Notes on Times Series Analysis

Shaowu Pan

May 30, 2019

Contents

Chapter 1

Preliminary

1.1 Linear model

The most general linear regression model assumed in this class is

$$y = \boldsymbol{x}\beta + \epsilon, \tag{1.1}$$

where $\boldsymbol{x}$ is called *independent variable/predictor/covariate/explanatory variable*, and y is called *dependent variable.* For example, the regression model can be written in scalar form as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \tag{1.2}$$

where $\begin{bmatrix} 1 & x & x^2 \end{bmatrix}$ consists of 3 covariates. But since the eq. (1.2) is linear in $\beta_i, i = 0, 1$, the model is still linear regression.

In time series, we are particularly interested in the following form of models.

$$x_{t+1} = \phi x_t + \epsilon. \tag{1.3}$$

1.2 Features of Time Series

In time series, the adjacent observations are dependent, which is also known as autocorrelated.

1.3 Review on Basic Statistics

1. Mean

$$\mu_X = \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_x(x) dx. \tag{1.4}$$

2. Variance

$$\sigma_X^2 = \mathbb{E}([X - \mu_X]^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx. \tag{1.5}$$

Standard deviation is $\sqrt{\sigma_X^2} = \sigma_X$.

3. Covariance

Consider two continuous R.V., X, and Y, and assume we know their joint PDF as $f_{XY}(x, y)$. Then the definition of covariance is simply,

$$\gamma_{XY} = \mathbb{E}([X - \mu_X][Y - \mu_Y]). \tag{1.6}$$

4. Correlation & Sample Correlation

$$\rho_{XY} = \frac{\gamma_{XY}}{\sigma_X \sigma_Y}. \tag{1.7}$$

Related concepts: positive correlation, negative correlation, uncorrelated. Note that correlation is only a second-order statistics.

While sample correlation is defined as

$$\hat{\rho_{XY}} = \frac{S_{XY}}{\sqrt{S_X^2}\sqrt{S_Y^2}} \tag{1.8}$$

5. Independence

Formal definition of independence is that $f_{XY}(x, y) \triangleq f_X(x) f_Y(y)$. Note that independence lead to uncorrelated random variables but not vice versa. Only for Gaussian distributions we have *uncorrelated* leads to *independence*.

6. Sample Mean and Variance

Given a bunch of sampled data $\{x_i\}_{i=1}^n$, then the sampled mean is given by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{1.9}$$

and corresponding the sampled variance is given by

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2. \tag{1.10}$$

Note that the above formula can also be interpreted as two estimators, which is essentially two R.V.. It can be proved that they are unbiased estimators, which means if we consider in a formal setting, $\{x_i\}_{i=1}^n$ is nothing but n random variables, then sampled mean and variance are R.V. as well. Integration over all its possible space would lead to the expectation which turns out to be *identical* to the true mean and variance.

7. Maximum likelihood estimator

   Estimator can also be constructed in a routine way, e.g., choose the parameter that maximizes the likelihood of data.

   It is easy to show that

   - $\overline{X}$ is also MLE of true mean,
   - $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})$ is MLE of true variance but **not unbiased** estimator[1],
   - the above MLE is asymptotically unbiased when $N \to \infty$.

8. Common distribution for statistic

   Given i.i.d. $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$,

   - mean of i.i.d Normal samples is still Normal: $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \sim \mathcal{N}(\mu_x, \frac{\sigma_x^2}{N})$
   - sum of square of N i.i.d. Normal samples is called $\chi^2(N)$ distribution, $Y = X_1^2 + \ldots + X_N^2 \sim \chi^2(N)$
   - standard Normal over the square root of an averaged N degree χ^2 is t-distribution: $t = \frac{X}{\sqrt{Y/N}} \sim t(N)$
   - ratio between two DOF averaged χ^2 is called F-distribution: $F = \frac{Y_1/N_1}{Y_2/N_2} \sim \mathcal{F}(N_1, N_2)$.

9. Hypothesis testing

   Given i.i.d. samples from $\mathcal{N}(\mu_x, \sigma_x^2)$, the main idea is to construct a Null hypothesis, so that one can derive the distribution for the QoI, then given the sample, we have confidence to claim something.

   - test mean with constant, when variance is known

     Null hypothesis: mean is μ_0, therefore the statistic, i.e., sample mean, follows the $\mathcal{N}(\mu_0, \sigma_a^2/N)$. In practice, we test it by $Z = (\overline{X} - \mu_0)/(\sigma_x/\sqrt{N}) \sim \mathcal{N}(0, 1)$. Note that since it is a double-side distribution, α would be split into two directions, so we reject the Null hypothesis when $|Z| > Z_{\alpha/2}$.
   - test mean with constant, when variance is unknown

     Null hypothesis: mean is μ_0. Let's construct the t-distribution! Consider the sample mean is Normal $(\overline{X} - \mu_0)/(\sigma_0/\sqrt{N}) \sim \mathcal{N}(0, 1)$, and $\sum_{i=1}^{N}\left(\frac{X_i - \overline{X}}{\sigma_0}\right)^2 \sim \chi^2(N-1)$. So according to the definition of t-distribution,

[1] remember, all estimator is R.V.

$$\frac{(\overline{X}-\mu_0)/(\sigma_0/\sqrt{N})}{\sqrt{\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma_0}\right)^2/N-1}} \sim t(N-1) \tag{1.11}$$

Note that it can be simplified as [2]

$$\frac{(\overline{X}-\mu_0)}{S_X/\sqrt{N}} \sim t(N-1) \tag{1.12}$$

Observation: So it is similar to the mean testing with known variance, just to replace σ_0 with S_X and change $\mathcal{N}(0,1)$ to $t(N-1)$.

- test variance with constant
  Null hypothesis: variance is σ_0^2.
  - *if mean is known*. And note that we have a hypothesis knowing the variance. Then naturally we have $(X_i-\mu_0)/\sigma_0 \sim \mathcal{N}(0,1)$. Then following the definition of χ^2, we have $\sum_{i=1}^{N}\left(\frac{X_i-\mu_0}{\sigma_0}\right)^2 \sim \chi^2(N)$. We will reject the null hypothesis, if the statistic is larger than $\chi^2_{N,\alpha}$ or for two sides case, smaller than $\chi^2_{N,1-\alpha/2}$ or larger than $\chi^2_{N,\alpha/2}$.
  - *if mean is unknown*, we would have one DOF decreased, so the statistic $\sum_{i=1}^{N}\left(\frac{X_i-\overline{X}}{\sigma_0}\right)^2 \sim \chi^2(N-1)$.
- test equality of variance from two Normal distribution
  Null hypothesis: $\sigma_1^2=\sigma_2^2$.
  Let's construct a F-distribution. Consider that

$$\frac{(N_1-1)S_{x,1}^2}{\sigma_1^2} \sim \chi^2(N_1-1) \tag{1.13}$$

$$\frac{(N_2-1)S_{x,2}^2}{\sigma_2^2} \sim \chi^2(N_2-1) \tag{1.14}$$

Therefore,

$$\frac{S_{x,1}^2}{\sigma_1^2}/\frac{S_{x,2}^2}{\sigma_2^2} \sim \mathcal{F}(N_1-1,N_2-1) \tag{1.15}$$

Since we assume $\sigma_1^2=\sigma_2^2$, so our statistic is $\frac{S_{x,1}^2}{S_{x,2}^2}$. The corresponding criterion is $\mathcal{F}_{N_1-1,N_2-1,\alpha}$ or double side $\mathcal{F}_{N_1-1,N_2-1,\alpha/2}$ and $\mathcal{F}_{N_1-1,N_2-1,1-\alpha/2}$.

[2] note that S_X is sampled standard deviation! Not sampled variance!

10. Random process

Definition 1. *Random process is an indexed sequence of random samples* $\{X_i\}_{i=1}^n$*. Usually we denote as*

$$\{X_t\}_{t=1}^n.$$

Definition 2. *Mean function of* $\{X_i\}_{i=1}^n$ *is*

$$\mu_t = \mathbb{E}(X_t).$$

Definition 3. *Autocorrelation of* $\{X_i\}_{i=1}^n$ *between index* t *and* u *is*

$$\gamma_{t,u} = \mathbb{E}(X_t - \mathbb{E}(X_t))(X_u - \mathbb{E}(X_t)).$$

Definition 4. *Weak stationary random process: a random process* $\{X_t\}$ *such that first two moments are independent of time origin. That is to say,*

- $\mu_t = \mu, \forall t$
- $\gamma_{t,t+k} = \gamma_k$*. So the auto-covariance is only a function of time lag* k*.*

Definition 5. *Strong stationary random process means the probability distribution is independent of time origin.*

Chapter 2

Linear regression

2.1 Basics of linear regression

2.1.1 Notation in linear regression

Given n pairs of *output observation* y_t and *input predictor variables* x_t, for simplicity, considering only one predictor, we assume the following relationship for the y_t and x_t,

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t, t = 1, 2, \dots, n, \tag{2.1}$$

where

$$\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

In general, the vector form is

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

where y is the column vector of all data, and $\mathbf{X}$ is the whole data matrix *contains* $p+1$ *features, i.e., constant one and* p *predictors*, ϵ is a column of random variables.

