Linear models

(An Introduction)

N. Torelli, L. Egidi, G. Di Credico

Spring 2020

University of Trieste

Introduction to linear models

Multiple linear model

Introduction to linear models

Linear regression model

Linear regression model is one of the basic tool 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
- a **response variable** (dependent variable).

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

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

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?

The main ingredients

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

A first formalization

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

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

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

- $g()$ 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

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

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

Simple linear model: a basic example

Heating consumption in a house depends on temperature?

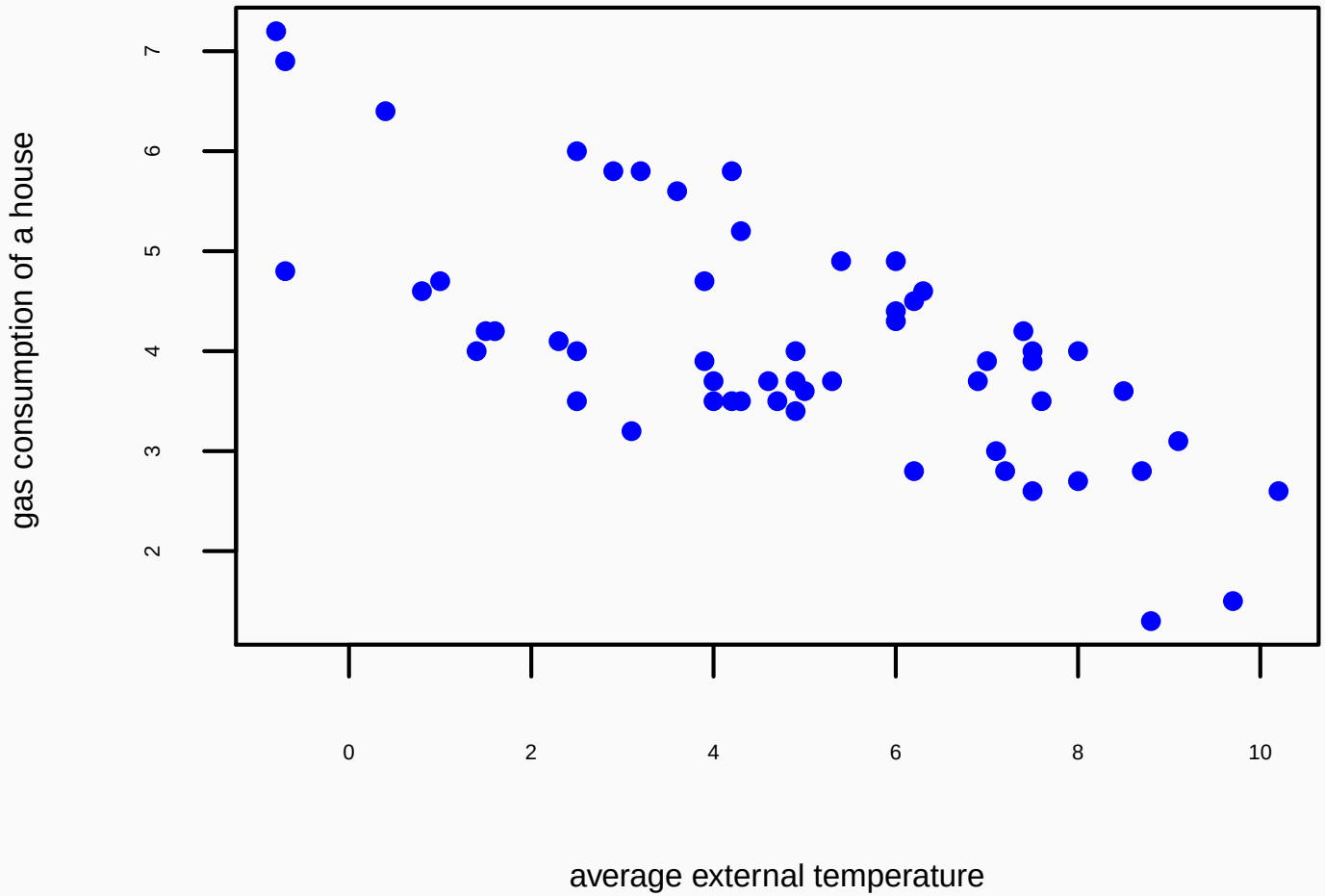
- We are interested in predicting the heating consumption in a house at some time
- We observe the weekly consumption of gas (Y, in thousands cube feet) over n weeks $(y_1, \ldots, y_i, \ldots, y_n)$
- A first rough prediction of the gas consumption is: $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$
- 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 $y_1, \ldots, y_n$, 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:

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

Simple linear model

A basic example: Heating consumption in a house 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 $\overline{y}$?

$$\begin{aligned} \text{gas consumption in week } i &= g(\text{temperature in the same week}) \\ y_i &= g(x_i) \end{aligned}$$

- 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 ϵ_i

$$\begin{aligned} \text{gas consumption in week } i &= g(\text{temperature in the same week;} \\ &\quad\ \text{non observed and less relevant factors}) \\ y_i &= g(x_i; \epsilon_i) \end{aligned}$$

- 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. μ_i is the population mean of y_i, conditional to the value of x_i.

