Multivariate Statistics

Multivariate statistics deals with 2 or more random variables and their co-movement.

Covariance between 2 variables is defined as the expected value of the product of the deviation from their means. Covariance is defined on the real number line.

$$cov(x, y) \ = \ E\left[\left(x - \mu_x\right)\left(y - \mu_y\right)\right]$$

Where μ_x, μ_y are the mean of x and y

Numerically, this is calculated like the variance

$$cov(x, y) \ = \ \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)\right]$$

Wikipedia lists a few facts on the covariance of linear combinations of random variables, all of which are key to understand. Most should be obvious from the definition.

1. $cov(x, a) \ = \ 0$ the covariance of a variable and a constant is 0
2. $cov(x, x) \ = \ var(x)$ as stated above.
3. $cov(x, y) \ = \ cov(y, x)$ order does not matter
4. $cov(ax, by) \ = \ ab\, cov(x, y)$ covariance scales similar to variance
5. $cov(a + x, b + y) \ = \ cov(x, y)$ covariance is invariant to the mean.
6. $cov(ax + by,\ cw + dv) \ = \ ac\, cov(x, w) \ + \ ad\, cov(x, v) \ + \ bc\, cov(y, w) \ + \ bd\, cov(y, v)$
7. $var\left(\sum_{i}^{n} a_i X_i\right) \ = \ \sum_{i}^{n}\sum_{j}^{n} a_i a_j cov\left(X_i, X_j\right)$ where X is a vector of random variables and a is a vector of constants

Correlation is a statistic, ranging between -1 and 1, that describes how strongly 1 variable moves with another.

There are 2 common correlation calculation methodologies used in risk management.

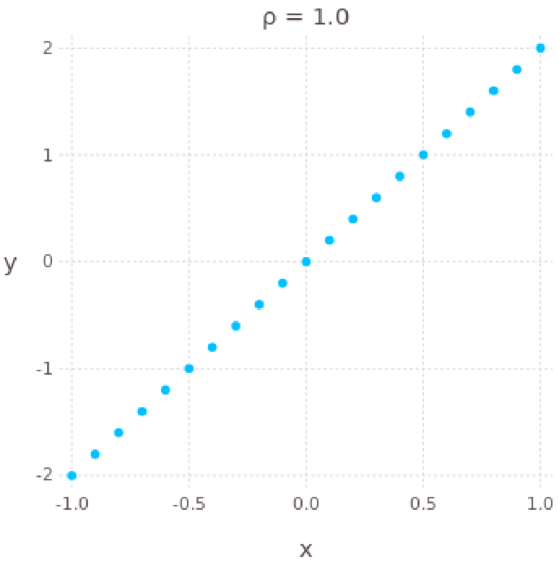
1. Pearson coefficient is the most widely used. It is linear measure.
2. Spearman's rank coefficient is a measure of monotonicity. It is less sensitive to outliers and does not depend on a linear relationship.

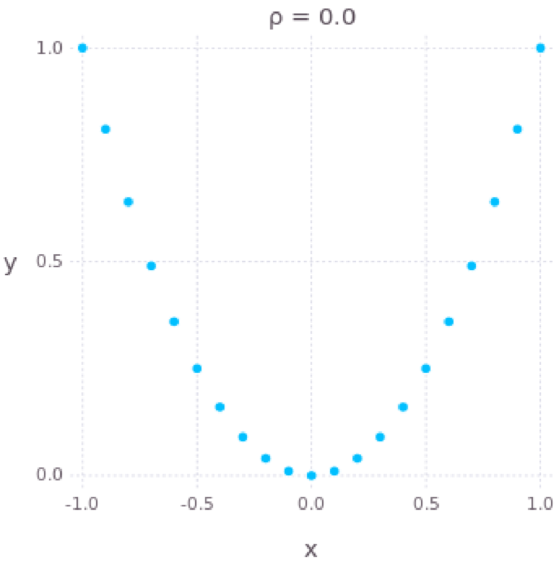
There are other measures of correlation and you are encouraged to research on your own.

Pearson Coefficient is a linear measure of correlation. It is the most common form of correlation taught. 99% of the time when someone uses the phrase “correlation,” they mean the Pearson Coefficient.

Pearson coefficient is the covariance scaled into the [-1,1] range.

