

# **Linear models**

## (An Introduction)

---

N. Torelli, G. Di Credico, V. Gioia

Fall 2024

University of Trieste

# **Introduction to linear models**

## **Multiple linear model**

## Introduction to linear models

---

THE MAIN IDEA IS THAT WE WANT TO ANALYZE A SPECIFIC VARIABLE AND FIND A LINEAR CORRELATION WITH ANOTHER VARIABLE (OR VARIABLES) AND "EXPLAIN" A CORRELATION

# Linear regression model

WE ASSUME THE RELATIONSHIP  
TO BE LINEAR

Linear regression model is one of the basic tools for statistical analysis.

Since pioneering works of Sir Francis Galton in the late XIX century, the main aim of regression models is to study the systematic influence of

- one or more **concomitant factors** (explanatory variables, regressors, covariates) on *THE VARIABLE WE BELIEVE INFLUENCES Y (RESPONSE VARIABLE)*
- a **response variable** (dependent variable).

THE RELATION BETWEEN THIS TWO  
IT'S AN ASYMMETRIC RELATION,  
BECAUSE WE WANT TO EXPLAIN Y USING  
OTHER VALUE.  
SO, THE ROLL OF THIS TWO IS DIFFERENT

The main goal of regression modelling is understanding whether and *how* the response variable (the phenomenon of interest) is related to the concomitant quantities.

THERE IS A RELATIONSHIP  
BETWEEN THE RESPONSE AND  
EXPLANATORY VARIABLES

THE RESPONSE VARIABLE CHANGES  
WHEN THE EXPLANATORY VARIABLES  
CHANGE

The basic regression model has been expanded in many directions in order to apply it in extremely complex situations and to large and complex data sets, but the basic aim remained the same.

WE INTERPRET A LINEAR REGRESSION MODEL AS THE  
EFFECT OF THE EXPLANATORY VARIABLES ON THE DEPENDENT  
VARIABLE, AND NOT VICEVERSA

# Aims of regression modelling

- prediction/forecast: regression modelling is a tool to provide a prediction of the phenomenon of interest, given the knowledge of the concomitant factors (e.g. for time reasons, costs, or because the concomitant factors are easier to measure)
- interpretation: which factors affect more the phenomenon of interest and how? Which is the direction of the relationship between the phenomenon of interest and a specific concomitant factor?

*UNDERSTAND THE EFFECT OF EACH EXPLANATORY VARIABLE ON Y, WHILE CONTROLLING FOR OTHERS*

## The main ingredients

- *Response variable* is the quantity of main interest (it can be quantitative or qualitative), let's denote it by  $Y$ ;
- *Explanatory variables* (also called, predictors or covariates) are the concomitant factors, let's denote them by  $X_1, \dots, X_{p-1}$ ;

WE HAVE TO MAKE SOME MODEL ASSUMPTION:

- LINEARITY BETWEEN Y AND X
- INDEPENDENCE BETWEEN EACH OBSERVATION
- HOMOSCEDASTICITY : VARIANCE OF ERRORS C IS CONSTANT ACROSS ALL VALUES OF X
- NORMALITY OF ERRORS

## A first formalization

A rough way to formalize the problem is by specifying a functional relationship:

$$Y = g(X_1, \dots, X_{p-1})$$

as an approximation of a possible “true”, yet unknown, relationship.

- $g(\cdot)$  is not known but in some case we can conjecture its shape by consideration on the nature and the characteristics of the phenomenon of interest
- or we can choose a very simple structure such as

$$Y = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + X_{p-1} \beta_{p-1}$$

SOMETIMES  
WE SAY  $C$ ,  
THAT IS SOME ERROR

P := NUMBER OF PREDICTORS

*EXPLANATORY VARIABLES*

*COEFFICIENTS (SLOPES)*

*RESPONSE VARIABLE*

*INTERCEPT TERM*

- usually involved variables are measured on a sample of  $n$  subjects  $(x_{i1}, \dots, x_{ip-1}; y_i), i = 1, \dots, n$ . We want to use these data to explore possible relationship between  $Y$  and the covariates.

FOR MULTIPLE PREDICTORS, REGRESSION CAN BE EXPRESSED COMPACTLY IN MATRIX NOTATION:

$$Y = X\beta + \epsilon \Rightarrow \begin{pmatrix} 1 \\ \vdots \\ n \end{pmatrix} = \begin{pmatrix} 1, 1, \dots, 1, p-1 \\ \vdots \\ n, 1, \dots, n, p-1 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ p-1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

WHERE:

- $Y$ :  $n \times 1$  VECTOR OF RESPONSE VARIABLES
- $X$ :  $n \times (p-1)$  MATRIX OF PREDICTORS (WITH A COLUMN OF ONES FOR THE INTERCEPT)
- $\beta$ :  $(p-1) \times 1$  VECTOR OF COEFFICIENTS
- $\epsilon$ :  $n \times 1$  VECTOR OF RANDOM ERRORS

ESTIMATION: USING THE ORDINARY LEAST SQUARES (OLS) METHOD, THE ESTIMATED COEFFICIENTS  $\hat{\beta}$  ARE

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

PREDICTION: PREDICTED VALUE  $\hat{Y}$  ARE

$$\hat{Y} = X\hat{\beta}$$

## EXPLAINING THE COMPONENTS

- INTERCEPT ( $\beta_0$ )

THE VALUE OF  $Y$  WHEN ALL PREDICTORS ARE ZERO.

FOR EXAMPLE, IN A HOUSE PRICE MODEL, IT COULD REPRESENT THE BASE PRICE OF A HOUSE WITH NO ADDITIONAL FEATURES.

- COEFFICIENTS ( $\beta_j$ )

THE CHANGE IN  $Y$  FOR A ONE-UNIT CHANGE IN  $X_j$ , HOLDING ALL OTHER VARIABLES CONSTANT.

FOR EXAMPLE, IF  $\beta_1 = 50$ , IT MEANS INCREASING  $X_1$  BY 1 UNIT ADD \$50 TO THE PREDICTED VALUE OF  $Y$ .