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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

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 μ of a constant increase of x is the same whatever is x

The simple linear regression model: Model assumptions

- The model is correctly specified:

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

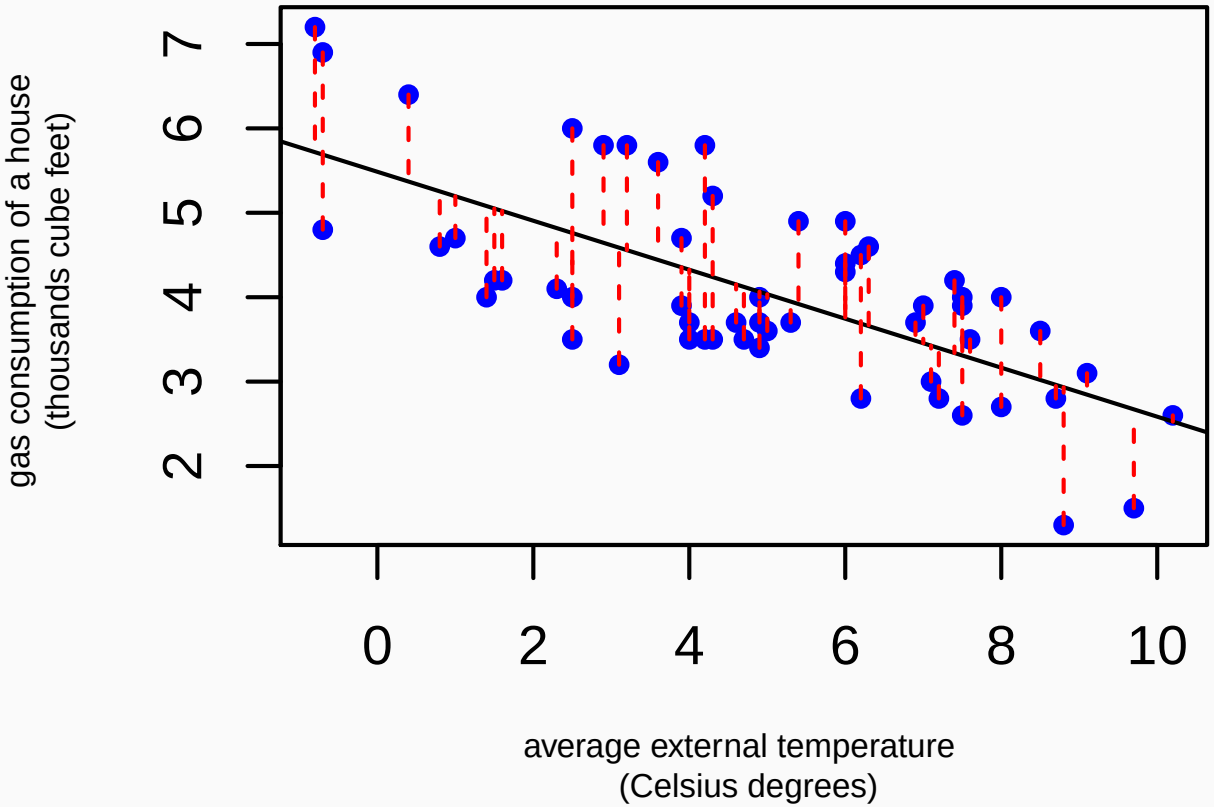
- systematic component: predictors are under the control of the researchers (non stochastic)
- stochastic component:
  - the error terms have zero mean
    - $\mapsto$ it does not include further sisthematic terms
    - $\mapsto$ $\mu_i = \beta_0 + \beta_1 x_i$
  - the error terms have constant variance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

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

With the nomality assumption

How to choose the best line? The least squares criterion

- Choose the line which minimizes the sum of squared residuals




```
## [1] 39.99487
```

How to choose the best line? The least squares criterion

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

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

- The minimization problem has the following solution:
  $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$ (slope)
  $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ (intercept)
  It is important to remember that if we assume also that the random component of the model is normally distributed than maximum likelihood estimation lead to the same solution (least squares is also the maximum likelihood solution).

In the gas consumption example the two estimated coefficients are:

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

##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?

  $\mapsto$ Test the usefulness of the whole model

- Does the explanatory variable X really affect the response variable?

  $\mapsto$ Test the significance of a single predictor

$\mapsto$ (in the simple linear regression model the two above are equivalent)

Estimating σ^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,

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

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

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:

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

- The whole variability of y may be decomposed as follows:
  - variability of y explained by the model:

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

  - residual variability (due to the chance):

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

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

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{Total SS}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Residual SS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}_{\text{Regression SS}}$$

Testing usefulness of the overall model: A first useful index

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

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

  - $\mapsto$ $0 \leq R^2 \leq 1$
  - $\mapsto$ 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.