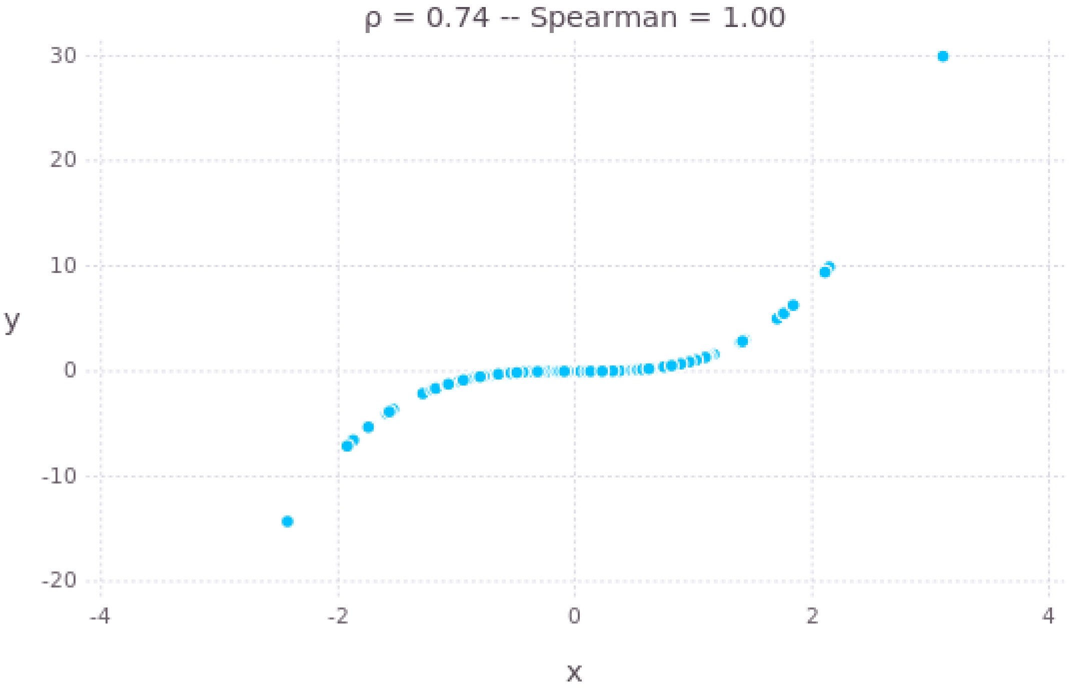
$$\rho_{i,j} = \frac{cov(X_i, X_j)}{\sigma_i \sigma_j}$$





Spearman Coefficient is a nonparametric measure of correlation. Spearman uses ranks to calculate the correlation. Because of this, the correlation is a measure of monotonicity.

Here $y = x^3$



Because the function is perfectly monotonically increasing, the Spearman coefficient is 1.

$$r_s = \rho_{R(x),R(y)} = \frac{cov(R(x),R(y))}{\sigma_{R(x)}\sigma_{R(y)}} = \frac{\sum_{i=1}^{n}\left(R\left(x_i\right) - \overline{R(x)}\right)\left(R\left(y_i\right) - \overline{R(y)}\right)}{\sqrt{\sum_{i=1}^{n}\left(R\left(x_i\right) - \overline{R(x)}\right)^2}\sqrt{\sum_{i=1}^{n}\left(R\left(y_i\right) - \overline{R(y)}\right)^2}}$$

Where

$\rho_{x,y}$ is the Pearson coefficient of x and y

$R(x)$ is the rank function

$\overline{R(x)}$ is the average of the rank function

The rank function assigns integer values from lowest to highest values in the input vector. Ties are broken by averaging the rank of all the tied values.

For example:

Value	Order	Rank
1.2	5	5
0.8	2	3
1.3	6	6
0.8	3	3
0.8	4	3
0.5	1	1

Multivariate Normal Distribution
The most commonly used multivariate distribution is the multivariate normal. It is crucial to understand the dynamics of this distribution.

