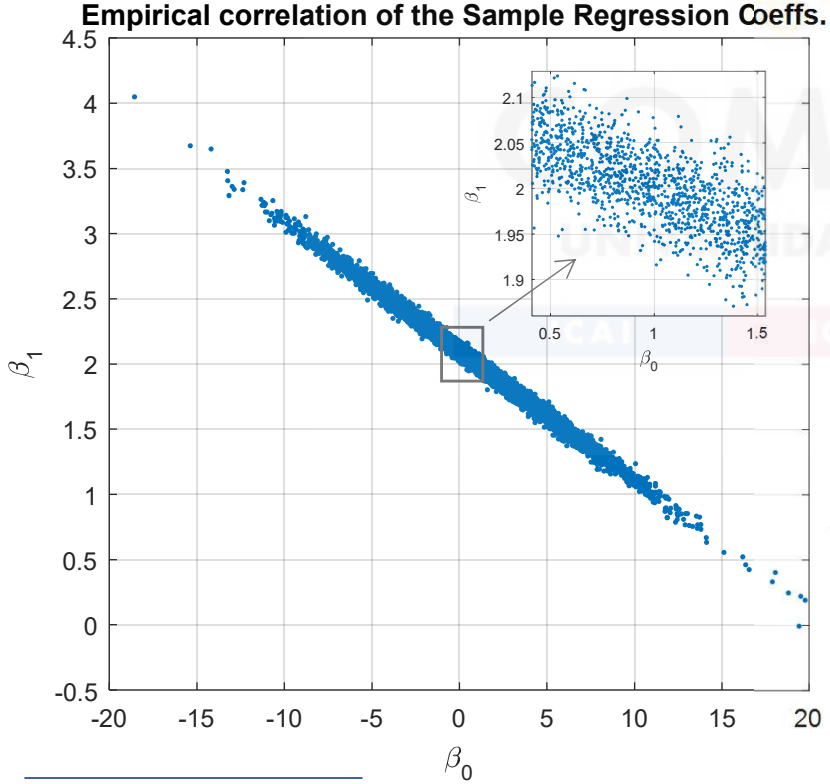
Show the sensitivity of the estimated coefficients to the sample used

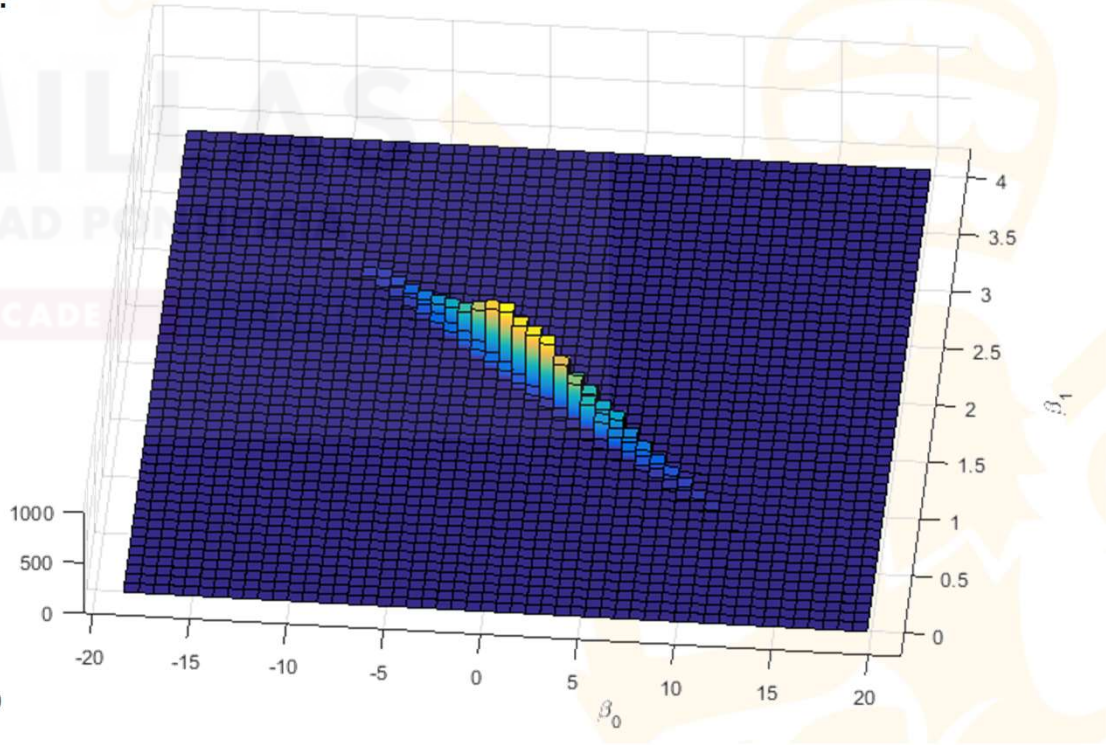
Simple Linear Regression
Accuracy in the coefficient estimation

- Joint sample distribution of the estimators
  - Illustrative example

```
n = 10; % sample size
m = 10000; % number of samples
x = normrnd(10,1,n,m);
noise = normrnd(0,1,size(x));
b0 = 1; b1 = 2;
y = b0 + b1 * x + noise;
```

- Given a sample, the estimations obtained for both parameters of the straight line are not independent
- In this example, if the slope estimation increases, then the intercept decreases

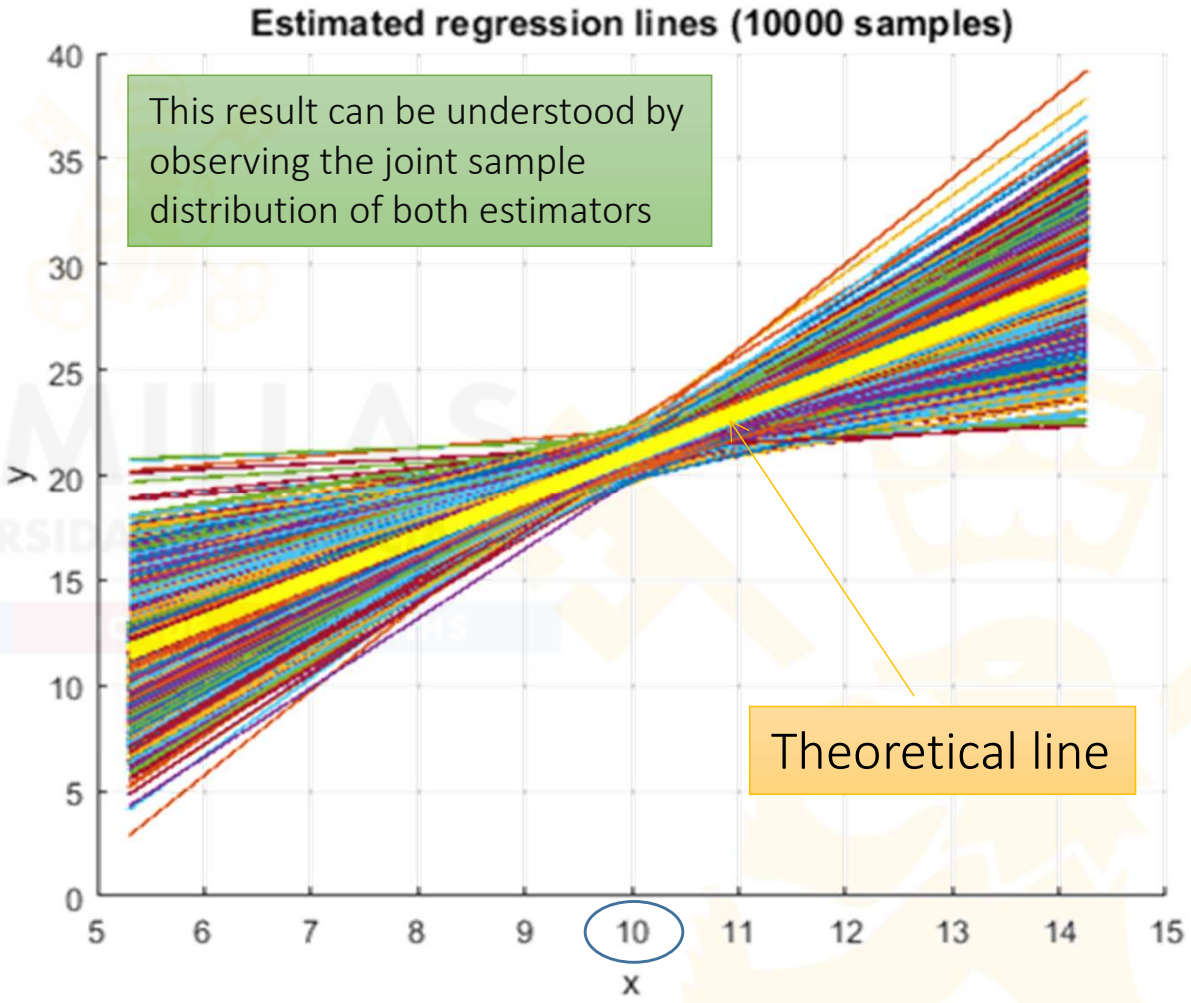




Simple Linear Regression
Accuracy in the coefficient estimation

- Sample distribution of the regression straight line
  - Illustrative example

```
n = 10; % sample size
m = 10000; % number of samples
x = normrnd(10,1,n,m);
noise = normrnd(0,1,size(x));
b0 = 1; b1 = 2;
y = b0 + b1 * x + noise;
```



- It is possible to compute analytically a confidence interval for the regression straight line
- This interval contains, with a predefined confidence level, the theoretical regression line
- The width of the confidence interval increases when moving away in x from its mean

Simple Linear Regression
Accuracy in the coefficient estimation

- Standard error of the estimators
  - Widely used as a variability measure of the sample distribution of the estimator to infer from its accuracy
  - It is possible to determine theoretically the standard deviation of the parameter estimators

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\epsilon \approx N(0, \sigma^2)$$

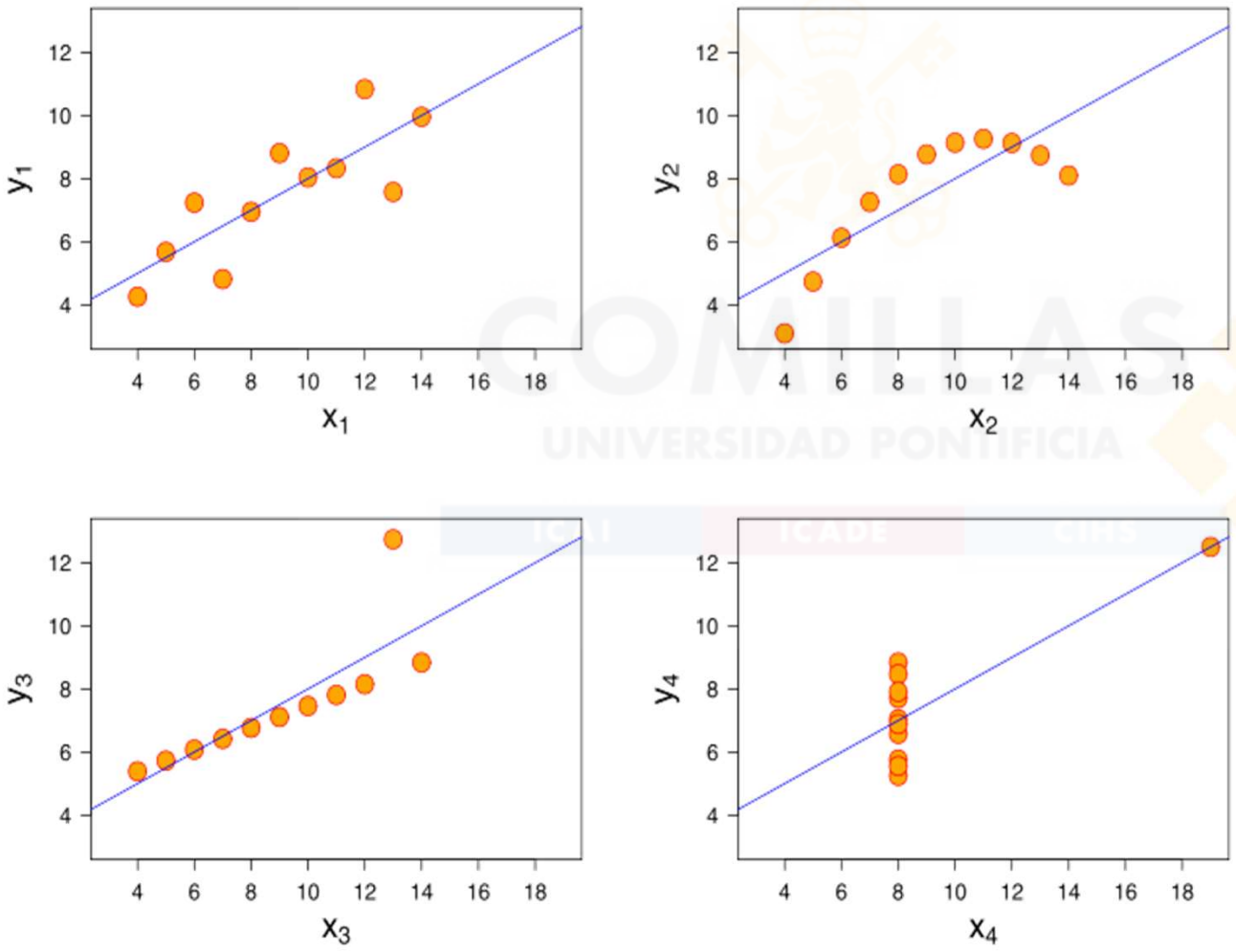
What causes a decrease in the standard error?

What happens if the variable X does not vary?

What happens if there is no noise?

Simple Linear Regression
Model quality: Anscombe's quartet

- The 4 sets of data have the same summary statistics (mean, variance, etc.). The linear fit has the same quality



- Only the first case makes sense with the proposed model
- The interpretation of the result is critical

Anscombe, F. J. (1973). "*Graphs in Statistical Analysis*". American Statistician 27 (1): 17–21. JSTOR 2682899

Simple Linear Regression
Goodness of fit

- Residual variance ("residual variability")
  - Unbiased estimator of the noise variance $\epsilon \approx N(0, \sigma^2)$
  - Measures the mean deviation of the output with respect to the theoretical line
  - It is the output variance not explained, although we were able to perfectly estimate the parameters of the regression line
  - Computed dividing the sum of the squared errors between the number of data – 2 (the degrees of freedom)

$$S^2_{RES} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{1}{n-2}\text{RSS} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Simple Linear Regression
Goodness of fit

- MSE: Mean Squared Error
  - Like the residual variance, measures the mean variability of the errors
  - Doesn't consider the number of estimated parameters in the denominator (biased version if n is small)
  - Not bounded, highly problem-dependent, difficult to interpret
  - RMSE (Root Mean Squared Error): if we compute the square root of the MSE, we obtain a value with the same units that the output (standard deviation of the residuals)

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

Simple Linear Regression
Goodness of fit

- Coefficient of determination
  - Measures the dispersion degree of the point cloud around the line in a normalized way
  - Allows to compare fits of different models and problems
  - Bounded between 0 (large dispersion) and 1 (data following the line)

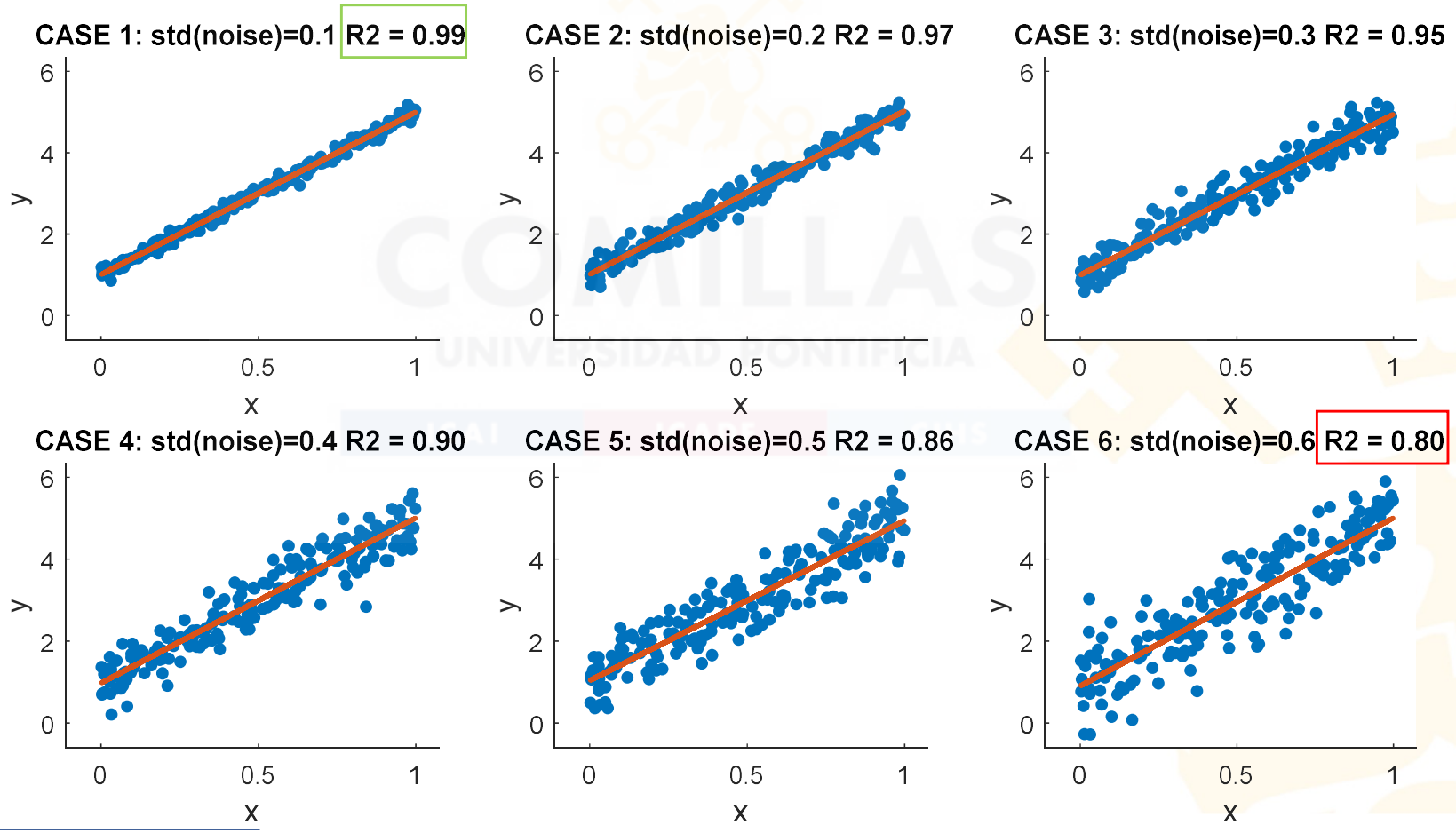
$$R^2 = 1 - \frac{S_{RES}^2}{S_Y^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In the case of a single input variable (simple model), it is the Pearson correlation coefficient squared.

If the non-explained variance (residual) is very similar to the variance of the variable Y, then the model doesn't contribute more than a mere constant.

Simple Linear Regression
Goodness of fit

- Coefficient of determination
  - Values > 0.8 are considered a reasonable fit
- Examples



Simple Linear Regression
Is the model significant?

- Can the slope be zero?
  - If the confidence interval of this parameter contains zero, the model is superfluous. It would be a constant
  - It can be stated as a hypothesis test

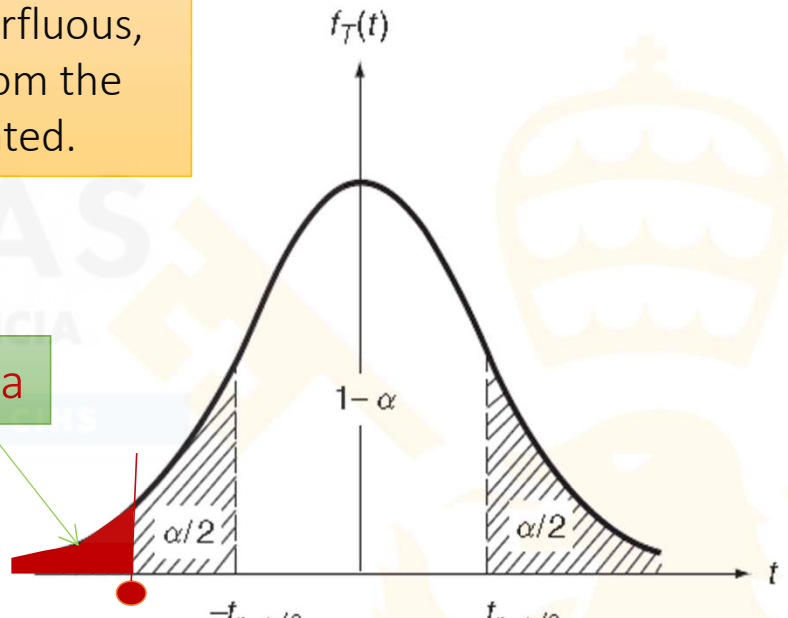
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

If this parameter is superfluous, it must be eliminated from the model, and the fit repeated.

  - If the null hypothesis is true, then

$$\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \approx t_{n-2}$$

p-value=2*area



Value of the sample test statistic

Acceptance region

Simple Linear Regression
Is the model significant?

- Can the intercept be zero?
  - If the confidence interval of the parameter contains zero, this parameter is negligible
  - In such a case, the model is simpler, i.e., one parameter less to estimate
  - It can be stated as a hypothesis test

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

If this parameter is superfluous, it must be eliminated from the model, and the fit repeated.

  - If the null hypothesis is true, then

$$\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \approx t_{n-2}$$

The same than for the slope

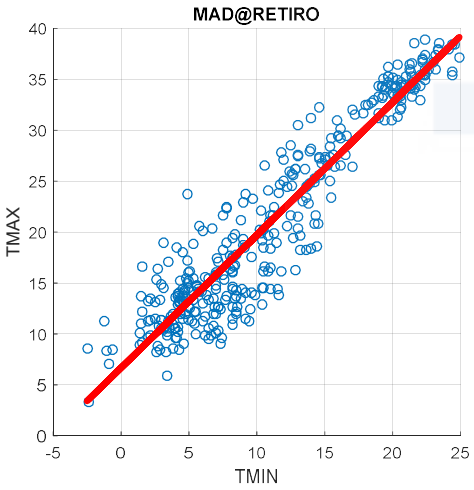
Deleting the parameter means forcing the regression line to cross the coordinate origin.

Simple Linear Regression
Interpreting the fitting model

- Usual information for diagnosis and model interpretation as a table
  - Estimated coefficients and accuracy
  - Quality of the data fitting the model (error measures)
- Temperature problem

6.6735 1.3043

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$




```
Linear regression model:
    TMAX ~ 1 + TMIN

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	6.6735	0.31475	21.202	1.5145e-65
TMIN	1.3043	0.024365	53.532	1.2589e-174

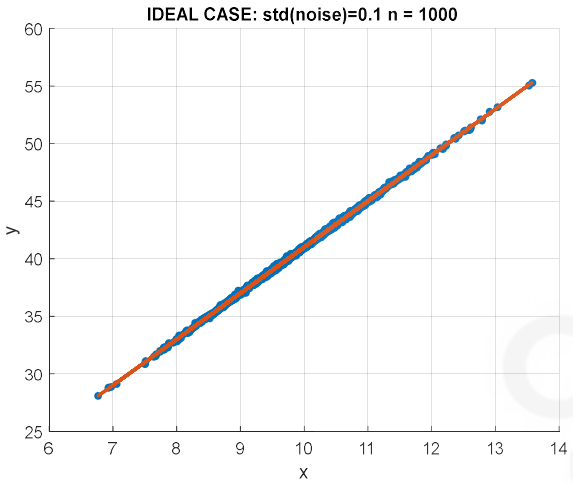
```
Number of observations: 366, Error degrees of freedom: 364
Root Mean Squared Error: 3.13
R-squared: 0.887,  Adjusted R-Squared 0.887
F-statistic vs. constant model: 2.87e+03, p-value = 1.26e-174
```

Simple Linear Regression
Residual analysis

- Once the model is fitted, it is important to check that residuals, what is not explained by the model, satisfy reasonably the basic assumptions
  - Normality: follow a normal distribution with zero mean and constant variance
  - Independence: are independent among them, with no clear internal structure
- Diagnosis of the model or residual analysis
  - Normality plot for the residuals (nonparametric test)
  - Plot the residuals versus the estimated values
  - Independence plot among residuals

Simple Linear Regression
Illustrative synthetic cases

- C1: IDEAL with small noise and many data




```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate        SE          tStat       pValue
                   ________    _________    ______    ____________

    (Intercept)    1.0129       0.031696    31.958    6.6123e-155
    x1             3.9991      0.0031641    1263.9              0


Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 0.0999
R-squared: 0.999,  Adjusted R-Squared 0.999
F-statistic vs. constant model: 1.6e.06, p-value = 0
```

```
n=1000;
x = normrnd(10,1,n,1);
stdnoise = 0.1;
b0 = 1; b1 = 4;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
mod = fitlm(x,y);
```

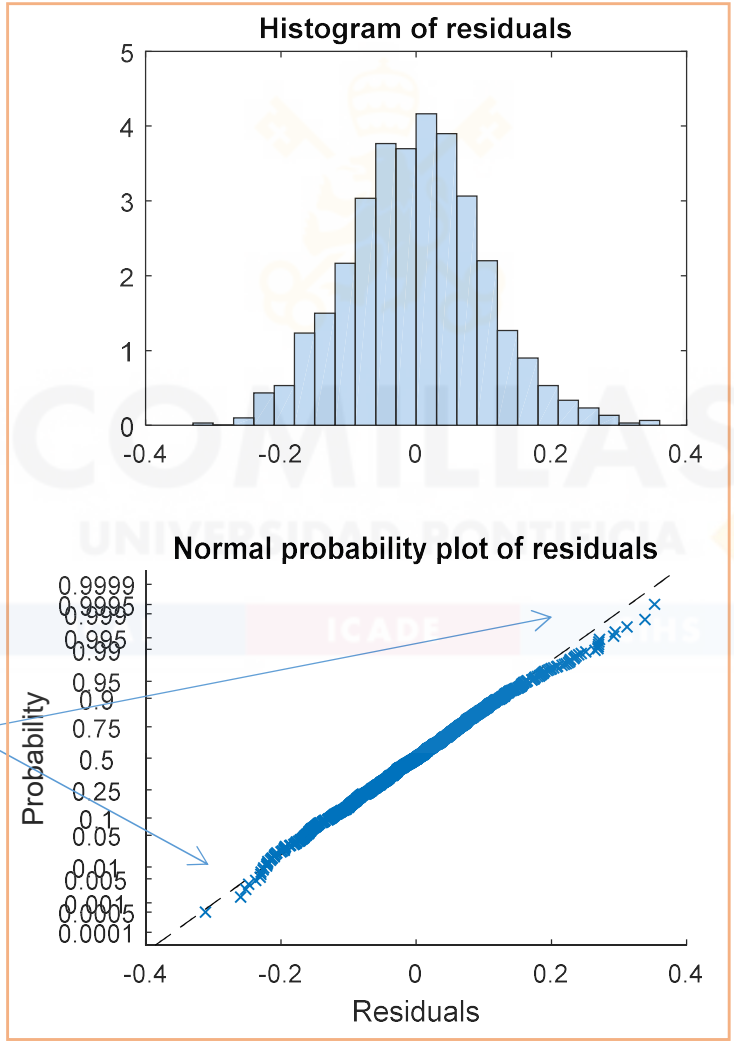
- All the coefficients are significant
- Coefficients estimated with high accuracy
  - Lower accuracy in the intercept
- R^2 is very high (good fit)

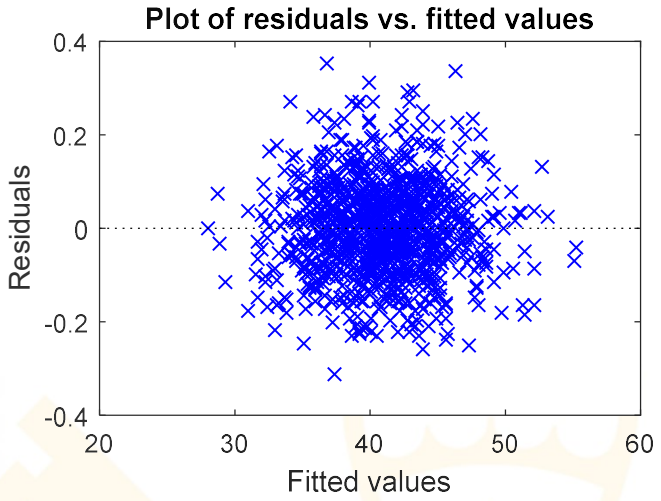
Simple Linear Regression
Illustrative synthetic cases

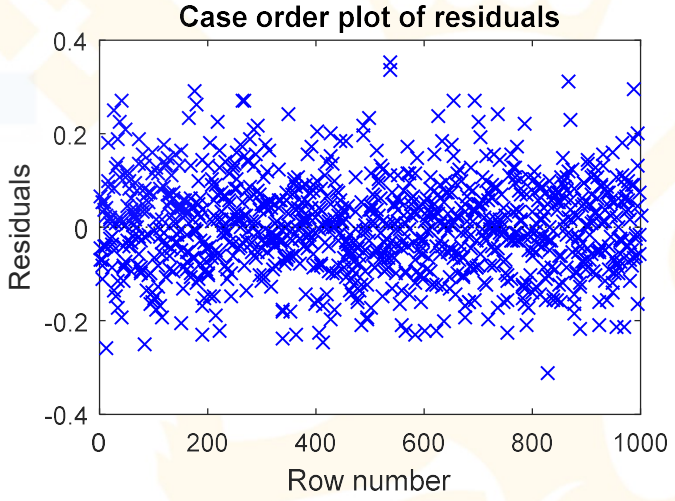
- C1: IDEAL with small noise and many data (residual analysis)

Residuals seem independent (do not change with the estimated value or with the position in the sample)

Some difference is found in the distribution tails, although it is a synthetic case with normal noise (it is common for this to happen)

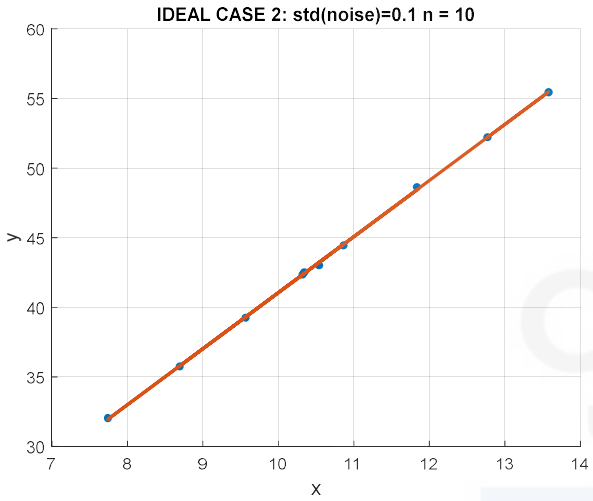






Simple Linear Regression
Illustrative synthetic cases

- C2: IDEAL with small noise and few data