Testing usefulness of the overall model: The F test

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

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

- The F statistic is the ratio of
  - $\mapsto$ the explained variability (as reflected by R^2) and
  - $\mapsto$ the unexplained variability (as reflected by $1 - R^2$)

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

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

- 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 if the explanatory variable X really affect the response variable
- A formal statistical test may be built to check:

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

- The t statistic is the ratio:

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

- 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

Inference: The Heating consumption example

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6324 -0.7119 -0.2047  0.8187  1.5327
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4862     0.2357  23.275  < 2e-16 ***
## x            -0.2902     0.0422  -6.876 6.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8606 on 54 degrees of freedom
## Multiple R-squared:  0.4668, Adjusted R-squared:  0.457
## F-statistic: 47.28 on 1 and 54 DF,  p-value: 6.545e-09
```

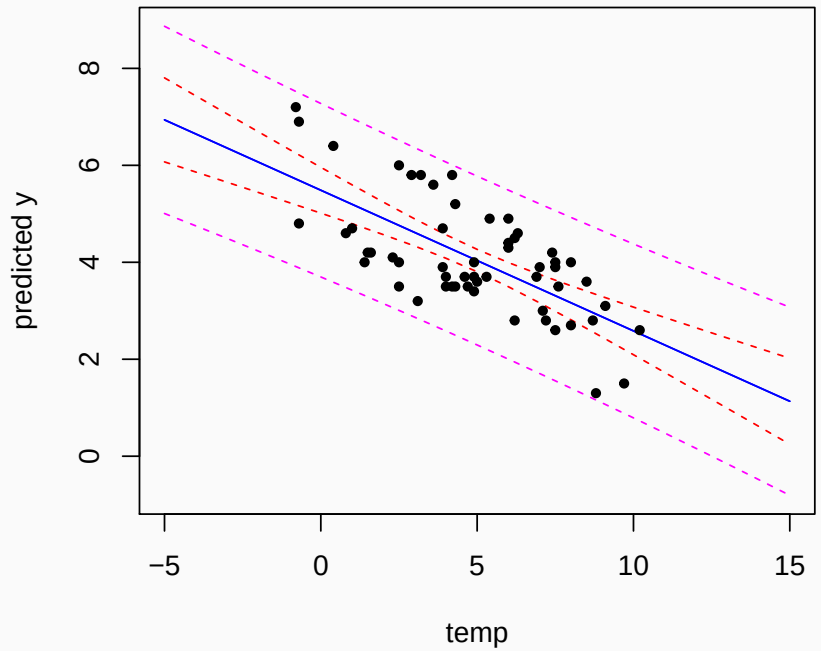
Prediction

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?

  $\mapsto$ Residuals are an estimate of the error components ϵ_i

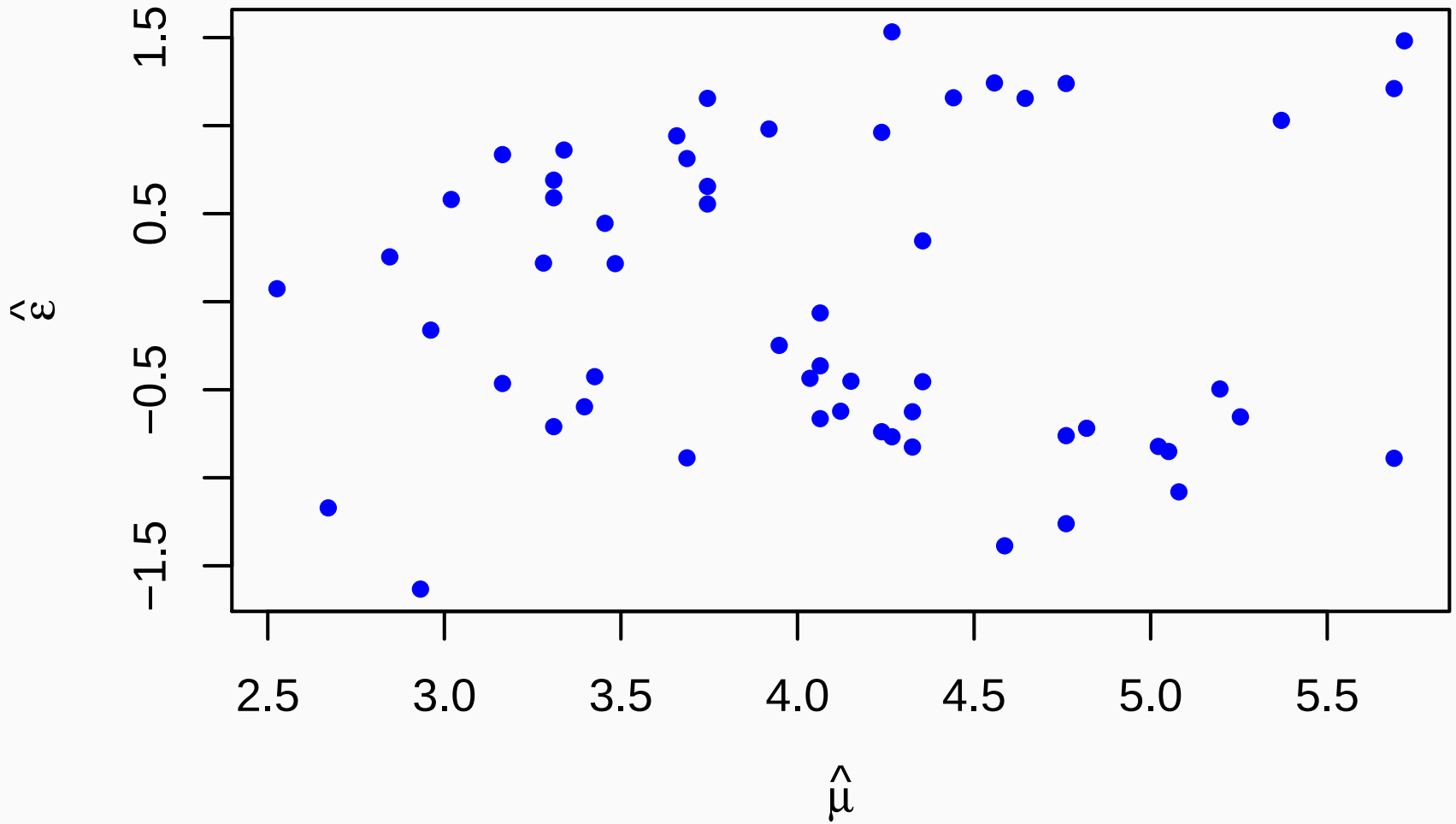
$$\begin{aligned} e_i = \hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= y_i - \hat{\mu}_i \quad i = 1, \ldots n \end{aligned}$$