Notation

$$X \sim N(\mu, \Sigma)$$

X is a $(Nx1)$ random vector
μ is the $(Nx1)$ vector of expected values
Σ is the (NxN) covariance matrix.

$$\Sigma_{i,j} = cov\left(X_i, X_j\right)$$

Σ is positive definite in the non-degenerative case. We will deal with this later.

$$w'\Sigma w \geq 0 \; \forall \, w \in \mathbb{R}^n$$

Note: $Var(w'X) = w'\Sigma w$

PDF

$$f(X) \; = \; \frac{e^{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)}}{\sqrt{(2\pi)^{k}|\Sigma|}}$$

If

$$X \sim N(\mu, \Sigma)$$

Then there exists

$$\mu \in \mathbb{R}^{n},\ L \in \mathbb{R}^{nxm}\ such\ that\ X \; = \; LZ \; + \; \mu$$

Where

$$Z_{i} \sim N(0, 1),\ \ i \; \in \; [1, m]$$

$$\Sigma \; = \; LL'$$

This gives us a way to simulate multivariate normal vectors. We will do this later.

Conditional Distributions – given a known value(s) of random values in the normal vector, what is the distribution of the remaining variables?

$$X \sim N(\mu, \Sigma)$$

Partition $X,\ \mu,\ \&\, \Sigma$ as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times 1 \\ (N-q) \times 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} q \times q & q \times (N-q) \\ (N-q) \times q & (N-q) \times (N-q) \end{bmatrix}$$

(credit to Wikipedia for nicely formatted matrices)

Given

$$X_2 = a$$

Then

$$X_1 \sim N(\overline{\mu}, \overline{\Sigma})$$

Where

$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}$$

Regression - OLS

Ordinary Least Squares, or OLS, regression is the most widely used form of regression analysis. There are 7 assumptions made for this model:

1. The model is linear in the parameters and error term. I.e. the model must be specified as such:

   $$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon$$

2. The error term, ϵ has 0 mean.
3. The regressors (X) or "independent variables" are uncorrelated with the error term.
4. The values of the error term are uncorrelated with each other

   $$E(\epsilon_i \epsilon_j) = 0 \, \forall \, i \neq j$$

5. The error term has a constant variance - -the variance does not vary over the sample.

   > Assumptions 4 + 5 mean the error term is identically and independently distributed (*iid*)

6. There is not a perfect linear dependence between the regressors.
7. The error term is normally distributed.

#7 can be broken, however, if assumed, then hypothesis testing along with confidence and prediction intervals can be calculated.

Assuming #7 and #4, and allowing an intercept term in your model of #1 (β_0), then #2 flows naturally. All expected value after accounting for the regressors is carried in the intercept and the error term will have 0 expected value.

Explore the relaxation of #3 on your own time.

We will deal with the relaxation of #4 and #5 later.

A violation of #6 usually means you need to reduce the number of variables in your analysis. We will discuss other methods later in this class. In the next section, we will see mathematically why this is an issue.

OLS

Construct the regressor matrix

$$X = [1 \ X]$$

Where 1 is a vector of 1s and X is a matrix of n observations of m dependent variables. The final matrix is then $(n, m + 1)$. Assume $m + 1 < n$.

$$Y = X\beta$$

This system is overdetermined and likely has no solution. We seek to approximate β such that

$$Y = X\widehat{\beta} + \epsilon$$

To find the smallest possible error, solve for $\widehat{\beta}$ so that

$$\epsilon = Y - X\widehat{\beta}$$

$$minimize\, S(\widehat{\beta}) = \epsilon'\epsilon$$

Recognize this as the sum of the squares of the errors.

If $\epsilon_i \sim N(0, \sigma^2)$ for all i, then this is the same as minimizing the variance of the errors

From a linear algebra perspective:

$$Y = X\beta$$

Multiply each side by $(X'X)^{-1}X'$

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\beta$$

$$(X'X)^{-1}X'Y = I_{m+1}\beta$$

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\beta}_i = \frac{cov(x_i, y)}{\sigma_y^2} = \frac{\sigma_{x(i)}\rho_{x(i),y}}{\sigma_y}$$

If all the assumptions 1-6 hold, then $\widehat{\beta}$ is the solution to the minimization problem. If assumption 7 holds, then this solution is also the maximum likelihood estimator.