- ERROR TERM ( $c$ )

REPRESENTS NOISE OR UNMEASURED FACTORS THAT AFFECT  $Y$ .  
THE MODEL ASSUMES  $c \sim N(0, \sigma^2)$

# Simple linear model: a basic example

## Heating consumption in a house depends on temperature?

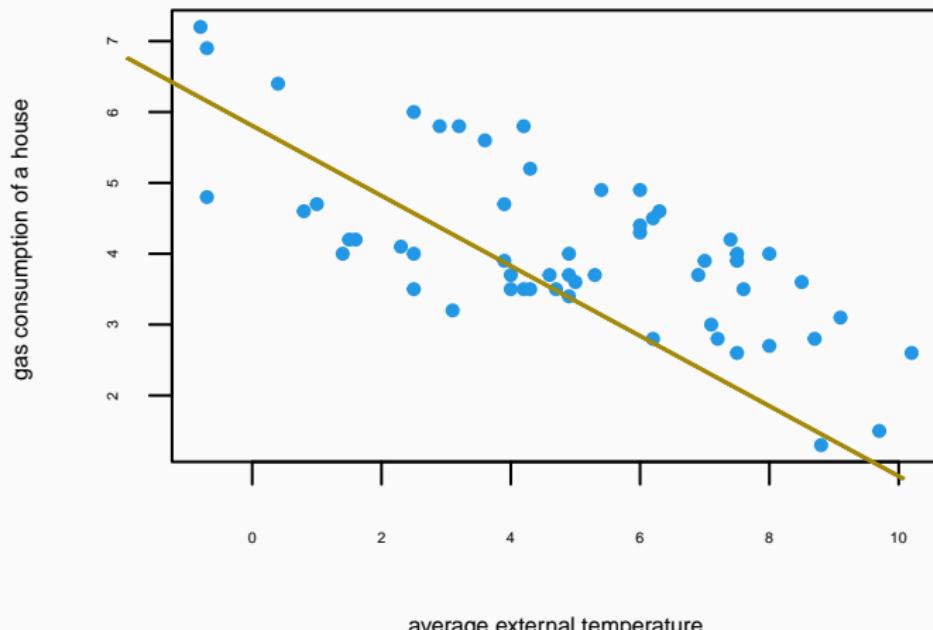
- We are interested in predicting the heating consumption in a house at some time
- We observe the weekly consumption of gas ( $Y$ , in thousands cube feet) over  $n$  weeks ( $y_1, \dots, y_i, \dots, y_n$ )
- A first rough prediction of the gas consumption is:  $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$
- It is sensible to think that the external temperature affects the heating consumption (we expect that the gas consumption decreases as the average external temperature increases).
- In addition to  $y_1, \dots, y_n$ , we observe the average external temperature ( $x_i$ , in Celsius degrees) registered for the same weeks of observation of  $y_i$ . Thus, our sample of data is:

$$(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

# Simple linear model: a basic example

## Heating house's consumption for varying temperature

Heating consumption in a house for varying temperature across 56 weeks



# The simple linear regression model: Model specification

- Can we better describe the conjectured relationship in order to make a more accurate prediction than  $\bar{y}$ ?

gas consumption in week  $i$  =  $g(\text{temperature in the same week})$

$$y_i = g(x_i) \quad E[Y_i | X=x_i] = \mu_i = g(x_i)$$

WE ARE PREDICTING, NOT EXPECTING

- Whatever  $g$  we assume, it will simply be an approximation and we should also take into account that the dependent variable is affected by a random error  $\epsilon_i$

gas consumption in week  $i$  =  $g(\text{temperature in the same week};$  UNTIL WE  
SAY  $\epsilon_i$   
non observed and less relevant factors)

$$y_i = g(x_i; \epsilon_i)$$

- In fact, we may say that, conditional to a given value of  $x_i$ , the expectation of  $y_i$  is:

$$\mu_i = g(x_i)$$

i.e.  $\mu_i$  is the population mean of  $y_i$ , conditional to the value of  $x_i$ .

ASSUMPTIONS ON ERRORS :

- $E[\epsilon_i] = 0$
- $\text{Var}(\epsilon_i) = \sigma^2, \forall i$  THE VARIABILITY IS THE SAME FOR ALL OBSERVATION
- $\epsilon_i \neq \epsilon_j, i \neq j$

## The simple linear regression model: Model specification

- Data suggest that the expected heating consumption can be modeled as a linear function of the external temperature
- A simple model: the straight line

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

or equivalently:

$$\mu_i = \beta_0 + \beta_1 x_i$$

- The model assumes a constant growth rate of gas consumption for decreasing values of temperature: the effect on  $\mu$  of a constant increase of  $x$  is the same whatever is  $x$ .

## The simple linear regression model: Model assumptions

- The model is correctly specified:

$$\begin{aligned}y_i = g(x_i; \epsilon_i) &= \beta_0 + \beta_1 x_i &+ \epsilon_i \\&= \text{systematic component} + \text{stochastic component}\end{aligned}$$

- systematic component: predictors are under the control of the researchers (non stochastic)
- stochastic component:
  - the error terms have zero mean
    - ↪ it does not include further systematic terms
    - ↪  $\mu_i = \beta_0 + \beta_1 x_i$
  - the error terms have constant variance and are uncorrelated
  - (useful, albeit to some extent not necessary, the error terms are normally distributed)

$$\epsilon_i \sim \text{i.i.d. } \mathcal{N}(0, \sigma^2)$$

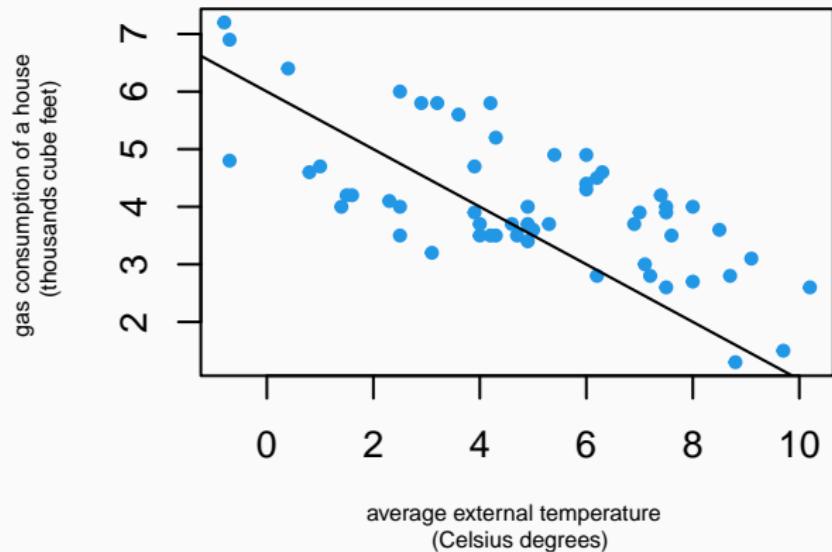
With the normality assumption

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

$Y_i$  and  $Y_j$  uncorrelated  $\forall i \neq j$ .