  $\mapsto$ 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



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
    - $\mapsto$ Positive residuals increase for increasing $\hat{\mu}$
    - $\mapsto$ 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

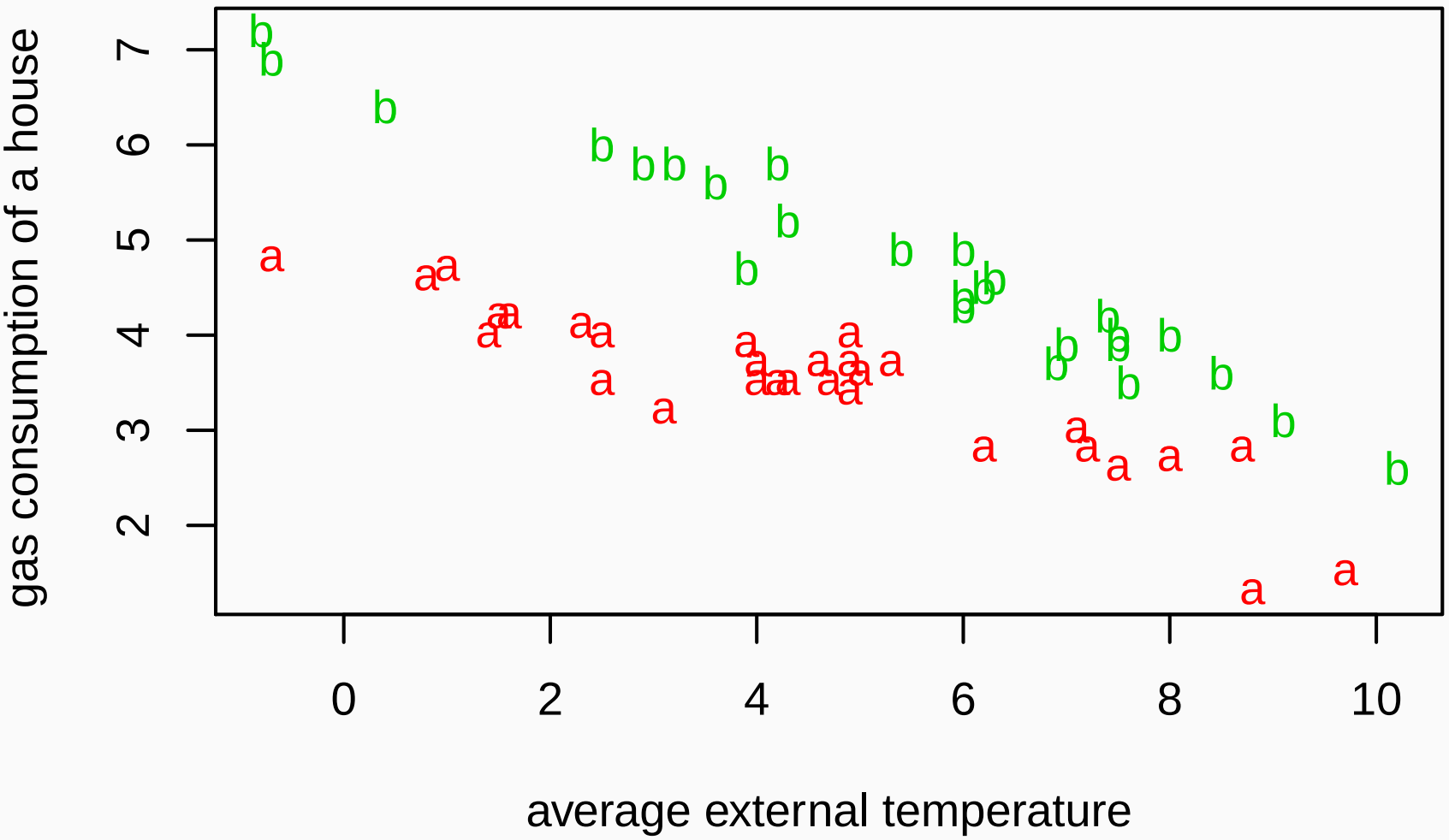
- In fact, we get to know that during the observation time, there has been an insulation intervention on the house so that in the last 30 weeks the house was insulated.
- Data at hand thus become:

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

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

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

The (multiple) linear regression model: Introduction



Model specification

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

  gas consumption in a week $=$ g(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*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \tag{1}$$

  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 stardard 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:

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

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

Qualitative predictors (factors)

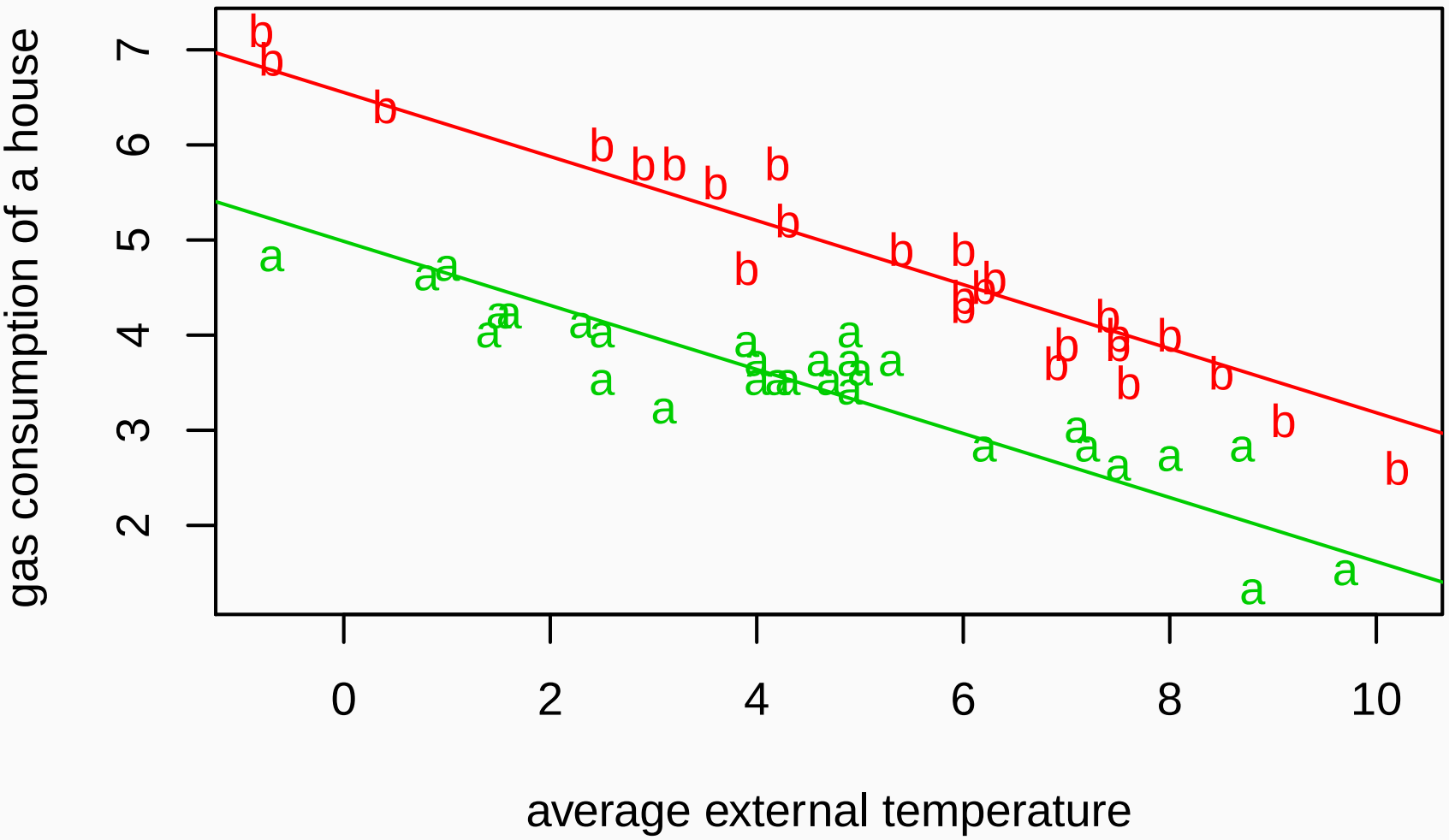
Model estimation via least squares easily extend to the multiple linear model:

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

The three coefficients are respectively:

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

Qualitative predictors (factors)