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \dots & \mathbf{x}^{(p)} \end{bmatrix},$$

where $\mathbf{x}^{(i)}$ is a column vector contains i-th predictor for *all* data as

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

where $i \in \{1, \dots, p\}$.

Note that we only treat y_t, i.e., output observation as *random variable* due to *random variable* residual error ϵ_t. The input predictors is not considered as *random variables.*

Moreover, we treat *constant predictor* as default predictor and we don't count it in the predictor. So p predictors would mean $p+1$ features in the optimization sense.

2.1.2 Ordinary least square estimation

The goal of *least square estimation* is to find *constant real numbers* that minimize the *variance of residual:*, if there is only one predictor,

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min \mathbb{E}_{y_t,x_t}(y_t - (\beta_0 + \beta_1 x_t))^2.$$

More generally, consider p predictors, the analytical solution is

$$\hat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top y, \tag{2.2}$$

where $\overline{x} = \frac{1}{N}\sum_{i=1}^N x_i$ and $\overline{y} = \frac{1}{N}\sum_{i=1}^N y_i$. $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$.

2.1.3 Statistical interpretation of linear regression

The key difference between linear regression and OLS is that, by *fully utilizing the assumption on residual error and considering that the linear transformation of a Gaussian distribution is still Gaussian*, the distribution of the *least-square estimator* $\hat{\beta}$ is naturally a random variables that follows *Normal distribution* due to that y is random variable.

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma_\epsilon^2(\mathbf{X}^\top \mathbf{X})^{-1}). \tag{2.3}$$

Here a more rigorous proof on why β is Gaussian is provided in the appendix A.1.

Back to the eq. (2.3), we are more interested in the *variance-covariance matrix*[1]

$$\mathbf{V}(\hat{\beta}) = \sigma_\epsilon^2(\mathbf{X}^\top \mathbf{X})^{-1} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_2) & \dots & Cov(\hat{\beta}_0, \hat{\beta}_p) \\ Cov(\hat{\beta}_2, \hat{\beta}_2) & Var(\hat{\beta}_2) & \dots & Cov(\hat{\beta}_2, \hat{\beta}_p) \\ \dots & \dots & \dots & \\ Cov(\hat{\beta}_p, \hat{\beta}_1) & \dots & \dots & Var(\hat{\beta}_p) \end{bmatrix} \in \mathbb{R}^{p+1\times p+1},$$

where on the *diagonal* is *variance* and *off-diagonal* is *covariance* between parameters. Usually, we use the diagonal component to estimate the variance and confidence interval of parameter $\hat{\beta}_i$.

2.1.4 OLS is the minimal variance linear unbiased estimator for Linear Regression

Though linear regression can be done in many different ways, the OLS solution, though most natural, is the *best linear unbiased estimator*, called *BLUE*.

The so-called *Gauss-Markov* Theorem states that OLS is the linear unbiased estimator with the minimal variance.

[1] note this is *nothing*, but just a matrix of second order statistical moment. You can certainly use any unbiased estimator to access them.

2.2 Summary of estimators in linear regression

In the most general linear regression expression as in eq. (2.1), the goal of linear regression is to estimate

1. σ_ϵ^2

   From appendix A.2, it is simply

$$\hat{\sigma_\epsilon^2} = \frac{SSR}{n-p-1}. \tag{2.4}$$

2. β

   Since we know the distribution is multivariate Normal distribution, so given

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma_\epsilon^2(\mathbf{X}^\top\mathbf{X})^{-1}), \tag{2.5}$$

   one can come up with unbiased estimator for the mean, simply as

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}, \tag{2.6}$$

$$\hat{\beta}_1 = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top y, \tag{2.7}$$

2.3 Measure the model significance

2.3.1 Coefficient of Determination

Coefficient of determination, sometimes called R^2, is a very helpful scalar that determine the model performance. It represents how much *variance* is explained by the model.

The definitions is motivated by the equality in *linear regression.*

$$\mathbf{SST} = \mathbf{SSR} + \mathbf{RSS}, \tag{2.8}$$

where $\mathbf{SST} = \sum_{i=1}^n (y_i - \overline{y})^2$, $\mathbf{SSR} = \sum_{i=1}^n (\hat{y}_i - \overline{y})^2$, $\mathbf{SSE} = \sum_{i=1}^n (\hat{y}_i - y_i)^2$.

Therefore, it is natural to define R^2 as

$$R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}} = \mathbf{1} - \frac{\mathbf{SSE}}{\mathbf{SST}} \tag{2.9}$$

Adjusted R^2

As number of variables increases, naturally R^2 increases, so it is likely to *overfit* the data. Instead, we can measure the *unexplained variance per degree of freedom versus the total variance per degree of freedom.*

$$R_a^2 = 1 - \frac{\mathbf{SSE}/(n-p-1)}{\mathbf{SST}/(n-1)} \tag{2.10}$$

To learn more about R^2 and adjusted R^2, there are two posts about how does it compare with AIC/BIC and can we use it for nonlinear regression.

2.3.2 Model utility test

Test for single parameter: t-test

To remove a parameter β_j, it is equivalent to do compare the OLS estimator $\hat{\beta}_j$ with 0 to see if the mean of $\hat{\beta}_j$, i.e., $\beta_j = 0$. Recall that we know how to *test mean with a constant, in the unknown variance situation.* So we do the following t-distribution test about $\hat{\beta}_j$. Note that in order to find such statistic, we just need a *standard Normal distribution*, and *a* χ^2 *distribution.* There is no restriction on what they are or whether they are related or not.

From appendix A.3, the resulting test statistic to *examine whether or not a parameter is indeed noise-centered around zero* is a t-distribution

$$t_j = \frac{\hat{\beta}_j}{\sqrt{v_j}\hat{\sigma}_\epsilon} \sim t(n-p-1). \tag{2.11}$$

Test for multiple parameters: F-test

Let's assume a null hypothesis as *all β's are zero, i.e., the model is no better than a mean-average random guess.* Then the test statistic is

$$f = \frac{\mathbf{SSR}/df_{SSR}}{\mathbf{SSE}/df_{SSE}} \sim \mathcal{F}(p, n-p-1) \tag{2.12}$$

2.4 Feature selection

Naturally, feature selection follows the previous section about model utility test.

Consider two models with overlapping parameters, without loss of generality, let's consider second one contains more parameter than first model,

1. $p_0 + 1$ parameters, $\hat{\beta}_0, \ldots, \hat{\beta}_{p_0}$, with $\mathbf{SSE}_0$,

2. $p_1 + 1$ parameters, $\hat{\beta}_0, \ldots, \hat{\beta}_{p_0}, \ldots \hat{\beta}_{p_1}$, with $\mathbf{SSE}_1$.

It is natural to consider whether the *extra predictors would be helpful or not*, so intuitively we isolate the variance explained by the *extra predictors* and normalize it by the total errors per degree of freedom.

Following eq. (2.12), recall the null hypothesis is that *extra predictors are useless*, the corresponding F statistic is

$$F = \frac{(\mathbf{SSE}_0 - \mathbf{SSE}_1)/(p_1 - p_0)}{\mathbf{SSE}_1/(n - p_1 - 1)}. \tag{2.13}$$

It is interesting to note the connection between single parameter significance test and multi parameter test: $\forall n \in \mathbb{N}, \mathcal{F}(1, n) = t^2(n)$.

2.4.1 Forward stepwise selection

It recursively *add* predictor from a *pool of predictors.*

1. suppose current model has k predictors $\rightarrow \mathbf{SSE}_k$
2. choose the $k+1$-th predictor that *maximize* the F-statistic $F = \frac{\mathbf{SSE}_k - \mathbf{SSE}_{k+1}}{\mathbf{SSE}_{k+1}/(n-k-2)}$.
3. if $F > \mathcal{F}_\alpha(1, n-k-1)$, then we reject the null hypothesis, so we include that predictor, otherwise, stop the process.

2.4.2 Backward stepwise selection

It recursively *remove* predictor from the *full predictor* models.

1. suppose current model has k predictors $\rightarrow \mathbf{SSE}_k$
2. choose the $k-1$-th predictors in the existing model that minimize the F-statistic $F = \frac{\mathbf{SSE}_{k-1} - \mathbf{SSE}_k}{\mathbf{SSE}_k/(n-k-1)}$.
3. if $F < \mathcal{F}_\alpha(1, n-k-1)$, then the null hypothesis is true, so we remove that predictor. Otherwise, stop the process.