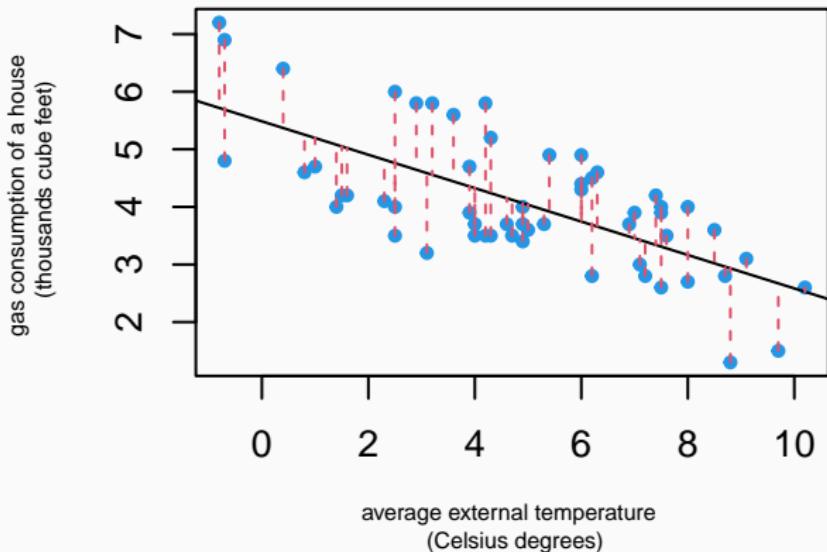
# The simple linear regression model: How to choose the best line?

- How to use data to estimate the unknown parameters  $\beta_0$  and  $\beta_1$ ?



## How to choose the best line? The least squares criterion

- Choose the line which minimizes the sum of squared residuals



## How to choose the best line? The least squares criterion

- Choose the line which minimizes the sum of squared residuals:

$$L(\beta_0, \beta_1) = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

THE RESIDUAL IT'S THE DISTANCE  
BETWEEN THE OBS POINT AND THE  
POINT ON THE LINE FOR A GIVEN X

- The minimization problem has the following solution:

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} \text{ (slope)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ (intercept)}$$

It is important to remind that if we assume also that the random component of the model is normally distributed, then maximum likelihood estimation lead to the same solution\*\* (least squares is also the maximum likelihood solution). => WHAT IS THE ADVANTAGE?  
THE ADVANTAGE IS THAT FROM THE LIKELIHOOD THEORY WE CAN DERIVE SOME SPECIFIC RESULTS ON  $\beta_0$  AND  $\beta_1$  AND COMPUTE CONFIDENCE INTERVALS OR DO HYPOTHESIS TESTING

In the gas consumption example the two estimated coefficients are:

```
## [1] 5.4861933 -0.2902082
```

LET'S FIND THE LEAST SQUARE ESTIMATOR

$$\begin{cases} \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^n (x_i (y_i - (\beta_0 + \beta_1 x_i))) = 0 \end{cases} \quad (1)$$

$$(1) \quad \sum_{i=1}^n y_i - \sum_{i=1}^n \beta_0 - \sum_{i=1}^n \beta_1 x_i = 0$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n \beta_1 x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$(2) \quad \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \beta_0 - \sum_{i=1}^n \beta_1 x_i^2$$

$$\sum_{i=1}^n x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^n x_i - \sum_{i=1}^n \beta_1 x_i^2 = 0$$

$$\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \sum_{i=1}^n \beta_1 x_i^2 = 0$$

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

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

$$\frac{\sum y_i - n \bar{y}}{\sum x_i - n \bar{x}}$$

IF YOU MULTIPLY  $\beta_1$  BY

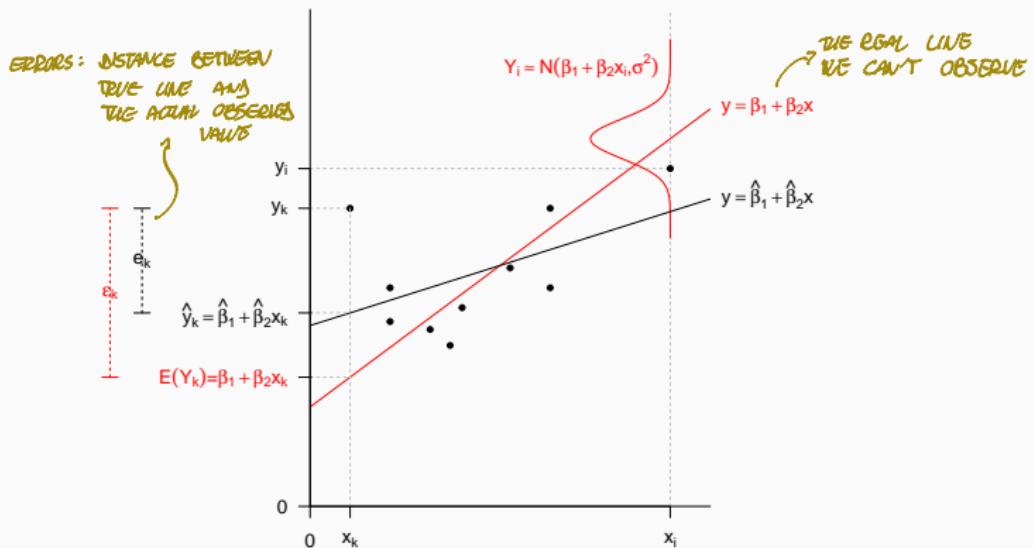
$$\frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(y)}}$$

$$\frac{\text{Cov}(x, y)}{\text{Var}(x)} \cdot \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(y)}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}} \cdot \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(y)}} = P_{xy} \cdot \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(x)}}$$

→ YOU OBTAIN THE CORRELATION COEFFICIENT

WITH THIS FORMULATION WE HIGHLIGHT THAT THE SIGN OF THE SLOP OF THE REGRESSION LINE IS GIVEN BY THE CORRELATION COEFFICIENT

# True and estimated relationship



# Inference

Statistical tests allow us to draw general considerations about the model, valid not only for the sample at hand

- Is the model useful somehow?
  - ↪ Test the usefulness of the whole model
- Does the explanatory variable  $X$  really affect the response variable?
  - ↪ Test the significance of a single predictor
  - ↪ (in the simple linear regression model the two above are equivalent)

## Estimating $\sigma^2$

The estimate  $\hat{\sigma}^2$  is  $\frac{1}{n} \sum_1^n e_i^2$  where  $e_i = (y_i - \hat{y}_i)$  are the residuals,

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

$$\hat{\sigma}^2 = \sum_i^n y_i^2/n - \bar{y}^2 - \hat{\beta}_1^2 \left( \sum_i^n x_i^2/n - \bar{x}^2 \right)$$

