Homework 2 (2023-24): Linear regression and Gaussian Linear ModelingDue date : 2024-03-15 23:55 (hard deadline)

- **M1 MIDS & MFA**
- **Université Paris Cité**
- Année 2023-2024
- Course Homepage

- Moodle

Objectives

This homework is concerned with Gaussian Linear Models. The objective consists of working with simulated data and visualizing/illustrating the main constructions and theorems from the sections of the Statistical Inference course dedicated to Gaussian Linear Models.

We start from the `whiteside` dataset from `MASS` package (`R`).

Fit a linear model with formula `Gas ~ poly(Temp, degree=2, raw=T) * Insul` to the `whiteside` data.

```r
lm2 <- lm(Gas ~ poly(Temp, degree=2, raw=T) * Insul,
          data=whiteside)
```

- Extract the coefficients vector ($\hat{\beta}$) and the model matrix (X).
- Extract the estimate $\hat{\sigma}$ of Gaussian noise standard deviation from the model.

You may rename the components of β and the columns of X up to your convenience.

```
# A tibble: 6 x 5
  term         estimate std.error statistic  p.value
  <chr>           <dbl>     <dbl>     <dbl>    <dbl>
1 Int           6.76      0.151      44.8   4.85e-42
2 Temp^1       -0.318     0.0630     -5.04  6.36e- 6
3 Temp^2       -0.00847   0.00662    -1.28  2.07e- 1
4 After        -2.26      0.220     -10.3   6.52e-14
5 After:Temp^1  0.180     0.0964      1.86  6.82e- 2
6 After:Temp^2 -0.00651   0.00997    -0.653 5.17e- 1
```

Simulate random data conforming to GLM with fixed design

Generate $N = 1000$ instances of the Gaussian Linear Model defined by

$$\begin{bmatrix} \vdots \\ Y \\ \vdots \end{bmatrix} = \begin{bmatrix} & X & \end{bmatrix} \times \widehat{\beta} + \widehat{\sigma} \times \begin{bmatrix} \vdots \\ \epsilon \\ \vdots \end{bmatrix}$$

where $\epsilon \sim \mathcal{N}(0, \text{Id}_{56})$ and $X, \widehat{\beta}$, and $\widehat{\sigma}$ have been extracted above from the linear fit to the `whiteside` data.

> 🔥 Caution
>
> Try to avoid unnecessary computations.

For each simulated instance, fit a linear model.

Now you should have N identically distributed, independent realizations of the response vector Y (and with some more work, realizations of prediction vectors $\widehat{Y}$). Denote the N independent realizations of the response vectors by $Y^*_1, \ldots, Y^*_N$, and denote the N realizations of the estimators $\widehat{\beta}^*_1, \ldots, \beta^*_N$ and $\widehat{\sigma}^*_1, \ldots, \widehat{\sigma}^*_N$. Use the same style of notation for predictions and residuals.

The Statistical Inference course tells us a lot of things about the distribution of the response vectors, the prediction vectors, the residuals, and so on. Those theoretical results are used by function `lm()`, method `summary.lm()`, and diagnostic plot methods `plot.lm()` (and also by `aov()`, `anova()`, `stepAIC()`, …)

Distribution of estimators of noise variance

Plot your sample of estimators of the noise variance $\widehat{\sigma}^*_1, \ldots, \widehat{\sigma}^*_N$. Compare with the theoretical density of the distribution of these estimators (histograms, CDF, quantile plots).

For $\alpha = 5\%, 1\%$, compute the N confidence regions for β (N ellipsoids), and compute the empirical *coverage* of your confidence regions (the number of times the true parameter belongs to the confidence region). How should this empirical coverage be distributed? What is its expectation? its departure from expectation?

Fluctuations of coefficients estimates (I)

According to the GLM, the distribution of the coefficients estimates $\widehat{\beta}^*_1, \ldots, \widehat{\beta}^*_N$ is known if the noise variance is known.

Visualize the empirical joint distribution of $(\widehat{\beta}^*_i[1], \widehat{\beta}^*_i[2])$. Compare with theoretical distribution.

Fluctuations of coefficients estimates (II) : Studentized statistics

If the noise variance is not known, according to the GLM theory, we can use the noise variance estimator to build confidence regions.

