Statistics: When assumptions break

Brussels Summer School of Mathematics 2023

V. Meurice

Université libre de Bruxelles

2023

Table of Contents

Table of Contents

Standard setup

Consider a sample of realisations of n random variables $X_1, \ldots, X_n$*, that are independent and identically distributed (i.i.d.) normal; i.e.* $X_i \sim \mathcal{N}(\mu, \sigma^2)$ *for all* $i = 1, \ldots, n$*, with* $\mu \in \mathbb{R}$ *and* $\sigma^2 < \infty$*. Then,...*

This is the standard setup of many basic results in probability and statistics. It is easy to skim past what it entails. However, it instantly imposes a set of fairly strict assumptions:

- Full independence
- Normality
- Homoskedasticity (same variance)
- Common mean
- Finite variance

Asymptotic case

Alternatively, we often rely on the Central Limit Theorem to relax the normality assumption:

Central Limit Theorem

For $X_1, \ldots, X_n$ i.i.d. (any distribution) with mean $\mathbb{E}[X_1] =: \mu$, letting $\bar{X} := \sum_{i=1}^{n} X_i$ be the empirical mean, there exists σ^2 such that

$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

as $n \to \infty$.

This theorem introduces approximate Gaussianity through the use of $\bar{X}$ in an asymptotic setup.

Back to reality

In real life, we might very well have non-normal, dependent and heteroskedastic data in a relatively small sample. What happens then?

In general, theoricians impose various assumptions to control their mathematical framework, allowing for rigorous results to use in applications. Relaxing those assumptions to make methods based on them more universal is an active process that we will briefly discuss here through different examples.

Table of Contents

Table of Contents

Structures of interest

Although it can be difficult to observe dependence in data, some particular structures can be guessed to exist based on the context.

Two of the most common cases are

- serial correlation, for time-dependent data;
- spatial correlation, for geographical data.

Serial correlation

Consider a variable X_t observed at various times $t \in \mathbb{N}$:

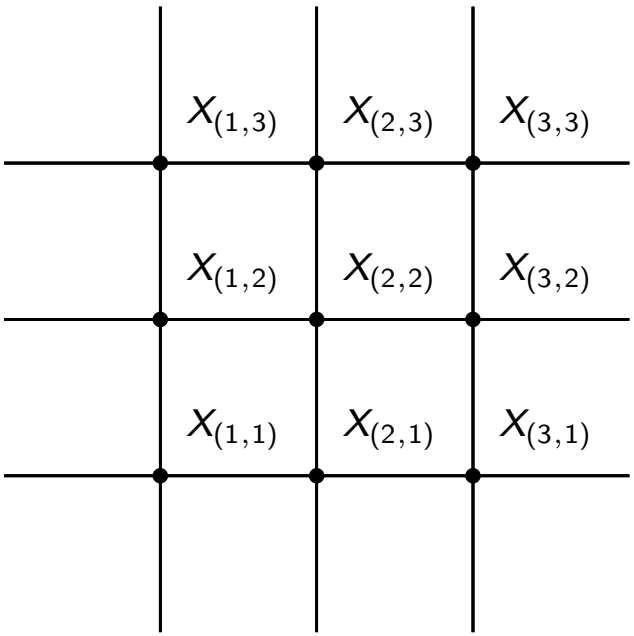
From a probability perspective, the sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ is called a (discrete) random process. The equivalent data sample is usually referred to as a time series.

It is natural for the variable in $t-1$ to be correlated with the one in t. By extension, any variable X_t will be somewhat correlated with any variable X_{t-s} for some $s \in \mathbb{Z}$.

This concept is called serial correlation or autocorrelation.

Spatial correlation

One can extend the index space to multidimensional cases too.
For example, consider a spatial process in two dimensions, i.e. data points indexed on a map:



It is once again natural to expect correlation between neighbouring points.

Example of consequences

The presence of correlation between data points can cause issues in a number of procedures, since it violates the independence assumption that is often required for valid inference.

We will consider a case that is well discussed in the literature: correlated error terms in regression models.

Table of Contents

Ordinary Least Squares regression I

Practitioners often want to study the link between variables of interest. The most widely used model for this problem is a linear one.