An unbiased estimate is

$$s^2 = \frac{n}{n-2} \hat{\sigma}^2$$

# Testing usefulness of the overall model: A first useful index

- $y$  varies in the population; its variability may be measured by:

$$\sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Total Sum of Squares (SS)}$$

OF THE MODEL

IT HELPS ASSES HOW WELL  
THE MODEL EXPLAINS THE DATA

- The whole variability of  $y$  may be decomposed as follows:

- variability of  $y$  explained by the model:

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{Regression SS}$$

↑ PREDICTED VALUE  
↓ MEAN OF THE OBS. VALUES

IT MEASURES HOW MUCH OF THE  
VARIABILITY IN  $y$  IS EXPLAINED BY  
THE RELATIONSHIP WITH THE  
PREDICTOR VARIABLES  $x_1, x_2, \dots, x_k$

↓  
A HIGH VALUE MEANS THE MODEL  
EXPLAINS A LARGE PROPORTION OF  $y$  VARIABILITY

- residual variability (due to the chance):

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

↑ RESIDUAL  
↓

IT MEASURES THE REMAINING VARIABILITY  
AFTER ACCOUNTING FOR THE EFFECT  
OF THE PREDICTORS

↓  
A SMALL VALUE INDICATES THAT THE MODEL  
FITS THE DATA WELL, LEAVING LITTLE  
UNEXPLAINED VARIABILITY

- It can be shown that, in the linear model:

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

Total SS

Residual SS

Regression SS

↑  
IT MEASURES HOW MUCH THE OBSERVED  
 $y$  VALUES DEVIATE FROM THE MEAN?

↓  
UNEXPLAINED VARIABILITY (RESIDUES)

↓  
VARIABILITY EXPLAINED BY THE MODEL

## Testing usefulness of the overall model: A first useful index

- If the model is good:  
    → Residual SS is small compared to Total SS  
    → Regression SS is the main portion of Total SS
- Coefficient of Determination:

$$R^2 = \frac{\text{RegressionSS}}{\text{TotalSS}} = 1 - \frac{\text{ResidualSS}}{\text{TotalSS}}$$

DON'T EXPLAIN ANYTHING      EXPLAIN EVERYTHING

- $0 \leq R^2 \leq 1$
- The lower  $R^2$  the worse the fitted model

In the heating consumption example: Total SS, Regression SS and  $R^2$  are respectively

```
## [1] 75.0142857 35.0194175 0.4668366
```

The model can explain about the 47% of the variability of  $y$ .

## HOW THE SUM OF SQUARES WORK IN PRACTICE

### 1. COMPUTE TSS :

- MEASURE TOTAL VARIABILITY IN  $Y$
- BASELINE COMPARISON BEFORE FITTING THE MODEL

### 2. COMPUTE REGRESSION SS :

- QUANTIFIES HOW MUCH OF TSS IS CAPTURED BY THE MODEL

### 3. COMPUTE RESIDUAL SS :

- MEASURE LEFTOVER VARIABILITY (ERRORS)

### 4. EVALUATE $R^2$ :

- ASSESS THE PROPORTION OF  $Y$ 'S VARIABILITY EXPLAINED BY THE PREDICTORS

## Testing usefulness of the overall model: The F test

- The quantities above may be used to build a formal statistical test:

$$H_0 : \mu_i = \beta_0 \quad \text{vs} \quad H_1 : \mu_i = \beta_0 + \beta_1 x_i$$

BASICALLY WE ARE ASSUMING  $\beta_1 \neq 0$

- The  $F$  statistic is the ratio of

- the explained variability (as reflected by  $R^2$ ) and
- the unexplained variability (as reflected by  $1 - R^2$ )

suitably adjusted according to the number of observations ( $n$ ) and the number of estimated parameters ( $p = 2$ ):

$$F = \frac{R^2}{(1 - R^2)} \frac{n - p}{p - 1}$$

↑ DEGREES OF FREEDOM  
OF  $R^2$  AND  $(1-R^2)$

- The larger the  $F$  statistic, the more useful the model.
- Under the assumption of gaussianity of the error term, the *probability distribution* of  $F$  is known and it allows us to define *critical values* and *p-values*.

## Testing single predictors

- It is of interest to test whether the explanatory variable  $X$  really affects the response variable
- A formal statistical test may be built to check:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

THE PREDICTOR  $X_1$  HAVE  
A SIGNIFICANT LINEAR  
RELATIONSHIP WITH  $Y$

- The  $t$  statistic is the ratio:

THE PREDICTOR  $X_1$  DOES NOT HAVE A  
SIGNIFICANT LINEAR RELATIONSHIP  
WITH THE RESPONSE VARIABLE  $Y$

$$t = \frac{\hat{\beta}_1}{\text{standard error}(\hat{\beta}_1)}$$

WHICH MEASURES THE  
VARIABILITY OF THE  
COEFFICIENT ESTIMATE  
ACROSS SAMPLES

- The larger the  $t$  statistic, the more evidence against  $H_0$
- Under the assumption of gaussianity of the error term, the *probability distribution* of  $t$  is known and it allows us to define *critical values* and *p-values*.
- Note that for simple regression the  $t$  test is equivalent to the  $F$  test

$$F = t^2$$

# Inference: The Heating consumption example

```
##  
## Call:  
## lm(formula = y ~ x)  
## Residuals: summary of the residuals (Y - \hat{Y})  
##      Min    1Q Median    3Q   Max  
## -1.6324 -0.7119 -0.2047  0.8187  1.5327  
## CLICK IF THEY ARE SYMMETRIC AND IF THE MEDIAN IS APPROXIMATELY ZERO  
##  
## Coefficients: most important only if it makes sense to set to zero the values of covariates (ex: x := sample mean)  
## only p <= 0.05 Estimate Std. Error t value Pr(>|t|) p-value associated with the t-test.  
## (Intercept) 5.4862     0.2357  23.275 < 2e-16 ***  
## x          -0.2902     0.0422 -6.876 6.55e-09 ***  
## --- negative relationship difference in Y by increasing X of 1 unit evaluate test statistic for H_0: \beta_1 = 0 A small p-value indicates that the predictor x_1 significantly contributes to explaining Y  
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (P = 2 * P(T > |t|))  
## how much the observed value of Y deviate from the fitted value on average small p \Rightarrow \beta_1 \neq 0  
## Residual standard error: 0.8606 on 54 degrees of freedom  
## Multiple R-squared: 0.4668, Adjusted R-squared: 0.457 \rightarrow we will see it later  
## F-statistic: 47.28 on 1 and 54 DF, p-value: 6.545e-09  
  
//  
\sqrt{\frac{\text{INTERCEPT\_Z-VALUE}}{\text{X\_T-VALUE}}}
```