Interpreting the model

- $\hat{\beta}_0$: expected response value when all the predictors are set zero (if it does make sense and 0 is in the range of observed predictors)

  $\mapsto$ 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:

  $\mapsto$ 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

  $\mapsto$ 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)

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

  $\mapsto$ What is the expected gas consumption when the external temperature is 5?

  - if the house is not insulated:

$$\begin{aligned} \hat{\mu} &= 5.486 - 0.2902x \\ &= 6.551 - 0.3367 \cdot 5 = 4.8675 \text{ thousand cube feet} \end{aligned}$$

  - if the house is insulated:

$$\begin{aligned} \hat{\mu} &= 5.486 - 0.2902x \\ &= 6.551 - 1.565 - 0.3367 \cdot 5 = 3.3025 \text{ thousand cube feet} \end{aligned}$$

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?

  $\mapsto$ Compute the *adjusted* R^2: a suitable adjustment of R^2 which penalises additional explanatory variables (descriptive)

  $\mapsto$ 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 β_1 and β_2 is not 0

- Does the $j-$th explanatory variable really affect the response variable?

  $\mapsto$ Test the significance of a single predictor: the t test

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

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 ***
## x           -0.33670    0.01776  -18.95   <2e-16 ***
## zafter      -1.56520    0.09705  -16.13   <2e-16 ***
## ---
## 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
```