```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     0.73986     0.23368    3.1661      0.013272
    x1               4.0311    0.021725    185.55     7.9653e-16

Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 0.115
R-squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 3.44e.04, p-value = 7.97e-16
```

```
n=10;
x = normrnd(10,1,n,1);
stdnoise = 0.1;
b0 = 1; b1 = 4;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
mod = fitlm(x,y);
```

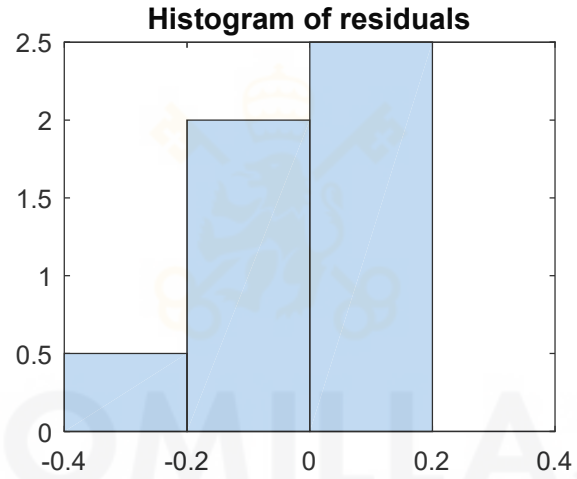
- R^2 is very high (good fit)
- The intercept has a p-value above 1%
  - With 1%, this parameter should be rejected

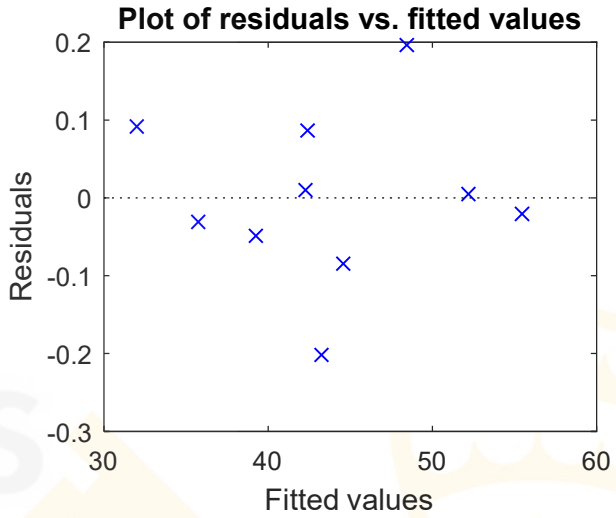
Simple Linear Regression
Illustrative synthetic cases

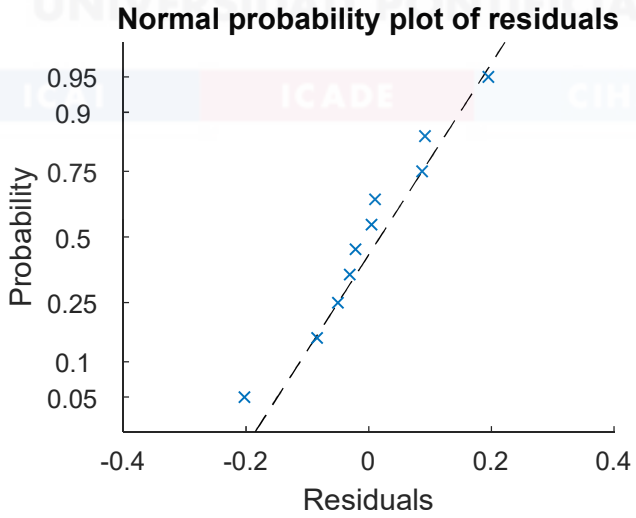
- C2: IDEAL with small noise and few data (residual analysis)

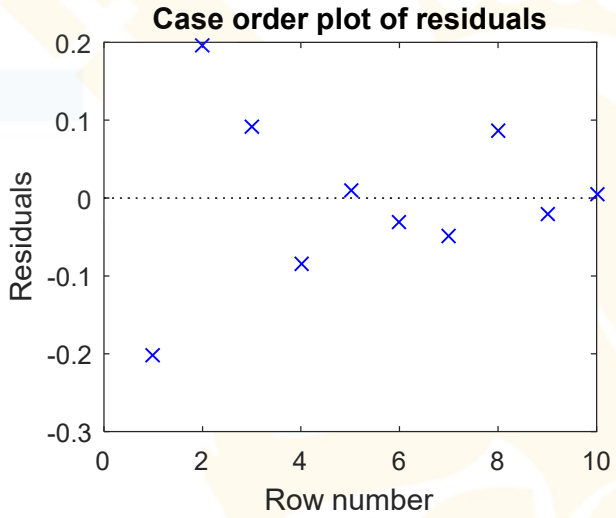
With few data, the nonparametric test plot doesn't shed light

They do not contradict the normality and independence hypotheses



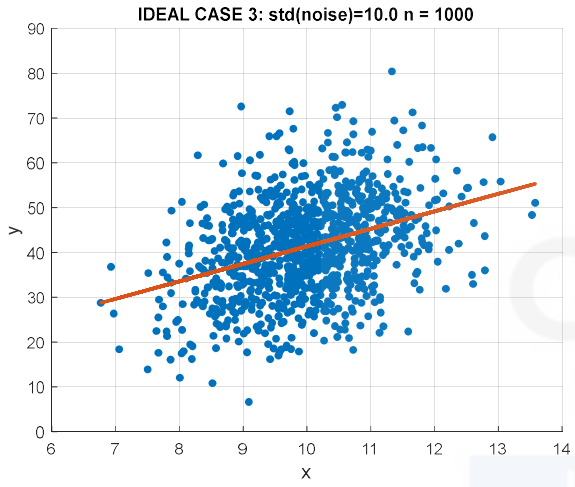






Simple Linear Regression
Illustrative synthetic cases

- C3: IDEAL with large noise and many data




```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)    2.2936       3.1696    0.72362       0.46947
    x1             3.9072      0.31641     12.348     1.0587e-32

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 9.99
R-squared: 0.133,  Adjusted R-Squared 0.132
F-statistic vs. constant model: 152, p-value = 1.06e-32
```

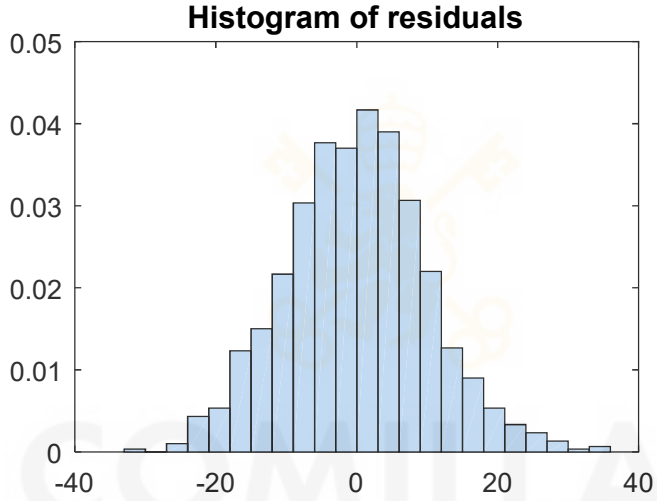
```
n=1000;
x = normrnd(10,1,n,1);
stdnoise = 10;
b0 = 1; b1 = 4;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
mod = fitlm(x,y);
```

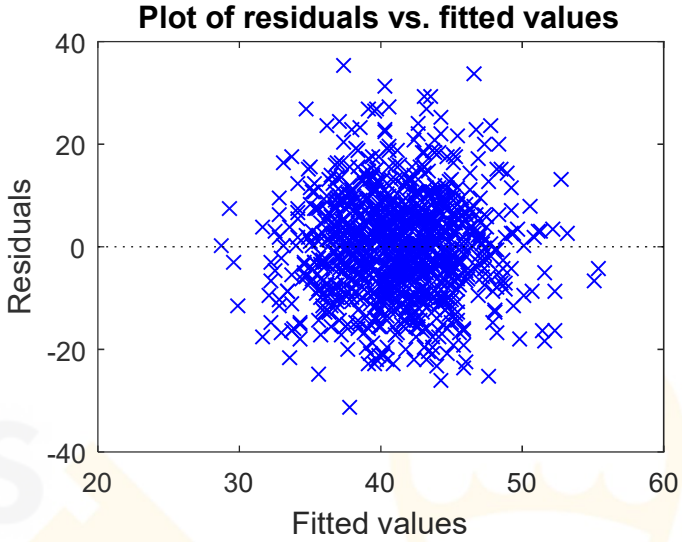
- It is difficult to see the regression line in the point cloud
- R^2 is very small (bad fit)
- The intercept can be zero (model must be simplified)

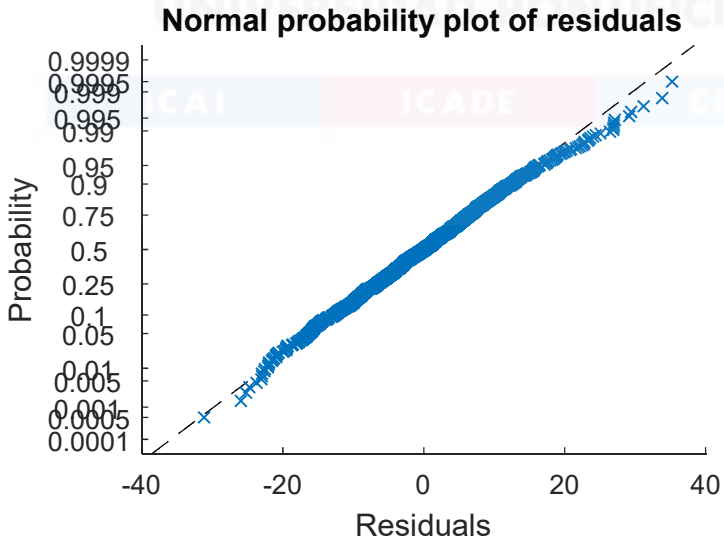
Simple Linear Regression
Illustrative synthetic cases

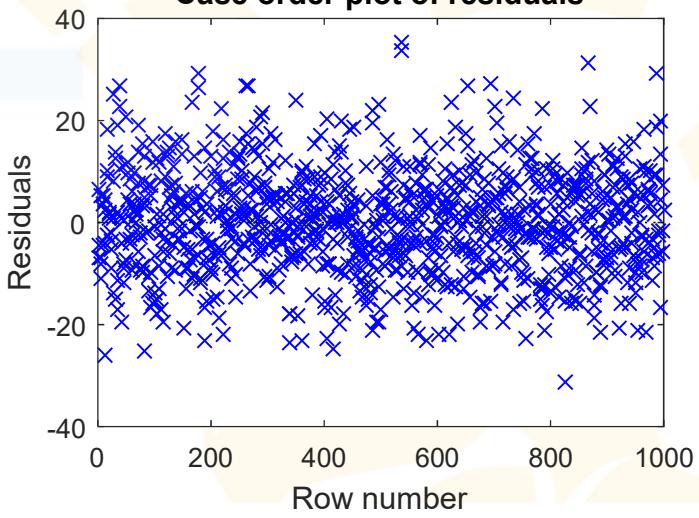
- C3: IDEAL with large noise and many data (residual analysis)

Residuals seem normal and independent



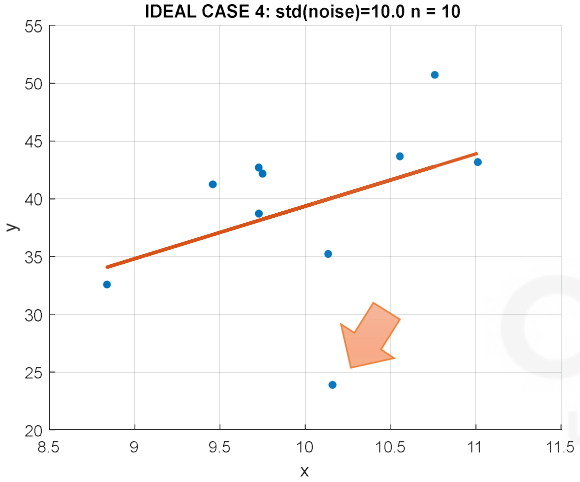






Simple Linear Regression
Illustrative synthetic cases

- C4: IDEAL with large noise and few data




```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    ______    ________    _______

    (Intercept)    -5.9901     36.784    -0.16285    0.87468
    x1              4.5355     3.6672     1.2368     0.25125

Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 7.16
R-squared: 0.161,  Adjusted R-Squared 0.0556
F-statistic vs. constant model: 1.53, p-value = 0.251
```

```
n=10;
x = normrnd(10,1,n,1);
stdnoise = 10;
b0 = 1; b1 = 4;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
mod = fitlm(x,y);
```

- Point cloud questions any complicated model
- R^2 shows that the fit is very poor
- High p-values show that both coefficients are superfluous (a constant is better)

Simple Linear Regression
Illustrative synthetic cases

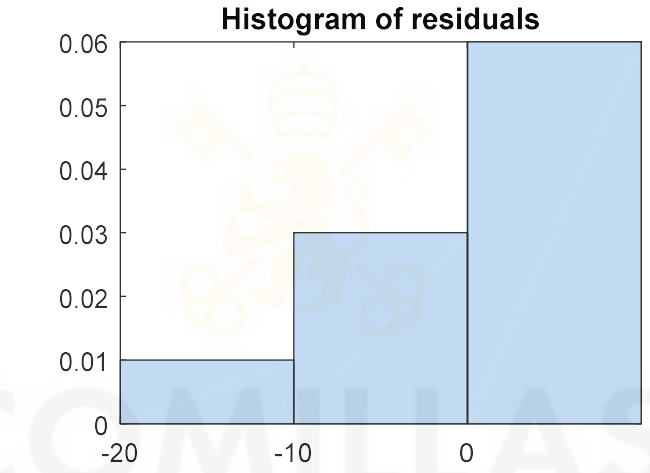
- C4: IDEAL with large noise and few data (residual analysis)

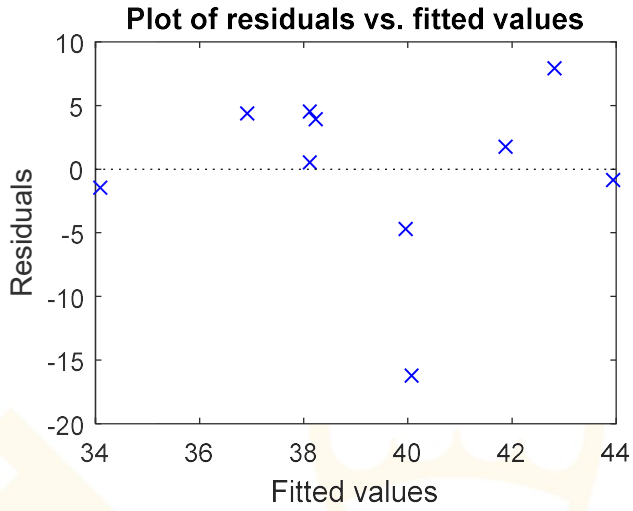
With few data, the nonparametric test plot doesn't shed light

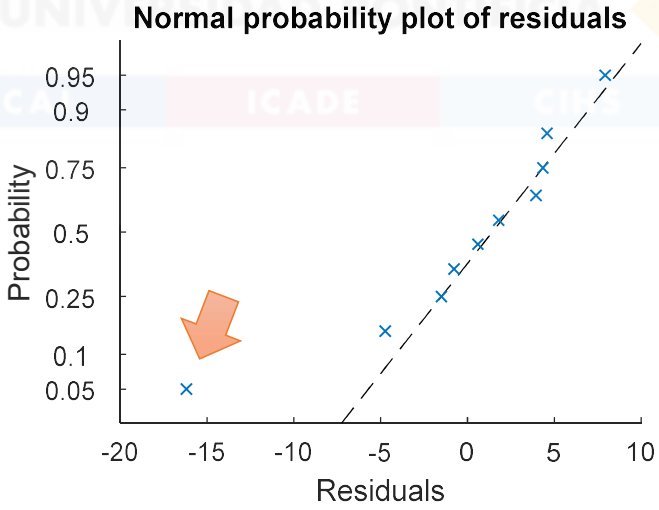
There is a suspicious point when comparing with the others (it can be an outlier)

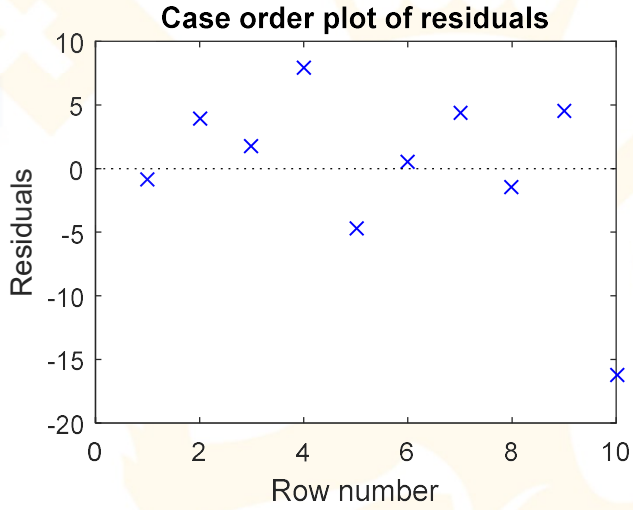
They do not contradict the normality and independence hypotheses

Besides, RMSE is lower than expected, a symptom of overfitting by using an over-parametrized model (in a real problem, this can't be known given that we don't have the error variance)



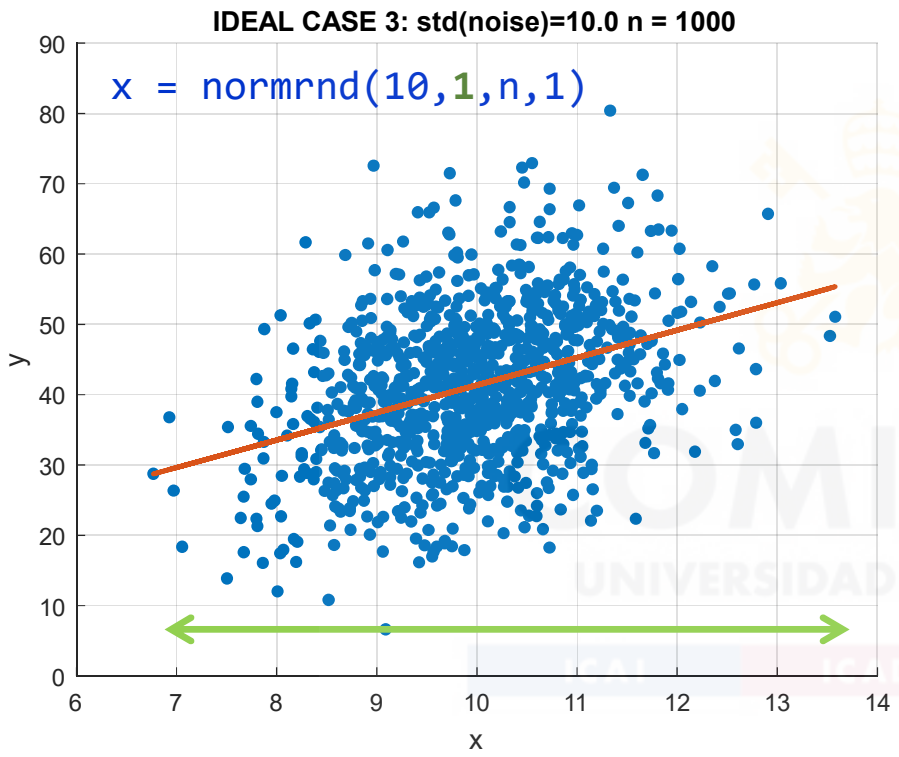


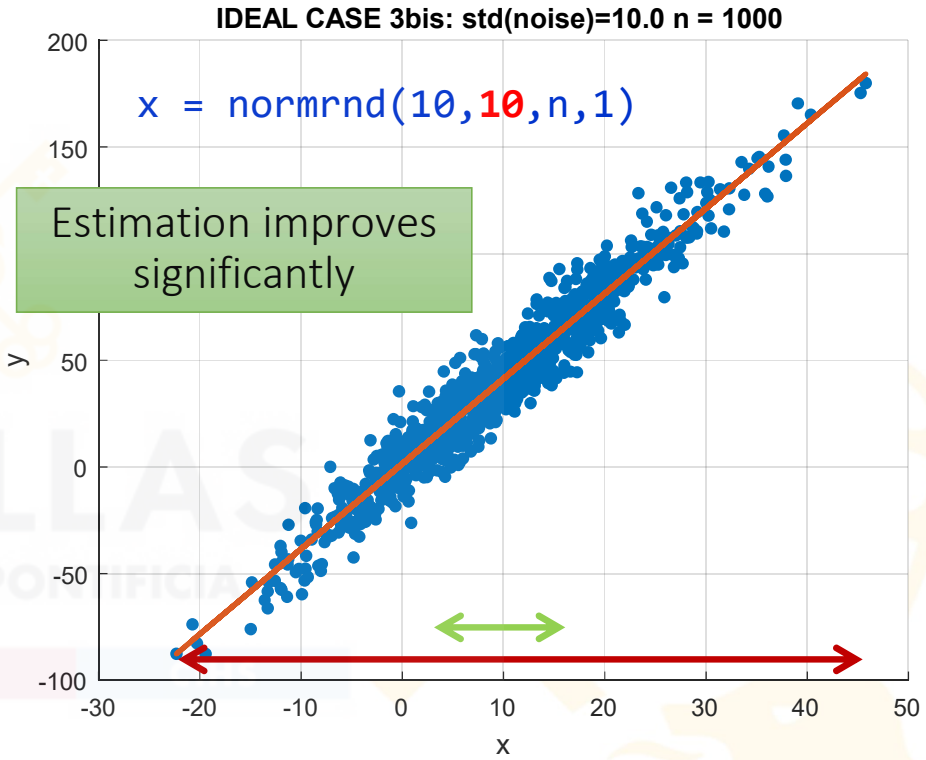




Simple Linear Regression
Illustrative synthetic cases

- C3bis: like C3 but with larger dispersion in x





	Estimate	SE	tStat	pValue
(Intercept)	2.2936	3.1696	0.72362	0.46947
x1	3.9072	0.31641	12.348	1.0587e-32

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 9.99
R-squared: 0.133, Adjusted R-Squared 0.132
F-statistic vs. constant model: 152, p-value = 1.06e-32

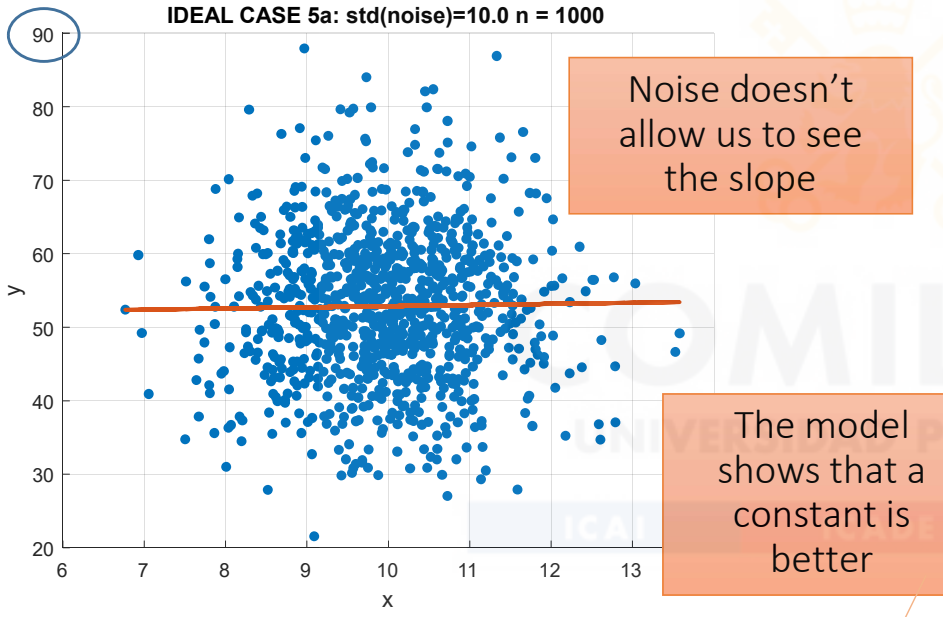
	Estimate	SE	tStat	pValue
(Intercept)	1.4587	**0.43989**	3.3161	0.00094567
x1	3.9907	**0.031641**	126.12	0

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 9.99
R-squared: 0.941, Adjusted R-Squared 0.941
F-statistic vs. constant model: 1.59e.04, p-value = 0

Simple Linear Regression
Illustrative synthetic cases

- C5a, C5b and C5c: IDEAL small slope with large and small noise

```
stdnoise = 10;
b0 = 50; b1 = 0.25;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
```



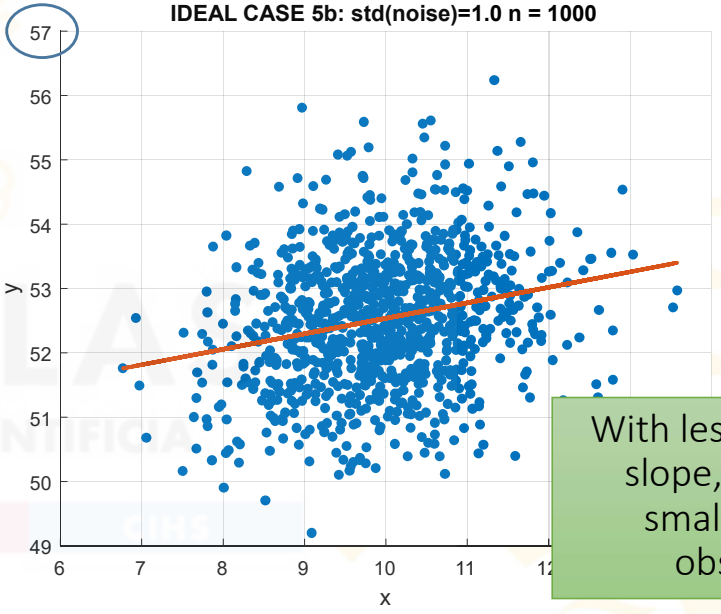
Noise doesn't allow us to see the slope

The model shows that a constant is better

	Estimate	SE	tStat	pValue
(Intercept)	51.294	3.1696	16.183	1.7461e-52
x1	0.15724	0.31641	0.49693	0.61934

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 9.99
R-squared: 0.000247, Adjusted R-Squared -0.000754
F-statistic vs. constant model: 0.247, p-value = 0.619

```
stdnoise = 1;
b0 = 50; b1 = 0.25;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
```



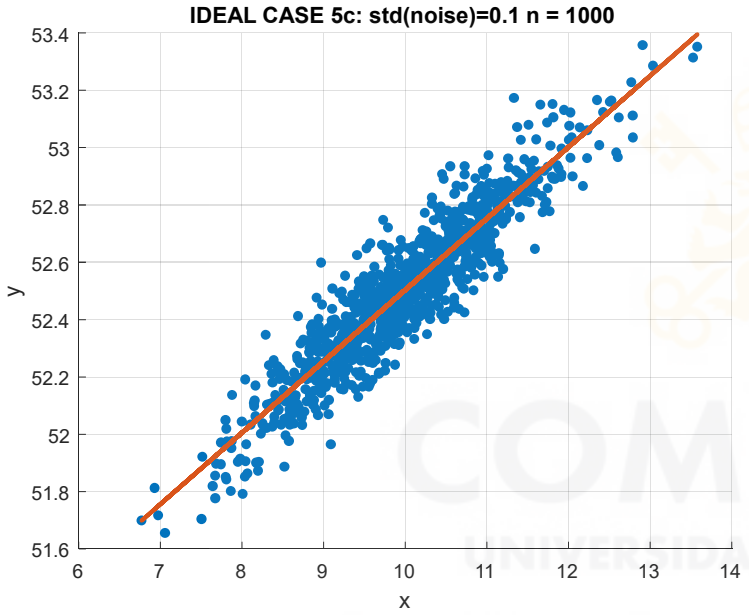
With less noise, the slope, although small, can be observed

	Estimate	SE	tStat	pValue
(Intercept)	50.129	0.31696	158.16	0
x1	0.24072	0.031641	7.6079	6.4292e-14

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 0.999
R-squared: 0.0548, Adjusted R-Squared 0.0539
F-statistic vs. constant model: 57.9, p-value = 6.43e-14

Simple Linear Regression
Illustrative synthetic cases

- C5a, C5b and C5c: IDEAL small slope with large and small noise




