# Metrics Summary

Jannes Klaas

March 2019

## 1 Basics

### 1.1 Variance

Variance is the expectation of the squared deviation of a random variable from its mean.

$$Var(X) = E[(X - \bar{X})^2] \tag{1}$$

Rearranging (1) we find that variance equals the expectation of the square minus the square of the expectation.

$$\boxed{Var(X) = E[X^2] - E[X]^2} \tag{2}$$

Variance is invariant to adding a constant:

$$Var(X + a) = Var(X) \tag{3}$$

If all values are scaled by a constant, the variance is scaled by the square of that constant:

$$Var(aX) = a^2 Var(X) \tag{4}$$

The variance of two random variables can be added:

$$\boxed{\begin{aligned} Var(aX + bY) &= a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y) \quad (5)\\ Var(aX - bY) &= a^2 Var(X) + b^2 Var(Y) - 2abCov(X, Y) \quad (6) \end{aligned}}$$

Notice how this makes diversification possible. If we have two uncorrelated assets with $Var(X) = Var(Y) = 1$, then holding a 50/50 portfolio has variance $0.5^2 + 0.5^2 = 0.5$, we halved the risk! But if the assets are perfectly correlated so that $Cov(X, Y) = 1$ then the portfolio variance is $0.5^2 + 0.5^2 + 2 * 0.5^2 = 1$ so we have not won anything by diversification.

## 1.2 Covariance

Covariance is a measure of the *joint* variability of two random variables. Notice the similarity between variance and covariance, both are the expectation of the multiple minus the multiple of the expectation:

$$Cov(X,Y) = E(X - \bar{X})E(Y - \bar{Y}) \tag{7}$$

Rearranged:

$$\boxed{Cov(X,Y) = E(XY) - E(X)E(Y)} \tag{8}$$

Correlation is just a scaled version of the covariance to achieve comparability:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{9}$$

So that $Corr(X,Y) \in [0,1]$

If $X$ is a multivariate matrix (e.g. prices of multiple assets in one matrix) then $Cov(X)$ is the covariance matrix of all rows in the vector.

$$\Sigma = Cov(X) \tag{10}$$

**Question**: Two variables have zero correlation, are they unrelated?

*Answer*: At least they are not linearly related, but there might still be some non linear relation.

**Question**: Assume we have a portfolio of stocks with returns $X \sim N(0,\Sigma)$ and a weight matrix $B$ so that our returns are $BX$. What is the new covariance of our portfolio?

*Answer*: This is an important result that you can derive from the variance scaling property.

$$Cov(BX) = BCov(X)B' = B\Sigma B' \tag{11}$$

Note how only the matrix notation is different, but basically the variance scales $B^2$ as in (4).

## 1.3 Conditional Variance

If $Y$ is an asset return and $X$ is a trading signal and they are both normally distributed have a correlation of $\rho$, then the conditional expectation and variance is:

$$E(Y|X) = \rho X \tag{12}$$

$$Var(Y|X) = 1 - \rho^2 \tag{13}$$

### 1.4 Statistical tests

Statistical tests test if a null hypothesis $H_0$ can be rejected in favor of an alternative hypothesis $H_1$. **Type I error** means falsely rejecting the null while **Type II error** means not rejecting the null hypothesis when it should have been. The **size** (sometimes written as $\alpha$) of a statistical test is the probability of making a type I error. A **p-value** is the probability of observing the data given that the null is correct. The size is thus the p-value we require to reject the null (the test is significant at $\alpha$ level). The **power** of a statistical test is the inverse of probability of making a type II error. That is, the probability that the study will detect an effect if there is one. While size can be set freely, power is dependent on factors such as the effect size and sample size.

## 2 Maximum Likelihood Estimation (MLE)

Key idea: Given a known density function $f$ with parameters $\theta$, choose $\theta$ to maximize the density of the data $Y|X$ under $f$. The likehood of $\theta$ is thus the density of $f$ at point $Y|X$ which is the product of the density of the individual samples.