Model checking

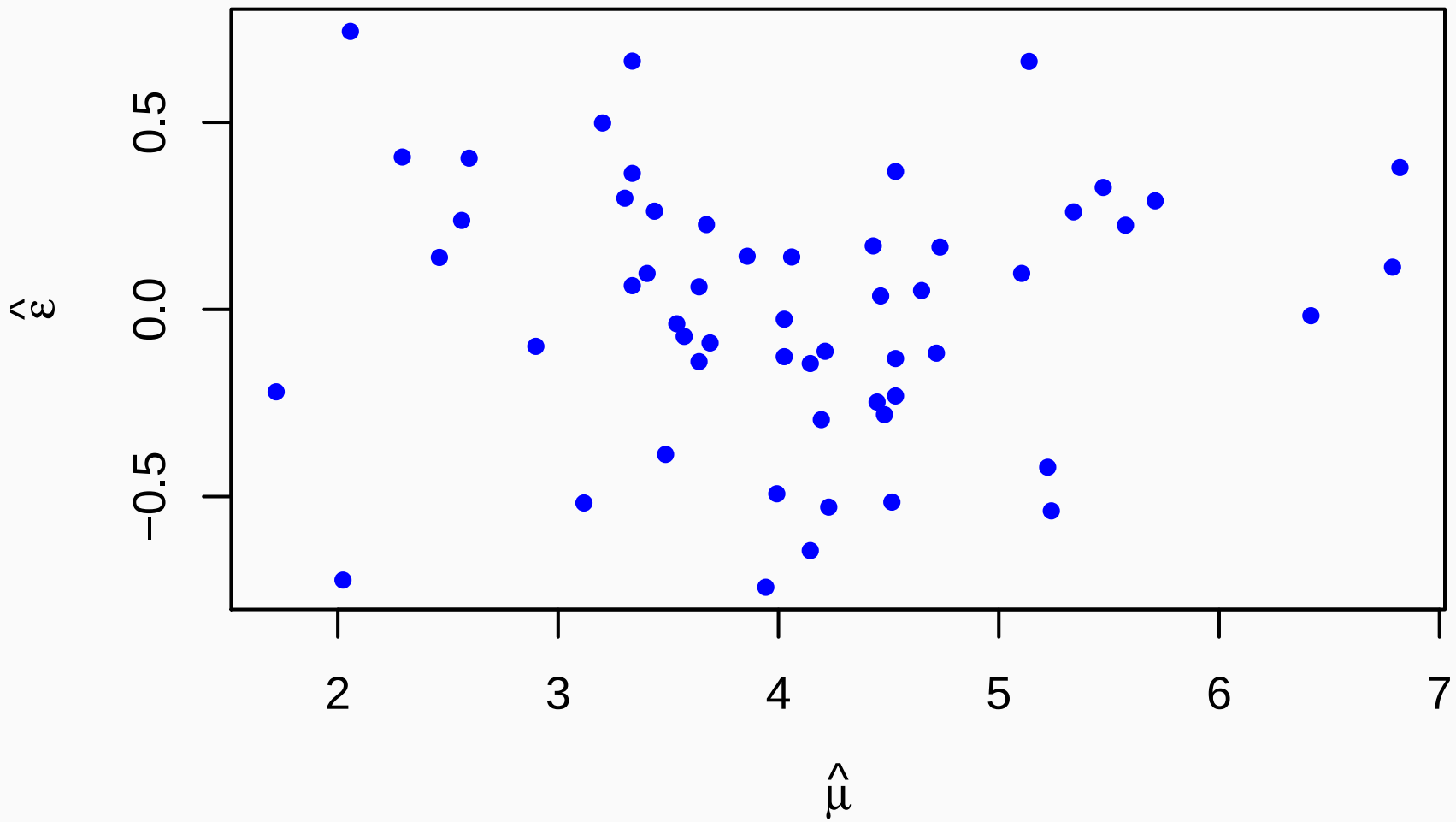
- The least squares linear model is the estimated linear model which best fits the data at hand but... *best* is not (necessarily) *good*.
- How to establish if the estimated model is a good one?