```
stdnoise = 0.1;
b0 = 50; b1 = 0.25;
y = b0 + b1 * x + normrnd(0,stdnoise,size(x));
```

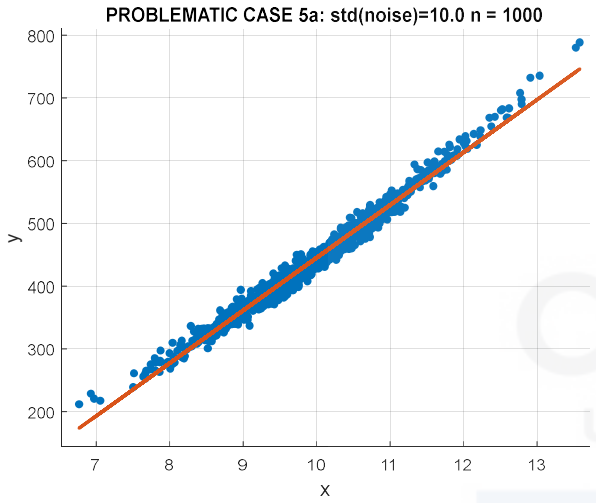
	Estimate	SE	tStat	pValue
(Intercept)	50.013	0.031696	1577.9	0
x1	0.24907	0.0031641	78.717	0

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 0.0999
R-squared: 0.861, Adjusted R-Squared 0.861
F-statistic vs. constant model: 6.2e.03, p-value = 0

- The lower the noise, the better the regression line can be observed
- The same happens with the model
  - No doubt about the existence of a straight line
  - Estimations of the coefficients improve
- Noise impacts approximately proportionally to the values of the deterministic term

Simple Linear Regression
Illustrative synthetic cases

- C6: PROBLEMATIC with small noise and many data




```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    _______    ________    ______

    (Intercept)    -394.89      3.5978    -109.76     0
    x1              84.026      0.35917    233.95     0

Number of observations: 1000, Error degrees of freedom: 998
Root Mean Squared Error: 11.3
R-squared: 0.982,  Adjusted R-Squared 0.982
F-statistic vs. constant model: 5.47e.04, p-value = 0
```

```
n=1000;
x = normrnd(10,1,n,1);
stdnoise = 10;
b0 = 1; b1 = 4;
y = b0 + b1 * x + b1 * x.^2 + normrnd(0,stdnoise,size(x));
mod = fitlm(x,y);
```

- The estimation of the coefficients has been done without problems
- R^2 confirms a good fit

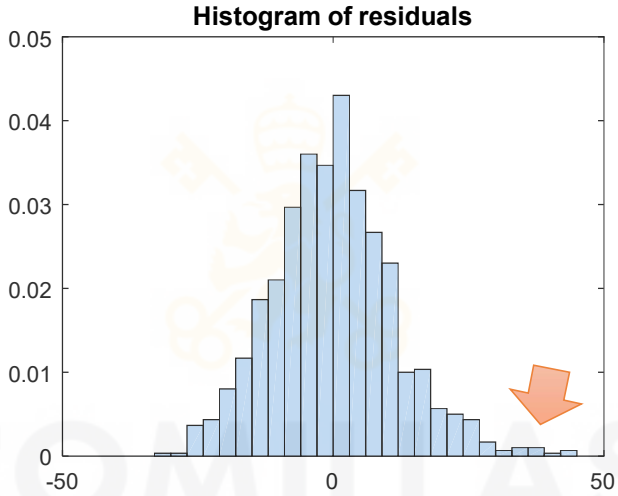
Simple Linear Regression
Illustrative synthetic cases

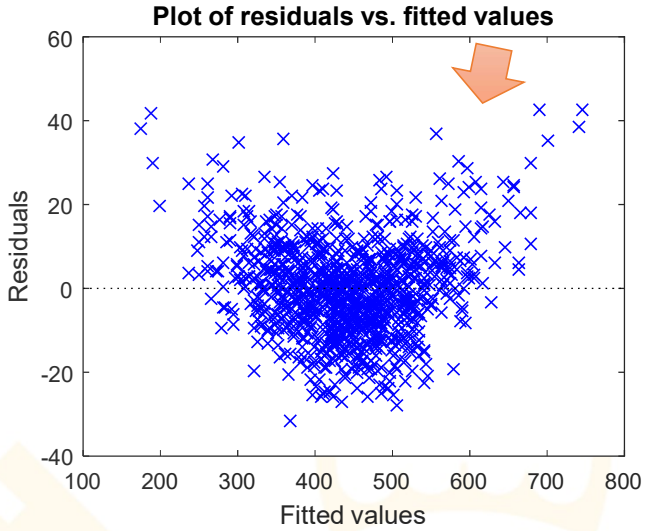
- C6: PROBLEMATIC with small noise and many data (residuals)

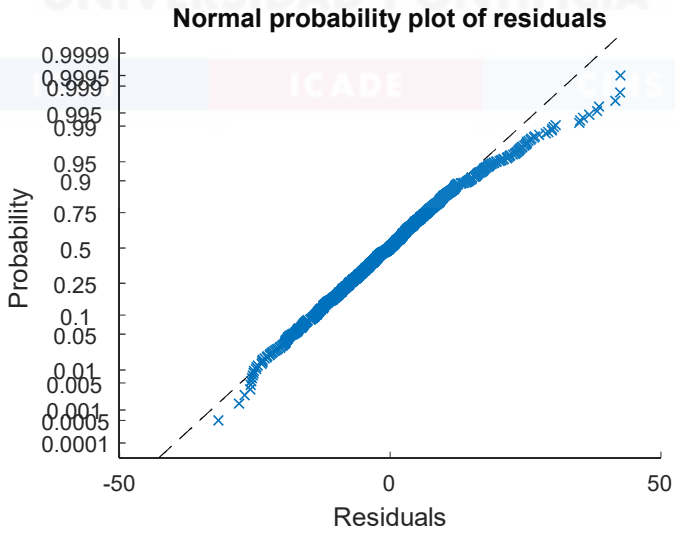
Residuals are far from normality, especially in the right tail (large positive errors)

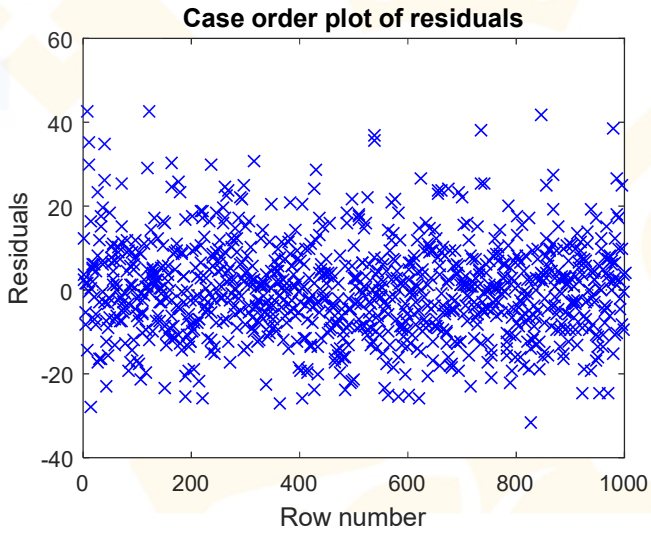
Residual variance doesn't seem constant

Residual analysis suggests that the model could be badly specified



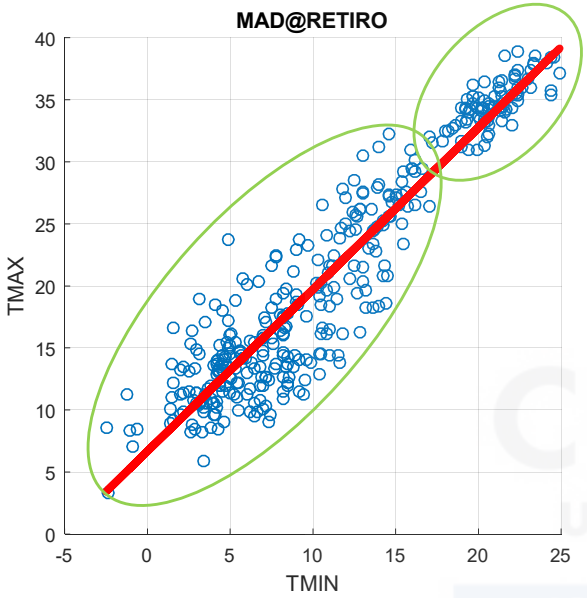






Simple Linear Regression
Illustrative real case

- Estimation of TMAX from TMIN in the Retiro Park




```
Linear regression model:
    TMAX ~ 1 + TMIN

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    _______    ____________

    (Intercept)    6.6735       0.31475     21.202     1.5145e-65
    TMIN           1.3043      0.024365     53.532    1.2589e-174


Number of observations: 366, Error degrees of freedom: 364
Root Mean Squared Error: 3.13
R-squared: 0.887,  Adjusted R-Squared 0.887
F-statistic vs. constant model: 2.87e.03, p-value = 1.26e-174
```

- Very small p-values show that both coefficients are significant
- Cloud suggests possible linear regression, but also a noise variance not constant. Two regions could be considered
  - In summer (high temperatures) the relation between TMIN and TMAX differs. There exist more correlation
- R^2 is reasonable, confirms that the goodness of fit is reasonable

Simple Linear Regression
Illustrative real case

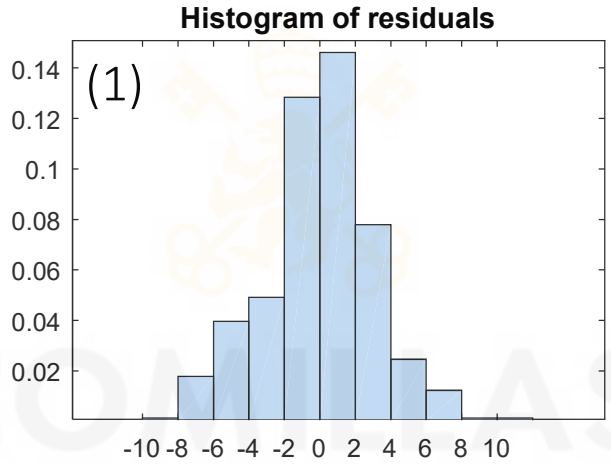
- Estimation of TMAX from TMIN in the Retiro Park
  - Residual analysis

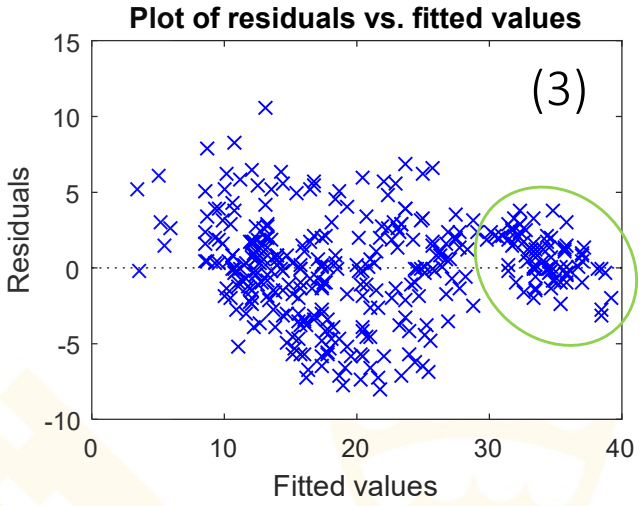
(1 and 2) The normality of the residuals can be questioned by observing the left tail

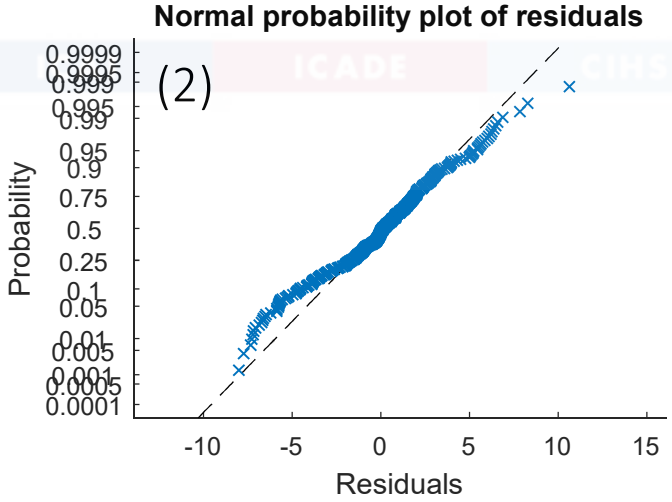
(3) Residual variance doesn't seem constant. For values TMAX>30, it seems to be reduced

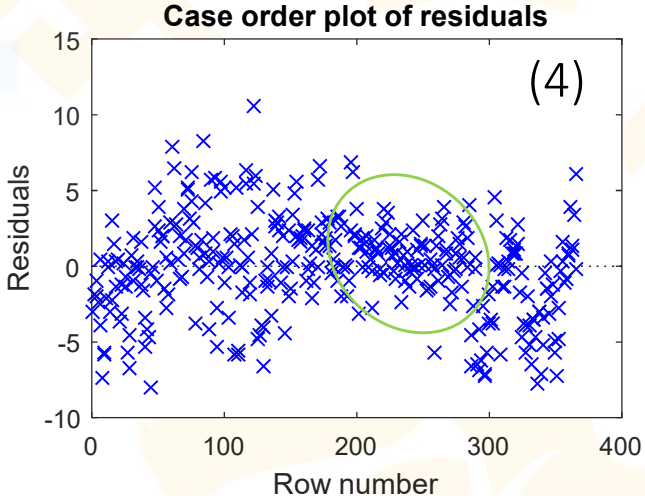
(4) Between days 200 and 270 (summer), residuals are small and positive. They don't seem independent

(4) A certain time dependence among residuals can be observed (autocorrelation)









Simple Linear Regression
Common steps

- Prepare the problem
  - Determine the input and output and data to be used (be careful with the outliers)
- Estimate the model parameters
  - Compute the estimated values, standard errors, and p-values
- Review the model
  - Verify that the p-values are small. If not, simplify the model and repeat the fitting
  - Analyze the R^2 and the RMSE
- Analyze the model residuals
  - Check that the normality of the residuals is acceptable
  - Verify that they are independent
- Use the model
  - Estimate the expected value of the output for a given input value

1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

4

Multiple Linear Regression

Multiple Linear Regression
Problem

- Estimate the maximum daily temperature in the Retiro Park from other maximum daily temperatures of other Spanish weather stations
- Objective
  - Detect possible wrong measures in the Retiro Park station
  - Fill gaps in the historical series of the Retiro

Multiple Linear Regression Statement

- The Multiple Linear Regression model allows to estimate the continuous random variable Y from a set of p input variables

Deterministic component ($p + 1$ coefficients)

Random component

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

HYPERPLANE OF POPULATION REGRESSION
(unknown theoretical relation)

Sample

$(x_1, y_1) \ldots (x_N, y_N)$

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$

$$\epsilon \approx N(0, \sigma^2)$$

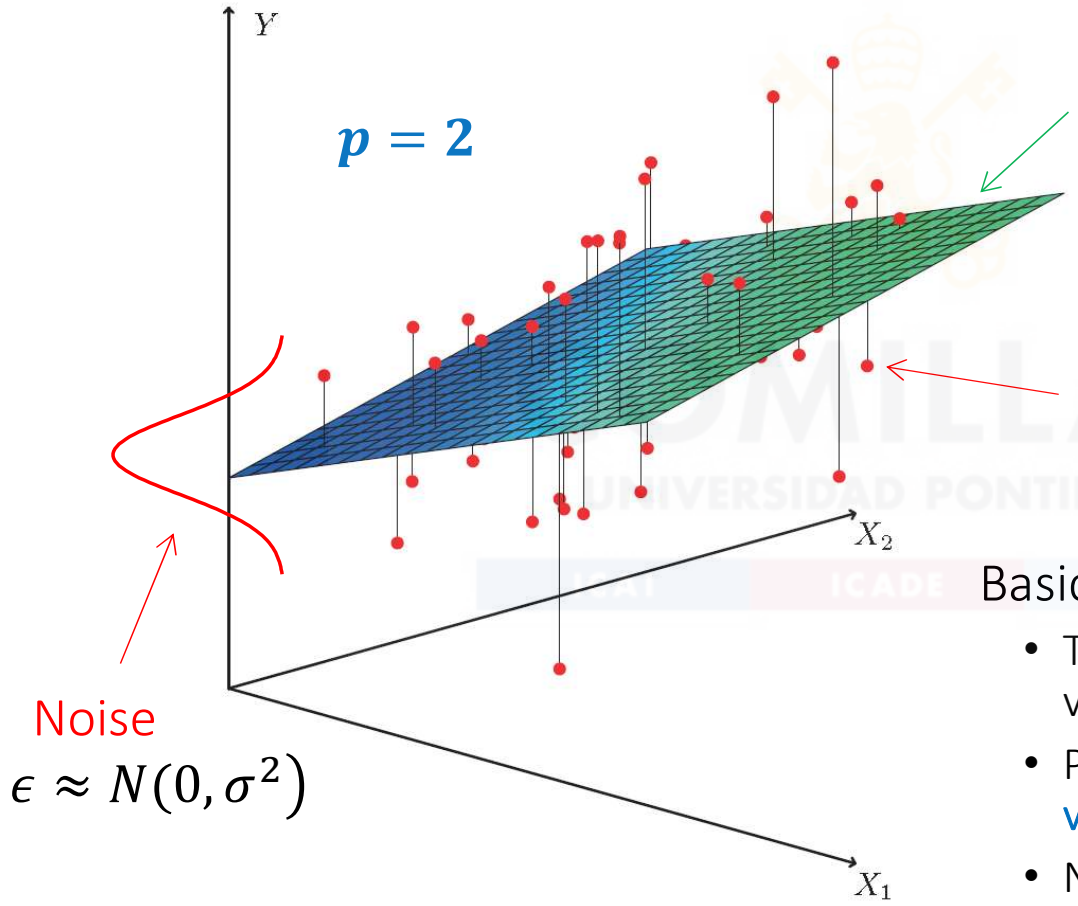
A random variable that considers the output variations with respect to the expected value of the deterministic component (error, noise)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

ESTIMATED REGRESSION LINE

Multiple Linear Regression
Statement

- The Multiple Linear Regression model assumes a theoretical functional relation with these basic characteristics



(Hyper)plane

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Cloud of points

$$(x_1, y_1) \ldots (x_N, y_N)$$

Noise

$$\epsilon \approx N(0, \sigma^2)$$

Basic assumptions

- There is a linear relation between the variables
- Possible noise has zero mean and constant variance
- Noise follows a normal distribution
- Noise has no structure. There is independence between errors

Multiple Linear Regression
Coefficient estimation

- Least squares approach (the common one):
  - The coefficients to determine will be those that minimize the sum of the squared errors

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2$$

  - As in the simple case, RSS is a quadratic function for all the coefficients, with a clear minimum if the problem is well defined (there is a linear relation)
  - Deriving with respect to each parameter and setting equal to zero we obtain the analytical expression of the estimated coefficients minimizing RSS

MATRIX NOTATION

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \Rightarrow \mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

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

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

$\boldsymbol{X} \in R^{n\times(p+1)}$

$\boldsymbol{y} \in R^{n}$

$\boldsymbol{\beta} \in R^{p+1}$

Multiple Linear Regression
Accuracy in the coefficient estimation

- Standard error of the estimators
  - It is possible to determine theoretically the joint distribution of the parameter estimators
  - From it, we can determine confidence intervals and do the hypothesis test
  - Usually, it is expressed in matrix notation

Variance-covariance matrix:

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

Noise variance affects the estimator's accuracy

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Normal multivariate with variance-covariance matrix

Centered in the theoretical value

Multiple Linear Regression
Accuracy in the coefficient estimation

- Standard error of the estimators
  - Example

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

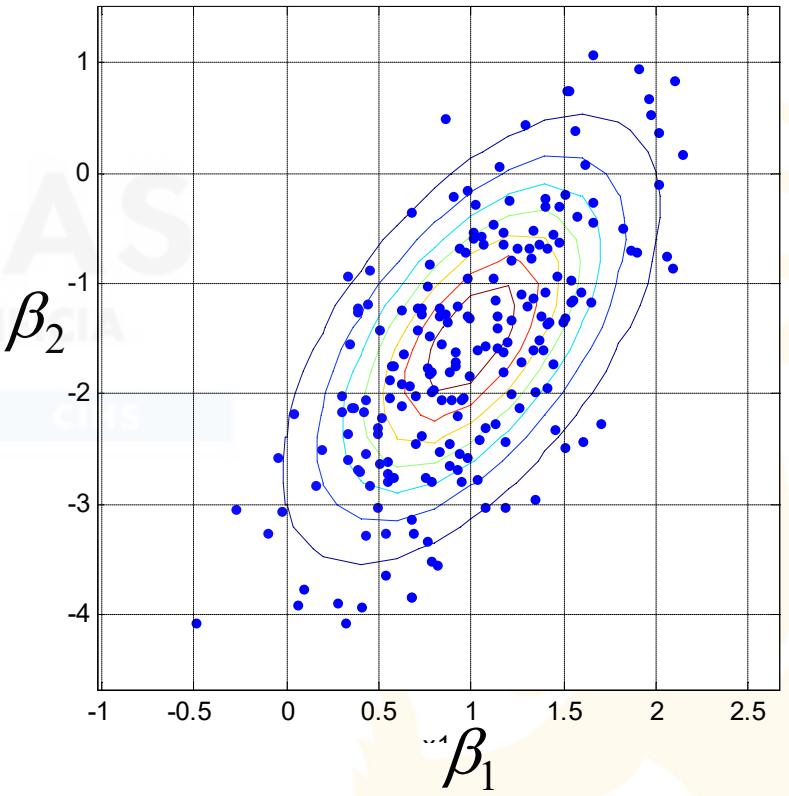
$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1.0 \\ -1.5 \end{pmatrix}$$

  - Variance-covariance matrix

$$\Sigma = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1.0 \end{pmatrix}$$

  - Correlation coefficient

$$\rho = \frac{0.3}{\sqrt{0.25 \cdot 1}} = 0.6$$



Multiple Linear Regression
Goodness of fit

- Same concepts that in simple linear regression
  - Residual variance

$$S^2_{RES} = \frac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}$$

In the simple case p=1

  - Root Mean Squared Error

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

  - Coefficient of determination $\boldsymbol{R^2}$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Insensitive to the number of variables

Multiple Linear Regression
Goodness of fit

- Adjusted coefficient of determination
  - Specific for multivariate regression
  - Improves the coefficient of determination $\boldsymbol{R^2}$ because it includes a penalty associated with the model complexity
    - Fixes the natural decreases that happen in R^2 when increasing the number of input variables due to possible overfitting (over-parametrization of the model)
  - Bounded between 0 (a large dispersion) and 1 (data in a hyperplane)
    - Interpreted as R^2

$$R^2_{AJU} = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}$$

Multiple Linear Regression
Is model significant?

- Is the model a constant?
  - It can be stated as a hypothesis test

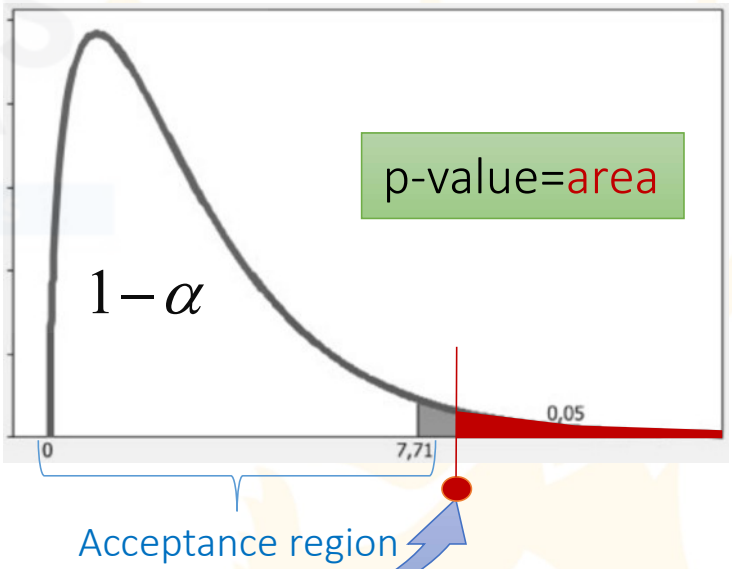
$$H_0: \beta_1 = \cdots = \beta_p = 0$$
$$H_1: \exists i / \beta_i \neq 0 \text{ (at least one)}$$

If all the coefficients (except the constant) can be zero, then the model is useless (there is no linear relation, and no input variable is significant)

  - If the null hypothesis is true, then

$$F = \frac{S^2_{EXP}}{S^2_{RES}} \approx F_{p,n-p-1}$$

The ratio between the explained variance and the non-explained variance (if H_0 is true, variances should be similar)



Value of the sample test statistic

Multiple Linear Regression
Is model significant?

- Can a specific input variable be deleted?
  - If the confidence interval of the associated coefficient contains zero, then it is superfluous in the model
  - It can be stated as a hypothesis test

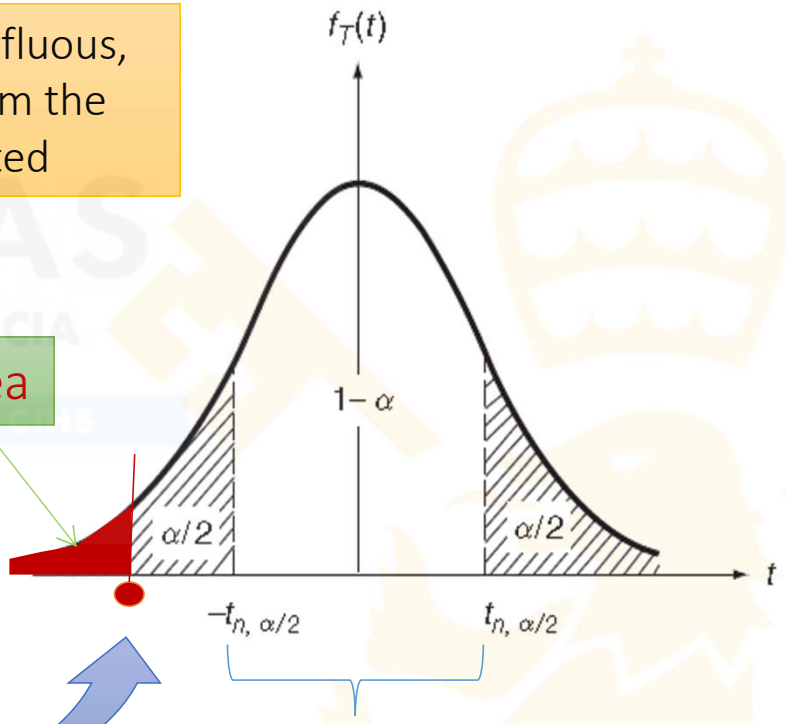
$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

If this parameter is superfluous, it must be eliminated from the model, and the fit repeated

  - If the null hypothesis is true, then

$$\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \approx t_{n-p-1}$$

p-value=2*area



Value of the sample test statistic

Acceptance region

Multiple Linear Regression
Interpretation of the fitted model

- The coefficients show the marginal contribution of each variable to the output (assuming the others do not change)
- Usual information for diagnosis and model interpretation as a table
  - Besides the simple linear regression parameters, the joint $\boldsymbol{F}$ test and the adjusted $\boldsymbol{R^2}$
- Example with two input variables

```
Linear regression model:
    y ~ 1 + x1 + x2
Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)      10.044    0.033416    300.57           0
    x1               1.9354    0.028088    68.904           0
    x2              0.98588    0.018831    52.353    2.32e-288

Number of observations: 1000, Error degrees of freedom: 997
Root Mean Squared Error: 0.877
R-squared: 0.898,  Adjusted R-Squared 0.898
F-statistic vs. constant model: 4.38e.03, p-value = 0
```

Multiple Linear Regression
Residual analysis

- Like the simple linear case
- Once the model is fitted, it is important to check that residuals, what is not explained by the model, satisfy reasonably the basic assumptions
  - Normality: follow a normal distribution with zero mean and constant variance
  - Independence: are independent among them, with no internal clear structure
- Diagnosis of the model or residual analysis
  - Normality plot for the residuals (nonparametric test)
  - Plot the residuals versus the estimated values
  - Independence plot among residuals, considering the different inputs

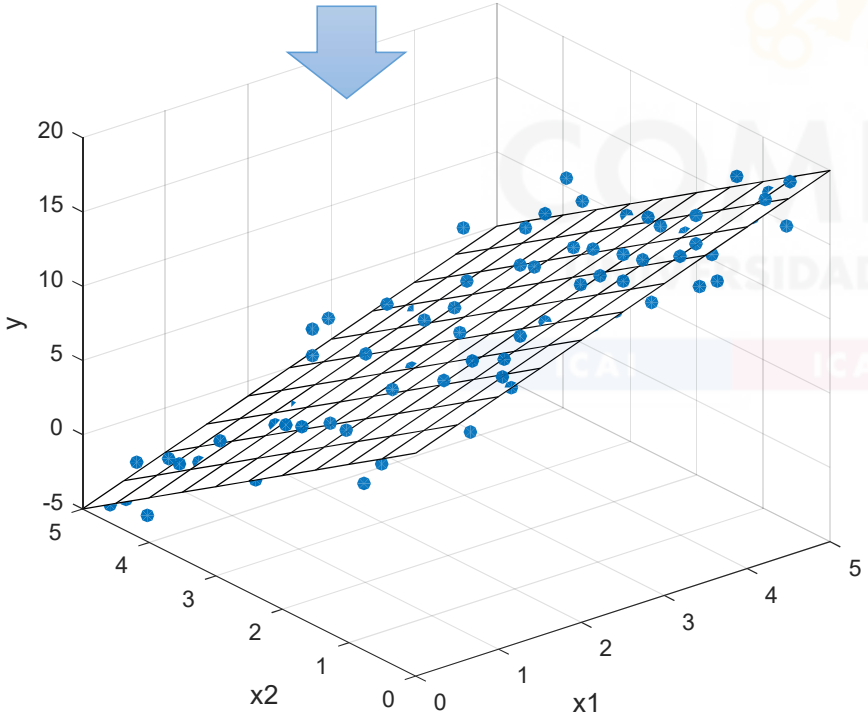
Multiple Linear Regression
Illustrative synthetic cases

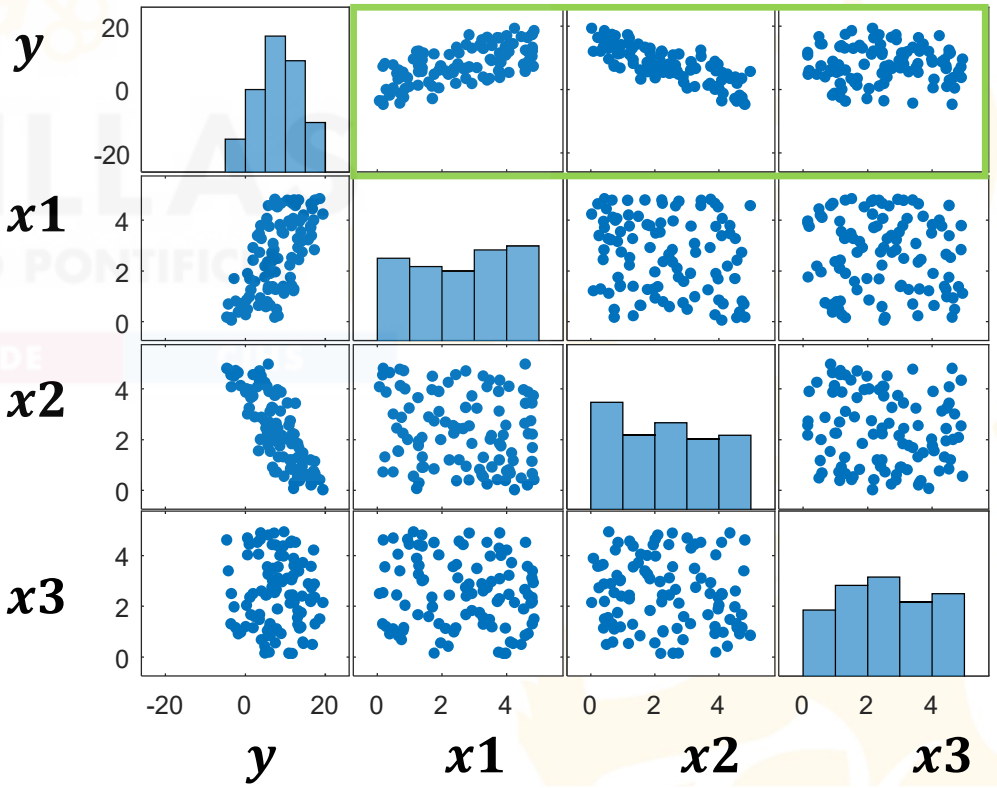
- C7: IDEAL independent input