Chapter 3

AR-1 model

Based on previous chapters about linear regression, we can now safely utilize the power of linear regression to do *first order autoregressive* models, which is known as "AR(1)" model.

3.1 Model setup

Before discussing about any *AR* or *ARMA* model, one necessary step to take is to remove the mean. So here we assume all the X_t is zero mean, i.e., $\mathbb{E}(X_t) = 0$. Also,

3.1.1 White noise

White noise is a special class of random process in that there is no correlation between each time instance. Formally it means that

$$\gamma_k(a_t) = \mathbb{E}(a_t a_{t+k}) \triangleq \gamma\delta_{k0} \tag{3.1}$$

Gaussian white noise

Among them, the most common one is Gaussian white noise.

Gaussian white noise can be formally defined as follows

$$\gamma_k(a_t) = \mathbb{E}(a_t a_{t+k}) = \left\{ \begin{array}{l} \sigma_a^2, k = 0 \\ 0, k \neq 0 \end{array} \right. . \tag{3.2}$$

One might interests in the *autocorrelation* of a_t, which is simply

$$\text{correlation between } a_t \text{ and } a_{t+k} = \rho_k(a_t) = \left\{ \begin{array}{l} 1, k = 0 \\ 0, k \neq 0 \end{array} \right. . \tag{3.3}$$

So there is no correlation between each time instance of them and since it is normally distributed, so uncorrelated means independent. Normally, we denote Gaussian white noise with variance σ_a^2 as $\text{NID}(0, \sigma_a^2)$.

3.1.2 AR-1 model: one-step ahead linear regression + white noise

Given a random process, $\{X_t\}_{t=1}^N$, First, we remove the mean from the data. So that $\overline{X}_t = 0$. Then we consider the following model to describe the data.

$$X_t = \phi_1 X_{t-1} + a_t, \tag{3.4}$$

where $a_t \sim \text{NID}(0, \sigma_a^2)$, where NID means *normally independently distribution.*

Least-square estimation of ϕ_1

Consider $\mathbf{y} = [X_2, \dots, X_N]$ and $\mathbf{X} = [X_1, \dots, X_{N-1}]$ and following the linear regression. We have the least square estimation follows the *multi-variate Normal distribution*,

$$\hat{\phi_1} \sim \mathcal{N}((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}, \sigma_a^2(\mathbf{X}^\top \mathbf{X})^{-1}). \tag{3.5}$$

However, we are still unsure about the σ_a^2, which will be inferred in the following.

Estimation of σ_a^2

Note that from the linear regression, we have a *unbiased estimator* for σ_a^2 as

$$\hat{\sigma}_a^2 = \frac{SSE}{N-1-1} = \frac{SSE}{N-2}, \tag{3.6}$$

which follows the χ^2 distribution as

$$(N-2)\frac{\hat{\sigma}_a^2}{\sigma_a^2} \sim \chi^2(N-2) \tag{3.7}$$

3.2 Independence check on a_t

One way to check whether or not the model assumption on residual being NID is to do the following three different statistical check.

1. *Scatter plots check*

   intuitive check by plotting a_t vs a_{t-k}.

2. *Portmanteau test*

   one need to first choose a integer K then consider the sample autocorrelation for j-th separation as

$$\hat{\rho}_j(a_t) \triangleq \hat{\rho}(a_t, a_{t-j}) = \frac{\sum_{t=j+2}^N a_t a_{t-k}}{\sum_{t=j+2}^N a_{t-k}^2}, \tag{3.8}$$

which can simply replaced by find the β_1 that uses a_{t-j} to predict a_t.

Then the test statistic is the following

$$Q = N \sum_{j=1}^{K} (\hat{\rho}_j(a_t))^2 \sim \chi^2(K-1) \tag{3.9}$$

3. *Bartlett band*

   simply check if any of $|\hat{\rho}_j| = |\hat{\rho}(a_t, a_{t-j})| > \frac{2}{\sqrt{N}}$. If so, then by $\alpha = 0.05$, we consider they are correlated.

3.3 AR(1) model is conditioned linear model

$$X_t | X_{t-1} \sim \mathcal{N}(\phi_1 X_{t-1}, \sigma_a^2), \tag{3.10}$$

where when N is large, we can safely replace ϕ_1 with $\hat{\phi}_1$ and σ_a^2 with $\hat{\sigma}_a^2$.

3.3.1 Notations

Several interesting notations are given here

- one-step ahead prediction based on $t-1$: $\hat{X}_t = \hat{X}_{t-1}(1) = \phi_1 X_{t-1}$
- error of one-step ahead prediction based on $t-1$: $e_{t-1}(1) = X_t - \hat{X}_t$

3.3.2 Prediction with unequal weights

Say we have c_o as over prediction per amount of over prediction, similarly c_u as under prediction penalty. We can easily arrive at the optimal prediction of future step as

$$k^* = F^{-1}(1-\alpha) \tag{3.11}$$

where $\alpha = \frac{c_o}{c_o + c_u}$ and $F(x)$ is the CDF of $\mathcal{N}(\phi_1 X_{t-1}, \sigma_a^2)$.

3.4 Property of the dynamical system of AR(1)

3.4.1 When $|\phi_1| < 1$, corresponding system is A.S.L. stable

$$X_t = \phi_1 X_{t-1} + a_t \tag{3.12}$$

$$X_t = \phi_1(\phi_1 X_{t-2} + a_{t-1}) + a_t \tag{3.13}$$

$$X_t = \phi_1^k X_{t-k} + \sum_{i=0}^{k-1} \phi_1^i a_{t-i} \tag{3.14}$$

Therefore, if $|\phi_1| < 1$, we can safely arrive at

$$\lim_{k\to\infty} \phi^k X_{t-k} = 0$$

Therefore

$$X_t = \sum_{i=0}^{\infty} \phi_1^i a_{t-i}$$

To see the system is A.S.L. stable, one can simply see that since *sum of independent normal distribution is still normal distribution*, so X_t is still a normal distribution and corresponding mean and variance are *naively summed together.*

Alternatively and generally, to determine whether or not a sum of random variables is stable or bounded, let's consider the variance and the mean of X_t, then applying Chebyshev inequality.

$$\mathbb{E}(X_t) = \mathbb{E}(\sum_{i=0}^{\infty} \phi_1^i a_{t-i}) = 0 \tag{3.15}$$

$$D(X_t) = \frac{\sigma_a^2}{1-\phi_1^2} \tag{3.16}$$

Consider Chebyshev inequality, we have

$$\mathbf{P}(|X_t| \geq a) \leq D(X_t)/a^2 = \frac{\sigma_a^2}{a^2(1-\phi_1^2)}. \tag{3.17}$$

Therefore, as $a \to \infty$, the probability for X_t to be unbounded is 0.

In addition to the above formula for variance and mean, we have corresponding formula for the covariance *decaying with respect to time* as

$$\mathbf{Cov}(X_t, X_{t+k}) = \frac{\phi_1^k}{1-\phi_1^2}\sigma_a^2. \tag{3.18}$$

The corresponding autocorrelation is

$$\rho_k(x_t) = \phi_1^k. \tag{3.19}$$

It is the first random process shown in time series analysis that is correlated and independent of t.

Therefore, AR(1) is a *weakly stationary random process.*

3.4.2 When $|\phi_1| > 1$, corresponding system is unstable, and non-stationary

It is easy to see and check.

3.4.3 When $|\phi_1| = 1$, corresponding system is called random walk

The system will still have strong dependency but X_t is considered marginally stable and non-stationary.

Note that the distribution of an approximation of X_t is, using the property of sum of i.i.d. normal distribution,

$$X_t^K = \sum_{i=0}^{K} \phi_1^i a_{t-i} \sim \mathcal{N}(0, \sum_{i=0}^{K} \phi_1^{2i}) = \mathcal{N}(0, K). \tag{3.20}$$

Clearly, we can see that $D(X_t) = \phi_1^2 D(X_{t-1}) + \sigma_a^2$. So it is not *stationary.*

Chapter 4

ARMA(n,m) model

4.1 Motivation: no more independent a_t's

ARMA model is the short name for *autoregressive moving average* model. The motivation for ARMA model comes from the deficiency of AR(1) models *when a_t is correlated with each other*.

Consider when the following AR(1) model is not enough,

$$X_t = \phi_1 X_{t-1} + \tilde{a}_t \tag{4.1}$$

while we find the residual $\tilde{a}_t$, has dependency on previous $\tilde{a}_{t-1}$

Well, notice that we still want to keep the NID property of $\tilde{a}_t$, so we can think of extending the AR models to contain moving average.

$$X_t = \phi_1 X_{t-1} + \tilde{a}_t - \theta_1 \tilde{a}_{t-1}, \tag{4.2}$$

where the residual $X_t - \phi_1 X_{t-1}$ has correlation with previous residual and *we can still keep to assume $\tilde{a}_t \sim NID(0, \sigma_a^2)$*. In the above model, there is one term for autoregressive between X and one term for moving average of residual, *so it is called AR(1,1) model.*

4.2 General form of ARMA(n,m) model

Based on previous motivation, it is natural to give the following definition on ARMA(n,m) model. Note that the negative sign would help us a lot in later.