Let $p = m + 1$

$$s^2 = \frac{\epsilon'\epsilon}{n-p}, \; \widehat{\sigma}^2 = \frac{n-p}{n}s^2$$

$$E(s^2 \mid X) = \sigma^2, \; E(\widehat{\beta} \mid X) = \beta$$

$$Var(\widehat{\beta} \mid X) = \sigma^2(X'X)^{-1}$$

The standard error of B is then

$$s.e.\,\hat{\beta} = s\sqrt{diag\left((X'X)^{-1}\right)}$$

Where diag() is the vector of the diagonal elements.

Implications of assumption #6. If the values of X are linearly dependent, then $X'X$ will not be full rank and $(X'X)^{-1}$ will not exist.

We can measure the amount of linear dependence by looking at the smallest eigenvalue of $(X'X)$. As this value tends to 0, the linear dependence among the regressors is increasing. This is called “multicollinearity.” The inverse becomes unstable, and as such so are the $\widehat{\beta}$ values. The standard errors of the coefficients rise.

The Durbin Watson statistic can be calculated with many statistical packages. It is a measure of serial correlation of the errors of a regression (breaking #4). Values near 2 indicate there is no serial correlation.

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) attempts to find parameter values to maximize the joint probability of the observed data.

In simple terms, the likelihood can be thought of as the product of the PDF values:

$$l \ = \ \prod_{i=1}^{n} f(Y \mid X; \ \beta)$$

Numerically, PDF values are < 1 – the CDF is the integral of the PDF and evaluated across the support equals 1. This causes problems as floating point math on computers loses precision as we start looking at smaller and smaller numbers.

We take the log of both sides to get

$$ll \ = \ \sum_{i=1}^{n} ln(f(Y|X; \beta))$$

MLE is then the maximization of this function.

For the normal distribution

$$l \ = \prod_{i=1}^{n} f(x_i \, ; \ \mu, \sigma) \ = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_i - \mu}{\sigma})^2}$$

$$ll \ = \ \sum_{i=1}^{n} ln\left(f\left(x_i\right)\right) \ = \sum_{i=1}^{n} \left[- \frac{1}{2} ln\left(\sigma^2 2\pi\right) - \frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \right]$$

$$ll \ =- \frac{n}{2} ln\left(\sigma^2 2\pi\right) \ - \ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i \ - \ \mu\right)^2$$

You can solve this maximization problem to show the MLE estimate for the Normal Distribution are

$$\hat{\mu} \ = \ \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \widehat{\sigma^2} \ = \frac{1}{n} \sum_{i=1}^{n} (x_i - \ \mu)^2$$

Notice the maximum likelihood estimator for the variance is NOT unbiased.

When fitting a regression model, you are fitting the likelihood to the error term.

$$\epsilon = Y - X\widehat{\beta}$$

We want to set the mean to 0.

$$ll = -\frac{n}{2} ln\left(\sigma^2 2\pi\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\epsilon_i - 0\right)^2$$

You can solve this problem to show that the solution for $\hat{\beta}$ is the same as the OLS solution.

Goodness of Fit

Often we have multiple models to choose between. Goodness of Fit statistics help with this process

R-Square, R^2, is the variance explained by the model as a fraction of the total independent variable's variance.

$$SS_{total} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

$$SS_{error} = \sum_{i=1}^{n}\left(\epsilon_i - \bar{\epsilon}\right)^2 = \sum_{i=1}^{n}\epsilon_i^2$$

$$SS_{model} = SS_{total} - SS_{error}$$

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

The biggest problem with $R^2$is that as variables are added, it continuously increases, regardless if they are explanatory.

In linear algebra terms, as you add additional values to X then

$$Y = X\beta$$

Becomes more deterministic. Once X is a square matrix, then the system can have a singular solution and $var(\epsilon) = 0$.

Adjusted R Square penalizes the metric for additional explanatory variables.

$$Adj\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{SS_{error}}{SS_{total}}\frac{n-1}{n-p-1}$$

Information Criteria

There are 2 different statistics widely used for model selection. Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). Both use the sum of the log likelihood functions, penalized for the number of parameters.

$$AIC = 2k - log(\hat{L})$$

In small sample sizes, AIC can tend to select models with higher numbers of parameters and overfitting of the model (like R^2). The following correction is made:

$$AICc = AIC + \frac{2k^2 + 2k}{n-k-1}$$

$$BIC = k\ ln(n) - log(\hat{L})$$

In each of these formulas,

$$k = p + d$$

Where d is the number of additional parameters fit for the distribution in the MLE process. If a model has q slope parameters, $p = q + 1$ (+1 for the intercept). Using the normal distribution, we have to account for the fitted variance, $d = 1,\ k = p + 1$

Time Series

One of the key assumptions in linear regression is the independence of the error

$$E(\epsilon_i \epsilon_j) = 0\ \forall\ i \neq j$$

When this is broken, the results from your regression are suspect. The beta values are consistent however the standard errors are not. Consistent means that as the number of observations goes to infinity, the estimate will converge to the true value. However, the estimate in small sample sizes is not “best,” it will have a much wider (larger variance) distribution.