```
n=100;
x=unifrnd(0,5,n,3); % 3 independent inputs
x1 = x(:,1); x2 = x(:,2); x3 = x(:,3);
% output: linear in x1 and x2 with Gaussian noise
stdnoise = 1.5;
y = 10 + 2*x1 - 3*x2 + normrnd(0,stdnoise,n,1);
```

Correlation matrix (y, x):

	y	$x1$	$x2$	$x3$
y	1.0000	0.6548	-0.8153	-0.0088
$x1$	0.6548	1.0000	-0.1776	-0.0963
$x2$	-0.8153	-0.1776	1.0000	-0.1204
$x3$	-0.0088	-0.0963	-0.1204	1.0000





Multiple Linear Regression
Illustrative synthetic cases

- C7: IDEAL independent input

```
%% Fit linear regression model
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,x3,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      10.017    0.54884     18.251     5.629e-33
    x1                2.031    0.10355     19.615     2.3691e-35
    x2               -2.9114   0.10645    -27.349     4.1698e-47
    x3              -0.19481    0.1109    -1.7565       0.082186


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.49
R-squared: 0.935,  Adjusted R-Squared 0.933
F-statistic vs. constant model: 463, p-value = 6.11e-57
```

Initial model

- Model is significant (overall F test)
- Adjusted R^2 is high, a good fit for the data
- All the model terms are significant, except variable $x3$ that could be superfluous (with a 95%, we should delete it)

A linear relation of y with $x1$ and $x2$ seems to exist

Multiple Linear Regression
Illustrative synthetic cases

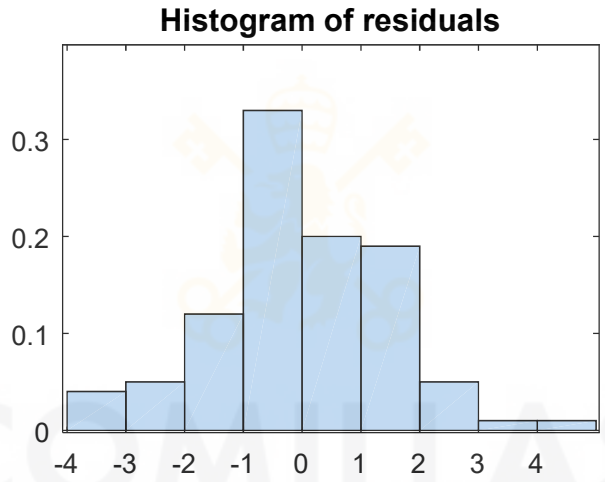
Initial model

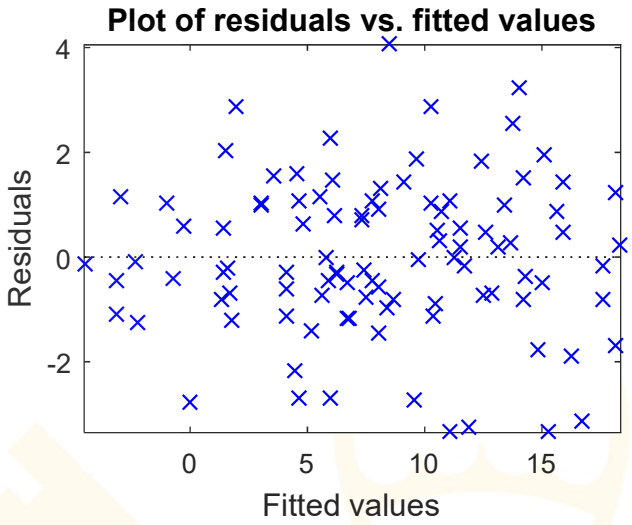
- C7: IDEAL independent input (residuals)

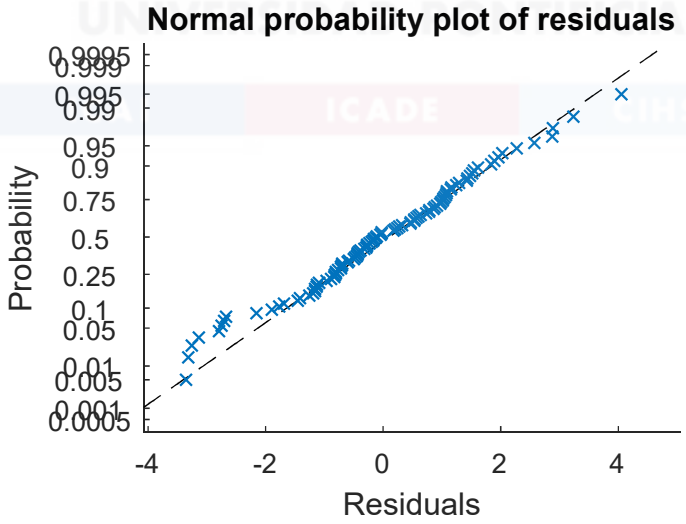
Residuals seem normal, although in the left tail, something seems remarkable but reasonable

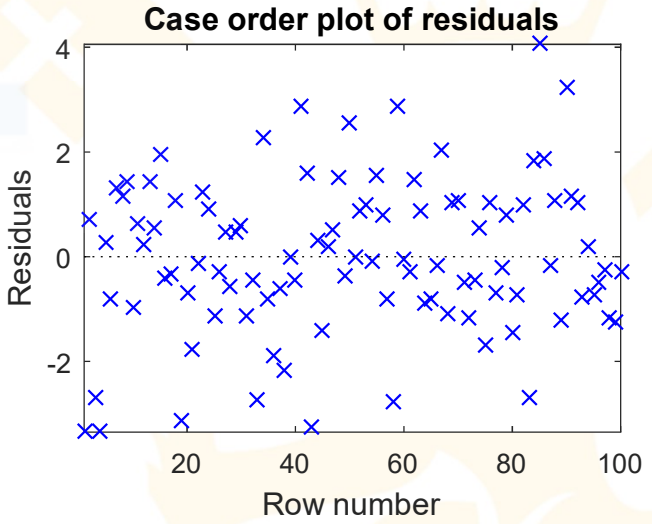
Residuals seem independent of the estimated value

Residual analysis does not question the initial model









Multiple Linear Regression
Illustrative synthetic cases

- C7: IDEAL independent input

```
%% Fit linear regression model
% With x1 and x2 input variables (includes intercept)

tb = table(x1,x2,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

Simplified model

```
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)     9.4061     0.42929     21.911    2.4243e-39
    x1              2.0529     0.10389      19.76    8.8546e-36
    x2             -2.8852     0.10653    -27.084     5.044e-47


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.51
R-squared: 0.933,  Adjusted R-Squared 0.932
F-statistic vs. constant model: 678, p-value = 9.45e-58
```

- The simplified model is significant (overall F test), now all its coefficients are significant
- RMSE is greater than with the initial model. There was some overfitting (the model was more complex than needed)
- Adjusted R^2 is like the previous one (slightly worsens)

Simplified model is better than the initial one

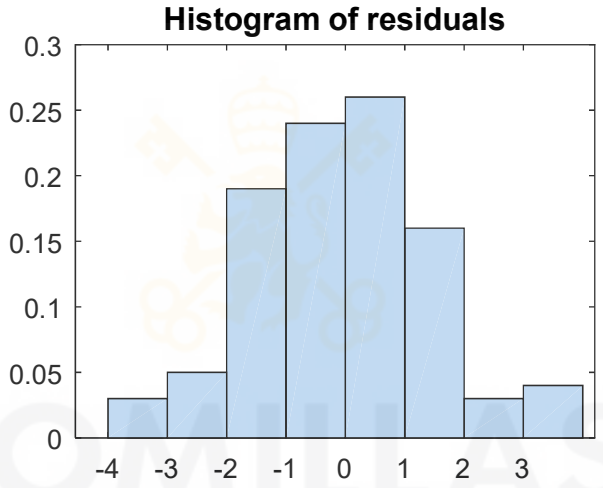
Multiple Linear Regression
Illustrative synthetic cases

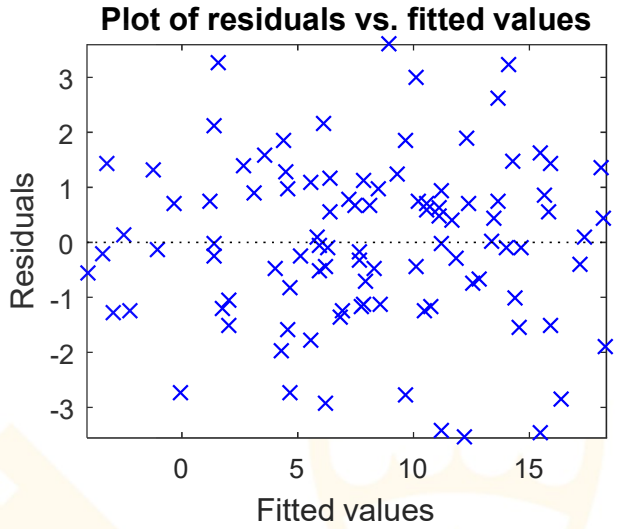
Simplified model

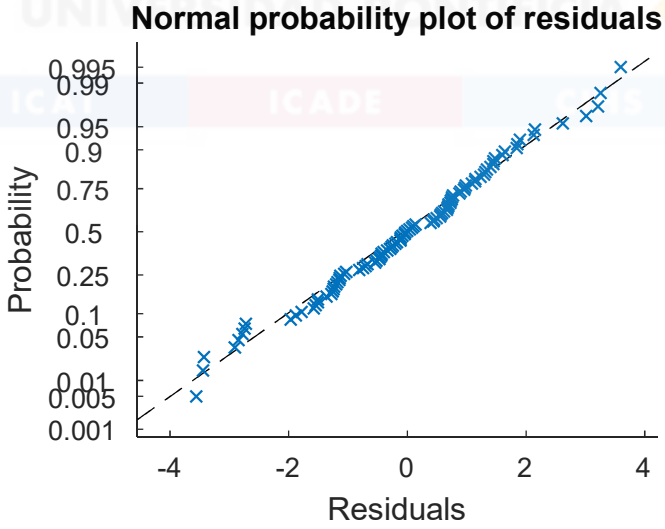
- C7: IDEAL independent input (residuals)

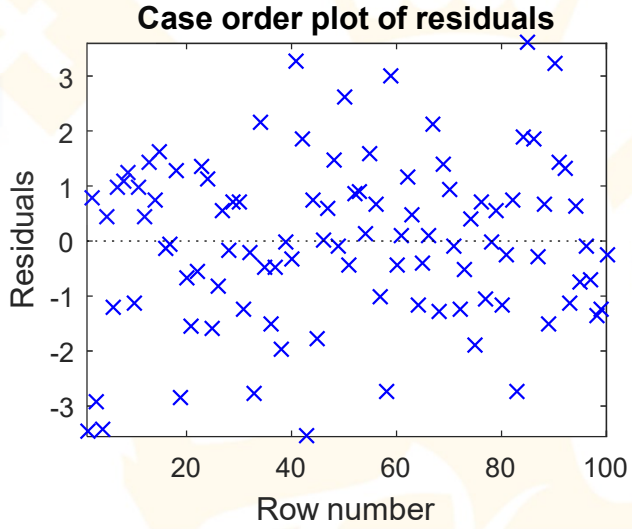
- Residuals seem to be normal
- Residuals seem independent of the estimated value

Residual analysis does not question the simplified model, confirming that it is the best option









Multiple Linear Regression
Illustrative synthetic cases

- C8: IDEAL strongly dependent input

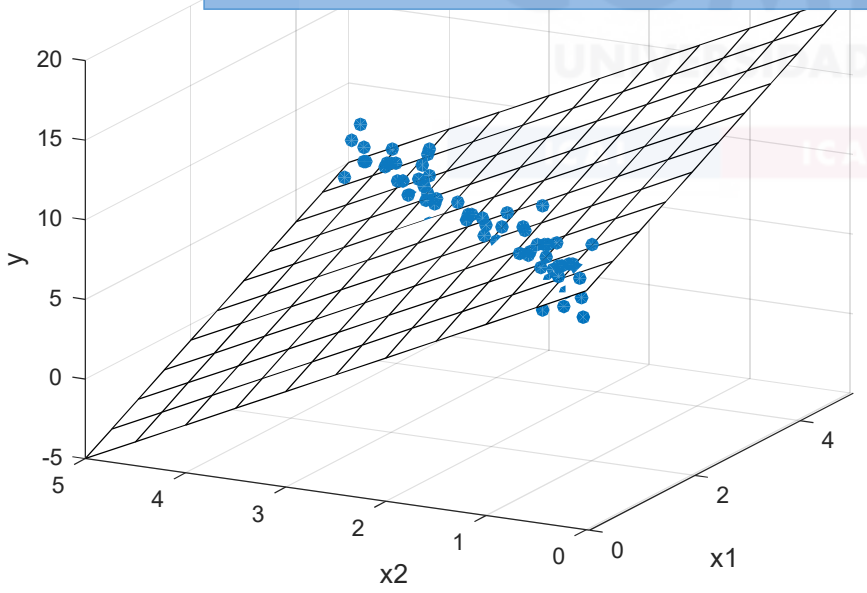
```
n=100;
x=unifrnd(0,5,n,2); % 3 independent inputs, positive
x1 = x(:,1); x3 = x(:,2);
x2 = x1 + normrnd(0,0.1,n,1); % correlated
% output: linear in x1 and x2 with Gaussian noise
stdnoise = 1.5;
y = 10 + 2*x1 - 3*x2 + normrnd(0,stdnoise,n,1);
```

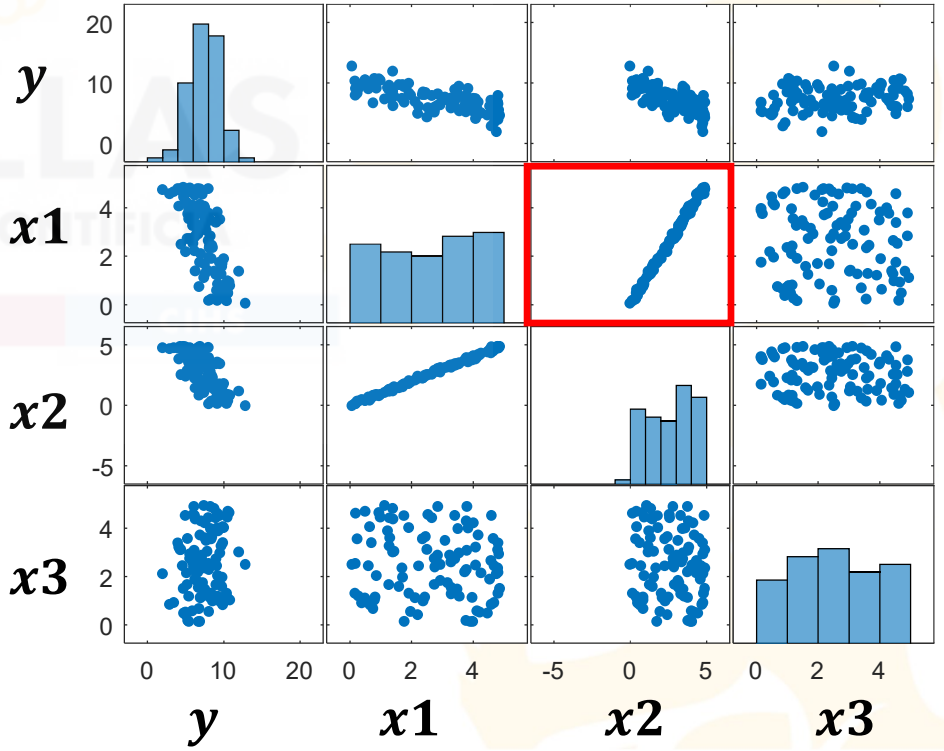
There is an inverse relation of y with $x1$ and y with $x2$

Correlation matrix $(y\ x)$:

	y	$x1$	$x2$	$x3$
y	1.0000	-0.7395	-0.7555	0.1769
$x1$	-0.7395	1.0000	0.9977	-0.0963
$x2$	-0.7555	0.9977	1.0000	-0.1088
$x3$	0.1769	-0.0963	-0.1088	1.0000

However, in the theoretical equation $x1$ is adding and $x2$ subtracting!!





Multiple Linear Regression
Illustrative synthetic cases

- C8: IDEAL strongly dependent input

```
%% Fit linear regression model
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,x3,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    ______    __________

    (Intercept)      9.7913     0.37239     26.293    1.1769e-45
    x1               4.1467      1.3203     3.1406     0.0022414
    x2               -5.152      1.3161    -3.9145    0.00016938
    x3             0.088221    0.096479    0.91441       0.36279


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.29
R-squared: 0.619,  Adjusted R-Squared 0.607
F-statistic vs. constant model: 52, p-value = 4.78e-20
```

- The model is significant (overall F test)
- Adjusted R^2 is not high; data fitting is normal
- Variable $x3$ is superfluous
- Now, as $x1$ and $x2$ are strongly correlated, coefficients are badly estimated (p-values are not so small)

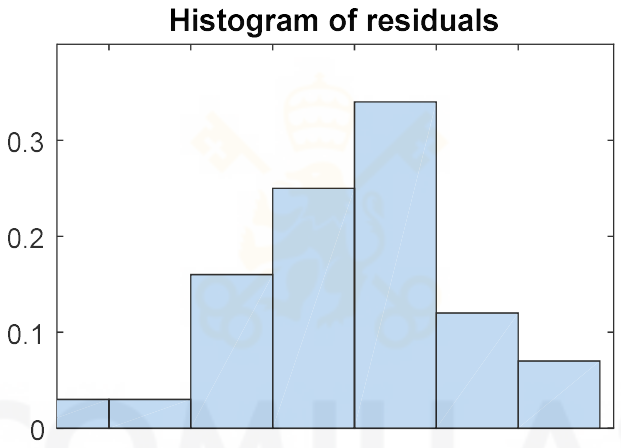
It seems to exist collinearity problems

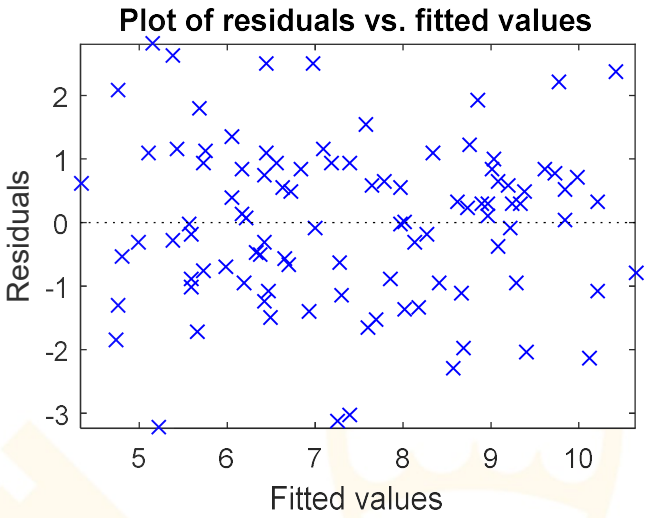
Multiple Linear Regression
Illustrative synthetic cases

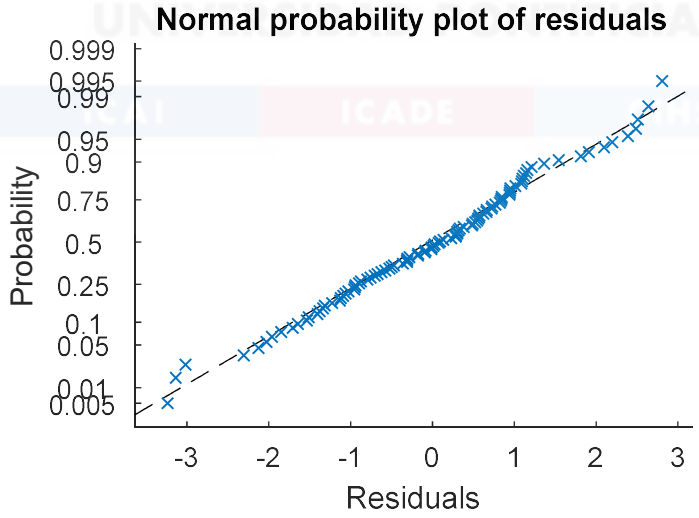
- C8: IDEAL strongly dependent input (residuals)

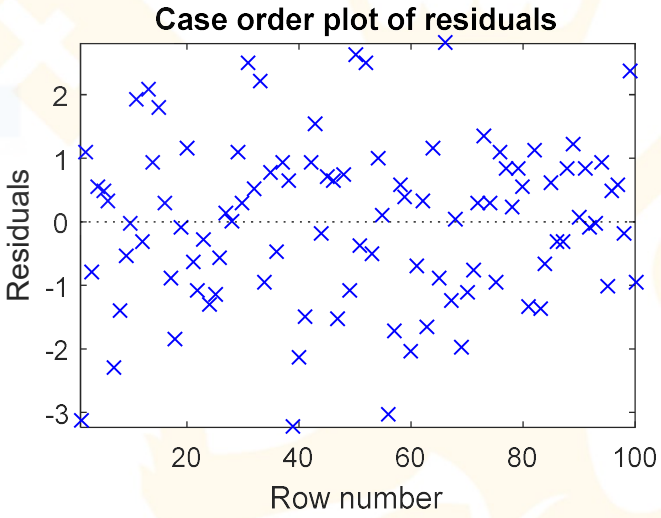
- Residuals seem to be normal
- Residuals seem independent of the estimated value

Residual analysis doesn't question the model







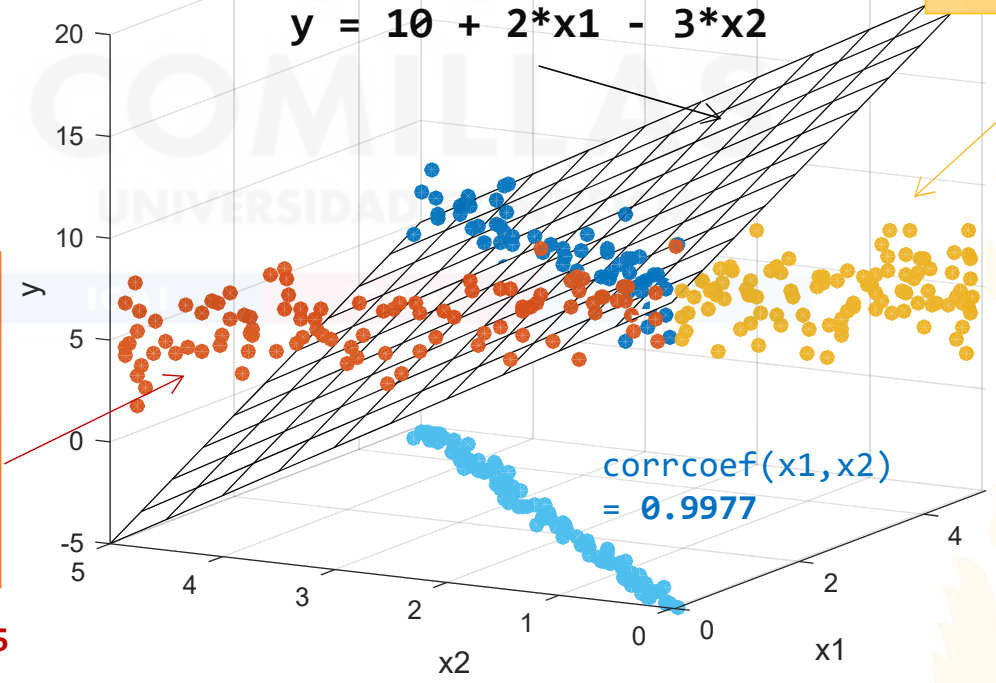


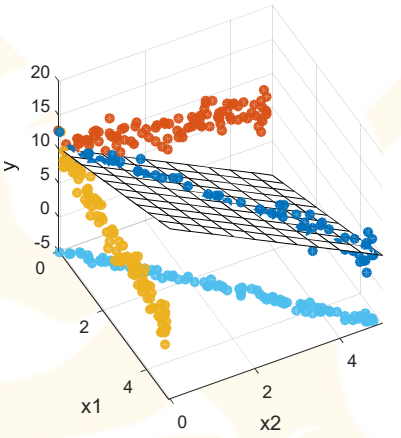
Multiple Linear Regression
Illustrative synthetic cases

- C8: IDEAL strongly dependent input
  - It is difficult to see the plane given that the sample is concentrated in a specific region of the input space due to the strong correlation among the input variables (although the variables change marginally between 0 and 5)

Data projection in the **plane $(x1, y)$** indicates that if $x1$ increases, output y decreases (inverse relation), shown by the linear correlation coefficient

corrcoef(x1,y) = **-0.7395**





Data projection in the **plane $(x2, y)$** indicates that if $x2$ increases, output y decreases (inverse relation), shown by the linear correlation coefficient

corrcoef(x2,y) = **-0.755**

Multiple Linear Regression
Collinearity

- It appears when there is a strong correlation among input variables

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

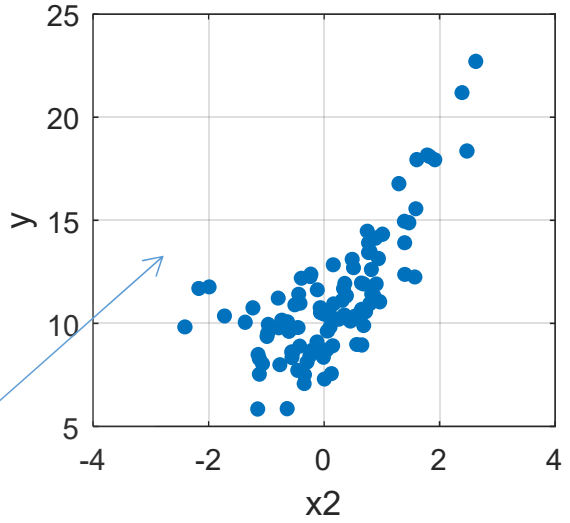
Low determination of the problem causes an inaccurate computation of the inverse

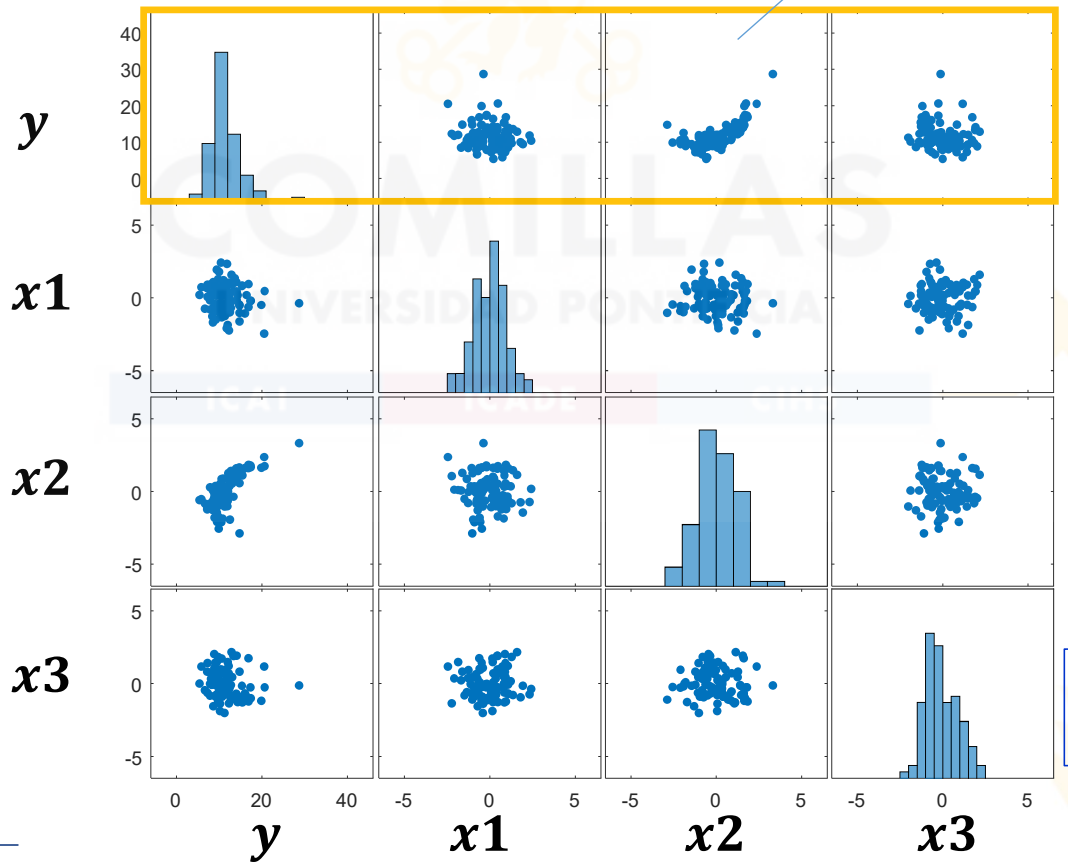
- Consequences
  - Model coefficient estimation can be arbitrary (it is not clear which variable explains the output)
  - It is an unstable model that can behave badly with new data
- How is it detected?
  - F test indicates that the model is clearly explanatory (with R^2 reasonable and small RMSE)
  - t individual tests do not show as clearly that, observed by the overall F test

Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data

```
n=100;
x=normrnd(0,1,n,3); % 3 independent inputs
x1 = x(:,1); x2 = x(:,2); x3 = x(:,3);
% output: quadratic in x2 with Gaussian noise
stdnoise = 1.5;
y = 10 + 2*x2 + 1*x2.^2 + normrnd(0,stdnoise,n,1);
```

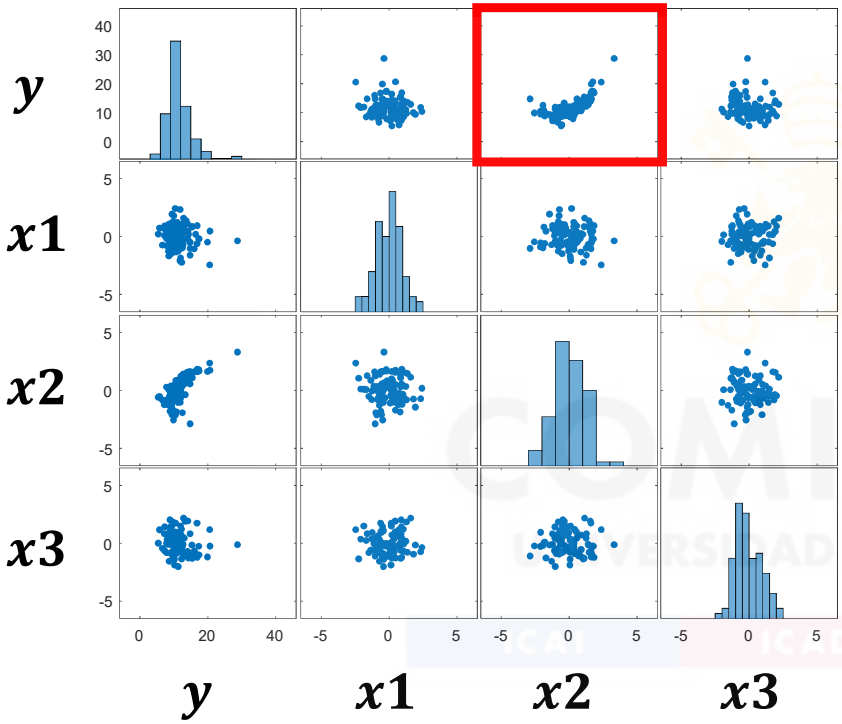





```
figure;
plotmatrix([y x]);
```

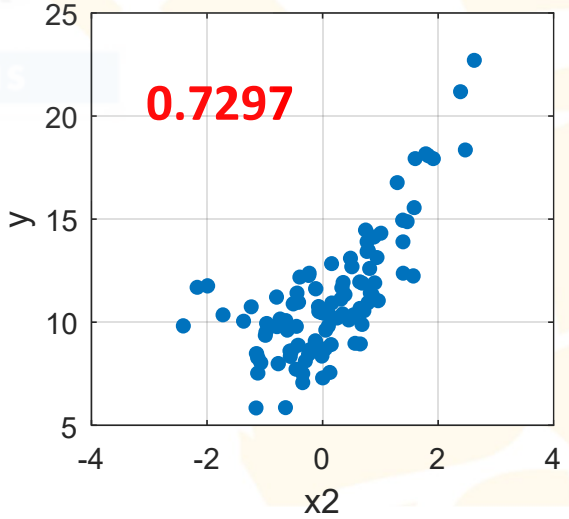
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data



Correlation matrix $(y\ x)$:

	y	$x1$	$x2$	$x3$
y	1.0000	-0.0119	0.7297	-0.0313
$x1$	-0.0119	1.0000	0.0436	-0.0499
$x2$	0.7297	0.0436	1.0000	-0.0513
$x3$	-0.0313	-0.0499	-0.0513	1.0000



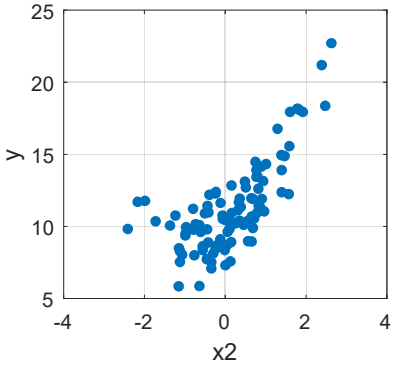
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data

```
%% Fit linear regression model
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,x3,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)      10.906    0.17901     60.924    1.5554e-78
    x1             -0.055654   0.18684    -0.29787      0.76645
    x2               2.1012    0.19866     10.577    8.6097e-18
    x3             -0.023184   0.16664    -0.13912      0.88964


Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.74
R-squared: 0.54,  Adjusted R-Squared 0.525
F-statistic vs. constant model: 37.5, p-value = 3.89e-16
```

Initial model



- Model is significant (overall F test)
- There are no significant coefficients (associated with $x1$ and $x3$), these variables must be eliminated from the model

There seems to exist a linear relation of y with $x2$

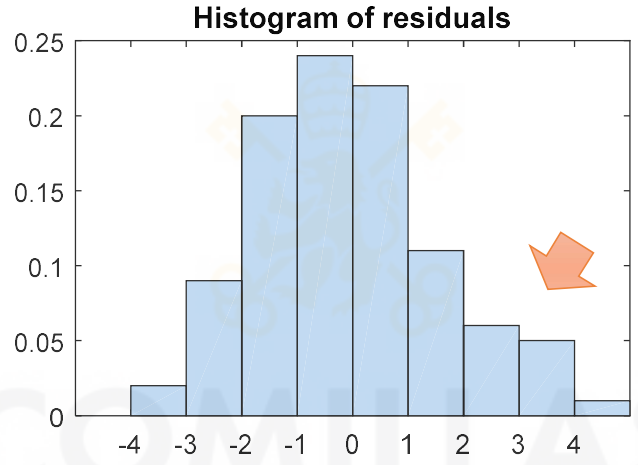
Multiple Linear Regression
Illustrative synthetic cases

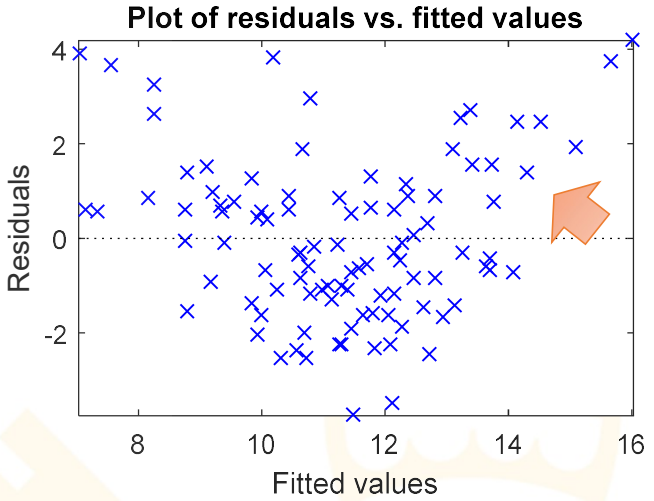
Initial model

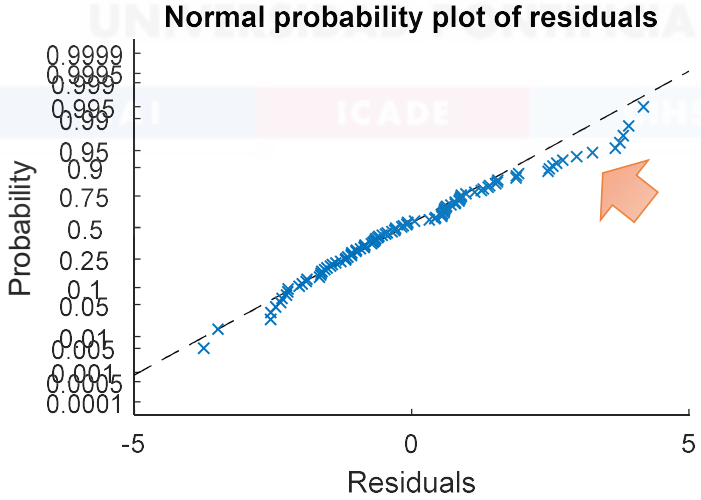
- C9: Quadratic with few data (residuals)

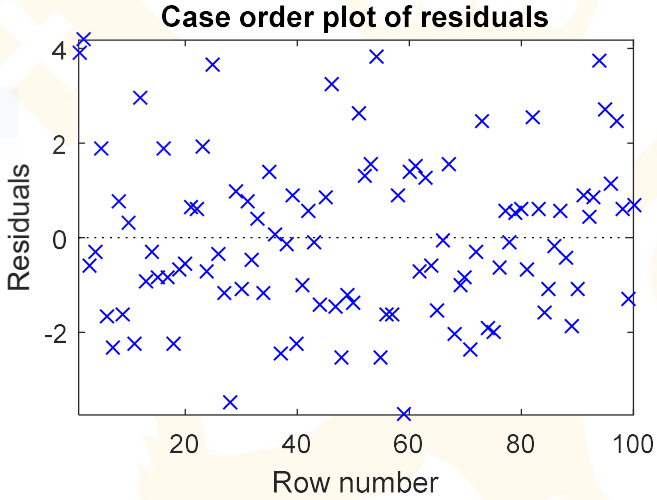
- Residuals are far from normality, especially in the right tail (large positive errors)
- Residuals are not independent of the estimated value

Residual analysis suggests that the model could be badly specified









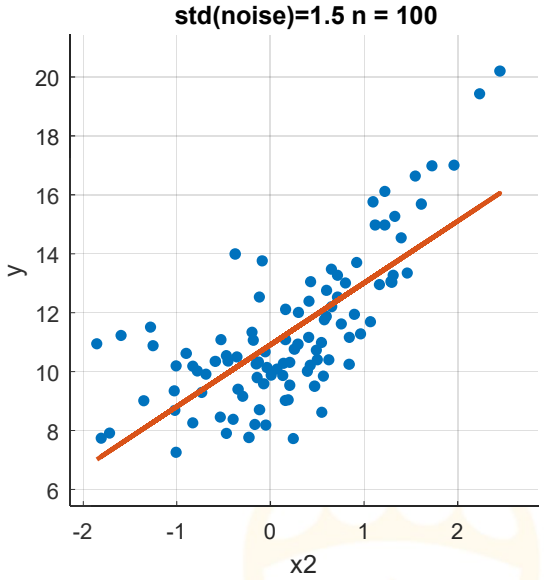
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data

```
%% Fit linear regression model
% With ALL the available input variables (includes intercept)

tb = table(x2,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

```
Linear regression model:
    y ~ 1 + x2

Estimated Coefficients:
                   Estimate      SE       tStat      pValue
                   ________    _______    ______    __________

    (Intercept)     10.91      0.17624    61.903    2.2657e-80
    x2              2.0984     0.19593     10.71    3.5098e-18


Number of observations: 100, Error degrees of freedom: 98
Root Mean Squared Error: 1.72
R-squared: 0.539,  Adjusted R-Squared 0.535
F-statistic vs. constant model: 115, p-value = 3.51e-18
```

Simplified model



- All the coefficients are significant
- Adjusted R^2 suggests the model is better than the former one (similar RMSE but with less variables)
- Point cloud questions the possible linear relation

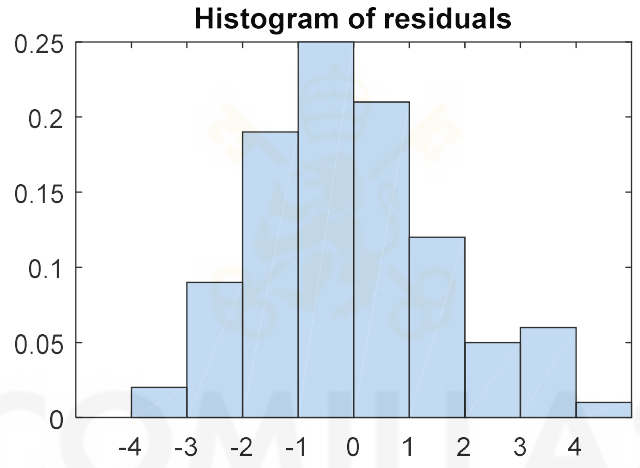
Multiple Linear Regression
Illustrative synthetic cases

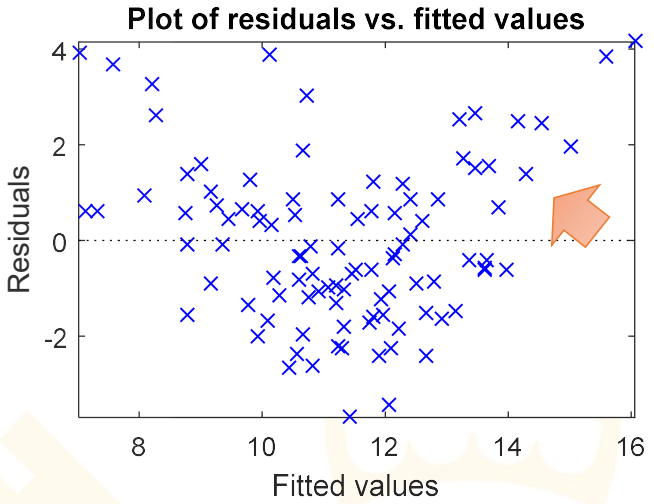
- C9: Quadratic with few data (residuals)

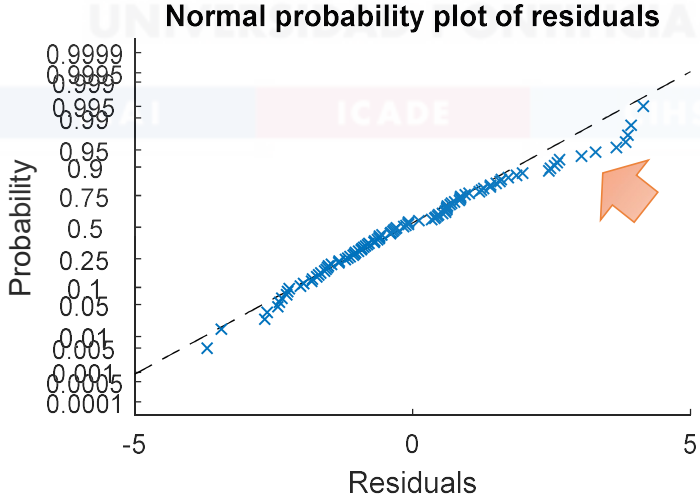
Simplified model

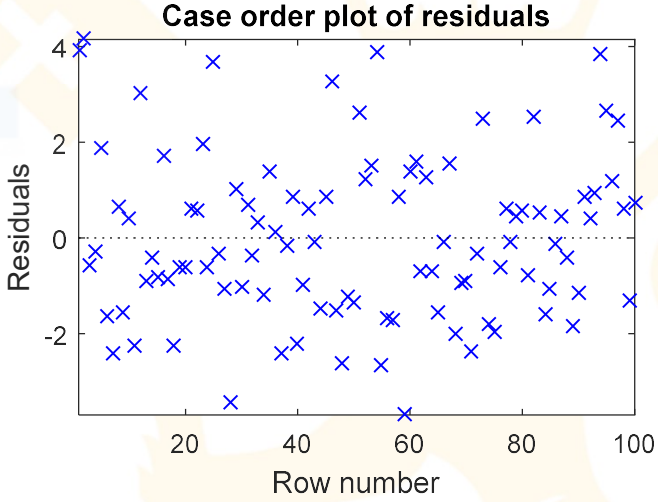
- It doesn't change. It is very similar to the not-pruned model
- Residuals seem to be far from normality, and residual variance doesn't seem constant

Residual analysis suggests that the simplified model could be badly specified









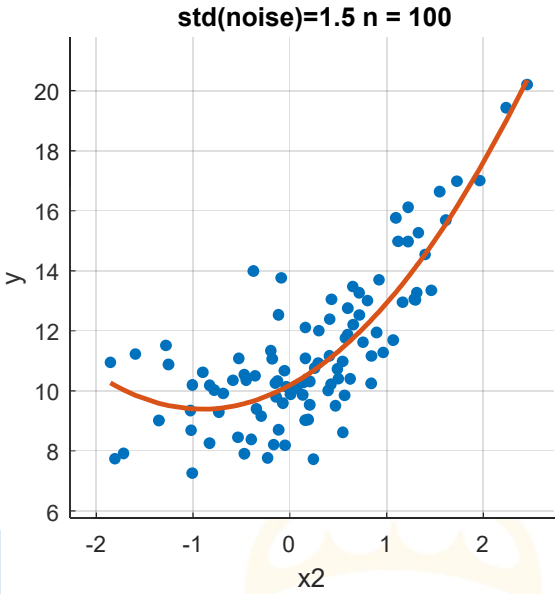
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data

```
%% Fit linear regression model
% Only with variable x2 (quadratic, includes intercept)

x2_2=x2.^2;

model = fitlm(table(x2,x2_2, y),'linear');
% model = fitlm([x2,y,'quadratic') produce the same result
disp(model);
```

```
Linear regression model:
    y ~ 1 + x2 + x2_2

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     10.187      0.17065    59.699    2.7381e-78
    x2              1.7627      0.16289    10.822    2.2777e-18
    x2_2            0.9727      0.12958    7.5067    2.9537e-11


Number of observations: 100, Error degrees of freedom: 97
Root Mean Squared Error: 1.38
R-squared: 0.709,  Adjusted R-Squared 0.703
F-statistic vs. constant model: 118, p-value = 1.07e-26
```

Improved simplified model



- All the coefficients are significant
- Adjusted R^2 has clearly improved with respect to the former model (RMSE is lower)

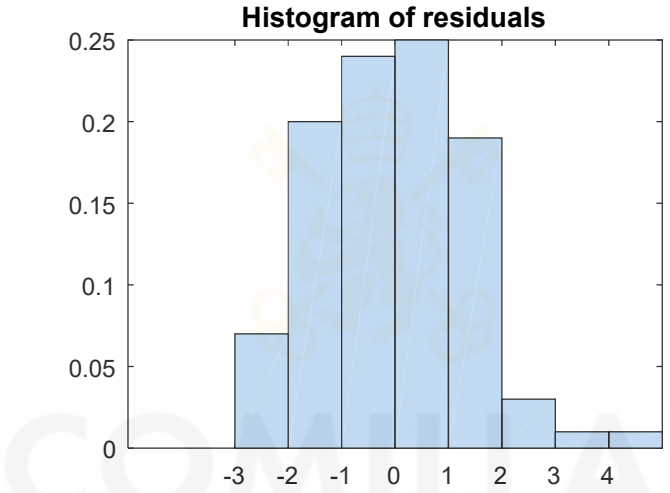
Multiple Linear Regression
Illustrative synthetic cases

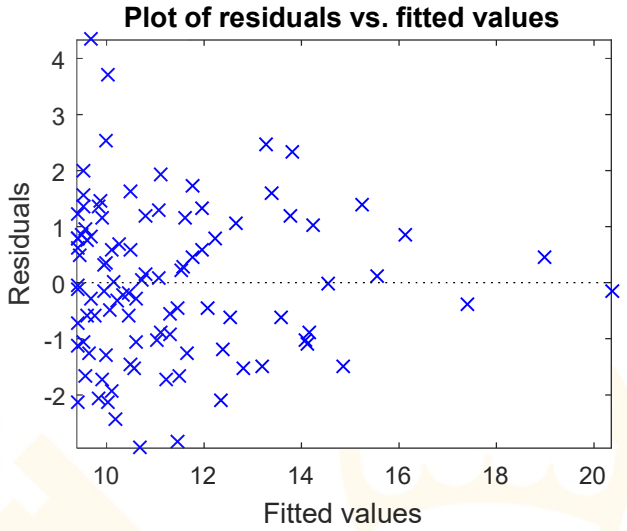
- C9: Quadratic with few data (residuals)

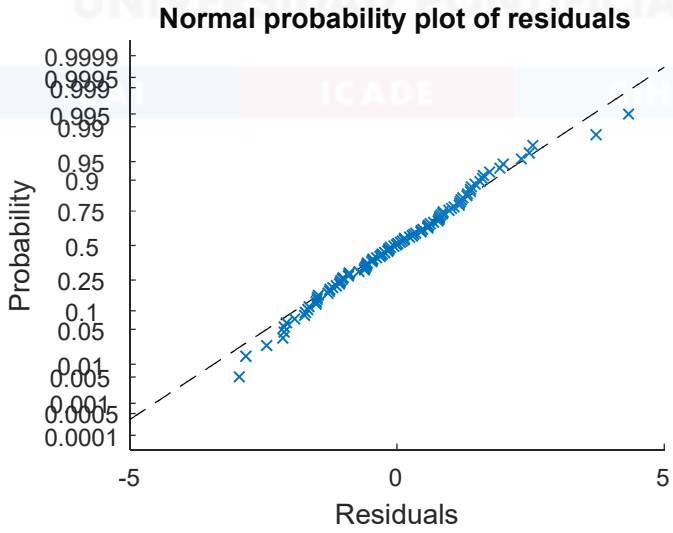
Improved simplified model

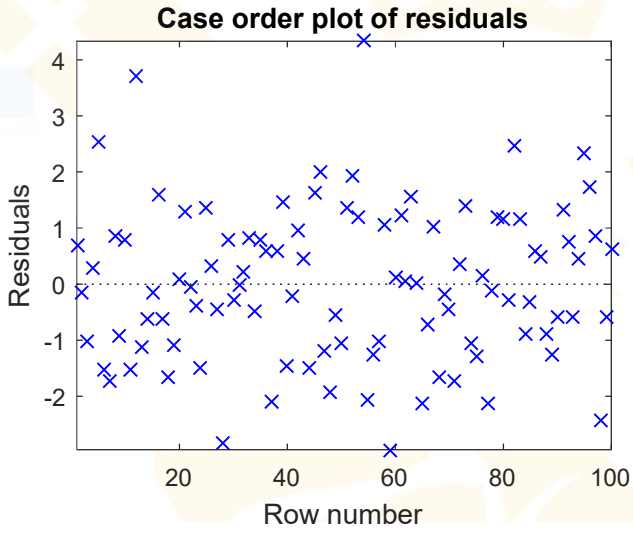
- They do not contradict the normality and independence assumption

Residual analysis suggests that the improved simplified model is reasonable









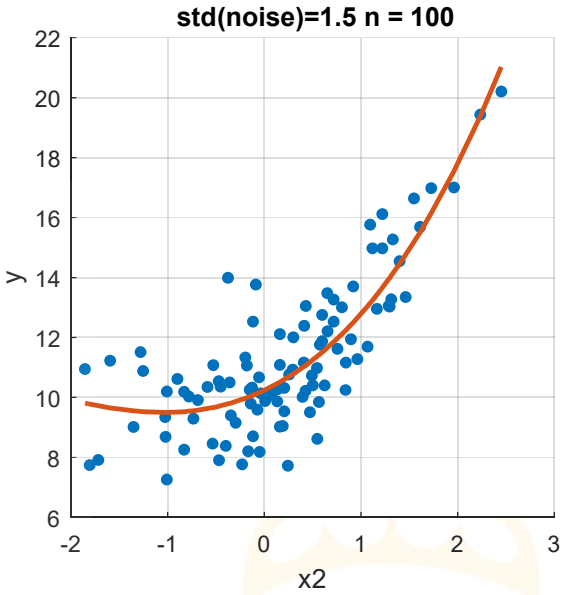
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with few data

```
%% Fit linear regression model
% Only with variable x2 (cubic, includes intercept)

x2_2=x2.^2;
x2_3=x2.^3;  % new term, cubic

model = fitlm(table(x2,x2_2,x2_3, y),'linear');
disp(model);
```

```
Linear regression model:
    y ~ 1 + x2 + x2_2 + x2_3

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ______    __________

    (Intercept)      10.241    0.17865    57.325    4.6085e-76
    x2               1.5375    0.27541    5.5827    2.195e-07
    x2_2            0.90382    0.14628    6.1786    1.559e-08
    x2_3            0.10717    0.10568     1.014      0.31311

Number of observations: 100, Error degrees of freedom: 96
Root Mean Squared Error: 1.38
R-squared: 0.712,  Adjusted R-Squared 0.703
F-statistic vs. constant model: 79, p-value = 7.95e-26
```

Over-parametrized simplified model



- The new cubic term $x2\_3$ is not significant, it is superfluous
- Adjusted R^2 and RMSE have not changed with respect to the improved simplified model

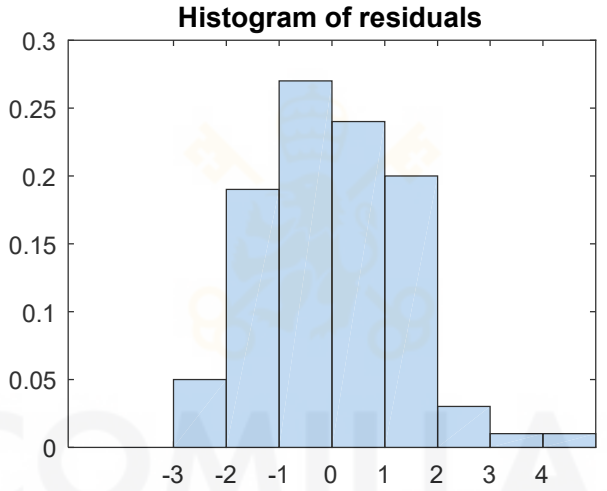
Multiple Linear Regression
Illustrative synthetic cases

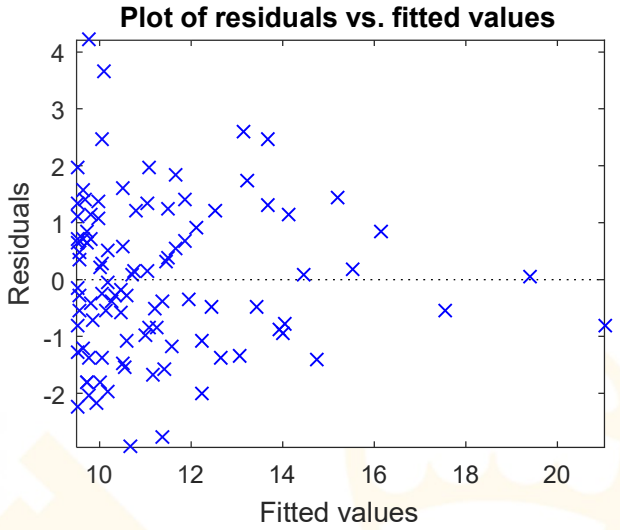
- C9: Quadratic with few data (residuals)

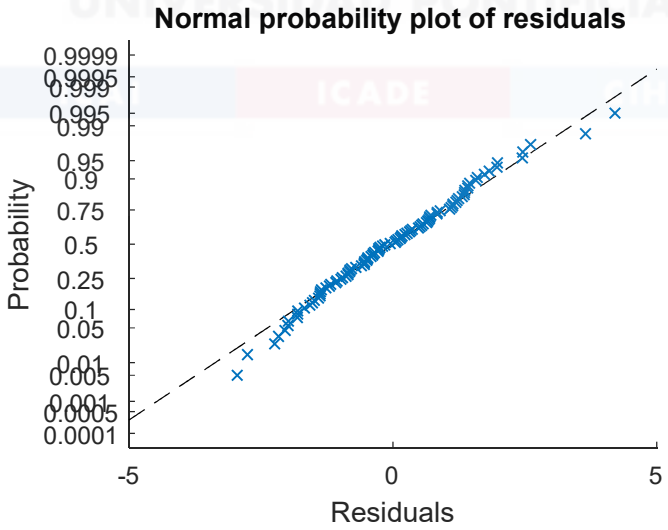
Over-parametrized simplified model

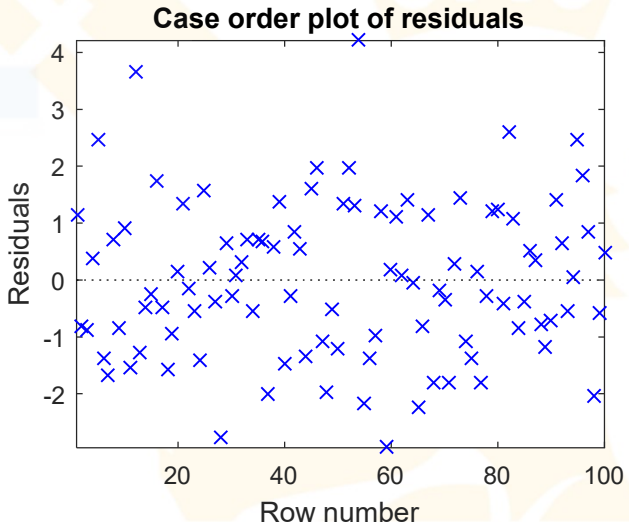
- They do not contradict the normality and independence assumption

Residual analysis suggests that the over-parametrized simplified model is reasonable





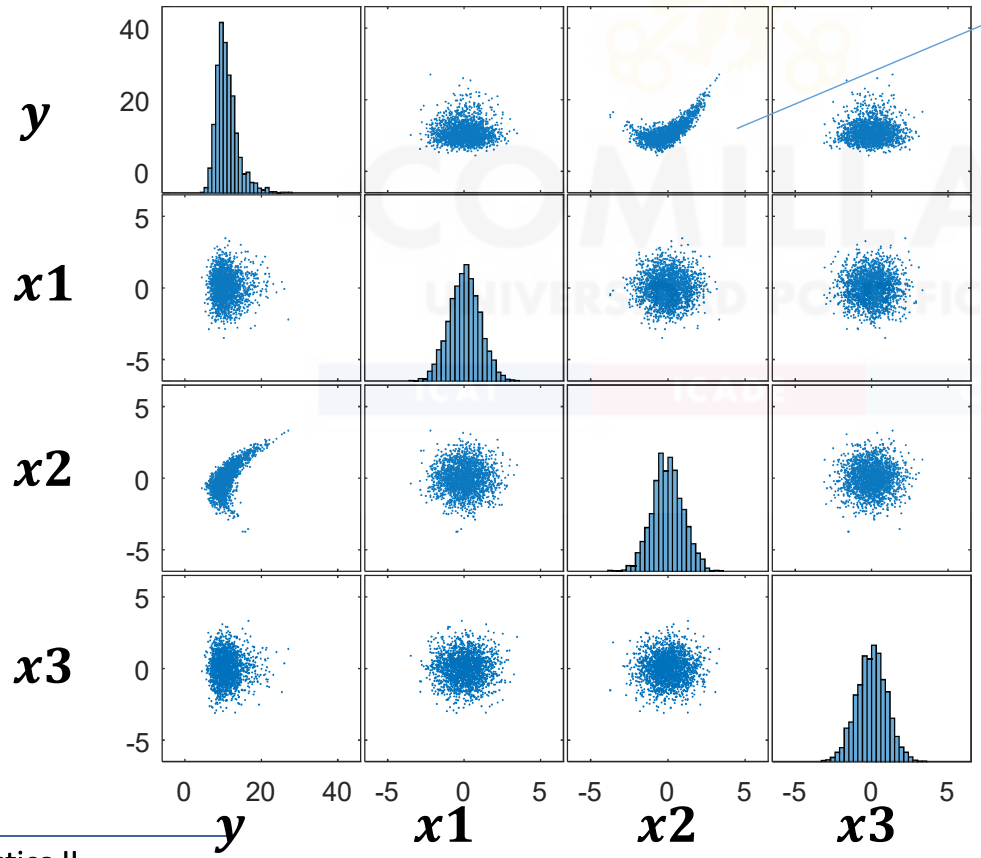


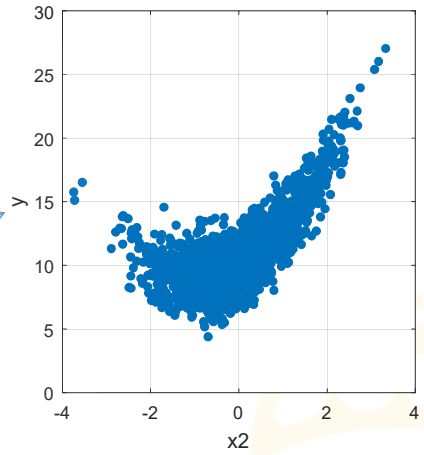


Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with small noise and many data