Definition 6. *ARMA(n,m) models follows*

$$X_t - \sum_{i=1}^{n} \phi_i X_{t-i} = a_t - \sum_{j=1}^{m} \theta_j a_{t-j}, \tag{4.3}$$

where $\phi_i \in \mathbb{R}, \theta_j \in \mathbb{R}$, $1 \leq i \leq n$, $1 \leq j \leq m$. n and m are called the order of AR part and MA part. a_t is a Gaussian white noise, i.e., follows the NID$(0, \sigma_a^2)$.

4.3 Least-square estimation of AR(n) model is simple

It is easy to get start with AR(n) model since it is still a linear regression. Clearly, for

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_n X_{t-n} + a_t \tag{4.4}$$

The least square solution to $\hat{\phi}_i$ is simply

$$\hat{\boldsymbol{\phi}} = (\mathbf{X}^\top \mathbf{X})^{-\mathbf{1}} \mathbf{X}^\top \mathbf{y}, \tag{4.5}$$

where

$$\mathbf{X} = \begin{bmatrix} X_{n-1} & X_{n-2} & \ldots & X_1 \\ \vdots & \vdots & \vdots & \vdots \\ X_{N-1} & X_{N-2} & \ldots & X_{N-n} \end{bmatrix} \tag{4.6}$$

$$\mathbf{y} = \begin{bmatrix} X_n \\ \vdots \\ X_N \end{bmatrix} \tag{4.7}$$

Note that for n previous embeddings, our data pairs becomes $N-n$ pairs. Remember that when doing AR(1) we only have $N-1$ pairs. Similarly, the estimation for $\hat{\sigma}_a^2$ is

$$\hat{\sigma}_a^2 = \frac{SSE}{N-n-n} = \frac{SSE}{N-2n} \tag{4.8}$$

4.4 Least-square estimation of ARMA(n,m) model would lead to nonlinear least-square regression

The difficulty of solving ARMA model lies in the MA part, not the AR part. To see this, check the ARMA(1,1) model, we have

$$X_t = \phi_1 X_{t-1} + a_t - \theta_1 a_{t-1}, \tag{4.9}$$

where a_{t-1} is

$$a_{t-1} = X_{t-1} - \phi_1 X_{t-2} + \theta_1 a_{t-2}. \tag{4.10}$$

Therefore it gives

$$X_t = \phi_1 X_{t-1} + a_t - \theta_1 (X_{t-1} - \phi_1 X_{t-2} + \theta_1 a_{t-2}), \tag{4.11}$$

$$X_t = (\phi_1 - \theta_1) X_{t-1} + \theta_1 \phi_1 X_{t-2} + a_t - \theta_1^2 a_{t-2}, \tag{4.12}$$

So ideally one can keep on replace a_{t-j} with a_{t-j-1} and assume the $|\theta| < 1$ to convert it back to a AR process.

We can simply leave it to either Matlab or python.

4.4.1 Model adequacy check

Still, we can check if the model a_t is correlated or not by the following three test.

1. Scatter plot
2. Portmanteau test

   $Q \sim \chi^2(K - n - m)$
3. Bartlett band

However, note that in ARMA model, computing a_t from observation data $\{X_t\}_{t=1}^N$ is not that straightforward. One would need to *recursively* computing the a_t, which would need to go back to first few a_t's. Usually we have two ways to deal with it

1. assume the first $\max(n, m)$ residuals are zer.
2. matlab has some mechanism to initialize the residuals

4.4.2 Model selection

In ARMA model, model selection is *equivalent* to selecting the model order, i.e., the n and m. Let's say the models are ARMA(n_1, m_1), ARMA(n_2, m_2) and $n_2 + m_2 > n_1 + m_1$.

Same as before, we can conduct the F-test.

$$F = \frac{\frac{\mathbf{SSE}_1 - \mathbf{SSE}_2}{s}}{\frac{\mathbf{SSE}_2}{N_0 - r}} \sim \mathcal{F}(s, N_0 - r) \tag{4.13}$$

where $s = n_2 + m_2 - (n_1 + m_1)$ is the difference in number of parameters, and $r = n_2 + m_2$, $N_0 = N - t_0 \approx N$, where $t_0 = \max(n, m)$.

Clearly, if F test is passed, then it means the *extra parameters are zero*, so we choose model 1. Otherwise, choose model 2.

ARMA$(n, n-1)$ strategy

However, the above selection only gives a criterion but not necessarily practical since there are so many different choices and the number could easily grow quadratically.

Fortunately, in many applications, it has been *empirically* shown that any stationary stochastic system, including random walk (which is not stationary), can be represented by an ARMA(n,$n-1$). Clearly, determining the order for this is much easier than an arbitrary ARMA model.

Note that in practice, we can also choose the ARMA$(2n, 2n-1)$ strategy, which is simply a faster version. So we will keep increase the number of n, and test the previous F test until we arrive at a stage where the number of extra predictors is satisfying the F-test.

Cross validation

Further, we can split the data into *three sets*, training set, validation set, and test set. Note that validation set is used to select a model.

Usually we can use *Root mean square error* (RMSE) or mean absolute error (MAE).

Chapter 5

Green function

Remember the time series we studied here considers noise naturally in our model. When expressing the time series in terms of *random shocks* or *Gaussian white noises*, Green function naturally comes out that measures the effect of historical random shock on current time step, or equivalently, measures the aftereffect of random shock at certain time step.

This chapter goes as follows: before everything is discussed, let's take a look at back shift operators. Then first we discuss the difference between linear regression and time series model. Note that although AR1 model heavily relies on the result from linear regression, it does not necessarily mean there is no thing more interesting than linear regression in AR1 model. Instead, AR1 model gives a dynamical system that is fundamentally different from linear regression, which would be emphasized.

5.1 Lag (Backshift) Operator

Given a time series, we define an operator B which transform the state back to its previous state. Note that we are assuming we have known all the time series, so the backshift operator is deterministic. Note that lag operator is seldom defined in the functional space due to the stochastic nature of the time series.

For example,

$$BX_t = X_{t-1}$$

$$B^2X_t = BX_{t-1} = X_{t-2}$$

$$B^jX_t = X_{t-j}$$

Similarly, since error e_t is the linear combination of state. Therefore, lag operator is also applicable to it as well.

$$B^je_t = e_{t-j} = X_t - \hat{X}_t,$$

where ˆ represents the expectation/prediction of that state.

5.1.1 Algebraic Equivalence

Note that lag operator follows the same rule as normal linear operator such as

1. scalar multiplicative: $B(aX_t) = aBX_t$
2. distributive: $B(X_t + Y_t) = BX_t + BY_t$
3. associative: $(a + bB)BX_t = (aB + bB^2)X_t = aX_{t-1} + bX_{t-2}$

5.1.2 Inverse Operator

Later we will see in most cases, we are interested in the inverse operator of $1 - \phi B$. Following the previous algebraic rule, we will have

$$(1 + \phi B + \phi^2 B^2 + \ldots + \phi^{t-1} B^{t-1})(1 - \phi B)X_t = (1 - \phi^t B^t)X_t = X_t - \phi^t X_0.$$

Note that **if $|\phi| < 1$, then we have the representation for the inverse operator of $(1 - \phi B)$ as infinite series**

$$(1 - \phi B)^{-1} = 1 + \phi B + \phi^2 B^2 + \ldots + \phi^{t-1} B^{t-1} + \ldots$$

5.1.3 Rational decomposition

If $|\lambda_1| < 1$ and $|\lambda_2| < 1$, we will have

$$\frac{1}{(1 - \lambda_1 B)(1 - \lambda_2 B)} = \frac{1}{\lambda_1 - \lambda_2} \left(\frac{\lambda_1}{1 - \lambda_1 B} - \frac{\lambda_2}{1 - \lambda_2 B} \right)$$

5.2 Characterization of dynamic aspect of ARMA

- Green function: *describe dynamics in terms of random shocks*, gives stability of the time series
- Inverse function: *describe dynamics in terms of past X_t*
- Auto-covariance function and autocorrelation function and their relationship on Green function, characterize the dependence between X_t, $X_{t-1}, \ldots$.

5.3 Green function

In this section, we will discuss Green function over several type of ARMA models.