### 2.1 Linear Regression MLE

Assuming a normal error distribution, the likehood equals:

$$\boxed{\begin{aligned} Y|X = x &\sim_{iid} N(x'\beta, \sigma^2) \qquad (14)\\ L(\theta, y) = f_{Y|X=x}(y;\theta) &= \prod_{i=1}^{n} f_{Y_i|X_i=x_i}(y;\theta_i) \qquad (15)\end{aligned}}$$

Because the product leads to tiny values for $L$ we maximize the log likehood:

$$\boxed{\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\theta, y)} \qquad (16)$$

### 2.2 Properties of MLE

1. In some cases no MLE exists. The optimizer will find no point where $\frac{d \log L(\theta,y)}{d\theta}) = 0$ for all $\theta$.
2. The $\hat{\theta}$ value found by (16) does not have to be unique, there might be local optima.
3. If the model $f$ is correct, MLE will converge on the correct parameters as sample size goes to infinity.
4. The distribution around the true parameter $\theta_0$ is $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I^{-1})$ where $I$ is the information, aka the covariance in likelihood gradient between parameters when $\hat{\theta} = \theta_0$.

5. **Information Equality**: If the model is correct, the likehood is fully differentiable and the support of $Y$ does not depend on $\theta$:

$$I = Cov(\frac{d \log L(\theta, y)}{d\theta})|_{\theta=\theta_0} \tag{17}$$

$$= -E(\frac{d^2 \log L(\theta, y)}{d\theta d'\theta}) \tag{18}$$

6. **Cramér-Rao Inequality**: Establishes a lower bound on the variance of an MLE estimator. Under the same assumptions as before, for large sample sizes $n$ and an unbiased estimator $\tilde{\theta}$, $Cov(\tilde{\theta}) - I^{-1} \geq 0$ (this is closely related to 4)

**Question**: What is maximum likelihood estimation, does it always exist and is it unique?

*Answer*: MLE is a method that maximizes the likehood of a parameterized density function. There might not always be an MLE solution and it might not always be unique because of local optima etc.

### 2.3 Common MLE's You Can Compute by Hand

1. **Normal Distribution**: $\mu = \bar{X}$ and $\sigma^2 = 1/N \sum_{i=1}^{N} (X_i - \bar{X})^2$
2. **Poisson Distribution**: $\lambda = \bar{X}$
3. **Bernoulli Distribution**: $p = \bar{X}$
4. **Exponential Distribution**: $\lambda = 1/\bar{X}$

## 3 The OLS Estimator

### 3.1 Basic Setup

Linear regression has two functions: Estimation and hypothesis testing. In systematic trading, the hypothesis testing is especially important.

$$\hat{y} = \alpha + \beta_1 x_1 + ... + \beta_n x_n \xrightarrow{Simplify} \hat{y} = \beta X \tag{19}$$

When we say linear regression, we pretty much always mean ordinary least squares (OLS) linear regression that solves:

$$\underset{\beta}{\text{argmin}} (y - \beta X)^2 \tag{20}$$

If only $X$ or $y$ is scaled, $\beta$ changes, too, but the p-values stay the same because the standard error also scales.

### 3.2 Assumptions

1. The relationship is actually linear. Linear regression might not find e.g. $y = X^2$
2. $X$ is fixed (little to no measurement error)
3. Constant variance (a.k.a. homoscedasticity), econonometrics has developed a large number of tools to deal with heterocedasticity.
4. Independence of errors: In finance, errors can be autocorrelated and again, there is a toolbox to deal with that.
5. Lack of perfect multicollinearity. If features are perfectly correlated, OLS will actually not compute. If they are somewhat correlated, the estimate will have significantly higher variance.

### 3.3 $\beta$ and $Var(\beta)$

You can think of the the OLS solution as $\hat{\beta} = X * y/(X^2)$, or formally:

$$\boxed{\hat{\beta} = (X'X)^{-1}X'y} \tag{21}$$

The residual $\epsilon$ of a linear regression is a vector and equals $\epsilon = (y - \hat{y})$. In linear regression $E[\epsilon] = 0$

$$\boxed{Var(\hat{\beta}) = (X'X)^{-1}Var(\epsilon)} \tag{22}$$

Having $\hat{\beta}$ and $Var(\hat{\beta})$ we then usually conduct a t-test (where the degrees of freedom equal sample size - 1) in which $H_0 : \beta = 0$. If we find the $\beta$ to be significant non zero, the feature is indeed predictive.