Time series models are built in a way to get back to the assumptions of linear regression for the errors.

1. Uncorrelated $E(\epsilon_i \epsilon_j) = 0\ \forall\ i \neq j$
2. 0 Expected value $E(e) = 0$
3. Constant variance $var(e) = \sigma^2$

Highly recommended reading:
https://mfe.baruch.cuny.edu/wp-content/uploads/2014/12/TS_Lecture1_2019.pdf

Autocovariance, Autocorrelation, and Partial Autocorrelation

Autocovariance is the covariance across time periods:

$$\gamma(n) = cov(y_t, y_{t-n})$$

The Autocorrelation Function (ACF) is the Person correlation across time periods

$$\rho(n) = \frac{\gamma(n)}{\sigma_t \sigma_{t-n}}$$

The Partial Autocorrelation Function (PACF) is the residual autocorrelation left in the series when $n - 1$ lags have been taken into account. That is, the autocorrelation of the residual of the series regressed on the $n - 1$ lags.

ACF and PACF are helpful for identifying the type of model and appropriate lags to use.

The structural Autocovariance matrix determined by a specific model is used during MLE fitting.

Moving Average processes, the independent variable is a function of past error terms.

MA(1)

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

innovations (random errors) have a 2 period impact.

$$E(y_t) = E(\mu + \epsilon_t + \theta\epsilon_{t-1}) = E(\mu) + E(\epsilon_t) + \theta E(\epsilon_{t-1}) = \mu$$

The expected value is the mean.

$$var(y_t) = var(\mu + \epsilon_t + \theta\epsilon_{t-1})$$

μ is constant so we can take that out of the equation

$$var(y_t) = var(\epsilon_t) + \theta^2 var(\epsilon_{t-1}) + 2\theta cov(\epsilon_t, \epsilon_{t-1})$$

$$var(y_t) = \sigma^2 + \theta^2\sigma^2 = (1 + \theta^2)\sigma^2$$

The variance of the series is different from the variance of the errors.

$$E(y_t|\Omega_{t-1}) = E(\mu + \epsilon_t + \theta\epsilon_{t-1}|\Omega_{t-1}) = \mu + 0 + \theta\epsilon_{t-1}$$

The autocovariance of MA(1)

$$\gamma(1) = E(y_t, y_{t-1}) = \theta\sigma^2$$

Because y_t does not depend on values beyond y_{t-1}

$$\gamma(k) = E(y_t, y_{t-k}) = 0 \, \forall \, k > 1$$

The ACF

$$\rho(1) = \frac{\gamma(1)}{(1+\theta^2)\sigma^2} = \frac{\theta}{1+\theta^2}$$

$$\rho(k) = 0 \, \forall \, k > 1$$

Reconfiguring the MA(1)

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$$

let

$$\widehat{y_t} = y_t - \mu = \epsilon_t + \theta\epsilon_{t-1}$$

$$\epsilon_t = \hat{y}_t - \theta\epsilon_{t-1}$$

$$\epsilon_{t-1} = \widehat{y_{t-1}} - \theta\epsilon_{t-2}$$

$$\epsilon_t = \hat{y}_t - \theta\left(\widehat{y_{t-1}} - \theta\epsilon_{t-2}\right) = \hat{y}_t - \theta\widehat{y_{t-1}} + \theta^2\epsilon_{t-2}$$