5.3.1 Concept of Green function: example AR1 model

Consider standard AR1 model

$$X_t = \phi_1 X_{t-1} + a_t,$$

where we can alternatively and recursively write

$$X_{t-1} = \phi_1 X_{t-2} + a_{t-1},$$

and so on. Then we can express X_t in terms of earlier terms which results in

$$X_t = \phi_1^t X_0 + a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \dots.$$

Now, we introduce a critical condition: $|\phi_1| < 1$, then as $t \to \infty$, we have simply

$$X_t = \sum_{j=0}^{\infty} \phi_1^j a_{t-j},$$

where $a_t \sim \mathcal{N}(0, \sigma_a^2)$ and it is i.i.d.. We call *Green function* as $G_j \triangleq \phi_1^j$.

In general, *any ARMA model can be written as*

$$X_t = X_t = \sum_{j=0}^{\infty} G_j a_{t-j},$$

where $G_0 = 1$ for any ARMA model.

5.3.2 Implication of Green function

- Green function enables us to express any *ARMA* model in terms of *MA* models. So any ARMA model is essentially a moving average model.
- The white noise process $\{a_t\}$ is transformed into a process $\{X_t\}$ by *linear filter*, which is essentially an infinite large matrix A, with $A\{a_t\}$.
- G_j determines how well the system correlates/remembers the input a_{t-j}.

5.3.3 Green function for ARMA(2,1)

Given

$$X_t = \phi_1 X_{t-1} \phi_2 X_{t-2} + a_t - \theta_1 a_{t-1},$$

we can rewrite it as

$$(1 - \phi_1 B - \phi_2 B^2) X_t = (1 - \theta_1 B) a_t,$$

where B is the backshift operator. Note that one can check B follows algebraic rules. So it is tempting to factorize it as

$$(1-\phi_1 B-\phi_2 B^2)=(1-\lambda_1 B)(1-\lambda_2 B),$$

where λ_1 and λ_2 are characteristic roots of the second order linear difference equation

$$\lambda^2-\phi_1\lambda-\phi_2=0,$$

with

$$\begin{aligned}\lambda_1+\lambda_2&=\phi_1,\\ \lambda_1\lambda_2&=-\phi_2.\end{aligned}$$

After some math, in general, the expression of G_j is

$$G_j=g_1\lambda_1^j+g_2\lambda_2^j \tag{5.1}$$

where g_1 and g_2 are constants determined by the moving average part.

In ARMA(2,1) model, we have

1. if $\lambda_1\neq\lambda_2$,
$$G_j=\frac{\lambda_1-\theta_1}{\lambda_1-\lambda_2}\lambda_1^j+\frac{\lambda_2-\theta_1}{\lambda_2-\lambda_1}\lambda_2^j. \tag{5.2}$$

2. if $\lambda_1=\lambda_2=\lambda$ are repeated roots then
$$G_j=(g_1+g_2 j)\lambda^j, \tag{5.3}$$
where $g_1=1$, $g_2=1-2\frac{\theta_1}{\phi_1}$.

3. if $\lambda_1=\overline{\lambda_2}$ are complex conjugate roots, then we have
$$G_j=2gr^j\cos(\beta+j\omega), \tag{5.4}$$
where
$$\begin{aligned}g&=\frac{1}{2}\sqrt{1+\left(\frac{\phi_1-2\theta_1}{\sqrt{-\phi_1^2-4\phi_2}}\right)^2},\\ r&=|\lambda_1|=|\lambda_2|,\\ \beta&=\tan^{-1}\frac{-\phi_1+2\theta_1}{\sqrt{-\phi_1^2-4\phi_2}}.\end{aligned}$$

5.3.4 Green function for Moving Average model

Note that Green function is defined for representing the effect of previous random shocks on current state. So for moving average model, which is already describing the effect of random shocks. So we have

1. MA(1): $G_0 = 1$, $G_1 = -\theta_1$, $G_j = 0, \forall j \geq 2$.
2. MA(2): $G_0 = 1$, $G_1 = -\theta_1$, $G_2 = -\theta_2$, $G_j = 0 \forall j \geq 3$.

5.3.5 Green function for general ARMA(n,m)

Distinct roots

In general, for ARMA(n,m) model, we have

$$G_j = g_1 \lambda_1^j + g_2 \lambda_2^j + \ldots + g_n \lambda_n^j$$

where λ_k are the roots of

$$\lambda_n - \phi_1 \lambda_{n-1} - \ldots - \phi_n = 0,$$

and g_k are determined through moving average part.

Repeated roots

if λ_1 are k repeated roots and λ_2 are j repeated roots, then

$$G_j = (c_1 + c_2 j + \ldots + c_{k-1} j^{k-1}) \lambda_1^j + (d_1 + d_2 j + \ldots + d_{l-1} j^{l-1}) \lambda_2^j + g_3 \lambda_3^j + \ldots + g_{n-k-l} \lambda_{n-k-l}^j$$

5.4 Wold's decomposition

Chapter 6

Stability & Invertibility

Stability of a ARMA model is very crucial. For me, I am focusing mostly on physical system, so I don't want my ARMA model found to be unstable. So first one need to know how to check if its ARMA model is unstable.

6.1 General principle for stability

For a general ARMA(n,m) model, we have the following three definition for stability.

1. A.S.L. stable $\sum_{j=0}^{\infty} |G_j| < \infty$.
2. Marginally stable, when $\lim_{j \to \infty} |G_j| < \infty$
3. Unstable if blows up

In general, we have the following result **Consider that growth of G_j is parameterized by λ^j, which means the stability is determined solely by, the roots of characteristic equation**

$$\lambda^n - \phi_1 \lambda^{n-1} - \phi_2 \lambda^{n-2} - \ldots - \phi_n = 0 \tag{6.1}$$

1. **If all the roots $|\lambda_j| < 1$, then it is A.S.L stable.**
2. if **one** root's module is equal 1 in module but other are smaller than 1, then it is **marginally stable.**
3. if **two** roots' module are equal 1 but they are different sign, then it is still **marginally stable**.
4. if we have repeated roots, then **even root module equal 1, it is unstable**.

6.2 Invertibility

The Green function carries the memory of a system in terms of past *random shocks*, a_t, and show how they affect X_t. However, it is not that helpful if we are not interested in the stochastic property of the system. Rather, the dynamics of ARMA system can also be represented by expressing X_t as a linear combination of past X_t. The coefficient is called inverse function as obtained by inverting the operator that yields Green function.

6.2.1 Convert MA(1) model to AR(∞) model

Starting with the simplest MA model, given $|\theta_1| < 1$, we have

$$\begin{aligned} X_t &= (1 - \theta_1 B)a_t \\ a_t &= \frac{1}{1 - \theta_1 B} X_t \\ &= (1 + \theta_1 B + \theta_1^2 B^2 + \ldots)X_t \\ X_t &= \sum_{j=1}^{\infty} -\theta_1^j X_{t-j} + a_t \\ X_t &= \sum_{j=1}^{\infty} I_j X_{t-j} + a_t \end{aligned}$$

So MA(1) model is equivalent to a AR(∞) model.

6.2.2 General ARMA(n,m) model

Clearly, the duality between I_j and G_j is obvious. So similarly, for a general ARMA(n,m) model, we have

$$\Phi(B)X_t = \Theta(B)a_t$$

To have the AR model converted, we simply derive

$$a_t = \frac{\Phi(B)}{\Theta(B)} X_t = X_t - \sum_{j=1} I_j X_{t-j}$$

The criterion is simply we need the roots $|v_j|$ of the following characteristic equation to be stable.

$$v^m - \theta_1 v^{m-1} - \ldots - \theta_m = 0$$

Chapter 7

Auto-covariance, Autocorrelation, Stationarity

7.1 Definition of Auto-covariance and Autocorrelation

For a **random process** which doesn't have to be stationary process, we have the following definition associated with it.

1. **Mean function**: $\mu_t = E(X_t)$
2. **Auto-covariance function**: $\gamma_{t,t+k} = E((X_t - \mu_t)(X_{t+k} - \mu_{t+k}))$
3. **Autocorrelation function**: $\rho_k = \frac{E((X_t-\mu_t)(X_{t+k}-\mu_{t+k}))}{E(X_t-\mu_t)^2} = \frac{\gamma_{t,t+k}}{\gamma_{t,t}}$

For stationary process, we can further simplify the result as

$$\gamma_{t,t+k} = \gamma_t$$

$$\rho_k = \gamma_k / \gamma_0$$

7.2 Expressing auto-covariance in Green function

Starting here, we simply assume X_t is zero mean, i.e., $\mu_t = 0$. Therefore we simplify the expression for autocorrelation.

7.2.1 Revisit on Green function

Recall that for general ARMA(n,m) model, we have Green function as

$$X_t = \sum_{j=0}^{\infty} G_j a_{t-j}$$

where $a_t \sim NID(0, \sigma_a^2)$.

Note that they are i.i.d, so $E(a_t a_{t+k}) = 0$ if $k \neq 0$.