Suppose that we want to estimate the effect of some variables $X_1, \ldots, X_k$ on a variable of interest Y.

A linear regression model will assume that one can write the relationship between $(X_1, \ldots, X_k)$ and Y as

$$Y = \beta_0 + \beta_1 * X_1 + \ldots + \beta_k * X_k + \varepsilon,$$

where $\beta_0, \beta_1, \ldots \beta_k$ are called the *regression coefficients* and ε is the *error term*.

Here, the β's are not known; we need to estimate them.

Ordinary Least Squares regression II

If we have n observations of each variable, we can write the equation like this:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

or equivalently in matrix terms:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

In this model, the error term ε is always (at least) assumed to have mean zero (so $\mathbb{E}[\varepsilon] = 0$).

Ordinary Least Squares regression III

Since we do not know $\boldsymbol{\beta}$, we need to estimate it. The most common way to do that is to use the *Ordinary Least Squares* (OLS) estimator:

$$\hat{\boldsymbol{\beta}}_{OLS} := (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

It is built to minimise the squared distance between the predicted values $\hat{Y}_i := \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{OLS}$ and the real ones Y_i.

Assume that the errors are independent with common variance $\sigma^2 < \infty$ (homoskedastic).

This estimator is then unbiased, i.e. $\mathbb{E}\left[\hat{\boldsymbol{\beta}}_{OLS}\right] = \boldsymbol{\beta}$, and has the smallest variance amongst other unbiased linear estimators (so it is the most precise one).

OLS under heteroskedastic or correlated errors

Many reasons can cause the errors to be correlated or heteroskedastic (for example, if there is a clear dependence structure in the data like described before). What happens then?

Under correlated or heteroskedastic errors,

- $\hat{\beta}_{OLS}$ is still unbiased (good ☺)
- $\hat{\beta}_{OLS}$ does not always have the smallest variance amongst unbiased linear estimators (bad ☹)

This means that on average, we will estimate the correct value for β, but that we lose (possibly a lot of) precision.

Consequences on tests and workarounds

Another procedure that usually accompanies the regression is to test whether some coefficient is equal to zero, i.e. testing

$$H_0 : \beta_i = 0 \quad \text{v.s.} \quad H_1 : \beta_i \neq 0$$

for some i (usually done for all i).
This test is once again invalidated by correlated or heteroskadstic errors.

The main problem in both cases is that correlated or heteroskedastic errors mess up the variance of $\hat{\beta}_{OLS}$.

- The typical workaround is to try and estimate said variance and plug it in to *nullify* the effect as much as possible.
- This is really difficult, partly because it usually involves making assumptions on the underlying dependence structure...

Table of Contents

Relaxing assumptions

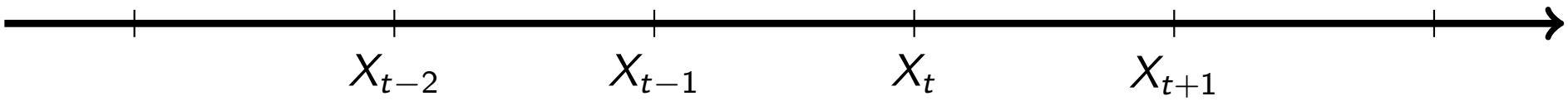
Estimating the covariance structure and plugging it in the procedure is really effective at mitigating negative effects of correlation, but is really hard to do.

As often when mathematicians need to deal with tricky assumptions, when deleting them altogether is too much of a challenge, they try to simply relax them instead.

What if we could mitigate the effect of autocorrelation by reaching some form of approximate independence, or at least *weaker* correlation?

Strongly stationary stochastic processes

Recall the stochastic (i.e. random) time process $\{X_t\}_{t\in\mathbb{N}}$ described earlier:



A first assumption that is widely used when studying such processes is *stationarity*.

Strongly stationary process

Let $F(X_{t_1}, \ldots, X_{t_p})$ represent the joint cumulative distribution function of X_t at times $t_1, \ldots, t_{t_p}$.
Then, $\{X_t\}_{t\in\mathbb{N}}$ is strongly stationary if

$$F(X_{t_1+\tau}, \ldots, X_{t_p+\tau}) = F(X_{t_1}, \ldots, X_{t_p})$$

for all τ and $t_1, \ldots, t_{t_p}$ and all p.