Indefinitely

$$\epsilon_t = \hat{y}_t + \sum_{i=1}^{\infty} (-\ \theta)^i \widehat{y_{t-i}}$$

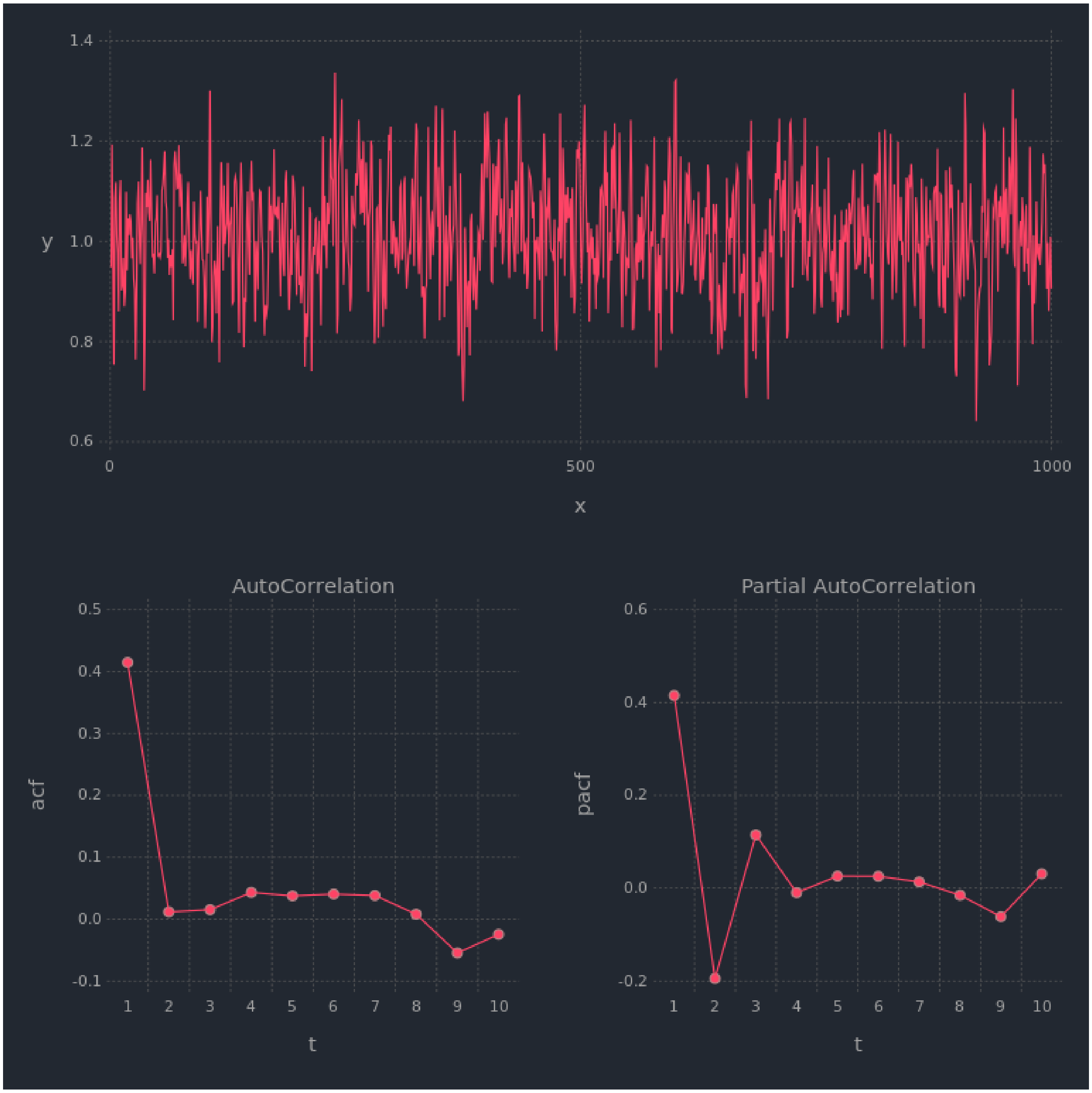
$$\hat{y}_t =- \sum_{i=1}^{\infty} \left((-\ \theta)^i \widehat{y_{t-i}} \right) + \epsilon_t$$

For y to be defined, $|\theta| < 1$

MA(q) processes extend the MA(1) into more lags.