7.2.2 General formula as a convolution

$$\gamma_k = \sum_{j=0}^{\infty} G_j G_{j+k} \sigma_a^2$$

7.2.3 AR(1)

$$\gamma_k = \sigma_a^2 \frac{\phi_1^k}{1 - \phi_1^2}$$

7.2.4 ARMA(2,1)

Distinct Real Roots

with $|\lambda_{1,2}| < 1$, and $\lambda_1 \neq \lambda_2$, we have

$$\gamma_k = d_1 \lambda_1^k + d_2 \lambda_2^k$$

where

$$d_1 = \left(\frac{g_1^2}{1 - \lambda_1^2} . \right)$$

$$d_1 = (\frac{g_1^2}{1 - \lambda_1^2} + \frac{g_1 g_2}{1 - \lambda_1 \lambda_2}) \sigma_a^2$$

$$d_2 = (\frac{g_2^2}{1 - \lambda_2^2} + \frac{g_1 g_2}{1 - \lambda_1 \lambda_2}) \sigma_a^2$$

where g_1 and g_2 are the two corresponding constants associated with Green function as

$$g_1 = \frac{\lambda_1 - \theta_1}{\lambda_1 - \lambda_2}$$

$$g_2 = \frac{\lambda_2 - \theta_1}{\lambda_2 - \lambda_1}$$

Complex roots

Note that for complex roots, the Green function for ARMA21 becomes

$$G_j = 2gr^j \cos(\beta + j\omega)$$

So the corresponding auto-covariance is

$$\gamma_k = \sigma_a^2 4g^2 r^k \sum_{j=0} r^{2j} \cos(\beta + j\omega)\cos(\beta + (j+k)\omega)$$

where $r = |\lambda|$ and $g = \frac{1}{2}\sqrt{1 + \left(\frac{\phi_1 - 2\theta_1}{\sqrt{-\phi_1^2 - 4\phi_2}}\right)^2}$, and $\beta = \tan^{-1}\frac{-\phi_1 + 2\theta_1}{\sqrt{-\phi_1^2 - 4\phi_2}}$.

Note that for complex roots, γ_k is not finite if $r \geq 1$. It only finite when $r < 1$.

7.2.5 MA(2)

For moving average model, it is obvious about the Green function. So we have

$$\begin{aligned}
\gamma_0 &= (1 + \theta_1^2 + \theta_2^2)\sigma_a^2 \\
\gamma_1 &= (-\theta_1 + \theta_2\theta_1)\sigma_a^2 \\
\gamma_2 &= (-\theta_2)\sigma_a^2 \\
\gamma_3 &= 0
\end{aligned}$$

Therefore, MA(n) model has $\gamma_{n+1} = 0$. It is a good criterion to test the order of MA model.

7.2.6 Some properties

Correlation strength of a_j and X_l is the Green function

Since we have access to X_t as a weighted sum of random shocks, let's consider an interesting relation between a_{t-l} and X_{t-k}.

$$\begin{aligned}
\mathbb{E}(a_{t-l}X_{t-k}) &= \mathbb{E}(a_{t-l}\sum_j G_j a_{t-k-j}) \\
&= \sum_j G_j \mathbb{E}(a_{t-l}a_{t-k-j}) \\
&= \sum_j \delta_{j,l-k} G_j \sigma_a^2 \\
\text{if } l \geq k,\ \mathbb{E}(a_{t-l}X_{t-k}) &= G_{l-k}\sigma_a^2 \\
\text{if } l < k,\ \mathbb{E}(a_{t-l}X_{t-k}) &= 0
\end{aligned}$$

Therefore, it depends on whether a_{t-l} or X_{t-k} comes first.

Autocovariance is a even function: $\gamma_k = \gamma_{-k}$

Due to stationary property, we have

$$\gamma_k = \mathbb{E}(X_{t-k}X_t) = \mathbb{E}(X_t X_{t+k}) = \gamma_{-k}$$

7.3 Dynamics of autocorrelation γ_k

Expressing the autocorrelation as a convolution process for ARMA model by using Green function is straightforward. But is there a dynamics that governs the autocorrelation γ_k? Let's take a took starting from the simplest case.

7.3.1 Example: ARMA(2,1) model

Recall we have assumed the time series X_t is zero mean and stationary. Note that

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t - \theta_1 a_{t-1} \\ &\text{consider multiply } X_{t-k} \text{ and have expectation on it} \\ \mathbb{E}(X_t X_{t-k}) &= \phi_1 \mathbb{E}(X_{t-1}X_{t-k}) + \phi_2 \mathbb{E}(X_{t-2}X_{t-k}) + \mathbb{E}(a_t X_{t-k}) - \theta_1 \mathbb{E}(a_{t-1}X_{t-k}) \\ \gamma_k &= \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \mathbb{E}(a_t X_{t-k}) - \theta_1 \mathbb{E}(a_{t-1}X_{t-k}) \end{aligned}$$

When $k = 0$,

$$\mathbb{E}(a_t X_{t-k}) = \mathbb{E}(a_t X_t) = G_0 \sigma_a^2 = \sigma_a^2$$

$$\mathbb{E}(a_{t-1} X_{t-k}) = \mathbb{E}(a_{t-1} X_t) = G_1 \sigma_a^2$$

Therefore,

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma_a^2 - \theta_1 G_1 \sigma_a^2$$

where $G_1 = \phi_1 - \theta_1$.

When $k = 1$, $\mathbb{E}(a_t X_{t-1}) = 0$ and $\mathbb{E}(a_{t-1} X_{t-1}) = G_0 \sigma_a^2 = \sigma_a^2$, therefore,

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 - \theta_1 \sigma_a^2$$

When $k > 1$, we simply have

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}$$

which implies the k-autocovariance γ_k follows the same long time trend as the original AR part in the ARMA model.

7.3.2 General dynamics of autocovariance γ_k

For a general ARMA(n,m) model, we have,

for $k \leq m$,

$$\gamma_k = \sum_{j=1}^{n} \phi_j \gamma_{k-j} + (1 - \sum_{l=k}^{m} \theta_l G_{l-k})\sigma_a^2$$

for $k \geq m+1$

$$\gamma_k = \sum_{j=1}^{n} \phi_j \gamma_{k-j}$$

The interpretation is that we need self-starting for first m autocorrelation. And for autocorrelation after m lag, it is completely self-sustaining with a AR(n) fashion.

7.4 Partial autocorrelation can determine order of AR model

Clearly, from previous subsection, we saw sample correlation for MA process can tell us the order of MA process clearly. However, for a pure AR process, still we don't have such indication tools. To cope with it, we introduce the concept called *partial autocorrelation*, ρ'_k which is modified from autocorrelation: ρ_k so as to provide a cut-off after n-th lag for AR(n) process.

Let's denote for an AR(k) process, the AR parameters are $\phi_{kl}, l = 1, 2, \ldots, n$.

$$X_t = \phi_{k1} X_{t-1} + \phi_{k2} X_{t-2} + \ldots + \phi_{kk} X_{t-k} + a_t$$

Then note that $i > m = 0$ from the above subsection, we have for i-th lag autocorrelation, $i = 1, 2, \ldots, k$,

$$\rho_i = \sum_{j=1}^{n} \phi_{kj} \rho_{i-j}.$$

The above expression is very important in that it reduces an AR process into a finite memory style, which means **if we consider a vector of sequentially autocorrelation, when the size of vector exceeds the order of AR model, its entry will be linear dependent!** This naturally leads to the idea of check the linear independence of autocorrelation to find the order of AR process n.

Let's consider the *even* property and consider iterate i over $\{1, \ldots, k\}$. We have the so-called **Yule-Walker** equations for AR(k) process.

$$
\begin{aligned}
\rho_1 &= \phi_{k1}\rho_0 + \ldots + \phi_{kk}\rho_{k-1} \\
\rho_2 &= \phi_{k1}\rho_1 + \ldots + \phi_{kk}\rho_{k-2} \\
&\vdots \\
\rho_k &= \phi_{k1}\rho_{k-1} + \ldots + \phi_{kk}\rho_0 \\
\iff \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} &= \begin{bmatrix} \rho_0 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_0 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix}
\end{aligned}
$$

Note that the necessary and sufficient condition for this to be an AR(k) process is that $\phi_{kk} \neq 0$. From Cramer's rule, we have

$$
\phi_{kk} \sim \begin{vmatrix} \rho_0 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_0 & \cdots & \rho_2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & \rho_k \end{vmatrix} \neq 0
$$

It is important to note that, we can still compute the autocorrelation of $k+1$ lag with gives ρ_{k+1}, which gives

$$
\rho_{k+1} = \phi_{k1}\rho_k + \phi_{k2}\rho_{k-1} + \ldots + \phi_{kk}\rho_1
$$

and the corresponding

$$
\phi_{k+1,k+1} = \begin{vmatrix} \rho_0 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & \rho_0 & \cdots & \rho_2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho_k & \rho_{k-1} & \cdots & \rho_{k+1} \end{vmatrix}
$$

Now I will show that the last row is a linear combination of first k rows with coefficient ϕ_{k1} to ϕ_{kk}. Consider such linear combination for the first column

$$
\phi_{k1}\rho_{k-1} + \ldots + \phi_{kk}\rho_0
$$

which is exactly the Yule-Walker equation for ρ_k, similarly for the rest first k columns, ρ_{k-1} to ρ_0. Therefore

$$
\textbf{partial autocorrelation: } \rho'_{k+1} \triangleq \phi_{k+1,k+1} = 0
$$

if the corresponding process AR order is smaller than $k+1$. It is

Chapter 8

Stochastic Trend and Seasonality

8.1 Motivations

In time series data, we usually see some trend, those trend can be *deterministic* or *stochastic*. In this chapter, we will only discuss about stochastic trend, which origins from *random shocks*, but later on purely due to the strong correlation ($\lambda = 1$) between history and current state.