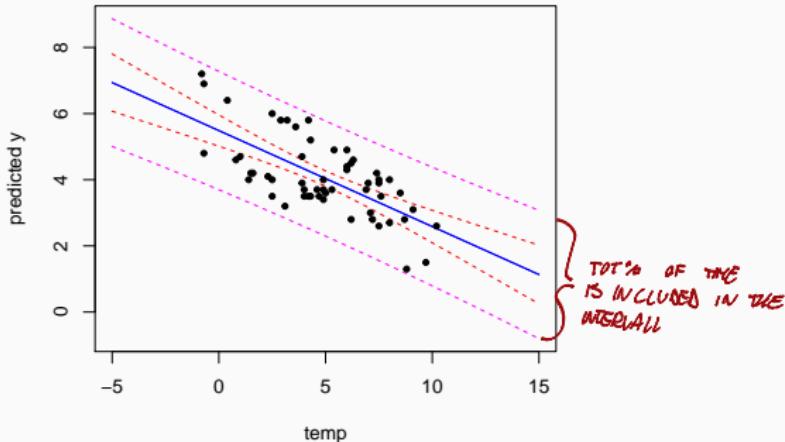
# Prediction

Confidence interval for the mean  $Y_0$  (red):

$$\hat{Y}_0 \pm t_{n-2,1-\alpha/2} \sqrt{s^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2} \right)}$$

Confidence interval for prediction of a new value  $\hat{y}_0$  (purple):

$$\hat{y}_0 \pm t_{n-2,1-\alpha/2} \sqrt{s^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i^n (x_i - \bar{x})^2} \right)}$$



## Model checking

- The least squares line is the line which best fits the data at hand but... *best* is not (necessarily) *good*.
- How to establish if the estimated model is a good one?
  - ↪ Residuals are an estimate of the error components  $\epsilon_i$

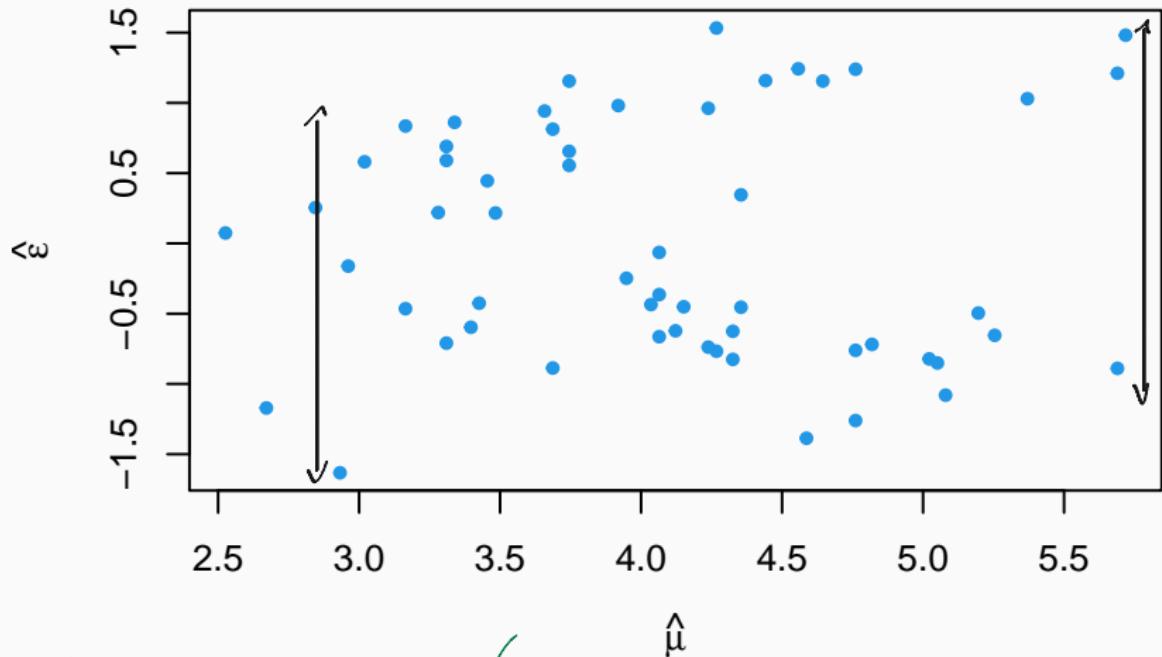
$$\begin{aligned} e_i &= \hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - \hat{\mu}_i \quad i = 1, \dots, n \end{aligned}$$

- ↪ Residual analysis allows to check if the model assumptions are met and if the model is good

## Model checking - Residual plots

- How should be the residuals in a good model?
  - $\mapsto$  both positive and negative (around zero)
  - $\mapsto$  small
  - $\mapsto$  have constant variability
  - $\mapsto$  scattered at random (if the model is well specified the amount of variability of  $y$  not explained by  $x$  must be due to the chance only  $\Rightarrow$  the residuals do not show any regularity)
- Residual plots:
  - $\mapsto \hat{\epsilon}_i$  vs  $\hat{\mu}_i$
  - $\mapsto \hat{\epsilon}_i$  vs each  $x_i$
  - $\mapsto$  Normal QQ-plot of  $\hat{\epsilon}_i$
- Other plots based on residuals
  - $\mapsto$  Leverages
  - $\mapsto$  Cook distances

## Model checking - Residual plots



- THE MEAN SHOULD BE AROUND ZERO ✓
- IS THE VARIANCE CONSTANT? NOT SO MUCH! SMALLER VARIANCE COMPARED TO LARGER VALUES
- THEY DON'T SPREAD OUT EVENLY BUT IT DOESN'T SEEM TO BE A PATTERN

## Model checking

- Although not too bad, the residual plot suggest some problems
  - The variability of the residuals is not constant
  - There seem to be two clusters of residuals
    - ↪ Positive residuals increase for increasing  $\hat{\mu}$
    - ↪ Negative residuals decrease for increasing  $\hat{\mu}$
  - It seems that for two groups of observations the estimated relationship between heating consumption and external temperature is different from the estimated one (as the residuals are still some function of the temperature via  $\hat{\mu}$ ).
- Have we forgotten anything relevant?