  $\mapsto$ Residuals are built as in the simple linear model and have the same interpretation (estimate of the error components ϵ_i)

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

  $\mapsto$ Residual analysis allows to check if the model assumptions are met and if the model is good

Model checking - Residual plots



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 $\hat{\mu}$)
  - $\mapsto$ Positive residuals for small and large $\hat{\mu}$
  - $\mapsto$ Negative residuals for intermediate values of $\hat{\mu}$
- Have we forgotten anything relevant?

The interaction term

- Going back to the scatterplot of the data...
  - $\mapsto$ the line corresponding to the observations before insulation tends to underestimate the gas consumption for small lower temperatures and overestimate it for higher temperatures
  - $\mapsto$ 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
  - $\mapsto$ *interaction term*: different relationship between Y and X for different values of d

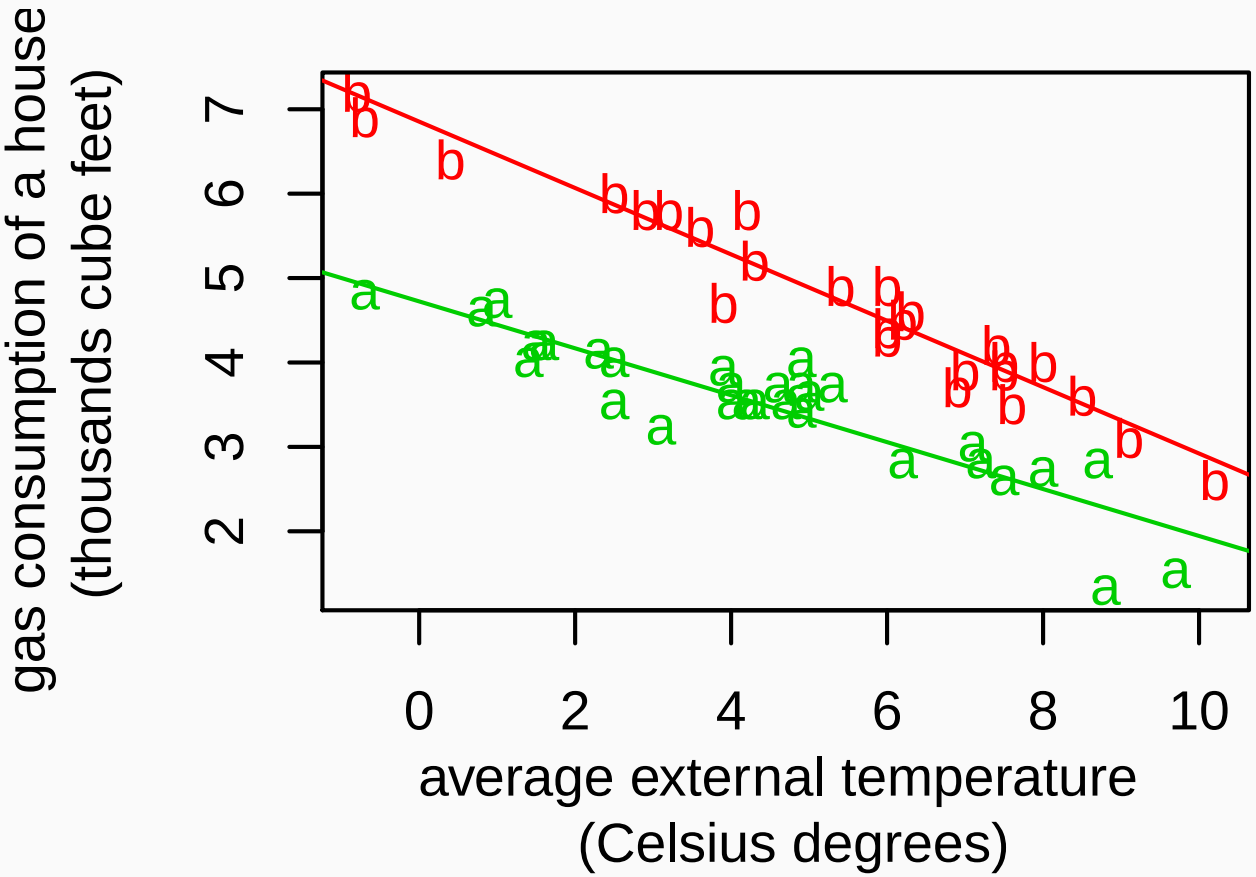
The interaction term: Formalization

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

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 (x_i \cdot d_i) + \epsilon_i \\