I have to admit the word trend is very inconsistent with stochasticity. But this concept is very interesting and reminds me of a chaotic dynamical system, which is a deterministic system, like Lorenz system. But why does every numerical run I have becomes unpredictable after long time? This is because the joint force between random shocks, i.e., the random, unpredictable effect and $\lambda = 1$, get into the system and does not die over time. But note the difference, stochastic trend is simply a linear system with forced stochastic, so it is a stochastic system. While Lorenz system is deterministic system.

8.2 Stochastic trend

Note that characteristic equations define a set of roots which describes the dynamics completely without the stochastic part. We define what we mean by trend and seasonality.

1. real roots represents trend: *constant, growth, decay*

2. complex roots represents seasonality

8.2.1 Stochastic constant trend

Look at the following example,

$$G_j = g_1 + \ldots + g_n \lambda_n^j \tag{8.1}$$

where $\lambda_1 = 1$ and $|\lambda_k| < 1, \forall k > 1$, and when $j \to \infty$, we will have G_j dominated by g_1 since other terms will decay.

Particularly, since G_j is not decaying out. We will have the a *constant trend*, which makes the data remain on the same level and note that those level changes are purely due to random noise.

To see this, let's consider the following case when time delay larger than L, we assume $G_j \approx g_1$,

$$\begin{aligned} X_t &= \sum G_j a_{t-j} \\ &= G_0 a_t + \ldots + G_L a_{t-L} + g_1(a_{t-L-1} + a_{t-L-2} + \ldots) \end{aligned}$$

Note that the second half describes the *long lasting effect* from the $\lambda = 1$, which further lead to the unbounded variance of the X_t. Therefore, it is not a stationary process but a marginally stable process. Also, note that the initial memory about where the simulation starts is forgotten. The later part will dominate and make this process more and more like a random walk process. Last, note that g_1 is determined by other roots and moving average part.

As a clear notation, from now on, we consider the following for describing stochastic trend,

$$\nabla X_t = X_t - X_{t-1} = (I - B)X_t \tag{8.2}$$

8.2.2 Stochastic linear trend

When there are two roots are close to one, let's consider a second order system. Due to the duplicity of the roots, we have

$$\text{when j is large, } G_j \approx g_1 + g_2 j, \tag{8.3}$$

where g_1 and g_2 are determined by other roots of ARMA model and moving average parts.

Specifically, for ARMA(2,1), we have

$$G_j = 1 + (1 - \theta_1)j. \tag{8.4}$$

Therefore G_j exhibits a linear trend in j, this is why we call it *stochastic linear trend.*

To give more insights into this, let's consider the following

$$\begin{aligned} X_t &= a_t + (1 + g_2)a_{t-1} + (1 + 2g_2)a_{t-2} + \ldots \\ X_{t+1} &= a_{t+1} + (1 + g_2)a_t + (1 + 2g_2)a_{t-1} + \ldots \\ X_{t+1} - X_t &= a_{t+1} + g_2(a_t + a_{t-1} + \ldots) \end{aligned}$$

Therefore, the difference between X_{t+1} and X_t is a stochastic constant trend. Note that this means the slope will not change suddenly. Therefore, the corresponding time series looks very smooth.

It is curious to me to further computing the second order difference,

$$\begin{aligned} Y_t &= X_{t+1} - X_t = a_{t+1} + g_2(a_t + a_{t-1} + \ldots) \\ Y_{t+1} &= X_{t+2} - X_{t+1} = a_{t+2} + g_2(a_{t+1} + a_t + \ldots) \\ Z_{t+2} \triangleq Y_{t+1} - Y_t &= X_{t+2} - 2X_{t+1} + X_t = a_{t+2} + (g_2 - 1)a_{t+1} = \text{stationary process!} \end{aligned}$$

So Z_t follows a *moving average* 1, i.e., MA(1) model.

8.2.3 Stochastic polynomial trend

If we have three roots close to one, we will see a quadratic trend in G_j. Note that this would numerically happen over an oversampling frequency is acting on a low frequency signals. The corresponding characteristic equation will have roots that is very close to 1.

In general, we call it a $n-th$ order polynomial trend in G_j if $n+1$ discrete real roots are close to one.

Due to the randomness from random shock a_t, we will see a stochastic nature.

8.3 Stochastic seasonality

If the roots are complex, the data will have seasonal or periodic pattern. If there are other roots but decaying, then long time later, the pattern G_j will be dominated by the roots close to 1. So G_j not decaying will most affect the *long time behavior of the system.*

8.3.1 A closer look at Green function

Consider we have first two roots in the AR characteristic equation being complex conjugate and $\lambda_1 = \overline{\lambda_2}$, $|\lambda| = 1$, while other roots are decaying real roots.

Then the Green function becomes

$$G_j = 2gr^j \cos(j\omega + \beta) + \lambda_3^j + \lambda_4^j + \ldots \to 2g\cos(j\omega + \beta) \tag{8.5}$$

which clearly shows that G_j is not decaying to zero as j goes ∞.

8.3.2 An operator view

Previously, for stochastic trend, we have operator ∇^d but for seasonality, what is the corresponding operator?

For seasonality, certainly we have for the complex conjugate, $r = 1$ and period p, so we have

$$(1 - \lambda_1 B)(1 - \lambda_2 B) = (1 - \phi_1 B - \phi_2 B^2) = 1 - 2\cos(\frac{2\pi}{p})B + B^2$$

Note that this operator would be helpful in extracting everything else into an stationary ARMA model. Moreover, since the above operator only has one parameter p, so a Table can be easily built for quick look.

8.3.3 Apply seasonal operator

Similar as before in the stochastic trend case, applying seasonal operator would lead to a simpler version of the remaining model and simply the cost is just a pre-processing of the data.

For example, we can reduce the cost of finding a ARMA(2,1) model into MA(1).

8.3.4 Example: ARMA(2,1)

Consider a ARMA(2,1) example, so we have

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t - \theta_1 a_{t-1} \tag{8.6}$$

where $\lambda_{1,2}$ are complex conjugate.

Note that from previous chapter, we have the equation for roots from ARMA coefficients as

$$\begin{aligned}\lambda_1, \lambda_2 &= \frac{1}{2}\left(\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}\right)\\ &= r(\cos(\omega) \pm \sin(\omega))\\ &= \frac{1}{2}\left(\phi_1 \pm +i\sqrt{-(\phi_1^2 + 4\phi_2)}\right)\end{aligned}$$

Therefore, $r \triangleq |\lambda_1| = \frac{1}{2}\sqrt{\phi_1^2 - \phi_1^2 - 4\phi_2} = \sqrt{-\phi_2}$. Also $r\cos(\omega) = \frac{1}{2}\phi_1$. Therefore $\cos\omega = \frac{\phi_1}{2r}$. So, $\omega = \cos^{-1}\frac{\phi_1}{2\sqrt{-\phi_2}}$.

If it is a seasonality, then $r \approx 1$, so we have

$$\omega = \cos^{-1}\frac{\phi_1}{2}$$

certainly the period is,

$$p = \frac{2\pi}{\omega} = \frac{2\pi}{\cos^{-1}\frac{\phi_1}{2}}$$

.

8.3.5 A strange example: $\lambda = -1$

Note that when $\lambda = -1$, it is easier to adapt it into the framework of seasonality because it simply shows a period of 2 of oscillation.

Chapter 9

ARIMA(p,d,q) model

9.1 Motivation

Recall that ARMA model is most suitable for dealing with stationary process. Since a lot of derivation have already assumed that the autocorrelation and autocovariance only depends on the time lag. So there is a statistically shift invariance in time, i.e., stationary.