### 3.4 The Gauss Markov Theorem

**OLS linear regression is the minimum variance unbiased estimator**. This is intuitive because a) OLS does not bias parameters (it has no priors, penalties, etc.) and b) OLS minimizes the error variance. However, there are good reasons to bias an estimator (e.g. regularization) and there are lots of non-linear relationships where OLS will fall short.

## 4 Heteroscedasticity

Heteroscedasticity means that the variance is not constant and usually dependent on some variable so that e.g. $y = \alpha + \beta x + \epsilon x$. Since $\epsilon$ is important to estimate $Var(\beta)$ ,see (22). Heteroscedasticity influences the variance estimates of the coefficients and can lead to faulty hypothesis tests. **Intuitively, we tackle covariance by multiplying errors with the square of X and dividing by the square of X, this removes the effect of X**, technically we multiply with $X'X$ and $(X'X)^{-1}$, but the intuition holds.

### 4.1 White Test for Heteroscedasticity

If $\epsilon$ is dependent on $X$ in some way, we can test for heteroscedasticity with an auxiliary regression on the *squared* error:

$$\boxed{\hat{\epsilon}_i^2 = \delta_0 + \delta_1 x_i + \delta_2 x_i^2 + \eta_i} \tag{23}$$

Where $H_0 : \delta_j = 0, \forall j > 0$. In practice we test the hypothesis with an $R^2$ and $\chi^2$ test.

### 4.2 White's covariance estimator

To estimate the error covariance robustly, we effectively have to apply our intuition and "multiply and divide by the square of X". Whites estimator is:

$$\hat{\Sigma}_{XX}^{-1} \hat{S} \hat{\Sigma}_{XX}^{-1} = n(X'X)^{-1}(X'\hat{E}X)(X'X)^{-1} \tag{24}$$

Where $\hat{E}$ is a matrix with the errors $\epsilon$ on its diagonal. **Using the heteroscedastic robust covariance estimator where it is not needed (homoscedastic data) leads to worse small sample properties. It e.g. distorts the test size so that a 5% test can reject the null 10% of the time even if it is true.**

## 5 Stationary Time Series

A process $\{y_t\}$ is strictly stationary if the joint distribution of $\{y_t, y_{t+1}, ..., y_{t+h}\}$ only depends only on $h$ and not on $t$.

### 5.1 Covariance stationary

A process is covariance stationary if

$$E[y_t] = \mu \text{ for } t = 1, 2, ... \tag{25}$$

$$V[y_t] = \sigma^2 < \infty \text{ for } t = 1, 2, ... \tag{26}$$

$$E[(y_t - \mu)(y_{t-s} - \mu)] = \gamma_s \text{ for } t = 1, 2, ..., s = 1, 2, ..., t-1 \tag{27}$$

Notice how covariance stationarity implies three types of stationarity, **mean stationarity**, (25), **variance stationarity** (26) and **autocovariance stationarity** (27). A **white noise** process is a covariance stationary process in which the autocovariance is zero.

### 5.2 Mean Stationarity

A process that is mean stationary must be **mean reverting**, speak if it diverges from the mean it sooner or later comes back to it. Common causes for a process *not* to be mean stationary are **unit roots** or **linear trends**. Trends can be

checked by regressing the time series against the time $t$. Unit roots can be checked e.g. with the Dickey Fuller test.

# 6 Autoregressive Models

## 6.1 AR(1)

An autoregressive model is simply an OLS model of a time series variable.

$$X_t = \phi_0 + \phi_1 X_{t-1} + \epsilon, \epsilon \sim iid(0, \sigma^2) \tag{28}$$

Assuming stationarity, it's moments are not time dependent.

$$E(X) = \phi_0 + \phi_1 E(X) + E(\epsilon) = \frac{\phi_0}{1 - \phi_1} \tag{29}$$

$$Var(X) = E(X_t^2 - E(X)^2) = \frac{\sigma^2}{1 - \phi_1^2} \tag{30}$$

$$Cov(X_t, X_{t-j}) = \frac{\phi_1^j}{1 - \phi_1^2}\sigma^2 \tag{31}$$

## 6.2 Stationarity

**The AR(p) process is stable if the roots of the lag polynomial lie outside the unit circle**, for an AR(1) that means $|\phi_1| < 1$. The general proof of condition for all AR models works by transforming the AR(q). For an AR(2), this means that $\phi_1 + \phi_2 < 1$, $\phi_1 - \phi_2 < 1$ and $|\phi_2| < 1$.

