LAB: Linear Regression on Whiteside data

2025-03-18

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2024
Course Homepage
Moodle

Introduction

The purpose of this lab is to introduce *linear regression* using base `R` and the tidyverse. We work on a dataset provided by the MASS package. This dataset is investigated in the book by Venables and Ripley. This discusssion is worth being read. Our aim is to relate regression as a tool for data exploration with regression as a method in statistical inference. To perform regression, we will rely on the base `R` function `lm()` and on the eponymous S3 class `lm`. We will spend time understanding how the *formula* argument can be used to construct a *design matrix* from a dataframe representing a dataset.

Packages installation and loading (again)

```
# We will use the following packages.
# If needed, install them : pak::pkg_install().
stopifnot(
  require("magrittr"),
  require("lobstr"),
  require("ggforce"),
  require("patchwork"),
  require("gt"),
  require("glue"),
  require("skimr"),
  require("corrr"),
  require("GGally"),
  require("broom"),
  require("tidyverse"),
  require("ggfortify"),
  require("autoplotly")

)
```

Besides the `tidyverse`, we rely on `skimr` to perform univariate analysis, `GGally::ggpairs` to perform pairwise (bivariate) analysis. Package `corrr` provide graphical tools to explore correlation matrices. At some point, we will showcase the exposing pipe `%$%` and the classical pipe `%>%` of `magrittr`. We use `gt` to display handy tables, `patchwork` to compose graphical objects. `glue` provides a kind of formatted strings. Package `broom` proves very useful when milking lienar models produced by `lm()` (and many other objects produced by estimators, tests, ...)

Dataset

The dataset is available from package `MASS`. `MASS` can be downloaded from `cran`.

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

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

The documentation of `R` tells us a little bit more about this data set.

> 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.

This means that our sample is made of 56 observations. Each observation corresponds to a week during heating season. For each observation. We have the average external temperature `Temp` (in degrees Celsius) and the weekly gas consumption `Gas`. We also have `Insul` which tells us whether the observation has been recorded `Before` or `After` treatment.

Temperature is the *explanatory* variable or the *covariate*. The target/response is the weekly Gas Consumption. We aim to *predict* or to *explain* the variations of weekly gas consumption as a function average weekly temperature.

The question is wether the treatment (insulation) modifies the relation between gas consumption and external temperature, and if we conclude that the treatment modifies the relation, in which way?.

Have a glimpse at the data.

```
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,~
```

Even though the experimenter, Mr Whiteside, decided to apply a *treatment* to his house. This is not exactly what we call *experimental data*. Namely, the experimenter has no way to clamp the external temperature. With respect to the Temperature variable (the explanatory variable) we are facing *observational* data.

Columnwise exploration

Question

Before before proceeding to linear regressions of `Gas` with respect to `Temp`, perform univariate analysis on each variable.

- Compute summary statistics
- Build the corresponding plots

Pairwise exploration

Question

Compare distributions of numeric variables with respect to categorical variable `Insul`

Covariance and correlation between `Gas` and `Temp`

Question

Compute the covariance matrix of `Gas` and `Temp`

Question

- Compute correlations (Pearson, Kendall, Spearman) and correlations per group
- Comment

Question

Use `ggpairs` from `GGally` to get a quick overview of the pairwise interactions.

Question

Build a scatterplot of the Whiteside dataset

Question

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

Question

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

Question

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

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

> **Question**
>
> Compute slope and intercept using elementary computations

> **Question**
>
> Overlay the scatterplot with the regression line.

Using `lm()`

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

- formula
- data

> **Question**
>
> Use `lm()` to compute slope and intercept. Denote the object created by constructor `lm()` as `lm0`.
>
> - What is the class of `lm0` ?
> -

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

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

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

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

> **Question**
>
> Plot a 95% confidence region for the parameters (assuming homoschedastic Gaussian noise).

Diagnostic plots

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

In order to obtain diagnostic plots as ggplot objects, use package `ggfortify` which defines an S3 method for class 'lm' for generic function `autoplot` (defined in package `ggplot`).

Question

What are the diagnostic plots good for?

The diagnostic plots are 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`: $\widehat{Y} = H \times Y = X \times \widehat{\beta}$
- `.resid`: $\widehat{\epsilon} = Y - \widehat{Y}$ residuals, $\sim (\mathrm{Id}_n - H) \times \epsilon$
- `.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

Question

Fitted against square root of standardized residuals.

Question

Hand-made normal qqplot for `lm0`

Question

Taking into account Insulation

Question

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`).

Question

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

Question

Display and comment the part of the summary of `lm1` concerning estimated coefficients.

Question

Comment the diagnostic plots built from the extended model using `autoplot()`. If possible, generate alternative diagnostic plots pipelining `broom` and `ggplot2`.

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

Question

Try understanding the computation of `.resid` in an `lm` object. Compare `.resid` with the projection of `Gas` on the linear subspace orthogonal to the columns of the design matrix.

Question

Understanding `.hat`

Question

Understanding `.std.resid`

- Estimate noise intensity from `residuals`
- Compare with the output of `glance()`

Question

Understanding column `.sigma`

Appendix

`S3` classes in `R`

Generic functions for `S3` classes

`methods(autoplot)` lists the `S3` classes for which an autoplot method is defined. Some methods are defined in `ggplot2`, others like `autoplot.lm` are defined in extension packages like `ggfortify`.

`S4` classes in `R`

The output of `autoplot.lm` is an instance of `S4` class

`tibbles` with list columns