```
n=2000;
x=normrnd(0,1,n,3); % 3 independent inputs
x1 = x(:,1); x2 = x(:,2); x3 = x(:,3);
% output: quadratic in x2 with Gaussian noise
stdnoise = 1.5;
y = 10 + 2*x2 + 1*x2.^2 + normrnd(0,stdnoise,n,1);
```





Correlation matrix $(y\ x)$:

1.0000	0.0103	0.6921	0.0126
0.0103	1.0000	0.0150	0.0029
0.6921	0.0150	1.0000	0.0198
0.0126	0.0029	0.0198	1.0000

```
figure;
plotmatrix([y x]);
```

Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with small noise and many data

```
%% Fit linear regression model
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,x3,y); % For simplicity as a table
model = fitlm(tb); % Fit model, output is the last column
disp(model);
```

Initial model

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3

Estimated Coefficients:
                    Estimate        SE         tStat        pValue
                   __________   ________   __________   __________

    (Intercept)         10.962   0.045501       240.91            0
    x1             -0.00025934    0.04548   -0.0057024      0.99545
    x2                  1.9615   0.045796       42.832   7.5517e-285
    x3                -0.003235   0.046349    -0.069797      0.94436


Number of observations: 2000, Error degrees of freedom: 1996
Root Mean Squared Error: 2.03
R-squared: 0.479,  Adjusted R-Squared 0.478
F-statistic vs. constant model: 612, p-value = 5.56e-282
```

- Model is significant (overall F test)
- Some coefficients are not significant (associated with $x1$ and $x3$). These variables must be eliminated from the model
- Due to having more data, the accuracy of the estimation improves dramatically
- It is evident that $x1$ and $x3$ are superfluous

There seems to exist a linear relation of y with $x2$

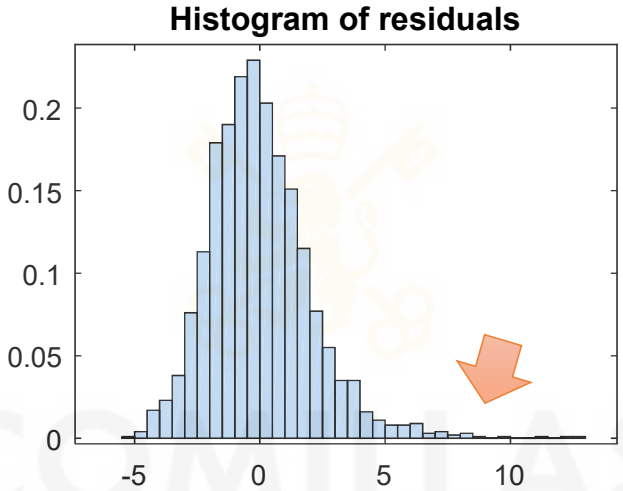
Multiple Linear Regression
Illustrative synthetic cases

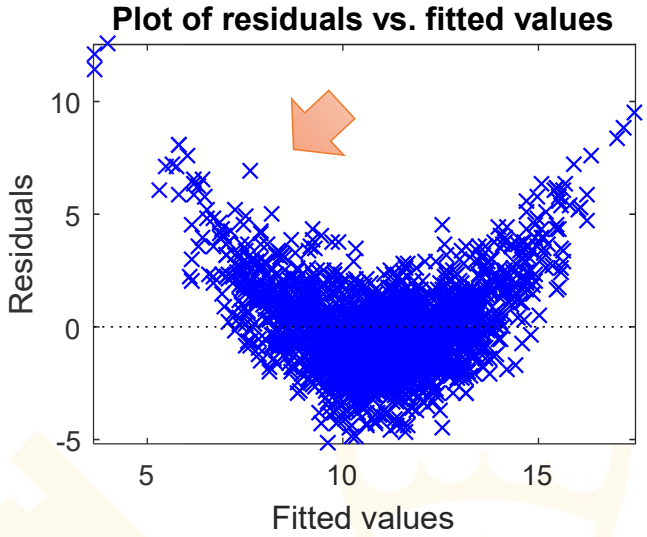
Initial model

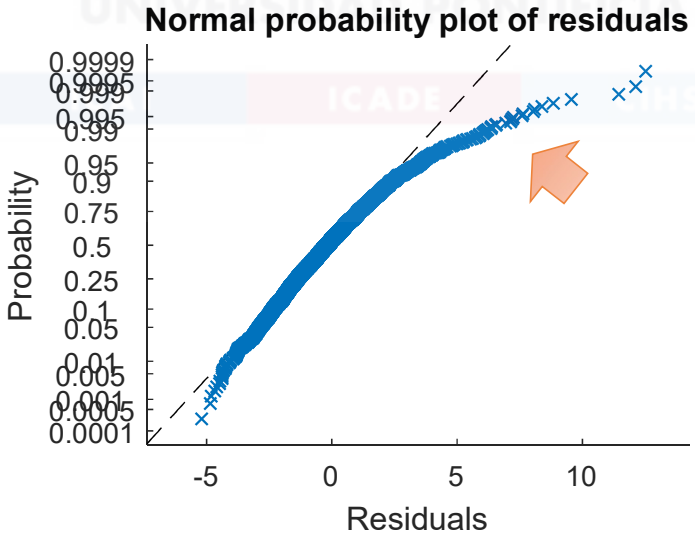
- C9: Quadratic with small noise and many data (residuals)

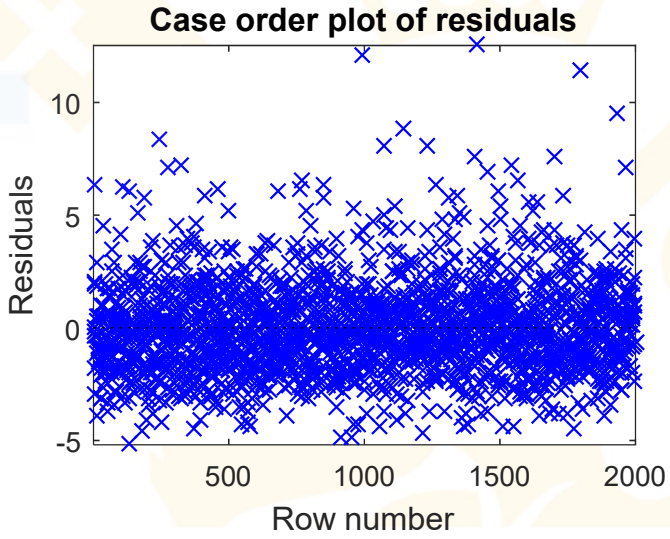
- What could be suspected with few data is evident with many data
- Nonnormality of residuals and residual variance doesn't seem constant

Residual analysis suggests that the initial model has important problems in its specification









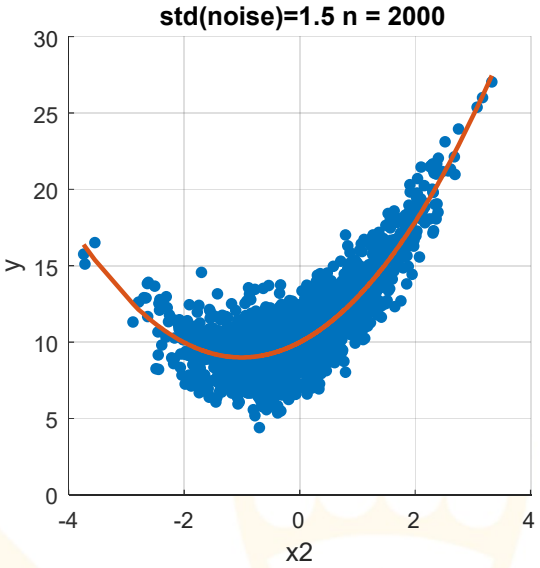
Multiple Linear Regression
Illustrative synthetic cases

- C9: Quadratic with small noise and many data

```
%% Fit linear regression model
% Only with variable x2 (quadratic, includes intercept)

x2_2=x2.^2;

model = fitlm(table(x2,x2_2, y),'linear');
% model = fitlm([x2,y,'quadratic') produce the same result
disp(model);
```

```
Linear regression model:
    y ~ 1 + x2 + x2_2

Estimated Coefficients:
                   Estimate      SE        tStat      pValue
                   ________    ________    ______    __________

    (Intercept)      9.9909    0.040157    248.79           0
    x2               1.9753     0.03321     59.48           0
    x2_2            0.98401    0.023215    42.387    1.4331e-280


Number of observations: 2000, Error degrees of freedom: 1997
Root Mean Squared Error: 1.48
R-squared: 0.726,  Adjusted R-Squared 0.725
F-statistic vs. constant model: 2.64e.03, p-value = 0
```

Improved simplified model



- All the coefficients are significant
- Adjusted R^2 has clearly improved with respect to the initial model (RMSE is lower)
- This model is the better one

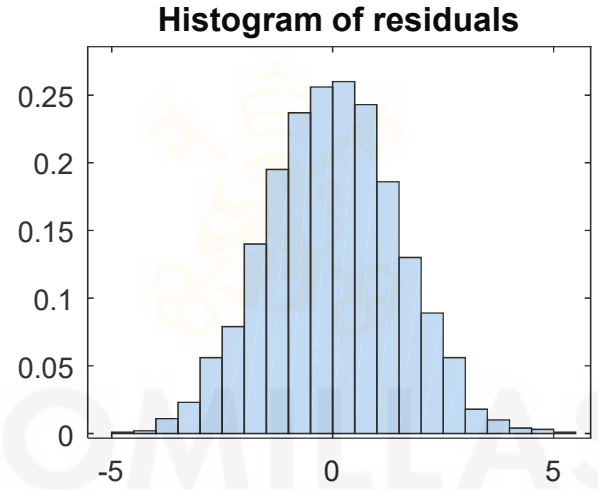
Multiple Linear Regression
Illustrative synthetic cases

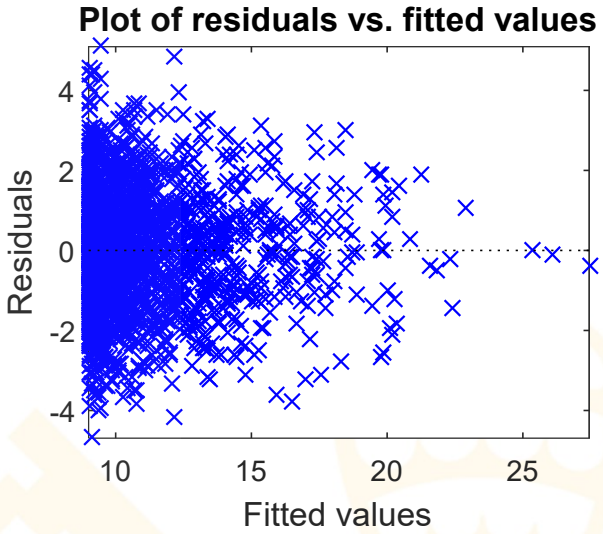
Improved simplified model

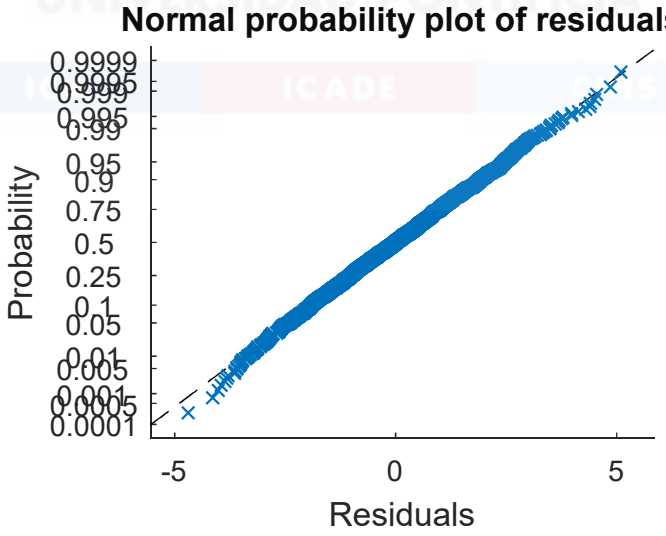
- C9: Quadratic with small noise and many data (residuals)

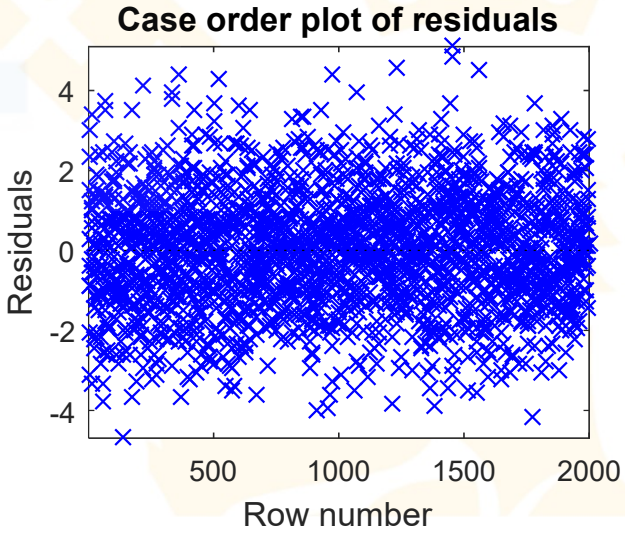
- What could be suspected with few data is evident with many data
- Normality and independence of residuals with this model can't be questioned

Residual analysis suggests that the improved simplified model is quite reasonable









1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

5

Other considerations

Other considerations in regression
Qualitative input variables

- In the regression model, we have assumed that inputs are quantitative
- In practice, often some inputs are qualitative
- Approach
  - Create a set of new dummy variables that represent the same information as the qualitative variable
  - Fit the regression model using the dummy variables instead of the qualitative one
- A dummy variable is an indicator variable that has two possible numerical values (typically 0 and 1)
- Two different cases
  - Input variables with two possible values
  - Input variables with more than two possible values

Other considerations in regression
Qualitative input variable with 2 levels

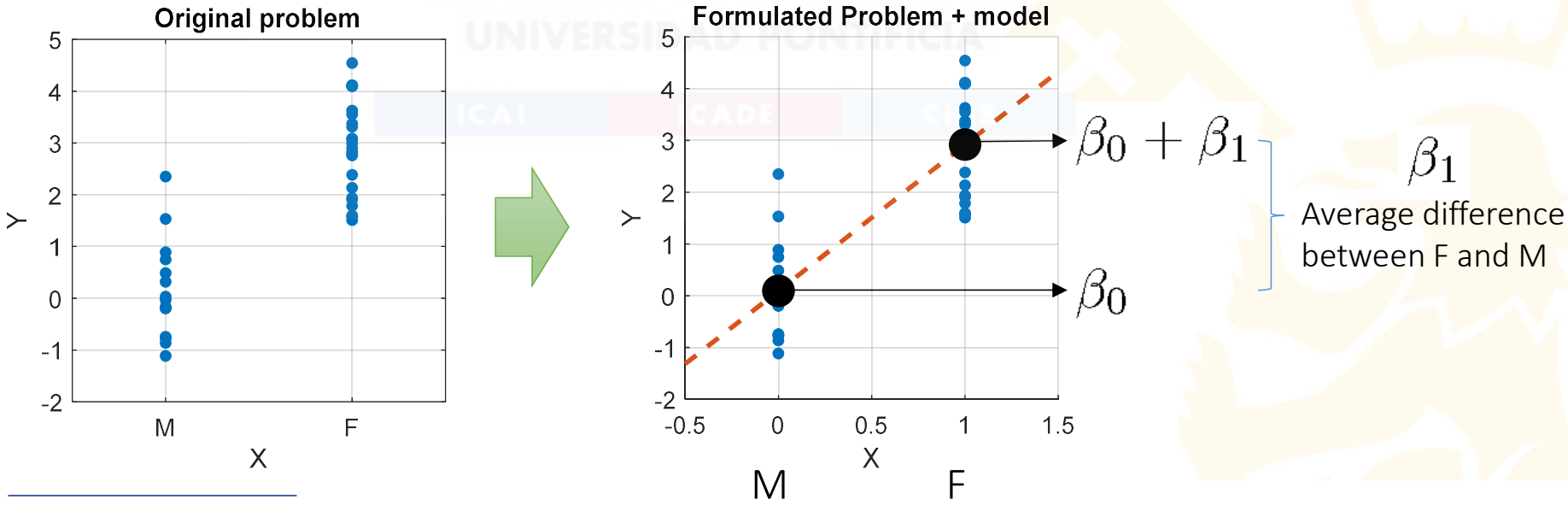
- If a qualitative variable has only two levels (or possible values), then incorporating it into a regression model is very simple
- Example

Dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

Regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$



Other considerations in regression
Qualitative input variable with 2 levels

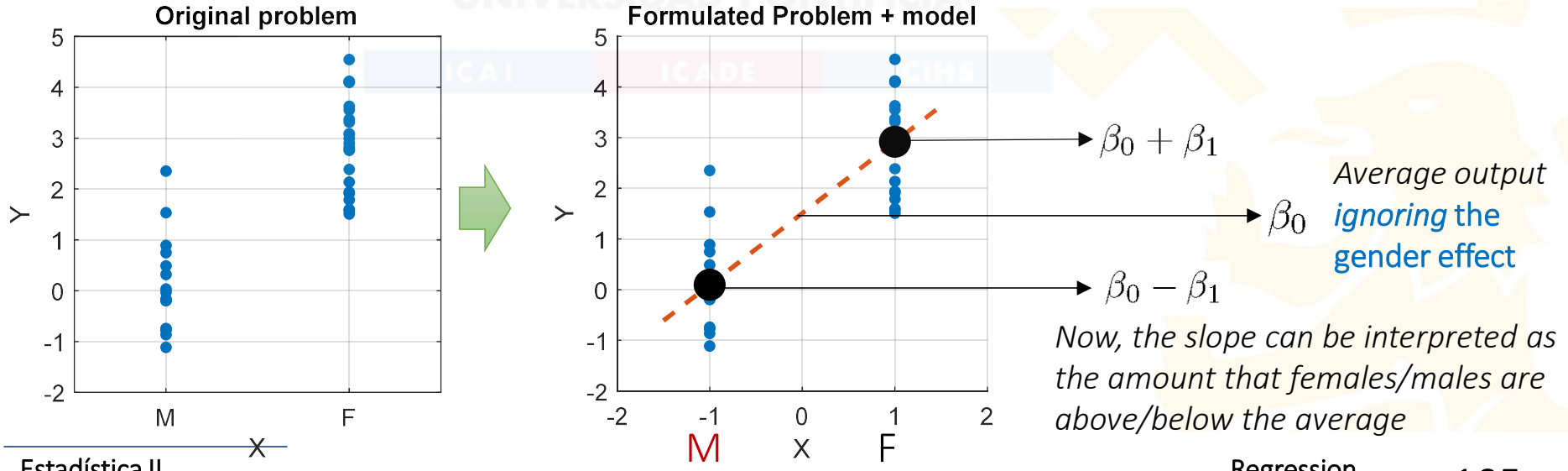
- Coding females as 1 and males as 0 is arbitrary
  - Although it does not affect the regression fit
  - It alters the interpretation of the coefficients
- A different coding (values -1 and 1)

Dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

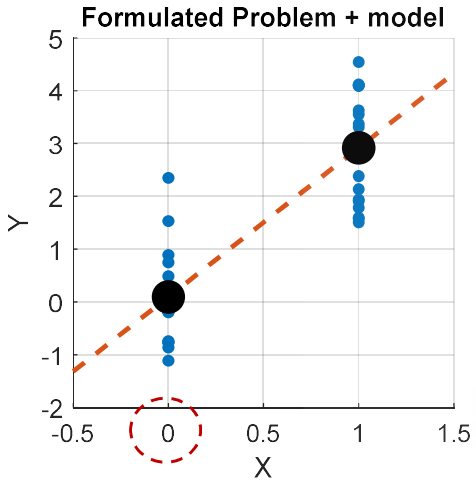
Regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

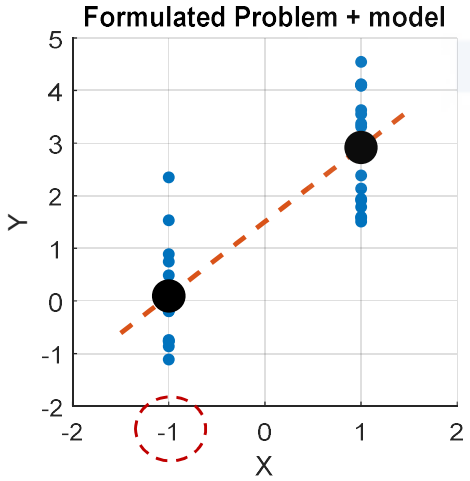


Other considerations in regression
Qualitative input variable with 2 levels

- Compare both models (the estimated coeffs. are different)



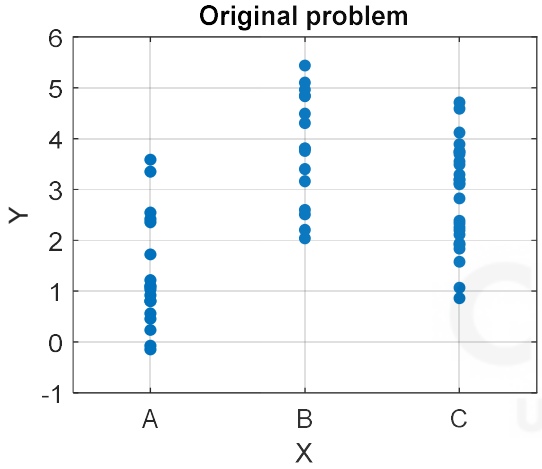

```
x is a dummy variable with values {0, 1}

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)    0.095147    0.23149    0.41102       0.68336
    x                2.8247    0.29885      9.452    1.5965e-11

Number of observations: 40, Error degrees of freedom: 38
Root Mean Squared Error: 0.926
R-squared: 0.702,  Adjusted R-Squared 0.694
F-statistic vs. constant model: 89.3, p-value = 1.6e-11
```

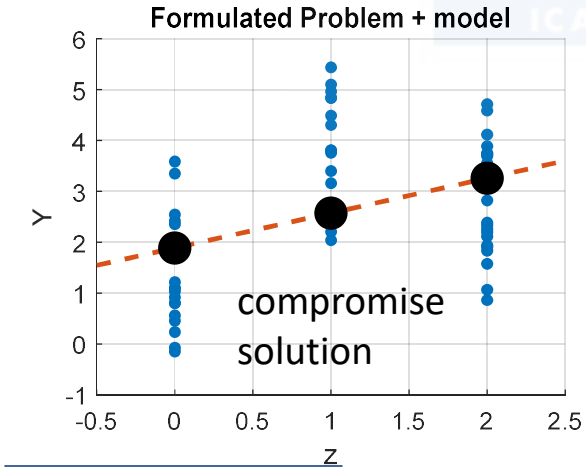



```
x is a dummy variable with values {-1, 1}

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)    1.5075      0.14943    10.089     2.6692e-12
    x              1.4124      0.14943     9.452     1.5965e-11

Number of observations: 40, Error degrees of freedom: 38
Root Mean Squared Error: 0.926
R-squared: 0.702,  Adjusted R-Squared 0.694
F-statistic vs. constant model: 89.3, p-value = 1.6e-11
```

Other considerations in regression
Qualitative input variable with >2 levels

- When a qualitative input has more than two levels, a single dummy variable cannot represent all possible values
- Wrong approach (example with 3 levels)





- Use only one quantitative variable with numeric values representing the different possible values of the original qualitative variable

Typical mistake when >2 levels

This model is quite bad because of the wrong modeling approach

```
z is a variable with values {0, 1, 2}

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    _______    ______    __________

    (Intercept)     1.8889      0.28259    6.6842    9.9212e-09
    z               0.68543     0.19982    3.4302     0.0011173

Number of observations: 60, Error degrees of freedom: 58
Root Mean Squared Error: 1.31
R-squared: 0.169,  Adjusted R-Squared 0.154
F-statistic vs. constant model: 11.8, p-value = 0.00112
```

Other considerations in regression
Qualitative input variable with >2 levels

- The correct approach consists of creating several dummy variables (for M levels, $M-1$ dummies are required)
- Example with 3 levels

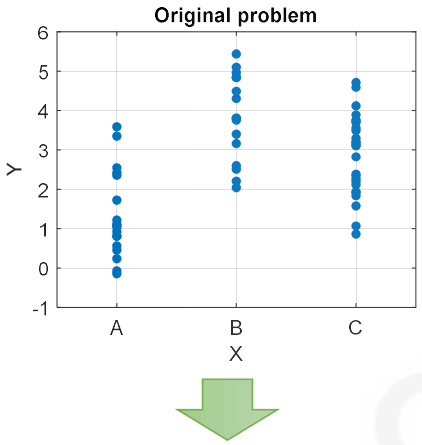
x	y
C	4.1174
C	1.9109
A	1.0326
C	3.5525
B	5.1006
A	2.5442

A	B	y
x1	x2	
0	0	4.1174
0	0	1.9109
1	0	1.0326
0	0	3.5525
0	1	5.1006
1	0	2.5442

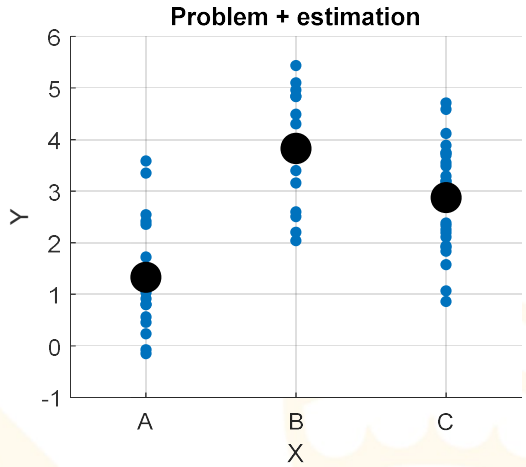
$$y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\epsilon_i = \begin{cases} \beta_0+\beta_1+\epsilon_i & \text{if } i\text{th observation is A} \\ \beta_0+\beta_2+\epsilon_i & \text{if } i\text{th observation is B} \\ \beta_0+\epsilon_i & \text{if } i\text{th observation is C} \end{cases}$$

