Linear regression, diagnostics, variable selection

2024-09-02

- **M1 MIDS & MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **!** **Objectives**

Linear Regression on Whiteside data

Packages installation and loading (again)

We will use the following packages. If needed, we install them.

```
stopifnot(
  require(tidyverse),
                  require(broom),
                  require(magrittr),
                  require(lobstr),
                  require(ggforce),
#                  require(cowplot),
                  require(patchwork),
                  require(glue),
                  require(DT),
                  require(viridis)
)
```

Dataset

```
whiteside <- MASS::whiteside # no need to load the whole package

cur_dataset <- str_to_title(as.character(substitute(whiteside)))
```

```
# ?whiteside
```

> Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

```
whiteside %>%
  glimpse
```

```
Rows: 56
Columns: 3
$ Insul <fct> Before, Before, Before, Before, Before, Before, Before, Before, ~
$ Temp  <dbl> -0.8, -0.7, 0.4, 2.5, 2.9, 3.2, 3.6, 3.9, 4.2, 4.3, 5.4, 6.0, 6.~
$ Gas   <dbl> 7.2, 6.9, 6.4, 6.0, 5.8, 5.8, 5.6, 4.7, 5.8, 5.2, 4.9, 4.9, 4.3,~
```

Start with columnwise and pairwise exploration

```
C <- whiteside %>%
  select(where(is.numeric)) %>%
  cov()

# Covariance between Gas and Temp

mu_n <- whiteside %>%
  select(where(is.numeric)) %>%
  colMeans()

# mu_n # Mean vector
```

$$C_n = \begin{bmatrix} 7.56 & -2.19 \\ -2.19 & 1.36 \end{bmatrix} \qquad \mu_n = \begin{bmatrix} 4.88 \\ 4.07 \end{bmatrix}$$

Use `skimr::skim()` to write univariate reports

Build a scatterplot of the Whiteside dataset

Build boxplots of `Temp` and `Gas` versus `Insul`

Build violine plots of `Temp` and `Gas` versus `Insul`

Plot histograms of `Temp` and `Gas` versus `Insul`

Plot density estimates of `Temp` and `Gas` versus `Insul`.

Hand-made calculatoin of simple linear regression estimates for `Gas` versus `Temp`

Overlay the scatterplot with the regression line.

Using `lm()`

`lm` stands for Linear Models. Function `lm` has a number of arguments, including:

- formula
- data

Including a rough summary in a report is not always a good idea. It is easy to extract a tabular version of the summary using functions `tidy()` and `glance()` from package `broom`.

For html output `DT::datatable()` allows us to polish the final output

Function `glance()` extract informations that can be helpful when performing model/variable selection.

`R` offers a function `confint()` that can be fed with objects of class `lm`. Explain the output of this function.

Diagnostic plots

Method `plot.lm()` of generic S3 function `plot` from base `R` offers six diagnostic plots. By default it displays four of them.

What are the diagnostic plots good for?

These diagnostic plots can be built from the information gathered in the `lm` object returned by `lm(...)`.

It is convenient to extract the required pieces of information using method `augment.lm`. of *generic function* `augment()` from package `broom`.

Recall that in the output of `augment()`

- `.fitted`: $\widehat{Y} = H \times Y = X \times \hat{\beta}$
- `.resid`: $\hat{\epsilon} = Y - \widehat{Y}$ residuals, $\sim (\text{Id}_n - H) \times \epsilon$
- `.hat`: diagonal coefficients of Hat matrix H
- `.sigma`: is meant to be the estimated standard deviation of components of $\widehat{Y}$

Compute the share of *explained variance*

Plot residuals against fitted values

Fitted against square root of standardized residuals.

TAF

Taking into account Insulation

Design a *formula* that allows us to take into account the possible impact of Insulation. Insulation may impact the relation between weekly `Gas` consumption and average external `Temperature` in two ways. Insulation may modify the `Intercept`, it may also modify the slope, that is the sensitivity of `Gas` consumption with respect to average external `Temperature`.

💡 Have a look at formula documentation (`?formula`).

Check the design using function `model.matrix()`. How can you relate this augmented design and the *one-hot encoding* of variable `Insul`?

Function `model.matrix()` allows us to inspect the design matrix.

In order to solve le Least-Square problems, we have to compute

$$(X^T \times X)^{-1} \times X^T$$

This can be done in several ways.

`lm()` uses QR factorization.

```
#matador::mat2latex(signif(solve(t(X) %*% X), 2))
```

$$(X^T \times X)^{-1} = \begin{bmatrix} 0.18 & -0.026 & -0.18 & 0.026 \\ -0.026 & 0.0048 & 0.026 & -0.0048 \\ -0.18 & 0.026 & 0.31 & -0.048 \\ 0.026 & -0.0048 & -0.048 & 0.0099 \end{bmatrix}$$

Understanding `.fitted` column

Understanding `.resid`

Understanding `.hat`

Understanding `.std.resid`

Understanding column `.sigma`