In non-mathy terms, the randomness structure does not evolve over time.

Mixing processes: intuition

Consider (once again) a time process $\{X_t\}_{t\in\mathbb{N}}$.

Intuitively, an event really far back in the past should have much smaller influence over what happens far into the future than, say X_t influences X_{t+1}.

Maybe we can formalise this and use it to our advantage?

α-mixing processes

Let $\{X_t\}_{t\in\mathbb{Z}}$ be a stochastic process.

α-mixing process

Define the mixing coefficients

$$\alpha(s) := \sup_{t\in\mathbb{Z}}\{|P(AB) - P(A)P(B)| : A \in \mathfrak{F}_{-\infty}^{t}, B \in \mathfrak{F}_{+\infty}^{t+s}\},$$

where $\mathfrak{F}_a^b$ is the σ-algebra generated by $\{X_a, X_{a+1}, \ldots, X_b\}$.
Then, $\{X_t\}_{t\in\mathbb{N}}$ is α-mixing if $\alpha(s) \to 0$ as $s \to \infty$.

In brief:

- The events A and B are the ones at least separated by s time increments
- $\alpha(s)$ measures the maximum correlation between those kinds of events
- $\alpha(s) \to 0$ as $s \to \infty$ means that events in the past have less and less influence over the future the more time passes, with the two being asymptotically uncorrelated

CLT for stationary processes

Ibragimov (1962)

This *asymptotic independence* property allows for some cool results, such as this one:

CLT for stationary processes

Let the univariate stochastic process $\{X_t\}_{t\in\mathbb{N}}$ be strongly stationary and α-mixing. Assume that for some $\delta > 0$,

(i) $\mathbb{E}\left[|X_1|^{2+\delta}\right] < \infty$ and

(ii) $\sum_{s=1}^{\infty} [\alpha(s)]^{\frac{\delta}{2+\delta}} < \infty$.

Then $\sigma^2 := \mathbb{E}\left[X_1 - \mathbb{E}[X_1]\right]^2 + 2\sum_{j=1}^{\infty} \mathbb{E}\left[(X_1 - \mathbb{E}[X_1])(X_j - \mathbb{E}[X_j])\right] < \infty$.
Morever, if $\sigma \neq 0$ and $\mathbb{E}[X_1] = 0$, we also have

$$(\sigma\sqrt{n})^{-1} \sum_{j=1}^{n} X_j \xrightarrow{d} \mathcal{N}(0,1).$$

as $n \to \infty$.

Table of Contents

A clever way to use asymptotic independence

Ibragimov & Müller (2010), Meurice & Preinerstorfer (2021)

A workaround for the specific covariance structure problem when testing $\beta = 0$ was developped as follows.

(i) Divide the sample into q blocks of the same size

(ii) Estimate $\hat{\beta}^{(i)}$ in each block ($i = 1, \ldots, q$)

(iii) Test whether the mean of the new sample $\left\{\hat{\beta}^{(1)}, \ldots, \hat{\beta}^{(q)}\right\}$ is equal to 0 (or a function of it)

If the errors are α-mixing and follow Ibragimov's CLT conditions, most of the observations in one block will be asymptotically *really far away* from most of any other block, making both blocks asymptotically uncorrelated.

Adding the result of Ibragimov's CLT, this means that the new sample will be approximately independent and normally distributed. **We got our basic conditions back!**

Table of Contents

Table of Contents

Classical t-test

In the previous example, we would want to test whether the βs' sample had mean zero or not. The classical procedure to do this is called the Student's one sample t-test.

Student's t-test

Consider a sample of n i.i.d. random variables $X_1, \ldots, X_n$ such that $X_i \sim \mathcal{N}(\mu, \sigma^2)\ \forall i$.
We would like to test

$$H_0 : \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0$$

for some chosen $\mu_0 \in \mathbb{R}$. Define the test statistic

$$T := \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{\hat{\sigma}},$$

where $\bar{X}$ is the sample mean and $\hat{\sigma}$ the sample standard deviation.
Then, one would reject H_0 at level α whenever $|T|$ exceeds the $1 - (\alpha/2)$ quantile of Student's t-distribution with $n - 1$ degrees of freedom.

t-test assumptions I

The classical t-test requires most usual assumptions, such as:

- Independence
- Normality
- Homoskedasticity (constant variance)

The test on $\left\{\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(q)}\right\}$ works, because

- Independence is approximated by the α-mixing process
- Normality is approximated through Ibragimov's CLT
- Homoskedasticity can actually be relaxed thanks to some result by Bakirov & Szekely (2006) (not discussed here)

t-test assumptions II

In general, what happens if we want to test the same hypotheses, have independence and homoskedasticity but lack normality?