## Multiple linear model

---

## The (multiple) linear regression model: Introduction

- In fact, we get to know that during the observation time, there has been an insulation intervention on the house so that in the last 30 weeks the house was insulated.
- Data at hand thus become:

WE USE A CODE TO  
ASSIGN ISOLATE = 0 AND  
NOT ISOLATE = 1

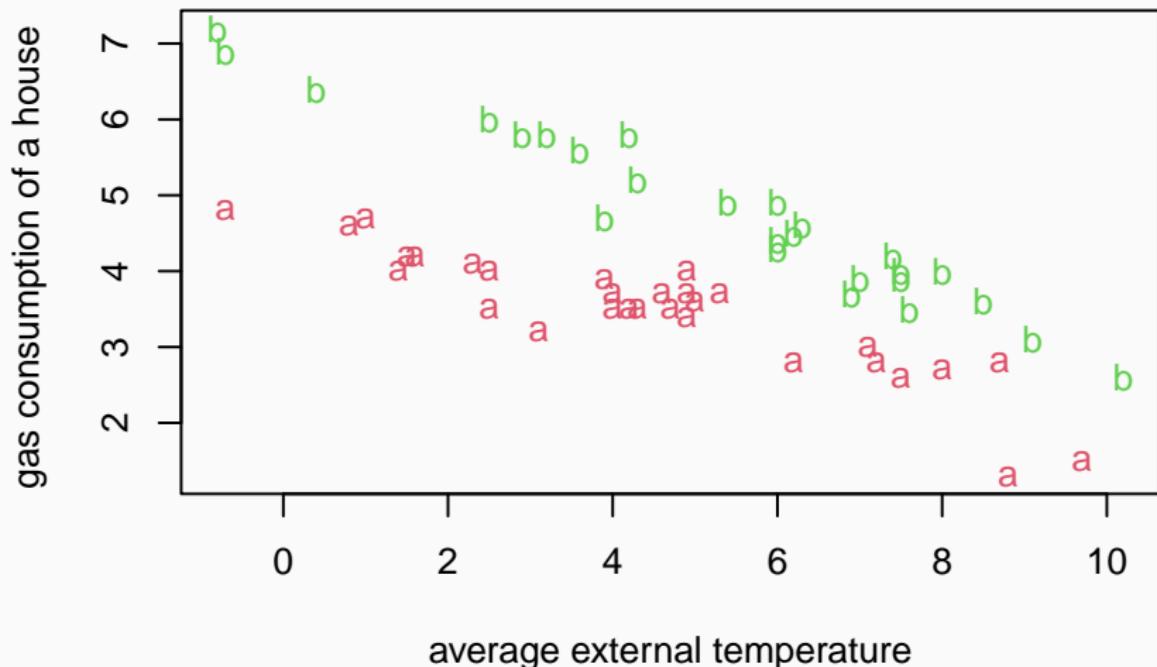
$$(x_1, z_1, y_1), (x_2, z_2, y_2), \dots, (x_i, z_i, y_i), \dots, (x_n, z_n, y_n)$$

with  $z_i$  = "before insulation" for  $i = 1, \dots, 26$   
and  $z_i$  = "after insulation" for  $i = 27, \dots, 56$

- It is sensible to expect that the isolation intervention has an impact on the mean heating consumption and that after the intervention the heating consumption is lower than before it.

# The (multiple) linear regression model: Introduction

WE CAN SEE A CLEAR DIFFERENCE  
BETWEEN THE BEFORE AND AFTER ISOLATE



# Model specification

WE UPDATE THE LINEAR REGRESSION

- In the light of the availability of the additional variable  $z$  the model can be specified as follows:

gas consumption in a week =  $g(\text{temperature in the same week, before/after intervention})$   
non observed and less relevant factors)

$$y_i = g(x_i, z_i; \epsilon_i)$$

- The natural extension of the straight line model is the following *(multiple) linear model*

Condition  
or not

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \quad (1)$$

or equivalently:

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

## Qualitative predictors

- The additional variable has a qualitative nature as it takes values "before intervention" and "after intervention"  $\Rightarrow$  the specified model does not make sense in the current form as  $z_i$  is not a number
- The standard way to overcome the problem is to introduce an auxiliary variable, an indicator variable (econometricians call it *dummy*):

$$d_i = \begin{cases} 0 & \text{if } z_i = \text{"before intervention"} \\ 1 & \text{if } z_i = \text{"after intervention"} \end{cases}$$

- The (1) then becomes:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i \\ &= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{before the intervention} \\ &= (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i \quad \text{after the intervention} \end{aligned}$$

*→ THE NEW INTERCEPT*  
*↳ EFFECT OF AN INTERVENTION*

- In other words the introduction of the indicator variable  $d$  gives rise to two parallel straight lines, one for each value of  $z_i$ :

WE WILL HAVE 2 LINES :  
• BEFORE INTERVENTION  
• AFTER INTERVENTION

WE EXPECT THE SAME SLOPE BUT DIFFERENT IN THE INTERCEPTION TO Y-AXIS

## Qualitative predictors (factors)

Model estimation via least squares easily extends to the multiple linear model:

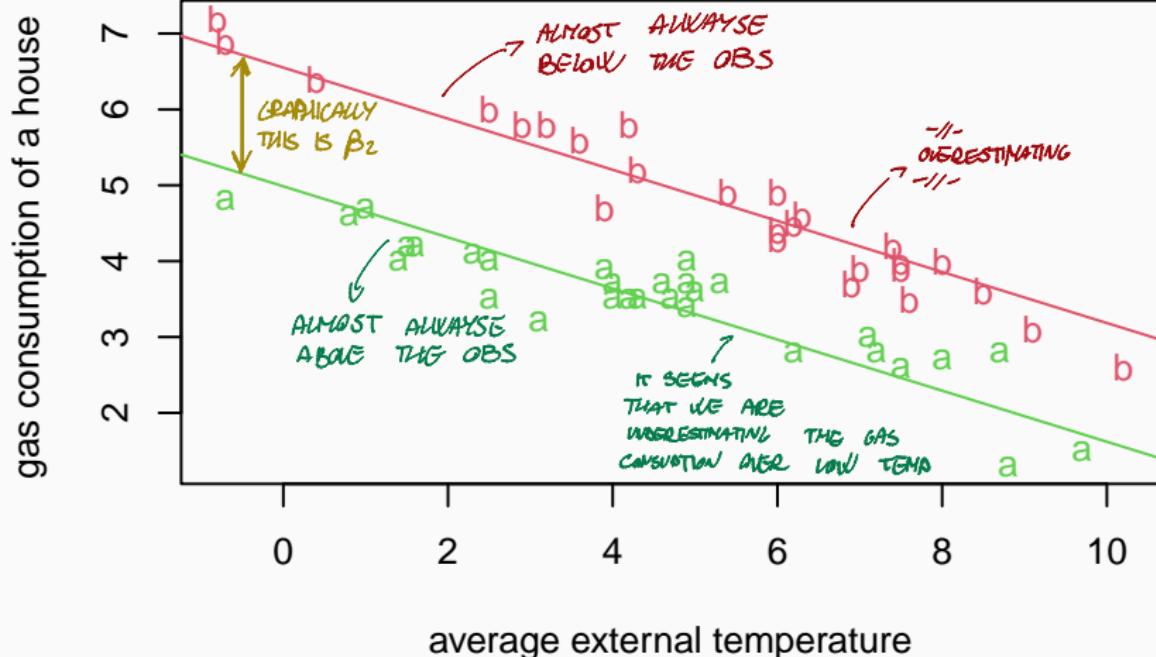
$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i + \beta_2 d_i))^2$$