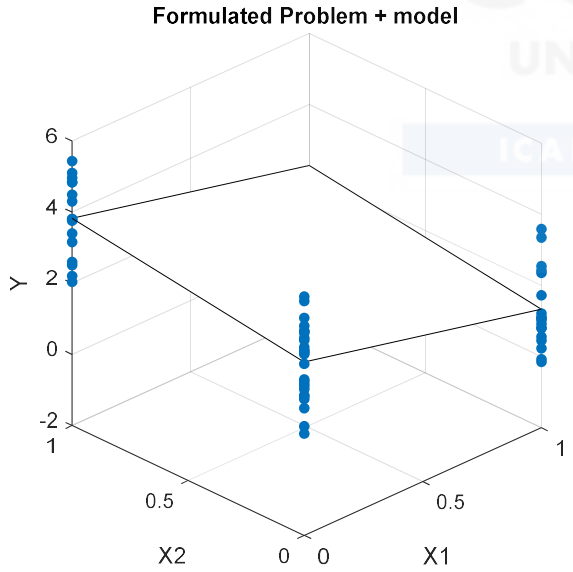
Other considerations in regression
Qualitative input variable with >2 levels

- The correct approach consists of fitting the regression model using the dummy variables instead of the qualitative one



The original problem has one input variable, but a multiple linear regression model is fitted to get the correct solution





x1 and x2 are dummy variables with values {0,1}

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	2.8741	0.20916	13.741	9.7126e-20
x1	-1.5423	0.32702	-4.7162	1.5957e-05
x2	0.95377	0.33888	2.8145	0.0066955

Number of observations: 60, Error degrees of freedom: 57
Root Mean Squared Error: 1.07
R-squared: 0.458, Adjusted R-Squared 0.439
F-statistic vs. constant model: 24.1, p-value = 2.64e-08

Other considerations in regression
Interactions between inputs

- The standard linear regression model provides interpretable results and works quite well on many real-world problems
- However, it makes several highly restrictive assumptions that are often violated in practice
  - One of the most important assumptions states that the relationship between the predictors and response is additive

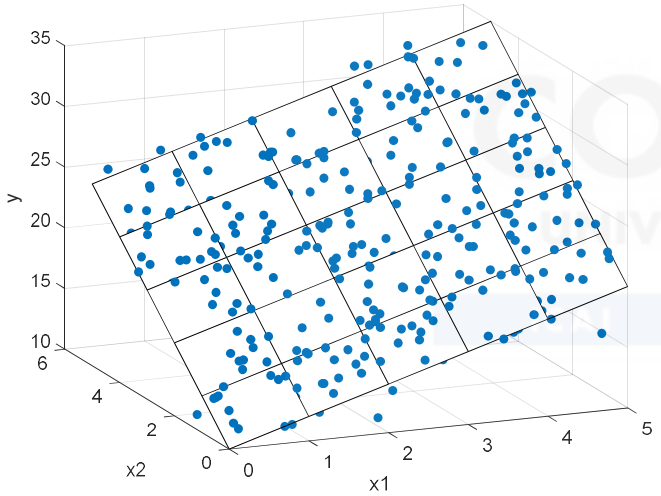
$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

The additive assumption means that the effect of changes in an input $\boldsymbol{X_i}$ on the response $\boldsymbol{Y}$ is independent of the values of the other predictors

Other considerations in regression
Interactions between inputs

- The interaction between inputs can have a synergy effect on the output
  - The combined effect or impact of inputs is greater than the sum of their individual effects on the output (e.g., a multiplicative effect)

WITHOUT Interaction

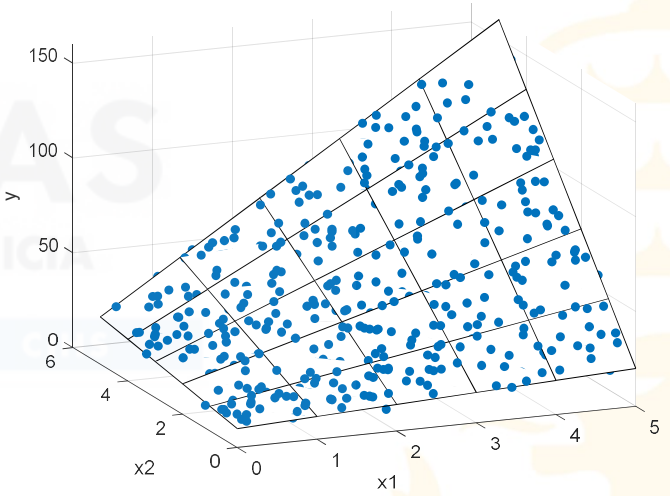


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Regardless of the value of X_2, a one-unit increase in X_1 will lead to a β_1-unit increase in Y

Note that the interaction effect has nothing to do with the correlation between inputs

WITH Interaction



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Interaction term

The whole is greater than the sum of the parts

Other considerations in regression
Interactions between inputs

- The standard way to allow for interaction effects is to include a third input variable, called an interaction term, which is constructed by computing the product of inputs

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Note that this model is equivalent to

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

where

$$\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$$

The effect of X_1 on Y is no longer constant: varying X_2 will change the impact of X_1 on Y

Other considerations in regression
Interactions between inputs

- The hierarchical principle states that:

> If we include interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

- In other words, if the interaction between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficient estimates have large p-values.
- The rationale for this principle is that if $X_1 \times X_2$ is related to the response, then whether the coefficients of X_1 or X_2 are exactly zero is of little interest.
- Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

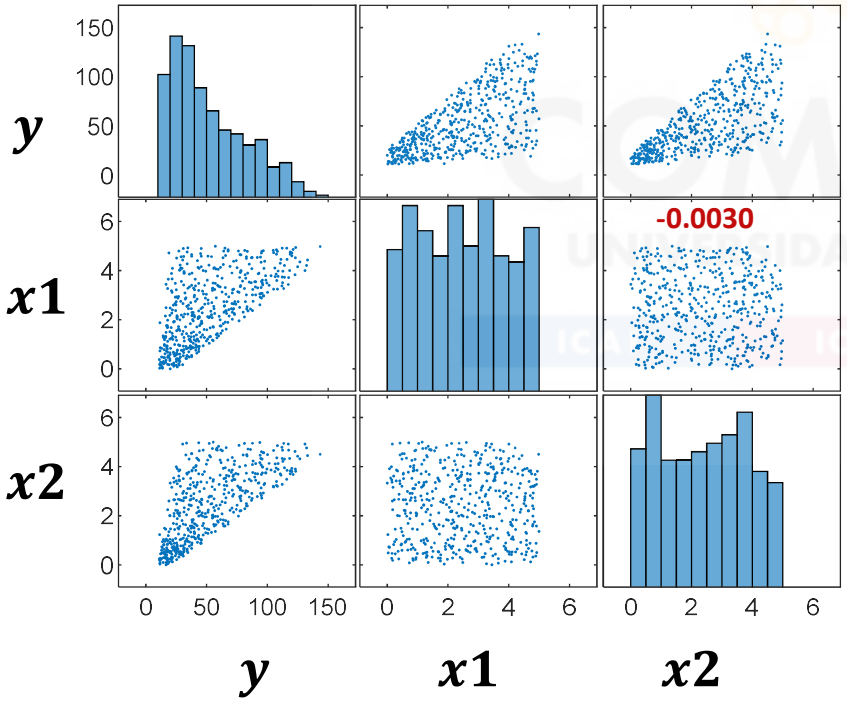
Other considerations in regression
Interactions: Illustrative cases

- C10: Interaction with small noise and many data

```
n=500;
x=unifrnd(0,5,n,2); % 3 independent inputs, uniforms
x1 = x(:,1); x2 = x(:,2);
% output: linear in x1, x2 & x1*x2 with Gaussian noise
stdnoise = 1.5; noise  = normrnd(0,stdnoise,n,1);
y = 10 + 2*x1 + 3*x2 + 5*x1.*x2 + noise;
```

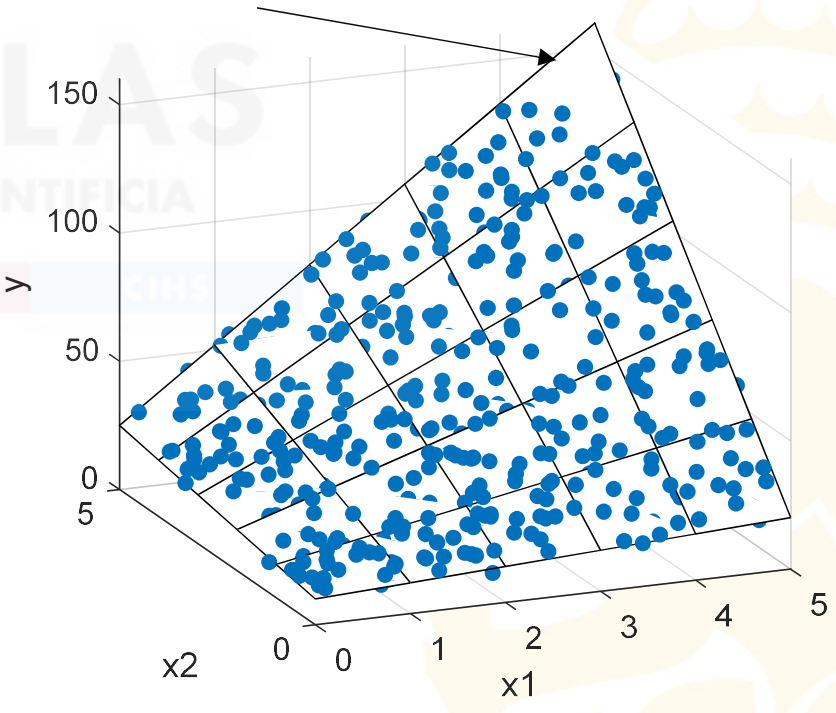
Correlation matrix $(y, x1, x2)$:

1.0000	0.6374	0.6965
0.6374	1.0000	-0.0030
0.6965	-0.0030	1.0000




```
figure; plotmatrix([y x]);
```

True deterministic term



Other considerations in regression
Interactions: Illustrative cases

- C10: Interaction with small noise and many data

```
%% Fit linear regression model WITHOUT INTERACTION
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,y); % For simplicity as a table
model = fitlm(tb,'y ~ 1+x1+x2');
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2


Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    ________    __________

    (Intercept)     -18.24      1.1759    -15.512     1.5311e-44
    x1              13.836      0.31586    43.803     9.139e-173
    x2               14.97      0.31291    47.842     3.7013e-188


Number of observations: 500, Error degrees of freedom: 497
Root Mean Squared Error: 9.95
R-squared: 0.894,  Adjusted R-Squared 0.894
F-statistic vs. constant model: 2.1e+03, p-value = 5.14e-243
```

Initial model (without interaction)

- Model is significant (overall F test)
- All the coefficients are significant

There seems to exist a linear relation of y with $x1$ and $x2$

- However, we know that there exists an interaction
- Compare RMSE with the standard deviation of the noise …

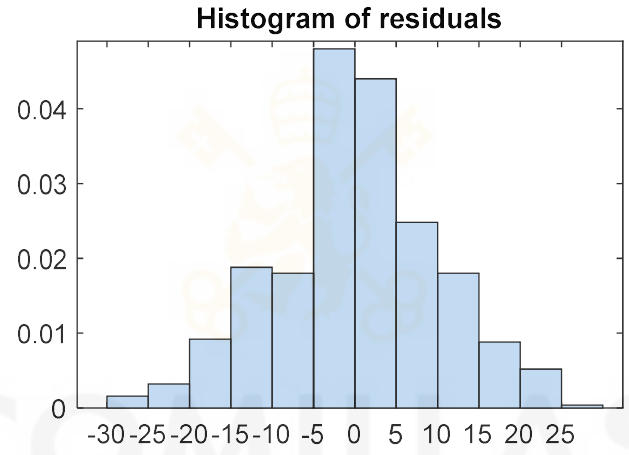
Other considerations in regression
Interactions: Illustrative cases

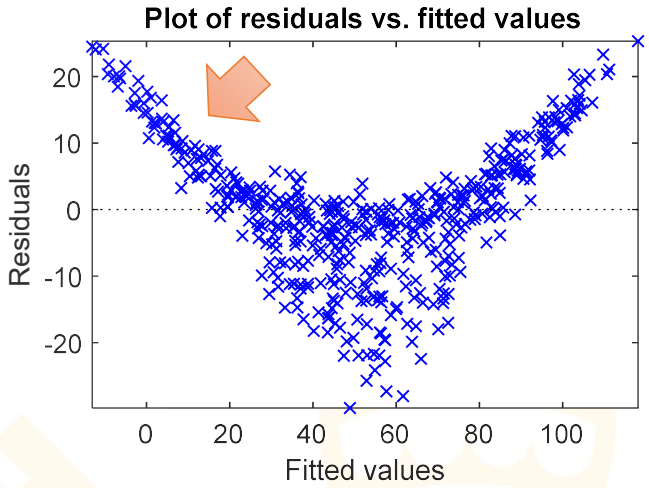
Initial model (without interaction)

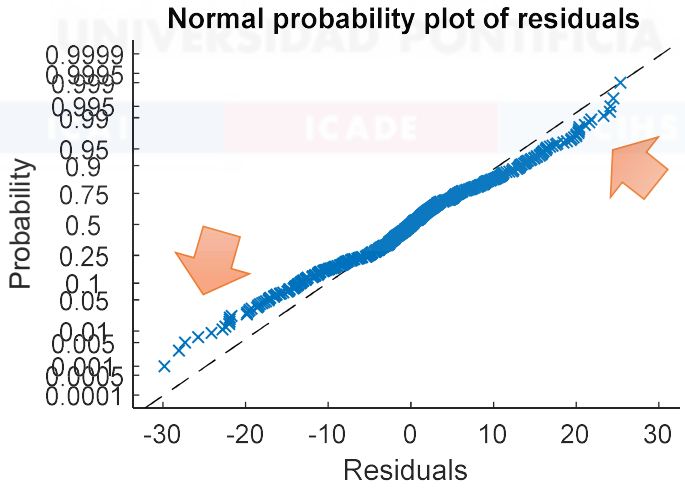
- C10: Interaction with small noise and many data (residuals)

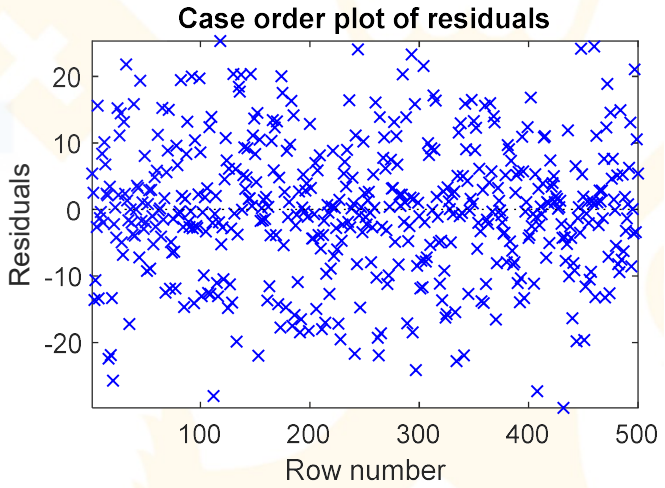
- Nonnormality of residuals and residual variance doesn't seem constant
- In fact, the plot "residuals vs fitted values" suggests some interaction

Residual analysis suggests that the initial model has important problems in its specification









Other considerations in regression
Interactions: Illustrative cases

- C10: Interaction with small noise and many data

```
%% Fit linear regression model WITH INTERACTION
% With ALL the available input variables (includes intercept)

tb = table(x1,x2,y); % For simplicity as a table
model = fitlm(tb,'y ~ 1+x1+x2+x1*x2');
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1*x2   <- note the difference


Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    _____    __________

    (Intercept)      10.341     0.26361    39.226    3.8461e-154
    x1               1.9521    0.094102    20.744     2.7696e-69
    x2               2.9048    0.094953    30.592    2.8824e-116
    x1:x2             5.019    0.034292    146.36              0

Means interaction (a "new" input with its slope)

Number of observations: 500, Error degrees of freedom: 496
Root Mean Squared Error: 1.5
R-squared: 0.998,  Adjusted R-Squared 0.998
F-statistic vs. constant model: 6.88e+04, p-value = 0
```

New model (with interaction)

- Model is significant (overall F test)
- All the coefficients are significant

There seems to exist a linear relation of y with $x1$ and $x2$

- Because of the correct specification of the model (including the interaction), the RMSE is very similar to the standard deviation of the noise

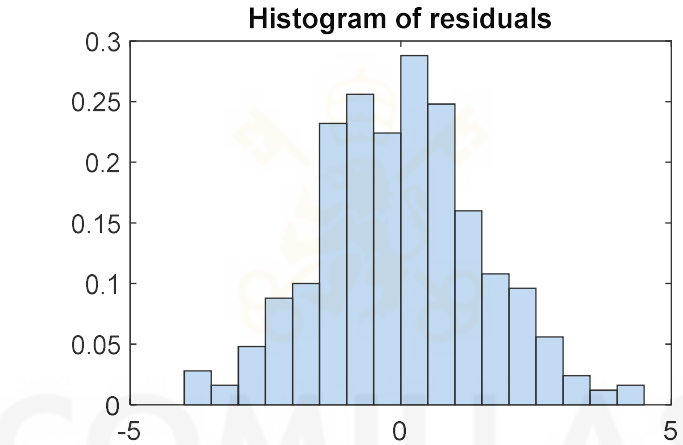
Other considerations in regression
Interactions: Illustrative cases

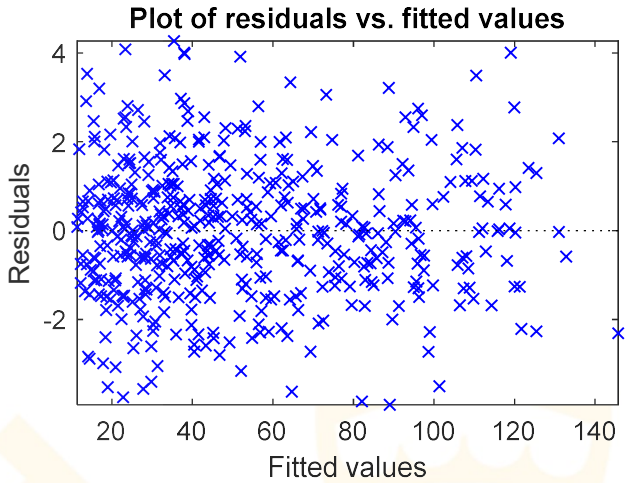
New model (with interaction)

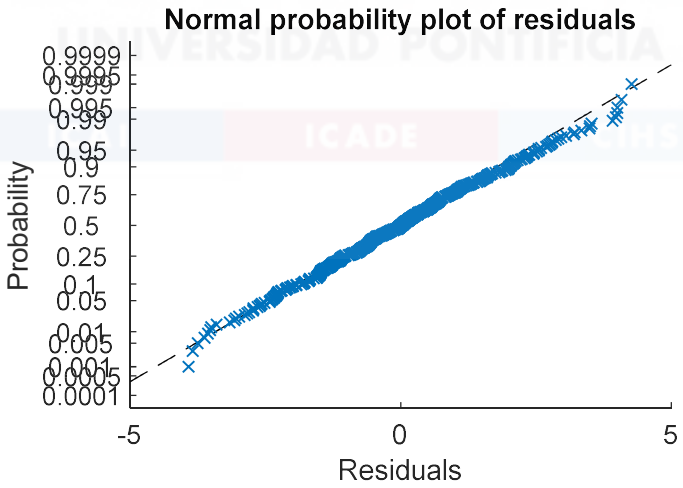
- C10: Interaction with small noise and many data (residuals)

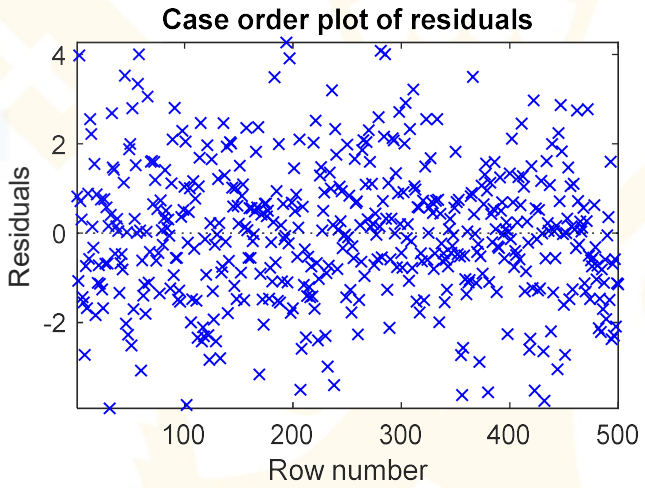
- Normality and independence of residuals with this model can't be questioned

Residual analysis suggests that the new model is very reasonable





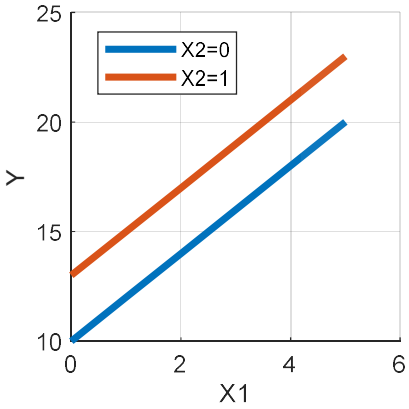




Other considerations in regression
Interactions between inputs

- The concept of interactions applies to
  - Quantitative inputs
  - Qualitative inputs (dummies, used in ANOVA)
  - A combination of quantitative and qualitative inputs
- An interaction between a qualitative variable and a quantitative variable has a nice interpretation

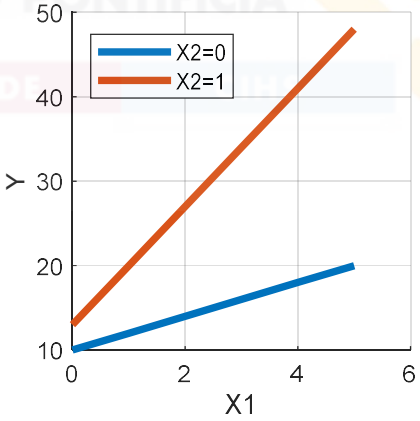
WITHOUT Interaction



- Two parallel lines, one for each value of X_2
- Two different intercepts

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

WITH Interaction

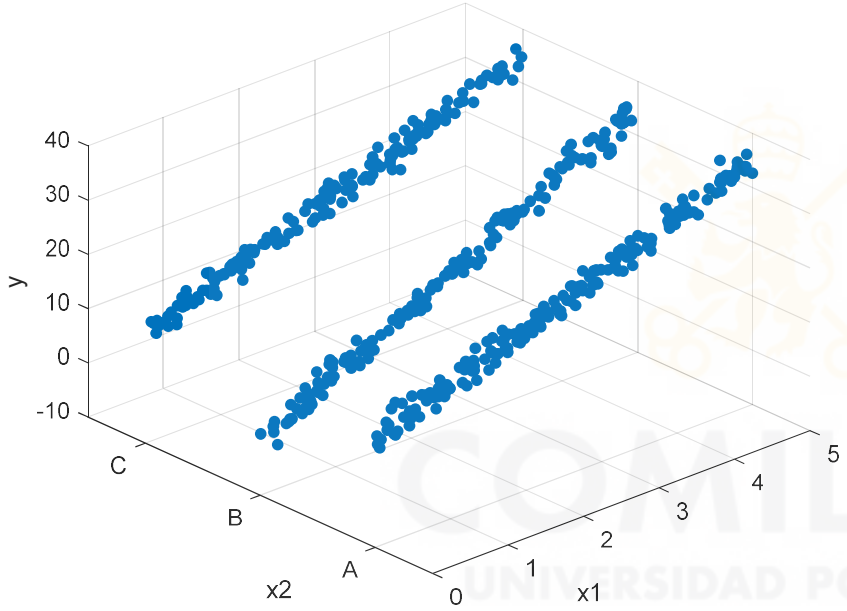


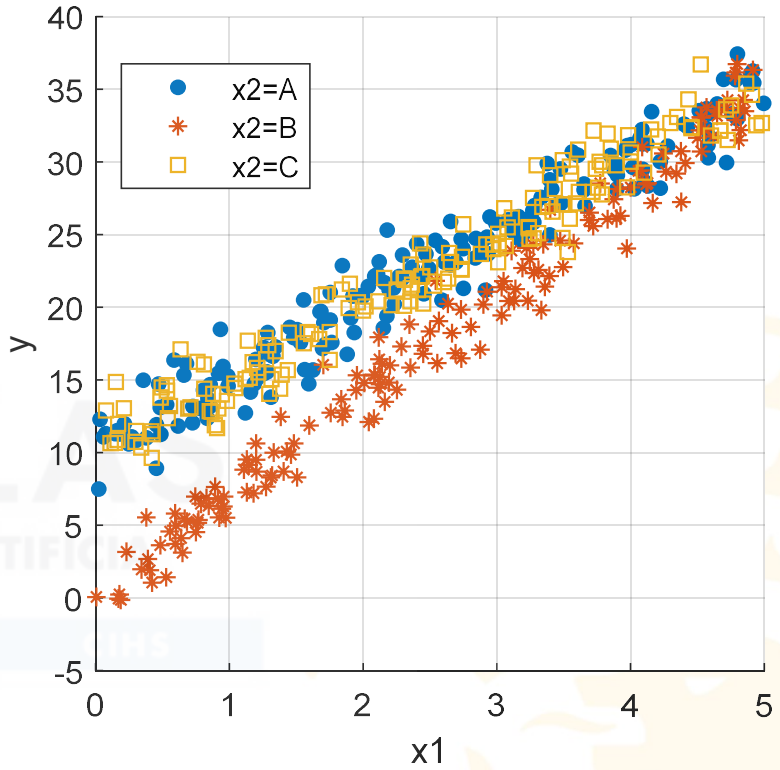
- Two different lines, with a different slope for each value of X_2
- Two different intercepts

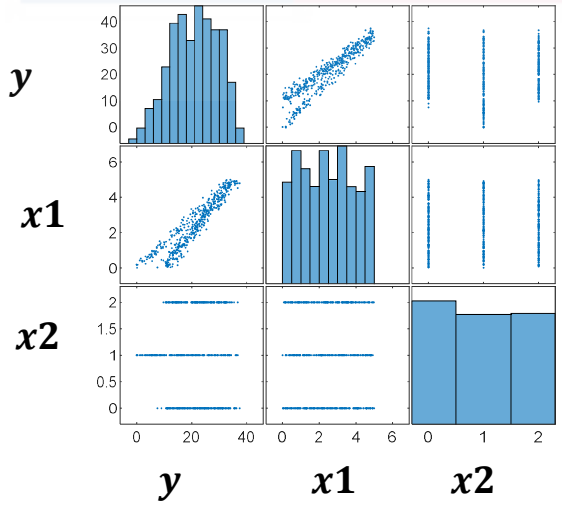
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Other considerations in regression
Interactions: Illustrative cases

- C11: Interaction between a qualitative & a quantitative variable





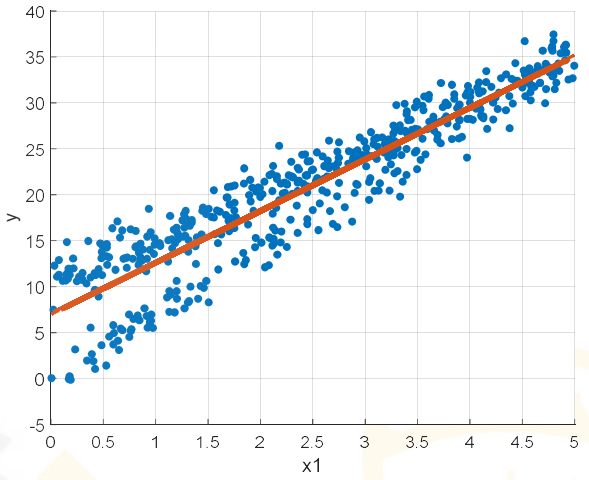


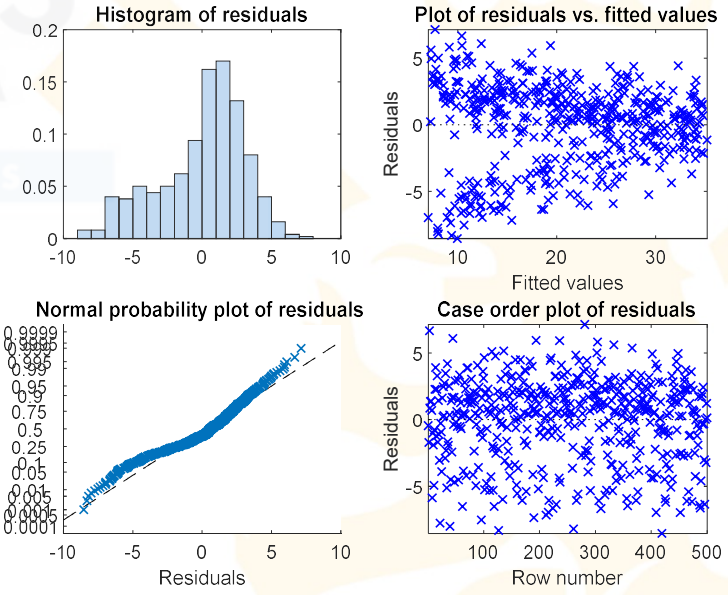
Other considerations in regression
Interactions: Illustrative cases

- C11: Interaction between a qualitative & a quantitative variable

```
%% Fit linear regression model WITHOUT INTERACTION
tb = table(x1,x2n,y); % x2n is qualitative {0,1,2}
model = fitlm(tb,'y ~ 1+x1+x2n');
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2n

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    ________    _________

    (Intercept)       7.134     0.32391      22.025    1.5841e-75
    x1                5.6173    0.098258     57.169    1.1047e-220
    x2n              -0.13783      0.168    -0.82043     0.41236


Number of observations: 500, Error degrees of freedom: 497
Root Mean Squared Error: 3.09
R-squared: 0.868,  Adjusted R-Squared 0.868
F-statistic vs. constant model: 1.63e+03, p-value = 2.6e-219
```





Very wrong approach
(using x2n, a qualitative version of x2)