&= \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{before the intervention, when } d_i = 0 \\
&= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \epsilon_i \quad \text{after the intervention, when } d_i = 1
\end{aligned}
$$

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

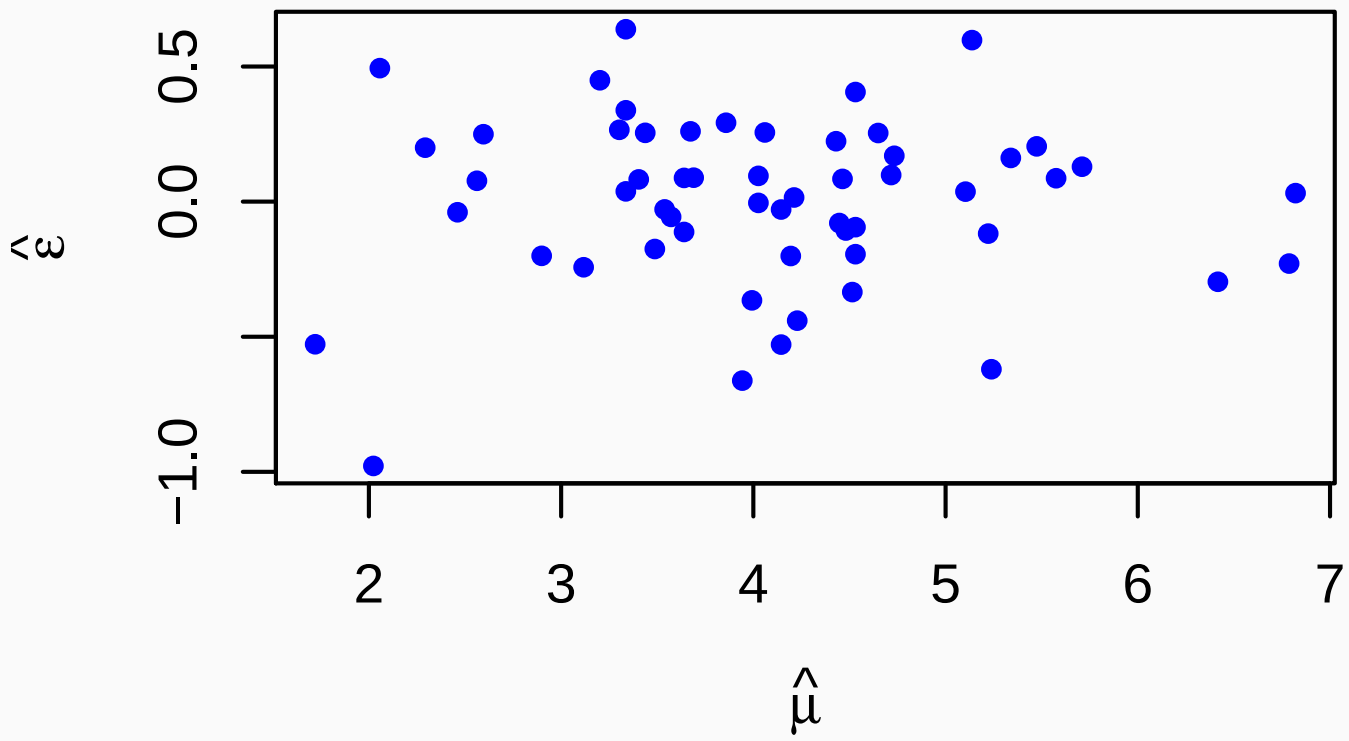
The estimated interaction model



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:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.85383    0.13596  50.409  < 2e-16 ***
## x           -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 ***
## ---
## 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



Generalization to multiple regression model

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

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

- The model is correctly specified:

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

- systematic component:
  - $\mapsto$ predictors are under the control of the researchers (non stochastic)
  - $\mapsto$ predictors are not collinear (i.e. not highly correlated)
- stochastic component:
  - the error terms have zero mean
    - $\mapsto$ do not include further sisthematic terms
    - $\mapsto$ this implies that $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_{ip-1}$