The three coefficients are respectively:

```
## [1] 6.551329 -0.336697 -1.565205
```

↳ THIS IS  $\beta_2$

## Qualitative predictors (factors)



## Interpreting the model

- $\hat{\beta}_0$ : expected response value when all the predictors are set at zero (if it does make sense and 0 is in the range of observed predictors)
  - ↪ If the external temperature is 0 degree, and before the isolation intervention ( $d_i = 0$ ), the expected consumption of gas is about 6.6 thousand cube feet
- $\hat{\beta}_j$  ( $j > 1$ ): expected change of  $y$  when the  $j$ -th predictor increases by 1 unit and all the other predictors are kept constant:
  - ↪ If the external temperature increases by 1 degree, the expected consumption of gas decreases by about 0.34 thousand cube feet, independently of the isolation intervention
  - ↪ If the house gets isolated ( $d_i$  passes from 0 to 1), the expected consumption of gas decreases by about 1.57 thousand cube feet independently of the external temperature

## Interpreting the model

- The estimated line can be used to get a prediction of  $y$  for any value of the predictors (in the range of observed values)

$$\hat{y} = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 d$$

→ What is the expected gas consumption when the external temperature is 5?

- if the house is not insulated:

$$\hat{\mu} = 6.551 - 0.3367 \cdot 5 = 4.8675 \text{ thousand cube feet}$$

- if the house is insulated:

$$\hat{\mu} = 6.551 - 1.565 - 0.3367 \cdot 5 = 3.3025 \text{ thousand cube feet}$$

# Inference

Statistical tests allow us to draw general considerations about the model, valid not only for the sample at hand

- Is the model useful somehow?
  - ↪ Compute the *adjusted R*<sup>2</sup>: a suitable adjustment of  $R^2$  which penalises additional explanatory variables (descriptive)
  - ↪ Test the usefulness of the whole model: the *F* test
    - $H_0 : \mu_i = \beta_0 \Leftrightarrow \beta_1 = \beta_2 = 0$  vs
    - $H_1$  : at least one between  $\beta_1$  and  $\beta_2$  is not 0
- Does the  $j$ -th explanatory variable really affect the response variable?
  - ↪ Test the significance of a single predictor: the *t* test

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

# Inference: The Heating consumption example

```
##  
## Call:  
## lm(formula = y ~ x + z)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.74236 -0.22291  0.04338  0.24377  0.74314  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.55133   0.11809   55.48 <2e-16 *** } THIS IS THE PROBABILITIES  
## x          -0.33670   0.01776  -18.95 <2e-16 *** } OF HAVING A VALUE DIFFERENT  
## zafter     -1.56520   0.09705  -16.13 <2e-16 *** } FROM THE ONE IT FOUND  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.3574 on 53 degrees of freedom  
## Multiple R-squared: 0.9097, Adjusted R-squared:  0.9063  
## F-statistic: 267.1 on 2 and 53 DF,  p-value: < 2.2e-16
```

AT LEAST 1 PREDICTOR  
IS STATISTICALLY  
SIGNIFICANT

## Model checking

BEFORE JUMPING TO CONCLUSION WE HAVE TO CHECK THE RESIDUALS

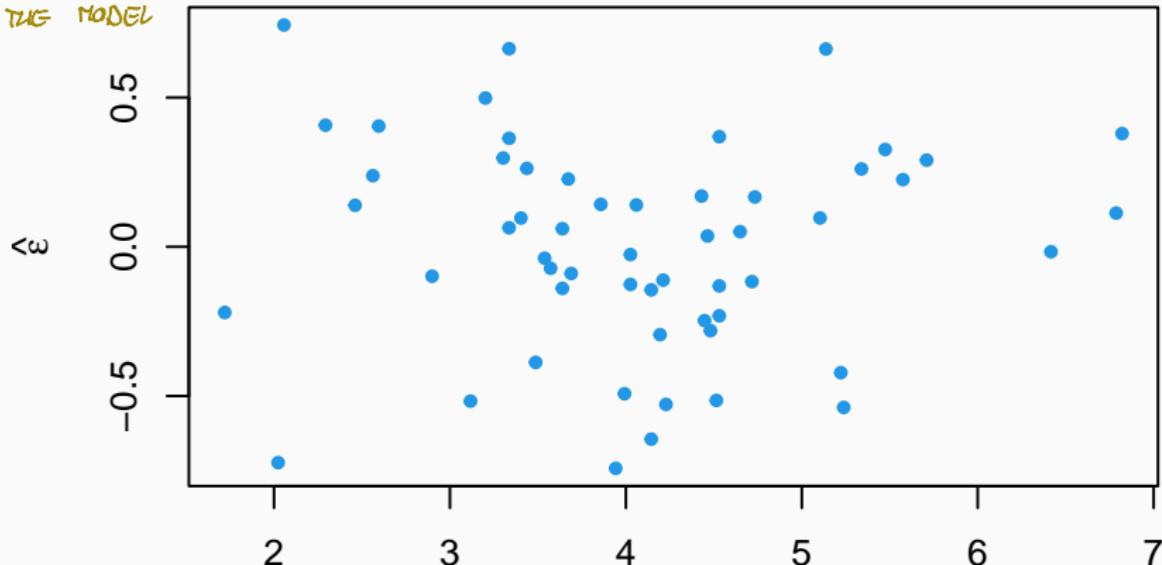
- The least squares linear model is the estimated linear model which best fits the data at hand but... *best* is not (necessarily) *good*.
- How to establish if the estimated model is a good one?
  - Residuals are built as in the simple linear model and have the same interpretation (estimate of the error components  $\epsilon_i$ )

$$\begin{aligned}\hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 d_i) \\ &= y_i - \hat{\mu}_i \quad i = 1, \dots, n\end{aligned}$$

- Residual analysis allows to check if the model assumptions are met and if the model is good

## Model checking - Residual plots

- 1)  $E[\hat{e}] \sim 0$  AND IT IS
- 2) WE DON'T HAVE NO MORE THE BAD VARIANCE
- 3) IT SEEMS EXIST A PATTERN, BUT WE DON'T HAVE ANY OTHER VARIABLE TO INCLUDE IN TO THE MODEL



IF WE CHECK THE PLOT WITH THE GREEN AND  $\hat{\mu}$  RED LINE, WE CAN SEE WE ARE NOT FITTING WELL ENOUGH THE DATA

## Model checking

- The “two clusters” problem is not present anymore
- The non-constant variability of the residuals is reduced
- The residual plot still suggests some problems: the residuals are still some function of the predictors (via  $\hat{\mu}$ )
  - ↪ Positive residuals for small and large  $\hat{\mu}$
  - ↪ Negative residuals for intermediate values of  $\hat{\mu}$
- Have we forgotten anything relevant?