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Example MA(1) Process:


1.4
1.2
y
1.0
0.8
0.6
0
500
1000
x
AutoCorrelation
0.5
0.4
0.3
0.2
acf
0.1
0.0
-0.1
1
2
3
4
5
6
7
8
9
10
t
Partial AutoCorrelation
0.6
0.4
pacf
0.2
0.0
-0.2
1
2
3
4
5
6
7
8
9
10
t

Autoregressive Process the independent variable is a function of its own lags

AR(1)
Easiest to work with if we center it on 0.

$$\widehat{y_t} = y_t - \mu$$

$$\widehat{y_t} = \beta\widehat{y_{t-1}} + \epsilon_t$$

Back substitute -

$$\widehat{y_t} = \sum_{i=1}^{\infty}\left(\beta^{i}\epsilon_{t-i}\right)$$

This is an special case of an infinite moving average where $\theta_i = \theta_j \ \forall\, i, j$
Also notice from above, the MA(1) is a special case of an infinite AR process.

Like the MA, y is only defined if $|\beta| < 1$

$$E(\widehat{y_t}) = E\left(\sum_{i=1}^{\infty}\left(\beta^{i}\epsilon_{t-i}\right)\right) = 0 \Rightarrow E(y_t) = \mu$$

Since μ is constant

$$var(y_t) = var(\widehat{y_t}) = var\left(\sum_{i=1}^{\infty}\left(\beta^{i}\epsilon_{t-i}\right)\right) = \sigma^2 \sum_{i=1}^{\infty}\left(\beta^{2i}\epsilon_{t-i}\right) = \frac{\sigma^2}{1-\beta^2}$$

If $|\beta| = 1$ then the variance of y is infinite. Greater than 1 and it is undefined as the variance cannot be negative.

In our original equation, we had an intercept term, C

$$E(y_t) = E(c + \beta y_{t-1} + \epsilon_t)$$

$$\mu = E(c) + \beta\mu \Rightarrow C = (1 - \beta)\mu \Rightarrow \mu = \frac{c}{1-\beta}$$

