Linear regression: ANOVA

2024-09-02

```r
theme_set(theme_minimal())
```

! Objectives

Variable/Model selection and ANOVA on Whiteside data

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

Challenge(s)

Comparing weekly average temperatures over two seasons

We address the following question: was the external temperature distributed in the same way during the two heating seasons? When we raise this question, we silently make modeling assumptions. Spell them out.

What kind of hypothesis are we testing in the next two chunks? Interpret the results.

```r
lm_temp <- lm(Temp ~ Insul, whiteside)

lm_temp |>
  tidy()
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    5.35      0.537      9.96 7.80e-14
2 InsulAfter    -0.887     0.734     -1.21 2.32e- 1
```

```r
lm_temp |>
  glance()
```

```
# A tibble: 1 x 12
```

```
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1    0.0263       0.00830  2.74      1.46   0.232     1  -135.  276.  282.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Display parallel boxplots, overlayed cumulative distribution functions and a quantile-quantile plot (QQ-plot) to compare the temperature distributions during the two heating seasons. Comment

Perform a Wilcoxon test to assess change of Temperature between the two seasons

Does Insulation matter?

- Does average Gas consumption change with Insulation?
- Does Gas consumption dependence on Temperature change with Insulation?

As we have to infer the *dependence on Temperature*, the questions turn tricky.

Compare Gas consumption before and after (leaving Temperature aside)

Draw a `qqplot` to compare Gas consumptions before and after insulation.

Compare ECDFs of Gas consumption before and after insulation.

Do Insulation and Temperature additively matter?

This consists in assessing whether the Intercept is modified after Insulation while the slope is left unchanged. Which models should be used to assess this hypothesis?

Draw the disgnostic plots for this model

Do Insulation and Temperature matter and interact?

Find the formula and build the model.

Do Insulation and powers of temperature interact?

Investigate formulae `Gas ~ poly(Temp, 2, raw=T)*Insul`, `Gas ~ poly(Temp, 2)*Insul`, `Gas ~ (Temp +I(Temp*2))*Insul`, `Gas ~ (Temp +I(Temp*2))| Insul`

Higher degree polynomials

Play it with degree 10 polynomials

Drying model exploration

Collecting the models a posteriori

Make a named list with the models constructed so far

Use `stepAIC()` to perform stepwise exploration

ANOVA table(s)

Use fonction `anova()` to compare models constructed with formulae

```r
formula(lm0)
```

```
Gas ~ Insul + Temp
```

Wikipedia on Analysis of Variance

WIKIPEDIA
The Free Encyclopedia

Analysis of variance

Analysis of variance (**ANOVA**) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among means. ANOVA was developed by the statistician Ronald Fisher. ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means. In other words, the ANOVA is used to test the difference between two or more means.

History

While the analysis of variance reached fruition in the 20th century, antecedents extend centuries into the past according to Stigler.[1] These include hypothesis testing, the partitioning of sums of squares, experimental techniques and the additive model. Laplace was performing hypothesis testing in the 1770s.[2] Around 1800, Laplace and Gauss developed the least-squares method for combining observations, which improved upon methods then used in astronomy and geodesy. It also initiated much study of the contributions to sums of squares. Laplace knew how to estimate a variance from a residual (rather than a total) sum of squares.[3] By 1827, Laplace was using least squares methods to address ANOVA problems regarding measurements of atmospheric tides.[4] Before 1800, astronomers had isolated observational errors resulting from reaction times (the "personal equation") and had developed methods of reducing the errors.[5] The experimental methods used in the study of the personal equation were later accepted by the emerging field of psychology [6] which developed strong (full factorial) experimental methods to which randomization and blinding were soon added.[7] An eloquent non-mathematical explanation of the additive effects model was available in 1885.[8]

Ronald Fisher introduced the term variance and proposed its formal analysis in a 1918 article on theoretical population genetics, *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.[9] His first application of the analysis of variance to data analysis was published in 1921, *Studies in Crop Variation I*.[10] This divided the variation of a time series into components representing annual causes and slow deterioration. Fisher's next piece, *Studies in Crop Variation II*, written with Winifred Mackenzie and published in 1923, studied the variation in yield across plots sown with different varieties and subjected to different fertiliser treatments.[11] Analysis of variance became widely known after being included in Fisher's 1925 book *Statistical Methods for Research Workers*.