Investigate and illustrate (histograms, CDF and quantiles plots) the distribution of the coefficients $\frac{1}{\widehat{\sigma}} A \times (\widehat{\beta}^* - \widehat{\beta})$ where A is a well-chosen matrix (that may depend on the design).

Regression of $\widehat{\beta}^*[6]$ with respect to all other estimated coefficients $\widehat{\beta}^*[1,..,5]$

The sample of N realizations of $\widehat{\beta}^*$: $\widehat{\beta}^*_1, \ldots, \widehat{\beta}^*_N$ may be considered as an instance of linear regression with respect to a *random design* where the response variable is $\widehat{\beta}^*[6]$ while the explanatory variables are $\widehat{\beta}^*[1], \ldots, \widehat{\beta}^*[5]$.

Compute the optimal regression coefficients. What is the distribution of $\widehat{\beta}^*[6] - \mathbb{E}\left[\widehat{\beta}^*[6] \mid \widehat{\beta}^*[1], \ldots, \widehat{\beta}^*[5]\right]$? Investigate graphically.

Diagnostic plots when the GLM assumptions hold

Pick 1 linear fit amongst the N linear fits performed on the simulated data. Draw the four diagnostic plots. Comment (briefly).

Overparametrized model

Define $\widehat{\theta} \in \mathbb{R}^6$ by zeroing the coefficients of $\widehat{\beta}$ corresponding to the quadratic terms (with respect to `Temp`)

Generate $N = 1000$ instances of the Gaussian Linear Model defined by

$$\begin{bmatrix} \vdots \\ Y \\ \vdots \end{bmatrix} = \begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix} \times \widehat{\theta} + \widehat{\sigma} \times \begin{bmatrix} \vdots \\ \epsilon \\ \vdots \end{bmatrix}$$

where $\epsilon \sim \mathcal{N}(0, \mathrm{Id}_{56})$.

Estimators of noise variance

Fit all N realizations with the same formula as above. Compute the new estimators of the noise variance.

Student's tests for coefficients

Perform student's tests for the coefficient vectors. How many times do you reject the null hypothesis concerning the coefficients of $\widehat{\beta}$ corresponding to the quadratic terms (with respect to `Temp`) if you choose a size/level equal to 5%. Plot the sample of the $|t|$ values for the coefficients of $\widehat{\beta}$ corresponding to the quadratic terms (with respect to `Temp`).

Plot the histogram of the empirical distrinbution of p-values. Compare with theoretical distribution of p values.

Fisher's test(s)

We aim at testing

- $H_0 : \widehat{\theta}[3] = \widehat{\theta}[6] = 0$ (null hypothesis, assuming that the third and the sixth coefficients represent `Temp^2` and `Temp^2:InsulAfter`)

versus

- $H_1 : \widehat{\theta}[3] \neq 0 \quad \text{or} \quad \widehat{\theta}[6] \neq 0$ (alternative)

Compute the Fisher statistics for the N simulated response vectors (when the null hypothesis is true). Plot the Fisher statistics and compare to the theoretical distribution under the the null hypothesis. If you choose a level/size equal to 1%, how many times do you reject the null hypothesis?

Compute the Fisher statistics for the N simulated response vectors (when the null hypothesis is not true). Plot the Fisher statistics and compare to the theoretical distribution under the the null hypothesis. If you choose a level/size equal to 1%, how many times do you reject the null hypothesis?

Again plot a histogram of the empirical distribution of p-values

Performance of `stepAIC`

Run `stepAIC()` on the overparametrized models you obtain. Describe graphically the distribution of outcomes, and the distribution of the AIC criteria for the selected models.

Departing from the Gaussian Linear Model assumptions

Replay the above described simulations but replace the Gaussian noise with Student's noise with three degrees of freedom `rt(N, df=3, ncp=0)` (with and without overparametrizatio). Recompute the Fisher statistics for testing H_0 against H_1. Visualize the distributions compare to the theoretical distribution of the Fisher statistics. If you choose a level/size equal to 1%, how many times do you reject the null hypothesis?

References

- Poly. S. Boucheron
- Poly. S. Coste

🎓 Grading criteria

Criterion	Points	Details
Spelling and syntax	20%	English/French ✒
Plots correction	25%	choice of `aesthetics`, `geom`, `scale` … 📈
Computing Statistics	30%	Aggregations, LR, PCA, CA, … 📈
DRY compliance	25%	DRY principle at W Wikipedia