Linear regression II

2024-09-02

```r
stopifnot(
  require(tidyverse),
  require(patchwork),
  require(httr),
  require(glue),
  require(broom)
)
old_theme <- theme_set(theme_minimal())
```

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

- Moodle

> ! **Objectives**

Linear fit using ordinary least squares (OLS)

- Perform linear regression of SAL_ACTUEL with respect to SAL_EMBAUCHE. Store the result in an object denoted by `lm_1`
- Inspect the numerical summary of `lm_1`
- Use `Environment` panel (Rstudio), to explore the structure of `lm_1`. Try to understand the signification of each element.

```r
datapath <- '../DATA'
fname <- 'Banque.csv'
fpath <- paste(datapath, fname, sep="/")
```

```r
if (!file.exists(fpath)) {
  baseurl <- 'https://stephane-v-boucheron.fr/data'
  download.file(url=paste(baseurl, fname, sep="/"),
                destfile=fpath)
  print(glue::glue('File {fname} downloaded at {fpath}!'))
} else {
  print(glue::glue('File {fname} already exists at {fpath}!'))
}
```

```
File Banque.csv already exists at ../DATA/Banque.csv!
```

```
bank <- readr::read_table(fpath,
    col_types = cols(
        SEXE = col_factor(levels = c("0", "1")),
        CATEGORIE = col_integer(),
        NB_ETUDES = col_integer(),
        SATIS_EMPLOI = col_factor(levels = c("non", "oui")),
        SATIS_CHEF = col_factor(levels = c("non", "oui")),
        SATIS_SALAIRE = col_factor(levels = c("non", "oui")),
        SATIS_COLLEGUES = col_factor(levels = c("non", "oui")),
        SATIS_CE = col_factor(levels = c("non", "oui"))
  )
)
```

- Make the model summary a dataframe/tibble using `broom::tidy()`
- Make model diagnostic information a dataframe/tibble using `broom::glance()`
- Preparing for diagnostic plots using `broom::augment()`

The output of `augment` may be described as adding 6 columns to dataframe `bank`. The six columns are built using items from `lm_1`. Can you explain their meaning and why they are relevant to diagnosing?

Let base `R` produce diagnostic plots

```
plot(lm_1, which = 1:6)
```

We will reproduce (and discuss) four of the six diagnostic plots provided by the `plot` method from base `R` (1,2,3,5).

- Reproduce first diagnostic plot with `ggplot` using the aumented version of `lm_1` (`augment(lm_1)`).
- Comment Diagnostic Plot 1.
- Compute the correlation coefficient between residuals and fitted values.
- Make your graphic pipeline a reusable function.
- What are *standardized residuals* ?
- Build the third diagnostic plot (square root of absolute values of standardized residuals versus fitted values) using `ggplot`.
- Why should we look at the square root of standardized residuals?

Make your graphic pipeline a reusable function.

- What is leverage ?

- Build the fifth diagnostic plot (standardized residuals versus leverage) using `ggplot`.

In the second diagnostic plot (the residuals qqplot), we build a quantile-quantile plot by plotting function $F_n^{\leftarrow} \circ \Phi$ where Φ is the ECDF of the standard Gaussian distribution while $F_n^{\leftarrow}$.

Build the second diagnostic plot using `ggplot`

Use package `patchwork::...` to collect your four diagnostic plots

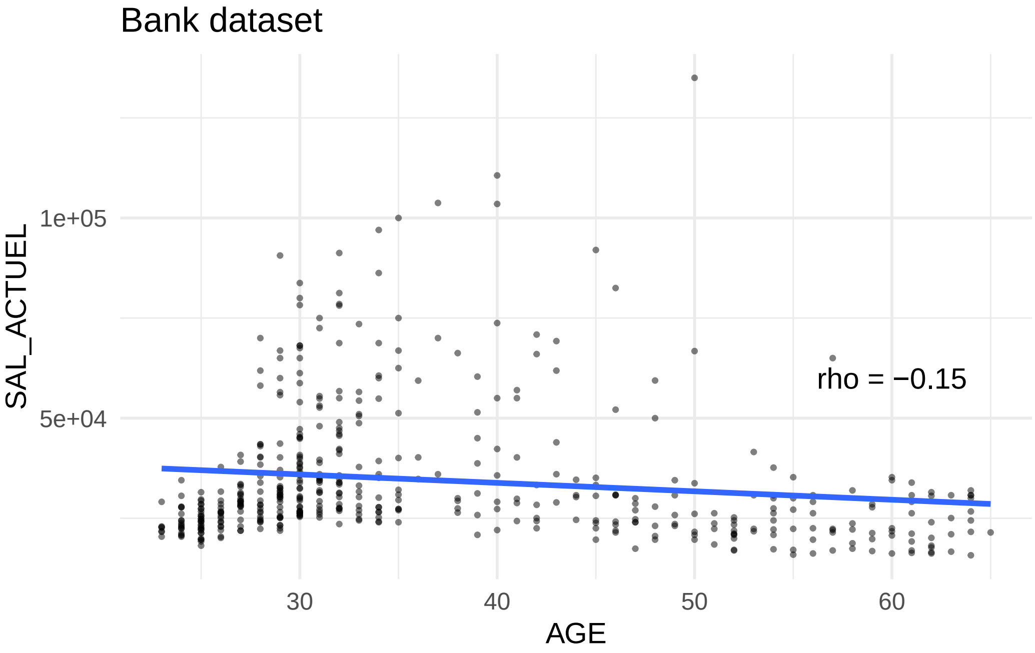
Plot actual values against fitted values for `SAL_ACTUEL`

Play it again with `AGE` and `SAL_ACTUEL`

Redo the above described steps and call the model `lm_2`.