## 6.3 Forecasting

Forecasts can be computed recursively.

$$E_t[y_{t+h}] = \phi_0 + \phi_1 E_[y_{t+h-1}] \tag{32}$$

Long run forecasts are computed with a sum:

$$\boxed{E_t[y_{t+h}] = \sum_{i=0}^{h-1} \phi_1^i \phi_0 + \phi_1^h y_t} \tag{33}$$

The **forecast variance** is driven only by $\epsilon$, so $Var_t(X_{t+1}) = Var(\epsilon_{t+1}) = \sigma^2$

# 7 Volatility Modeling

## 7.1 Realized variance

Realized variance $RV$ is a simple measure of the variance of a security return that day. Given $n$ intraday return, the realized variance equals.

$$\boxed{RV = \sum_{i}^{n} r_i^2} \tag{34}$$

Note that this approach **assumes that intraday returns have a mean of zero**. In general, daily returns are assumed to be mean zero as well. The daily variance given a daily return is thus $\sigma^2 = r^2$

## 7.2 ARCH

Arch models are by and large simple autoregressive models:

$$r_t = \epsilon_t \tag{35}$$
$$\epsilon_t = \sigma_t e_t \tag{36}$$
$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 \tag{37}$$
$$e_t \sim N(0,1) \tag{38}$$

Returns are mean zero and a random process (35), the random process consists of the volatility and some random term (36). The volatility is dependent on some long run level $\omega$ and the past $\epsilon_{t-1}^2$, (37). We can estimate how it depends on past returns by estimating the best fitting $\alpha_1$ value. **ARCH models can have a non-zero mean return as well which is then fit with an AR process, but that is less interesting, so we will omit it here**.

1. Conditional and unconditional mean of $\epsilon_t$ is always zero
2. The conditional variance is $\sigma_t^2$
3. The unconditional variance is $\omega/(1-\alpha_1)$
4. The jth autocovariance is $\alpha_1^j V[\epsilon_t^2]$
5. The jth autocorrelation is $\alpha_1^j$

## 7.3 GARCH

GARCH just adds a lagged variance term to an ARCH model

$$r_t = \epsilon_t \tag{39}$$
$$\epsilon_t = \sigma_t e_t \tag{40}$$
$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{41}$$
$$e_t \sim N(0,1) \tag{42}$$

The unconditional variance is now $\omega/(1-\alpha_1-\beta_1)$, other than that most things stay the same.

## 7.4 The big happy ARCH family

In principle there is nothing stopping us from making volatility dependent on more terms. A few notable examples are listed below:

1. GJR-GARCH adds $\gamma_1 \epsilon_{t-1}^2 I_{[\epsilon_{t-1}<0}$ to capture the fact that negative shocks have a larger impact on volatility.
2. TARCH works like GJR-GARCH but regresses on absolute instead of square values.
3. EGARCH regresses log volatility

Equally, you could use alternative error distributions, fat tails, asymmetric distributions, ...

## 7.5 Value at Risk

$VaR(\alpha)$ expresses "On $\alpha$% of days we expect to loose more than $VaR(\alpha)$". **Conditional VaR** makes the VaR conditional on the recent past: "Given that the last week was like X, we expect the VaR to be Y". **Unconditional VaR** computes the VaR using a much longer time horizon and is not concerned with the recent past.