If we have a big sample, the classical Central Limit Theorem will mean $\bar{X}$ will be approximatively normal (which is what we need)

But what if we have a relatively small sample?

Table of Contents

Non-parametric tests

Non-parametric methods do not rely on assumptions regarding a specific (parametric) distribution, usually allowing for a more universally valid use.

The drawback is that they are typically less effective and powerful than parametric methods when those are applicable.

One of the most basic non-parametric options to test the mean value of a given sample is called a sign test.

Sign test I

Sign tests (including *the* sign test) rely only on some variant of the sign function sgn of the observations:

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

In any i.i.d. sample of median 0 (equivalent to the mean for symmetric distributions) of size n, the number of positive observations (i.e. sum of positive signs) n_+ follows a Binomial law $n_+ \sim Bin(n, 1/2)$.

Sign test II

One sample sign test

Suppose we observe a (small) sample $X_1, \ldots, X_n$ i.i.d. with some arbitrary symmetric continuous distribution.
We would like to test

$$H_0 : \mu = \mu_0 \quad \text{v.s.} \quad H_1 : \mu \neq \mu_0$$

for some chosen $\mu_0 \in \mathbb{R}$.
Let n_+^* be the number of observations strictly greater than μ_0.
Then, one would reject H_0 at level α whenever $n_+^* \notin [b_{n,\alpha/2}, b_{n,1-(\alpha/2)})]$, where $b_{n,p}$ is the p-quantile of the $Bin(n, 1/2)$ distribution.

Rank-based tests I

An other family of non-parametric methods for various location test problems is the one of rank-based tests.

Considering (as always) a sample $X_1, \ldots, X_n$, the rank of each observation $R(X_i)$ is the position in the sample after ordering it.

For example, say $X_1 = 6$, $X_2 = 10$ and $X_3 = 4$.
Then, $R(X_1) = 2$, $R(X_2) = 3$ and $R(X_3) = 1$.

Rank-based tests II

In any i.i.d. (continuous) sample, the rank of each observation $R(X_i)$ follows a $Uniform\{1, \ldots, n\}$ distribution.

Equivalently, the vector of all n ranks is distributed uniformly over the $n!$ permutations of $\{1, \ldots, n\}$.

A basic example of how this is useful is as follows. Suppose you have two samples, $X_1, \ldots, X_n$ with median μ and $Y_1, \ldots, Y_n$ with median ν. We would want to test

$$H_0 : \mu = \nu \quad \text{v.s.} \quad H_1 : \mu \neq \nu.$$

Wilcoxon's Rank sum test (intuition)

Without going into details, the rank sum test of Wilcoxon works as follows:

(i) Put both samples X and Y into one big common sample;

(ii) Compute the ranks of all observations in this new common sample;

(iii) Under the null (i.e. if $\mu = \nu$), on average, the ranks of the X's should be similar to those of the Y's;

(iv) If the ranks of either sample average much greater values than the other's, reject H_0.

Table of Contents

Conclusions

Classical assumptions are really useful, but can be violated easily.

Most assumptions can be relaxed to some extent. Mild dependence can be dealt with in big samples, as does non-normality.

Bigger workarounds are required in small samples, as asymptotic results do not apply anymore.

References

- Ibragimov, I. A. (1962). Some Limit Theorems for Stationary Processes. *Theory of Probability & Its Applications, 7*(4), 349–382. https://doi.org/10.1137/1107036
- Ibragimov, R., & Muller, U. K. (2010). t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics, 28*(4), 453–468. https://doi.org/10.1198/jbes.2009.08046
- Meurice, V., & Preinerstorfer, D. (2021). On a blocked t-test procedure for regression coefficients under possibly dependent errors. *Master thesis*.