- `ggplot` programming : write a function with arguments `df`, `varx` and `vary` where `varx` and `vary` are two strings denoting numerical columns in `df`, that outputs a ggplot object made of a scatterplot of columns `vary` and `vary`, a linear regression of `vary` against `varx`. The ggplot plot object should be annotated with the linear correlation coefficient of `vary` and `varx` and equipped with a title.

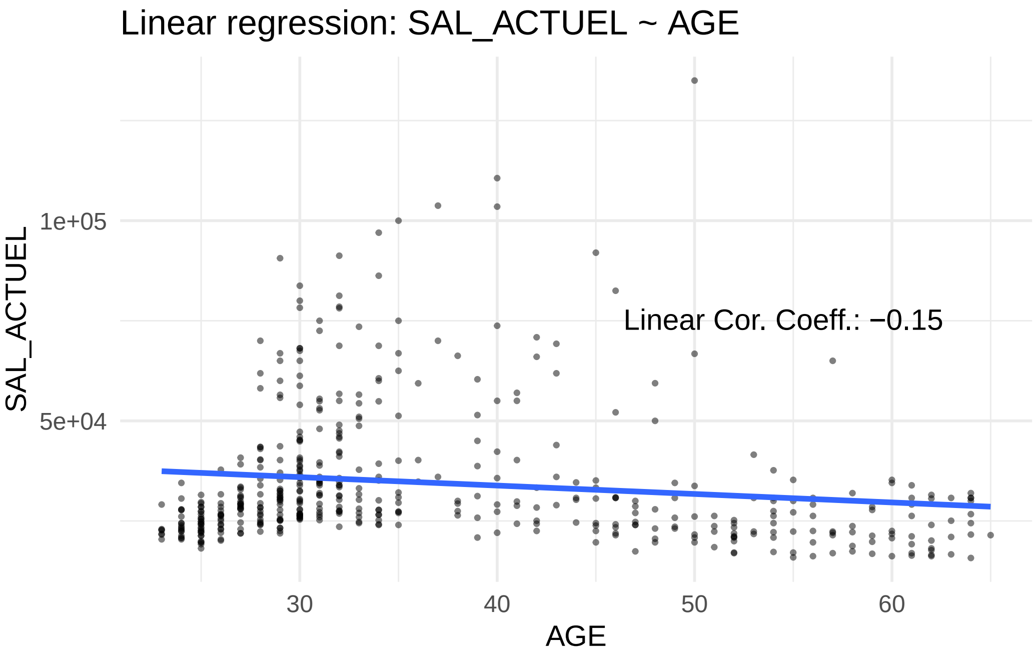
```
bank %>%
  ggplot() +
  aes(x=AGE, y=SAL_ACTUEL) +
  geom_point(alpha=.5, size=.5, ) +
  geom_smooth(method="lm", formula= y ~ x, se=F) +
  annotate(geom="text", x=60, y=60000,
           label=str_c("rho = ",
                       round(cor(bank$SAL_ACTUEL, bank$AGE), 2))) +
  ggtitle("Bank dataset")
```



```
ggplot_lin_reg <-  function(df, varx, vary){
  rho <- round(cor(df[[varx]], df[[vary]]), 2)
  posx <- sum(range(df[[varx]])*c(.25 , .75))
  posy <- sum(range(df[[vary]])*c(.5 , .5))

  df %>%
    ggplot() +
    aes(x=.data[[varx]], y=.data[[vary]]) +
    geom_point(alpha=.5, size=.5, ) +
    geom_smooth(method="lm", formula= y ~ x, se=F) +
    annotate(geom="text", x=posx, y=posy,
           label=glue("Linear Cor. Coeff.: {rho}")) +
    ggtitle(glue("Linear regression: {vary} ~ {varx}"))
}

ggplot_lin_reg(bank, "AGE", "SAL_ACTUEL")
```



Inspect rows with high Cook's distance

Discuss the relevance of Simple Linear Regression for analyzing the connection between `SAL_ACTUEL` and `AGE`

Compute the Pearson correlation coefficient for every pair of quantitative variable? Draw corresponding scatterplots.

Predictive linear regression of `SAL_ACTUEL` as a function of age `AGE`

To perform linear fitting, we choose 450 points amongst the 474 sample points: the 24 remaining points are used to assess the merits of the linear fit.

Randomly select 450 rows in the `banque` dataframe.

Function `sample` from base `R` is convenient. You may also enjoy `slice_sample()` from `dplyr`. Denote by `trainset` the vector of of selected indices. Bind the vector of left behind indices to variable `testset`. Functions `match`, `setdiff` or operator `%in%` may be useful.

- □ Linear fit of `SAL_ACTUEL` with respect to `AGE`, on the training set. Call the result `lm_3`.
- □ How do you feel about such a linear fit? (Use diagnostic plots)

Inspecting points with high Cook's distance

- □ Use `lm_3` to predict the values of `SAL_ACTUEL` as an affine function of `AGE` on the testing set `testset` (`broom::augment()` with optional argument `newdata` may be useful). Compare the data frame with the one obtained from `augment(lm_3)`.
- □ Compare training error and testing error
- □ Analyse residuals (prediction errors) on the testing set. Compare with training set

Expectations under Gaussian Linear Modelling Assumptions

$$\begin{pmatrix} Y \end{pmatrix} = \begin{pmatrix} \mathbb{Z} \end{pmatrix} \times \beta + \sigma \begin{pmatrix} \epsilon \end{pmatrix}$$