1. **1996 RiskMetrics:** Uses an exponentially weighted moving average to compute volatility:
$$\sigma_{t+1}^2 = (1-\lambda) r_t^2 + \lambda \sigma_t^2 \tag{43}$$
Notice how this is the same as a GARCH model in which $\alpha = (1-\lambda)$ and $\beta = \lambda$, see (41). It then assumes returns are **normally distributed** and computes the VaR as the $\alpha$ quantile of a normal distribution with mean zero and standard deviation $\sigma$
$$VaR_{t+1} = -\sigma_{t+1} \Phi^{-1}(\alpha) \tag{44}$$
2. **2006 RiskMetrics:** Adds long memory, Student distribution, residuals scale correction, lagged correlations. It is a more complicated model but generally yields better results.
3. **Historical Simulation:** Uses an **empirical distribution**, by sampling returns over some lookback window, sorting them, and then finding the $\alpha$ quantile of that distribution.
4. **Weighted Historical Simulation:** Again, an empirical distribution is constructed, but returns are exponentially weighted so that older returns contribute less. To achieve this, we can not just sort the returns and take

the quantile but have to compute the density function as the weighted average number of points below a support point:

$$\hat{G}_t(r) = \sum_{i=1}^{t} w_i I_{[r<r_i]} \tag{45}$$

5. **Filtered Historical Simulation:** Computes an empirical distribution of **devarianced shocks**. It first computes the daily volatility $\hat{\sigma}_t$ using an ARCH model. It then devariances returns $\hat{e}_t = \hat{\epsilon}_t/\hat{\sigma}_t$. All $\hat{e}_t$ values are then sorted, the $\alpha$ quantile is computed and scaled by the ARCH predicted $\hat{\sigma}_{t+1}$
6. **CaViaR:** directly **forecasts the quantile** rather than parameters of the distribution. Its inputs are the previous quantile value $q_t$ as well as $HIT_t$ which indicates if there was a VaR exceedance.

$$q_{t+1} = \omega + \gamma HIT_t + \beta q_t \tag{46}$$

# 8 Vector Autoregression

VAR models enable to model multiple time series together which can improve forecasting and and analysis of the system.

## 8.1 Basic Properties

A VAR is really just an AR with matrices, intuitively it is easier to remember the AR properties.

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \epsilon_t \tag{47}$$

| | AR | VAR |
|---|---|---|
| Mean | $\phi_0/(1-\phi_1)$ | $(I-\Phi_1)^{-1}\Phi_0$ |
| Variance | $\sigma^2/(1-\phi_1^2)$ | $(I-\Phi_1\otimes\Phi_1)^{-1}\text{vec}(\Sigma)$ |
| sth Autocovariance | $\gamma_s\phi_1^s V[y_t]$ | $\Phi_1^s V[y_t]$ |

Table 1: AR(1) and VAR(1) properties

$$\boxed{\text{VAR(1) is stationary if } |\lambda_i| < 1 \text{ where } \lambda_i \text{ are eigenvalues of } \Phi_1} \tag{48}$$

## 8.2 Granger Causality

**A variable $x$ does *NOT* granger cause $y$ if the forecast of $y$ does not change when conditioned on $x$:** $E[y_t|x_{t-1}, y_{t-1}, ...] = E[y_t|y_{t-1}, ...]$. We test this via a restricted VAR, where we force the variable mapping $x$ to $y$ to be zero. We can then compare the restricted and unrestricted var for performance.

## 8.3 Impulse response function

Plots the response of $y$ if $x$ (or $y$ itself) experiences a *unit shock* of one standard deviation. Makes interpretation of VAR easier.

## 8.4 Cointegration

Cointegration is the VAR version of unit roots. $x_t$ **and** $y_t$ **are cointegrated if both are unit roots and there exists a vector** $\beta$ **with both elements non-zero so that** $\beta_1 x_t - \beta_2 y_t \sim I(0)$. We can find cointegration through eigenvalues of $\Phi_1$. If only one eigenvalue is 1, they are cointegrated. If both are 1, they are two independent unit roots.

The **Engle-Granger test** tests for cointegration constructs an OLS estimate $y_t = \alpha + \beta x_t + \epsilon_t$ and then performs and Adjusted Dickey fuller test for a unit root in $\epsilon$

## 8.5 Error correction models

If two variables $x_t$, $y_t$ are cointegrated, then an error correction model forecasts $\Delta x_t$ and $Deltay_t$ dependent on $x_{t-1}$ and $y_{t-1}$. Analogy: A drunk person with dog on a leash walk through a park. They both move randomly but always find back to another. An error correction model captures how.

If $\delta y_t = \pi y_{t-1} + \epsilon_t$ then we can decompose $\pi$ into $\pi = \alpha\beta'$ where $\alpha$ measures the speed of convergence and $\beta$ contain the cointegrating vectors.