Autocovariance

$$\gamma(k) = \beta\gamma(k - 1) = \frac{\sigma^2\beta^k}{1-\beta^2}$$

Autocorrelation

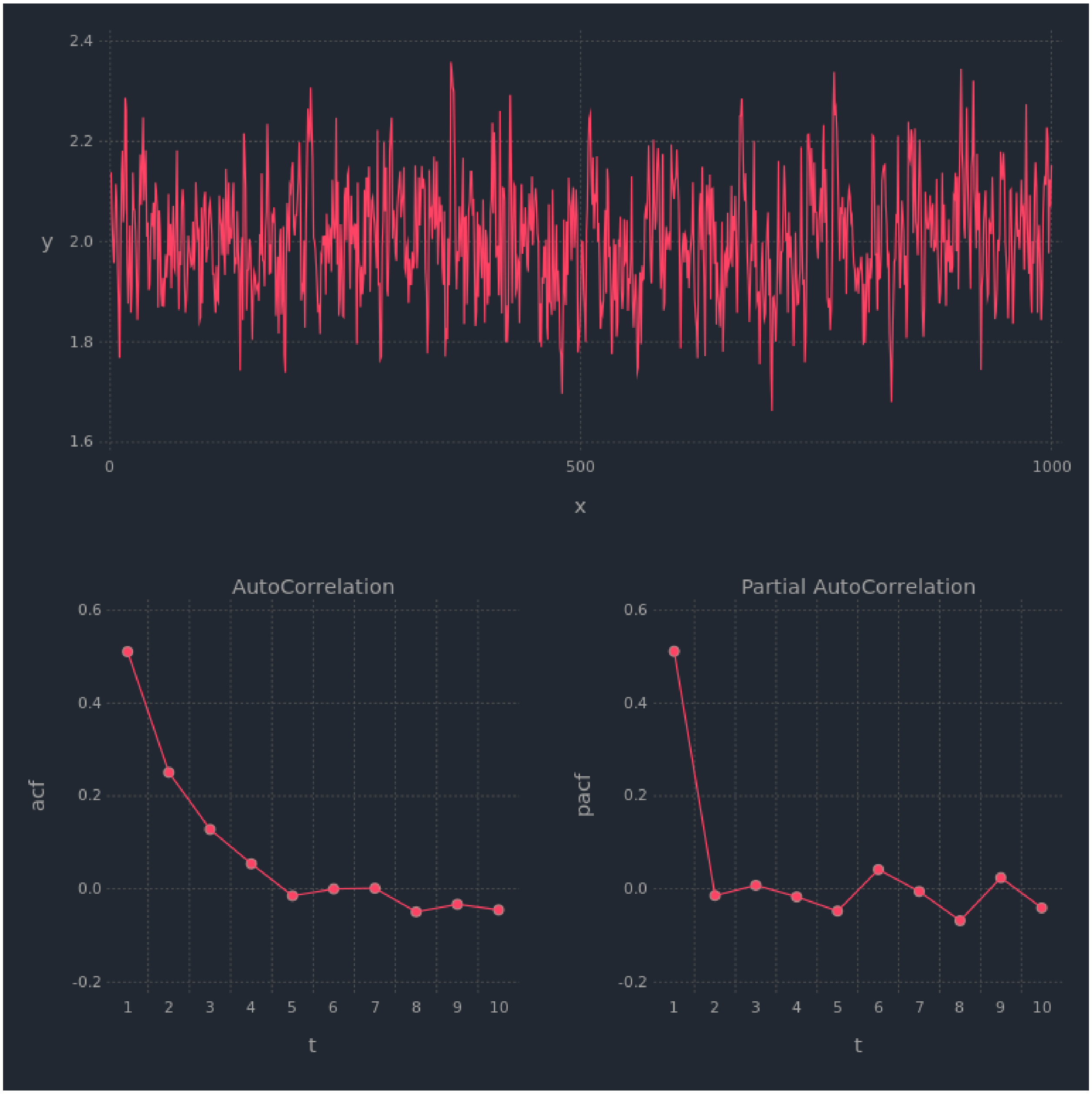
$$\rho(k) = \beta^k$$

Here, too, we can see that if $|\beta| = 1$ the variance is infinite, and if it is greater than 1, the variance is undefined.

β determines the persistence of the correlation through time. A beta near 0 will quickly decay to 0, while a beta near 1 (or -1) will persist long into the future.

Beta = 1 is a random walk.

Example AR(1)



AR and MA processes can be combined into ARMA(p,q) models – a combination of AR(p) + MA(q).

When we include additional regressor (exongenous) variables, we call this an ARMAX.

If those other regressors are themselves serially correlated, this breaks other assumptions and we then have to estimate their ARMA process (with cross error effects) in a simultaneous equations model called VARMAX. V is for Vector. Usually you see these are VAR, vector autoregression, models with the MA component.

There are multiple ways to fit AMRA models. All are involved. Most statistics software has model fitting for these.

For the MLE technique, we have to add the Autocovariance matrix to the likelihood function.

$$ll \ =- \frac{1}{2}\left(n * ln(2\pi) + ln(|\Gamma(\theta, \beta|) \ + X'\Gamma(\theta, \beta)^{-1}X\right)$$

Where

$$X \ = \ Y - \ \mu$$

$$\Gamma(\theta, \beta)$$

Is the Autocovariance matrix of size (nxn) implied by the values of the MA (θ) and AR (β) parameters.

We are finding the max of the function above for μ, θ, β.

GARCH Models

Another assumption that can break OLS regression is the constant variance of the errors. If the variance of the error term changes overtime, this is called heteroskedasticity (homoscedasticity is constant variance). One class of model to account for this the GARCH Model where we apply the ARMA framework to the variance of the error term

$$\epsilon_t = y_t - \ \mu$$

$$\epsilon_t = \sigma_t z_t \, where \ z \sim N(0, 1) \ iid$$

Here, z becomes our constant and uncorrelated error process

The GARCH(p,q) model is then

$$\sigma_t^2 = \ \omega \ + \sum_{i=1}^{p}\left(\alpha\epsilon_{t-i}^2\right) + \sum_{i=1}^{q}\left(\beta\sigma_{t-i}^2\right)$$

We can get the long run average variance by taking the expectation of both sides

For the GARCH(1,1) model

$$\sigma^2 = E\left(\sigma_t^2\right) = E\left(\omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2\right)$$

$$\sigma^2 = \omega + \alpha\sigma^2 + \beta\sigma^2 = \frac{\omega}{1-\alpha-\beta}$$

This puts bounds on alpha and beta. Like the ARMA(1,1) coefficients, the must be <1. Also their sum must be less than 1 $\alpha + \beta < 1$.