## The interaction term

- Going back to the scatterplot of the data...
  - ↪ the line corresponding to the observations before insulation tends to underestimate the gas consumption for small lower temperatures and overestimate it for higher temperatures
  - ↪ the line corresponding to the observations after insulation tends to overestimate the gas consumption for small lower temperatures and underestimate it for higher temperatures
- Data show that not only the intercepts but also slopes might be different before and after the insulation intervention
  - ↪ *interaction term*: different relationship between  $Y$  and  $X$  for different values of  $d$

## The interaction term: Formalization

- The linear model with interaction of the predictors may be formalized as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (x_i \cdot d_i) + \epsilon_i$$

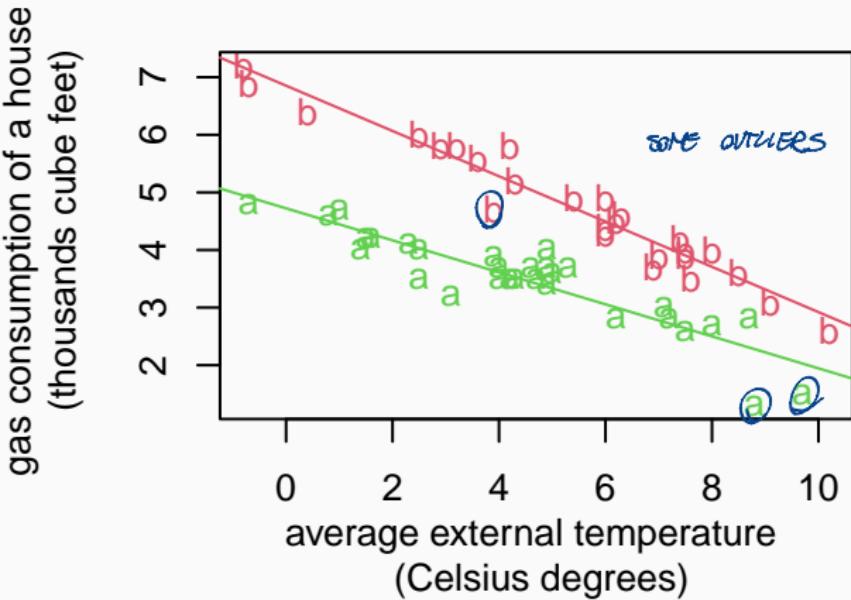
=  $\beta_0 + \beta_1 x_i + \epsilon_i$  before the intervention, when  $d_i = 0$

=  $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \epsilon_i$  after the intervention, when  $d_i = 1$

- The model is estimated by the least squares criterion, as before

NOW WE HAVE 2 REGRESSION LINES WITH  
DIFFERENT SLOPES AND DIFFERENT  
INTERCEPTION

## The estimated interaction model

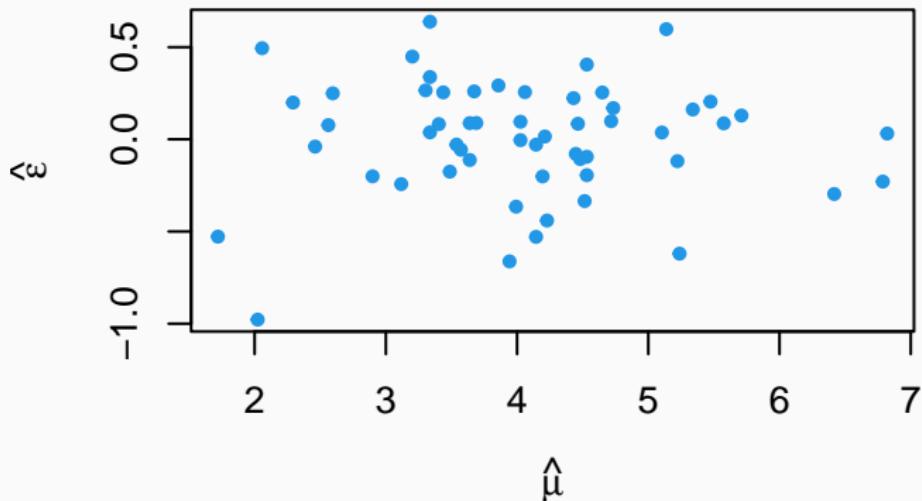


# The estimated interaction model

```
##  
## Call:  
## lm(formula = y ~ x * z)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -0.97802 -0.18011  0.03757  0.20930  0.63803  
##  
## Coefficients:  
## ETSI wieder raus Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.85383   0.13596 50.409 < 2e-16 ***  
## x z=0       -0.39324   0.02249 -17.487 < 2e-16 ***  
## zafter      -2.12998   0.18009 -11.827 2.32e-16 ***  
## x:zafter     0.11530   0.03211  3.591 0.000731 ***  
## --- 'x ≠ 0'  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.323 on 52 degrees of freedom  
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235  
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

## Model checking - Residual plots

- 1)  $E[\hat{\epsilon}] \approx 0$  IT SEEKS LIKE
- 2) WE HAVE A GOOD VARIABILITY WITH SOME OUTLIER
- 3) THE PATTERN SEEKS DISAPPEARED



## Generalization to multiple regression model

Given a response variable  $Y$  and  $p - 1$  predictors  $X_1, \dots, X_{p-1}$ , observed on a sample of  $n$  subjects, the multiple linear model is specified as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_{p-1} x_{ip-1} + \epsilon_i$$

- The model is correctly specified:

$$\begin{aligned} y_i = g(x_{i1}, \dots, x_{ip-1}; \epsilon_i) &= \beta_0 + \beta_1 x_i + \dots + \beta_{p-1} x_{ip-1} + \epsilon_i \\ &= \text{systematic component} + \text{stochastic component} \end{aligned}$$

- systematic component:
  - ↪ predictors are under the control of the researchers (non stochastic)
  - ↪ predictors are not collinear (i.e. not highly correlated)

- stochastic component:
  - the error terms have zero mean
    - ↪ do not include further systematic terms
    - ↪ this implies that  $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i + \dots + \beta_{p-1} x_{ip-1}$
  - the error terms have constant variance
  - the error terms are independent
  - (optional: the error terms are normally distributed)