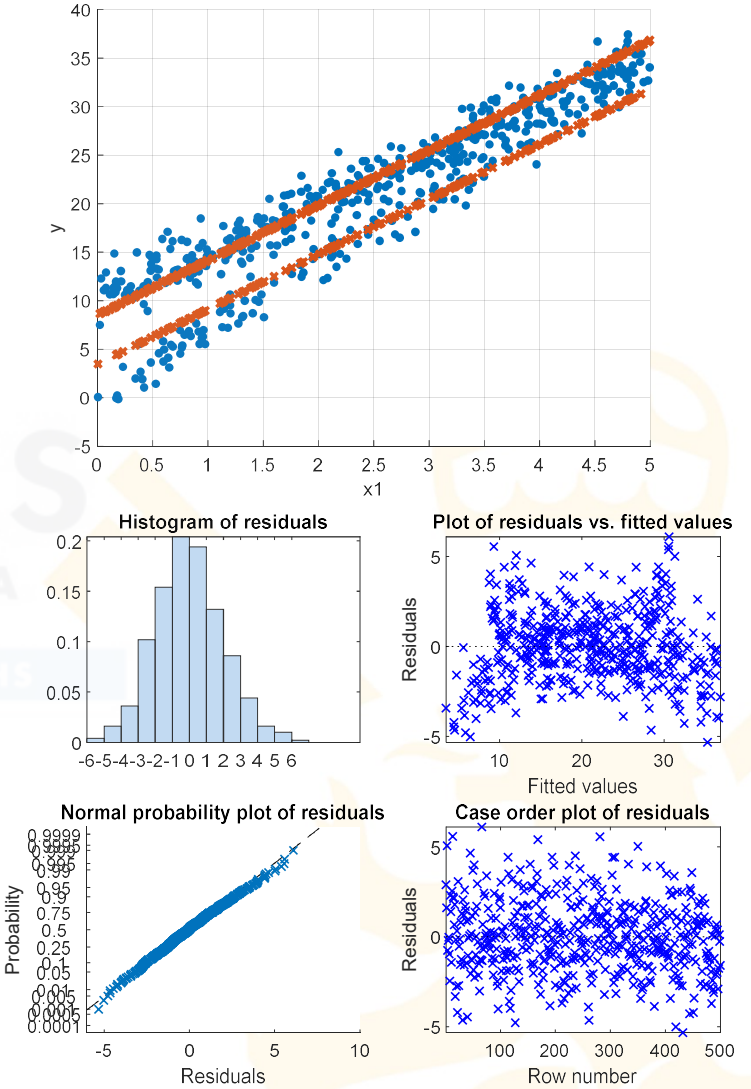
Other considerations in regression
Interactions: Illustrative cases

- C11: Interaction between a qualitative & a quantitative variable

```
%% Fit linear regression model WITHOUT INTERACTION
tb = table(x1,x2,y);
model = fitlm(tb,'y ~ 1+x1+x2');
disp(model);
```

```
Linear regression model:
    y ~ 1 + x1 + x2

Estimated Coefficients:
                   Estimate        SE        tStat        pValue
                   ________    ________    ________    __________

    (Intercept)       8.5387     0.21697      39.355    1.1305e-154
    x1                5.6679    0.063799       88.84    9.3672e-307
    x2_B              -5.092     0.21857      -23.297    1.2024e-81
    x2_C             -0.10329     0.21816     -0.47344       0.63611


Number of observations: 500, Error degrees of freedom: 496
Root Mean Squared Error: 2.01
R-squared: 0.945,  Adjusted R-Squared 0.944
F-statistic vs. constant model: 2.82e+03, p-value = 5.63e-311
```

Model without interaction
(but using x2 as qualitative)



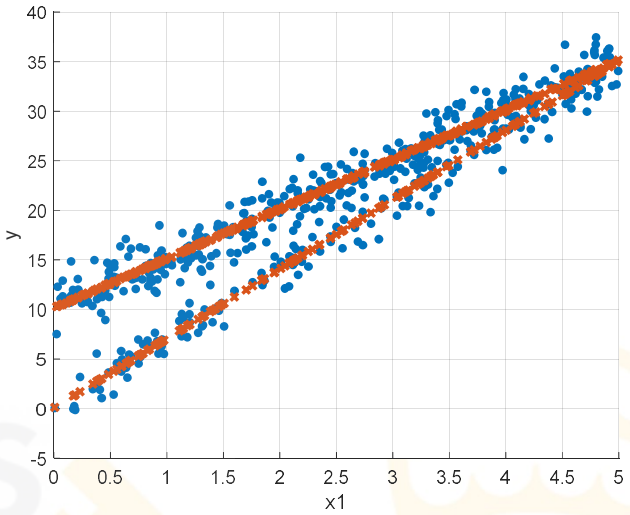
Other considerations in regression
Interactions: Illustrative cases

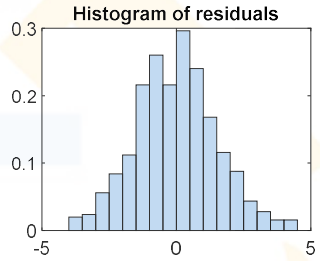
- C11: Interaction between a qualitative & a quantitative variable

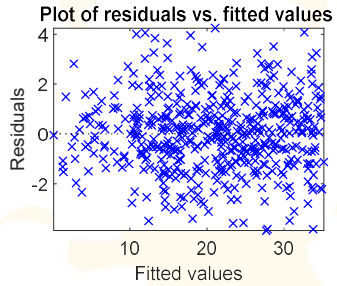
```
%% Fit linear regression model WITH INTERACTION
tb = table(x1,x2,y);
model = fitlm(tb,'y ~ x1*x2');
disp(model);
```

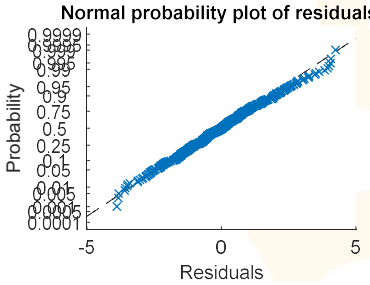
```
Linear regression model:
    y ~ 1 + x1*x2

Estimated Coefficients:
                   Estimate        SE         tStat        pValue
                   ________    ________    ________    __________

    (Intercept)      10.161     0.22441       45.28    5.9674e-178
    x1               5.0071    0.079133      63.275    4.3865e-239
    x2_B            -10.087     0.32869      -30.688    1.5761e-116
    x2_C           -0.064836     0.33004    -0.19645        0.84434
    x1:x2_B            1.9922     0.11425      17.437    2.0609e-53
    x1:x2_C         -0.023527     0.11758     -0.2001        0.84149


Number of observations: 500, Error degrees of freedom: 494
Root Mean Squared Error: 1.5
R-squared: 0.969,  Adjusted R-Squared 0.969
F-statistic vs. constant model: 3.1e+03, p-value = 0
```

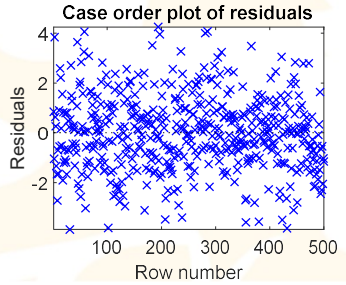
Model with interaction











1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

6

Quiz

Quiz
Question 1

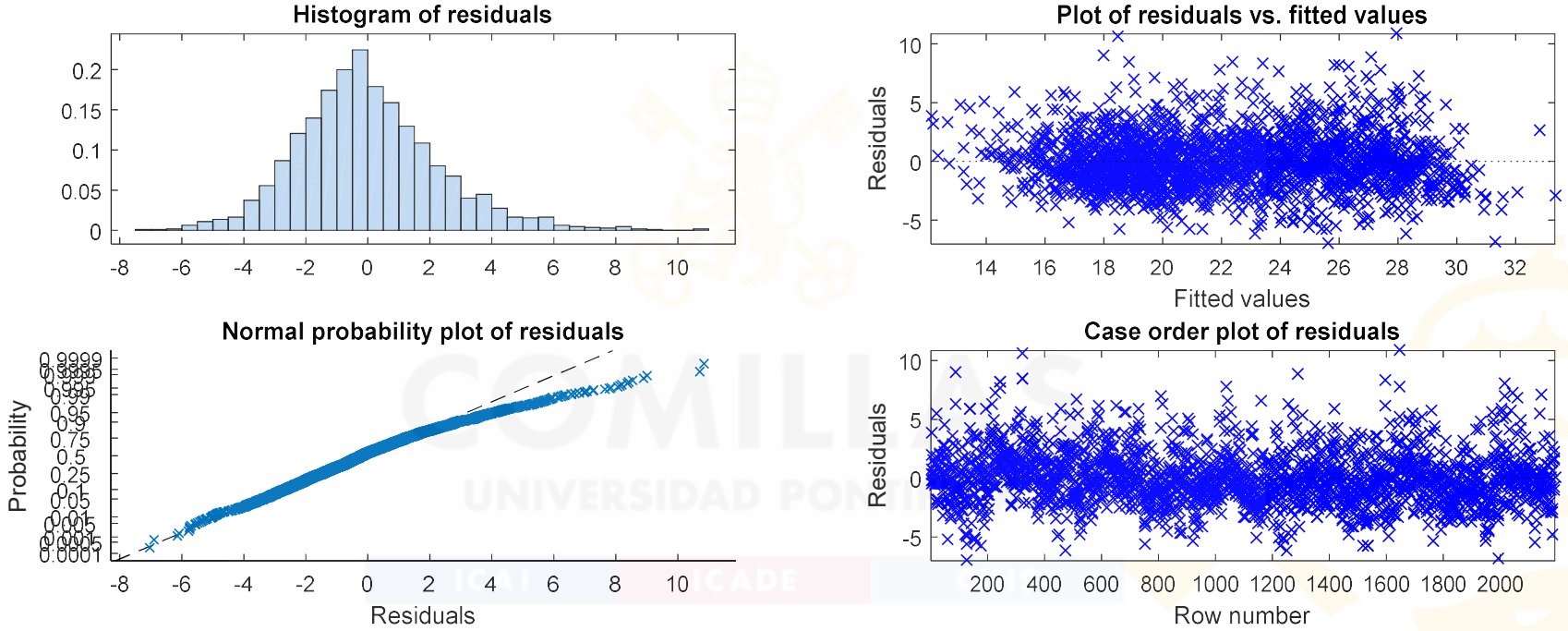
- Dado el siguiente modelo de regresión lineal, se puede afirmar que

```
Estimated Coefficients:
                   Estimate        SE        tStat       pValue
                   ________    _________    _______    __________

    (Intercept)    -0.20236     0.041135    -4.9195    8.9158e-07
    TMIN421        -0.02195     0.003138     -6.995    2.9497e-12
    TMAX3919       0.017952    0.0037182     4.8283    1.4123e-06
    TMAX3948        0.99871    0.0022468     444.51             0


Number of observations: 5844, Error degrees of freedom: 5840
Root Mean Squared Error: 0.668
R-squared: 0.994,  Adjusted R-Squared 0.994
F-statistic vs. constant model: 3.33e+05, p-value = 0
```

A. La variable TMAX3919 es relevante en el modelo, su coeficiente asociado no debe considerarse nulo.
B. Ninguno de los coeficientes del modelo es significativo.
C. La constante puede ser nula, por lo que debe de eliminarse del modelo.

Quiz
Question 2

- Dado el siguiente análisis de los residuos de un modelo de regresión lineal se puede afirmar que



A. El principal problema que arroja el análisis es la falta de normalidad en los residuos, pudiéndose asumir que su varianza es constante.
B. No hay evidencias suficientemente fuertes como para rechazar la hipótesis de normalidad de los residuos, debiendo considerarse como normales. Sin embargo, es difícil aceptar que la varianza de los residuos se mantiene constante.
C. El modelo ajustado es perfectamente válido al ser los residuos normales, independientes y tener varianza constante.

Quiz
Question 3

- Se han ajustado los dos modelos indicados para estimar a partir de una muestra de datos la misma variable de salida. Un análisis de la bondad del ajuste de estos modelos permite afirmar que

M1

```
Estimated Coefficients:
                   Estimate       SE          tStat     pValue
                   ________    _________    _______    ______

    (Intercept)    -6.0439      0.097901    -61.735    0
    TMAX3967       0.70318     0.0040576     173.3     0


Number of observations: 7305, Error degrees of freedom: 7303
Root Mean Squared Error: 3.11
R-squared: 0.804,  Adjusted R-Squared 0.804
F-statistic vs. constant model: 3e+04, p-value = 0
```

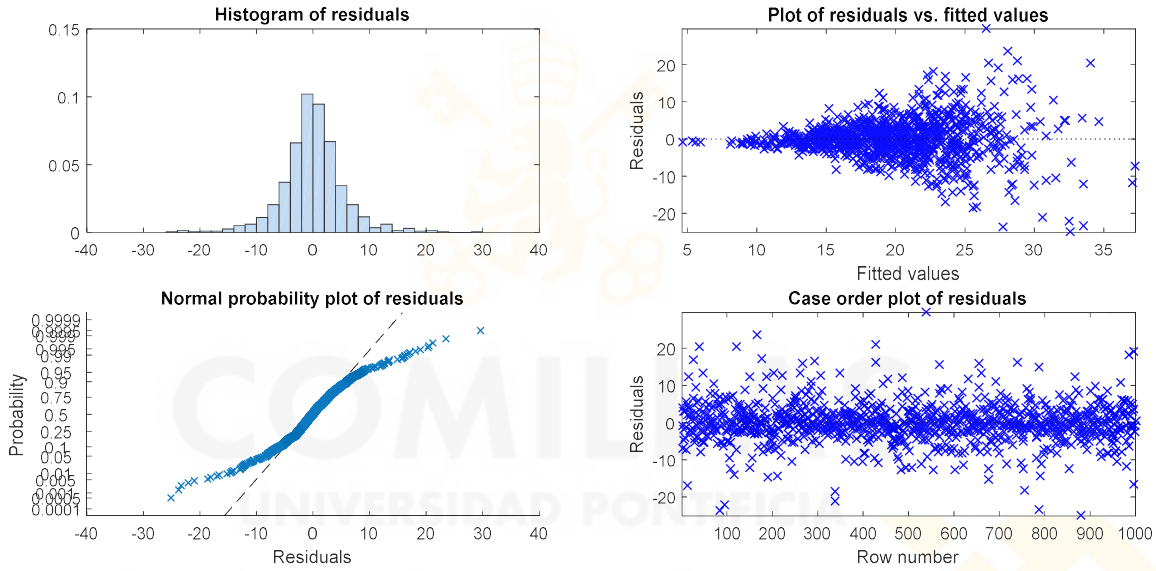
M2

```
Estimated Coefficients:
                   Estimate       SE          tStat     pValue
                   ________    _________    _______    ______

    (Intercept)    -8.5819      0.18098     -47.419    0
    TMAX3954       0.95436     0.0090493     105.46    0


Number of observations: 7305, Error degrees of freedom: 7303
Root Mean Squared Error: 4.43
R-squared: 0.604,  Adjusted R-Squared 0.604
F-statistic vs. constant model: 1.11e+04, p-value = 0
```

A. El modelo M1 es peor que el modelo M2 ya que tiene un R2 menor y un RMSE menor.
B. El modelo M1 es mejor que el modelo M2 ya que tiene un RMSE menor y un coeficiente de determinación mayor.
C. No se pueden comparar ya que utilizan variables de entrada diferentes.

Quiz
Question 4

- Se ha ajustado un modelo de regresión lineal. En vista del análisis de sus residuos se puede afirmar que

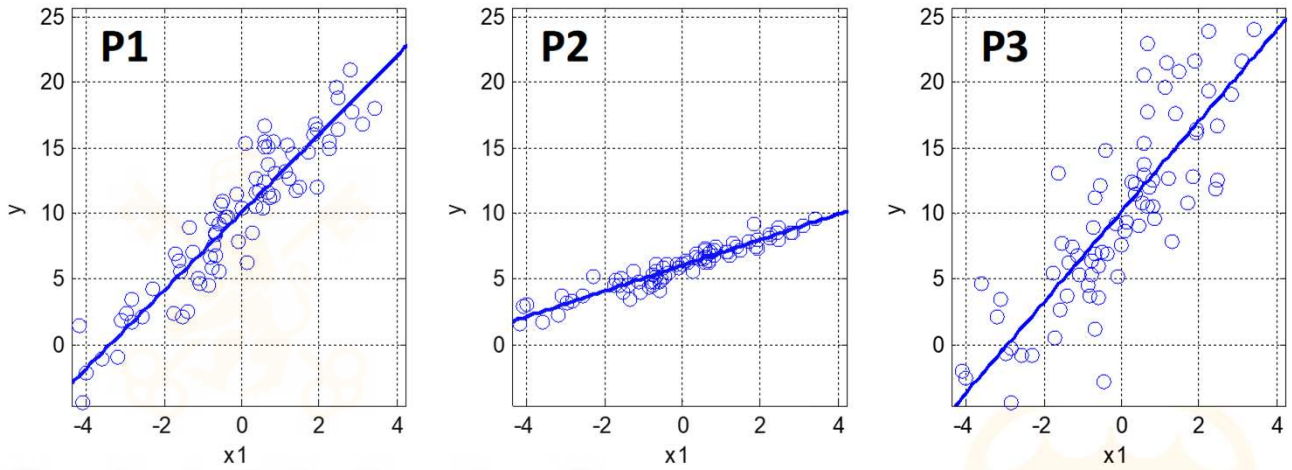


A. No se puede rechazar la hipótesis de varianza de los residuos constante al no existir evidencias claras, siendo incuestionable la normalidad de los mismos.
B. No hay evidencias suficientemente fuertes como para rechazar que la varianza de los residuos se mantiene constante. Sin embargo, es difícil aceptar que los residuos son independientes.
C. Hay evidencias suficientemente fuertes que permiten rechazar la hipótesis de normalidad de los residuos. Además, es difícil aceptar que la varianza de los residuos se mantiene constante.

Quiz
Question 5

- En la siguiente figura se muestra el modelo de regresión lineal simple obtenido en tres problemas diferentes. Considerando la información de los modelos M1 y M2, se puede afirmar que

A. El modelo M1 se ha estimado en el problema P1, mientras que el modelo M2 es el obtenido en el problema P2.
B. El modelo M1 se ha estimado en el problema P1, mientras que el modelo M2 es el obtenido en el problema P3.
C. El modelo M1 se ha estimado en el problema P3, mientras que el modelo M2 es el obtenido en el problema P1.



M1

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	10.148	0.46327	21.906	4.6025e-35
x1	3.4702	0.23865	14.541	6.4825e-24

Number of observations: 80, Error degrees of freedom: 78
Root Mean Squared Error: 4.14
R-squared: 0.731, Adjusted R-Squared 0.727
F-statistic vs. constant model: 211, p-value = 6.48e-24

M2

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	10.104	0.22842	44.236	5.2722e-57
x1	2.9911	0.11767	25.419	1.7259e-39

Number of observations: 80, Error degrees of freedom: 78
Root Mean Squared Error: 2.04
R-squared: 0.892, Adjusted R-Squared 0.891
F-statistic vs. constant model: 646, p-value = 1.73e-39

Quiz
Question 6

- Se ha ajustado un modelo de regresión lineal múltiple. En vista de los resultados obtenidos se puede afirmar que

```
Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    _______    __________

    (Intercept)      20.21      0.26398     76.558     5.505e-148
    x1               4.963     0.090359     54.925    5.4357e-121
    x2_S            -19.975      0.38389    -52.033    1.1123e-116
    x1:x2_S          1.8995      0.13308     14.273     3.6931e-32

Number of observations: 200, Error degrees of freedom: 196
Root Mean Squared Error: 1.38
R-squared: 0.987,  Adjusted R-Squared 0.986
F-statistic vs. constant model: 4.78e+03, p-value = 5.84e-183
```

A. El modelo se ha ajustado con dos variables de entrada: x1 (categórica con dos posibles valores), x2 (categórica con un único valor), incluyendo interacción entre las mismas.
B. El modelo se ha ajustado con dos variables de entrada: x1 (continua) y x2 (categórica con dos posibles valores), incluyendo interacción entre las mismas.
C. El modelo se ha ajustado con dos variables de entrada: x1 (continua) y x2 (categórica con dos posibles valores), sin incluir interacción entre las mismas.

Quiz Answers

- Q1-A
- Q2-A
- Q3-B
- Q4-C
- Q5-C
- Q6-B

1. Introduction
2. Constant Model
3. Simple Linear Regression
4. Multiple Linear Regression
5. Other considerations
6. Quiz
7. Real examples

7

Real examples

Real cases
Center of gravity estimation for aircraft

Contents lists available at ScienceDirect

Aerospace Science and Technology

www.elsevier.com/locate/aescte

Preliminary development of an onboard weight and balance estimator for commercial aircraft

F.A.V. Chaves *, M.A.R. Silvestre, P.V. Gamboa

University of Beira Interior, Aerospace Sciences Department, Edifício II das Engenharias, Calçada Fonte do Lameiro, n.° 1, 6201-001 Covilhã, Portugal

A novel approach to onboard in-flight weight and balance estimation systems is presented. Data from an Airbus A320 fleet from an airline were used to assess the feasibility of the approach. Simple flight mechanics in combination with statistics allowed for the identification of weight and center of gravity position using cruise angle of attack, Mach number and elevator deflection values. The good agreement between the theoretical model and the obtained values for the lift curve slope as a function of Mach as well as the standard error of the estimate for center of gravity position and cruise flying weight indicate that the method is sound. The major implication of this work is that the development of onboard and in-flight weight and balance systems can be significantly simpler than previous literature suggested. The impact of this paper could be immediate for airlines since all the tools required to implement the system as described are readily available. This could have an effect in operating costs, safety and environment.

© 2017 Elsevier Masson SAS. All rights reserved.

Real cases
Center of gravity estimation for aircraft

3.3. Center of gravity position estimation

$$x_{CG} = 24.956C_L - 1.692\delta_{HT} + 25.928$$

Multiple linear regression was used to obtain x_{CG} from the airplane's cruise trim horizontal tail equivalent deflection position and the lift coefficient

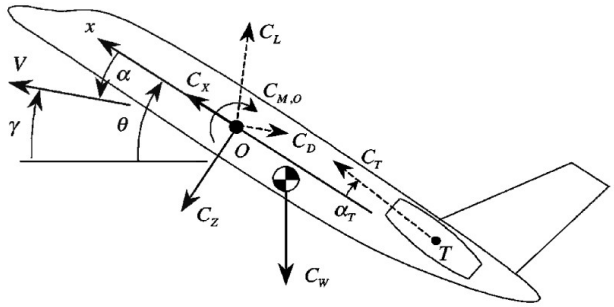


Fig. 2. Non-dimensional forces and moments [18].

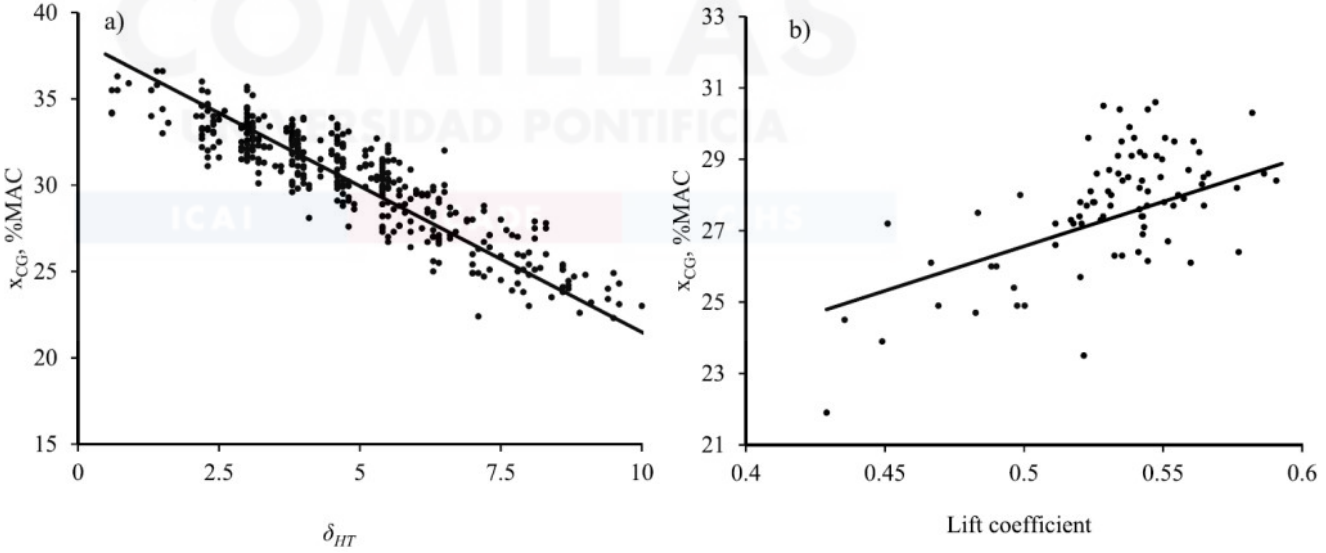


Fig. 10. a) Data points dispersion and prediction line with constant lift coefficient ($C_L = 0.5$) and 10 b) Data points dispersion and prediction line with constant δ_{HT} ($\delta_{HT} = 7$ deg).

Real cases
Center of gravity estimation for aircraft

A.1. Multiple linear regression assumptions check for center of gravity determination

A random scatterplot of the regression standardized residual against regression standardized predicted value provided evidence of linearity, Fig. A.1. The assumption of normality was tested via

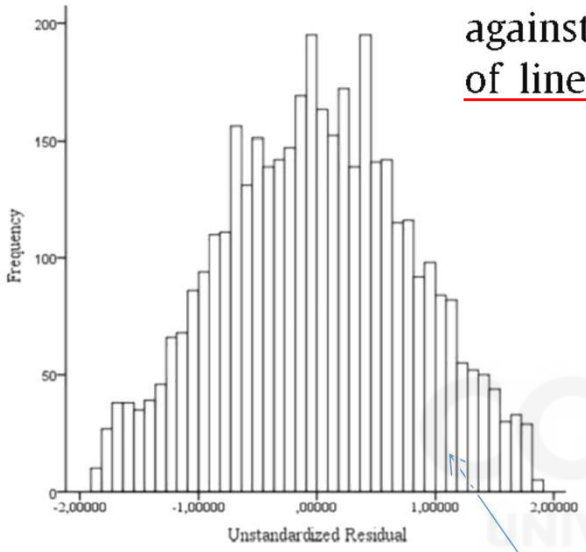


Fig. A.4. Histogram of unstandardized residual for multiple linear regression model.

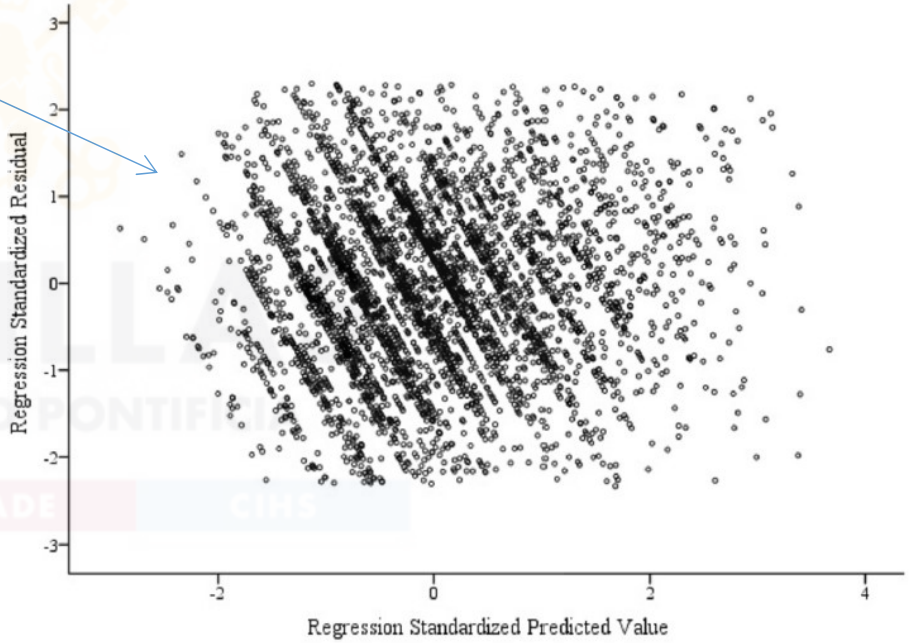


Fig. A.1. Scatter plot for regression standardized predicted value vs regression standardized residual.

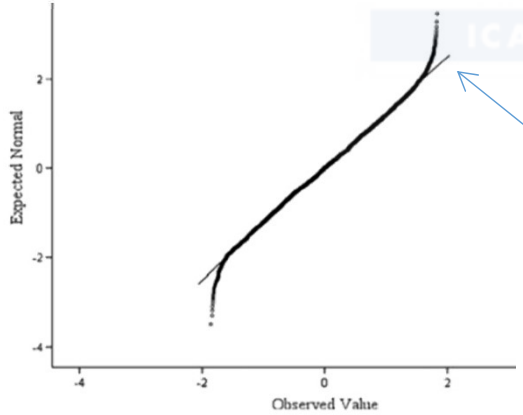


Fig. A.3. Normal Q-Q plot of unstandardized residual for multiple linear regression model.

liers) of the residuals, Fig. A.2. Q–Q plot showed some deviation from normal at the tails, but the histogram of unstandardized residuals suggested normality was reasonable, Figs. A.3 and A.4.

A

Dictionary

Dictionary

- **Bias** – sesgo
- **Bound** - cota
- **Dummy variable** – variable ficticia
- **Fit** - ajuste
- **Goodness of fit** – bondad de ajuste
- **Intercept** – ordenada en el origen
- **Joint distribution** – distribución conjunta
- **Least squares** – Mínimos cuadrados
- **Mean squared error** – error cuadrático medio
- **Sample** – muestra
- **Scatterplot-** nube de puntos
- **Slope** - pendiente
- **Straight line** – línea recta

Thank you for your attention

Eugenio Sánchez Úbeda