Autoregressive Integrated Moving Average (ARIMA) model is designed for dealing with non-stationary process that contains stochastic trend. The key observation is stated in the previous chapter that *by differencing the non-stationary process by some iterations, we will end up with back to a stationary process*. Therefore, the original non-stationary process can be obtained by summing the stationary process. Also, later we will show that system with seasonality is not stationary as well.

Mathematically, this is how it works and makes sense. The general expression for ARMA(p+d,q) model is

$$\begin{aligned} \text{Given } \Phi(B)X_t &= \Theta(B)a_t \\ \phi(B)\nabla^d X_t &= \Theta(B)a_t \end{aligned}$$

$$\begin{aligned} \phi(B)\omega_t &= \Theta(B)a_t \text{ This is a ARMA(p,q) model} \\ \nabla^d X_t &= \omega_t \text{ This is simply deterministic integration} \end{aligned}$$

where $\Phi(\cdot)$ is a $p+d$-order polynomial, $\phi(\cdot)$ is a p order polynomial and $\Theta(\cdot)$ is a q-order polynomial. We call this process as ARIMA(p,d,q) process. Note that when $p=0$, we call integrated moving averaging (IMA) model.

9.2 Implementation of ARIMA model

The steps to implement ARIMA Model is:

1. use ARMA model on raw data, if there are d roots close to 1 is found, apply ∇^d on the original data, to get new time series ω_t
2. apply ARMA Model on ω_t

Maybe you wonder why not using ARMA model directly. Indeed, using ARMA model directly would require estimate AR coefficient while if we directly apply ∇^d, we don't need to find those parameter anymore. The resulting model that has reduced parameter to determine is called *parsimonious model*.

Chapter 10

Deterministic trends and seasonality

10.1 Motivation

Sitting in front of the data, we have prior knowledge either from domain knowledge or physics, to model such time series. To incorporate them, we use them separately in a deterministic way since we are very confident about it.

For example, consider a toy process as

$$X_t = 1.16t + 0.76X_{t-1}$$

which is clearly has a deterministic trend $1.16t$, although whether it is physical is quite questionable.

Note that stochastic linear trend has nothing to do with the fact that the model is linear in time.

10.2 General idea: study the residual

In general there are two approaches, first is two step and second is integrated approach.

For simplicity, we consider two step to model the time series. First simply run a regression model and predict the dynamics of the residuals as ARMA model. It should be noted that it is similar to what I have for data-driven closure model, where X_t is modeled as a NARMAX model using polynomial or neural network instead of a linear one in ARMA model.

$$\begin{aligned} y_t &= \beta_0 + \beta_1 t + X_t \\ X_t &= ARMA(n, m) \end{aligned}$$

Note that the residual from linear regression is being studied as a time series using ARMA model to predict. Note that study the autocorrelation is helpful in determining whether or not should we use ARMA model.

Note that data is not centered in this case and two step approach is much easier to carry out for simple problems. Also, two step result can be used as an initial condition for integrated approach.

10.3 Difference with stochastic trend/seasonality

When the trend/seasonality in the data is not really clear and it changes from data to data, then this *trend/seasonality* is probably a stochastic trend/seasonality.

10.4 Examples

10.4.1 Deterministic seasonal trend

First filtering with a deterministic seasonal trend model with ω and ϕ_k as model parameter and n as hyperparameter,

$$y_t = \alpha_0 + \sum_{k=1}^{n} \beta_k \sin(k\omega t + \phi_k) + X_t,$$

while in many application $n = 1$ or $n = 2$ is enough.

We expand the expression as

$$\sin(k\omega t + \phi_k) = c_k \sin(k\omega t) + \sqrt{1 - c_k^2} \cos(k\omega t)$$

10.4.2 Linear/exponential and seasonal trend

For a linear growing but still oscillation trend, one possible formulation is

$$y_t = \sum_{k=1}^{l} \alpha_k e^{r_k t} + \sum_{k=1}^{i} \beta_k e^{b_k t} \sin(k\omega t + \phi_k) + X_t$$

$$X_t \text{ follows ARMA model}$$

Note that when $r_k \ll 1$, the exponential term reduces to a linear term, so it can capture deterministic linear term. Certainly, polynomial trend is not a problem as well. That is the reason we use exponential function. Usually, this setup of hybrid model requires two stages. First is to estimate the deterministic part and second all the residuals that cannot be explained come to ARMA part.

Chapter 11

Forecasting

In this section, I will briefly talk about the forecasting issues. The basic Golden rule to remember is the conditional expectation, or conditional distribution. I will follow the classnotes to give examples one by one.

11.1 One step ahead prediction AR(1) model

Note that

$$X_t = \phi_1 X_{t-1} + a_t$$

We denote one step ahead prediciton at time $t-1$ is $\hat{X}_{t-1}(1)$. It means that everything before and including $t-1$ is known.

Clearly, given $t-1$, we know X_{t-1}. Therefore

$$\hat{X}_{t-1}(1) = \mathbb{E}_{t-1} X_t = \mathbb{E}_{t-1}(\phi_1 X_{t-1} + a_t)$$

Clearly, in the second, term a_t is a future term so conditional expectation is full expectation which is 0. Therefore it is $\sim \mathcal{N}(\phi_1 X_{t-1}, \sigma_a^2)$.

11.2 l step ahead prediction

In general, we denote $\hat{X}_t(l)$ as l step ahead prediction at time t, as

$$\hat{X}_t(l) = \mathbb{E}(X_{t+l} | X_t, X_{t-1}, \ldots)$$

Given the above formula, we iterate over $l = 1, 2, 3, 4, \ldots$. So we can find a recursive expression for each of them.

Next, we define the l step ahead prediction error as

$$e_t(l) = X_{t+l} - \hat{X}_t(l)$$

Naturally, we have $X_{t+l} = \hat{X}_t(l) + e_t(l)$

11.2.1 Expression from Green function

The general formula in Green function expression would be

$$X_t = \sum_{j=0} G_j a_{t-j}$$

Therefore $e_t(l) = a_{t+l} + G_1 a_{t+l-1} + \ldots + G_{l-1} a_{t-1}$.

$$E(e_t(l)) = 0$$

and

$$\text{Var}(e_t(l)) = \sigma_a^2(1 + G_1^2 + \ldots G_{l-1}^2)$$

.

Clearly the corresponding predictive distribution would be

$$X_{t+l}|X_t, X_{t-1}, \ldots \sim \mathcal{N}(\hat{X}_t(l), \sigma_a^2(1 + G_1^2 + \ldots + G_{l-1}^2))$$

11.2.2 Eventual forecasts

When l goes to ∞, the autoregressive parameters will control the forecasts as

$$\hat{X}_t(l) = \phi_1 \hat{X}_t(l-1) + \phi_2 \hat{X}_t(l-2) + \ldots \phi_n \hat{X}_t(l-n)$$

For marginally stable system, the eventual forecast will be non-zero. For stable or unstable, it is obvious.

11.3 Exponential smoothing

The idea is very simple. So there is no stochastic model here. Only

$$\hat{X}_t(1) = \sum_j \lambda(1-\lambda)^j X_{t-j}$$

where $\lambda = 1 - \theta$. When θ is large, each term decays slowly so it would be long memory. Otherwise it is short memory.

Note that $\hat{X}_t(1) = \hat{X}_{t-1}(1) + \lambda(X_t - \hat{X}_{t-1}(1))$.

Appendix A

Proof in linear regression

A.1 Linear Least square estimator follows a Multi-variate Gaussian distribution

It is widely known that *if the data is generated by a Gaussian, the least square estimator is a unbiased estimator with a multi-Variate Gaussian/Normal (MVN) distribution.* Note that it is applicable to any linear least square estimation, including *linearly reconstruction a signals.*

We know from intuitive that the transformation is a Gaussian, but how to rigorously prove it? Here I provide a strict proof starting from the PDF.

Theorem 1. *Given data $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$, where $\mathbf{X}$ and β is constant and $\epsilon_{n\times 1} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n})$, if $\mathbf{X}$ is full rank, the following estimator for β follows a multi-variate normal distribution*

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{Y} \sim \mathcal{N}(\beta, \sigma^2((\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}))$$

Proof. Note that $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ can be equivalently written as $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$, which further states the *probability* around $\mathbf{y}$ with interval $d^n\mathbf{y}$ as

$$f_{\mathbf{Y}}(\mathbf{y})d^n\mathbf{y} = \frac{1}{\sqrt{(2\pi)^n}\sigma^n}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)}d^n\mathbf{y} \tag{A.1}$$

where $d^n\mathbf{y} = dy_1 dy_2 \dots dy_n$.

Consider that $\forall \mathbf{A} \in \mathbb{R}^{n\times n}$, $(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{A} = \mathbf{I}$, therefore plug that into the above we have

$$\begin{aligned}(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta) &= (\mathbf{y}-\mathbf{X}\beta)^\top((\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{A})^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{A}(\mathbf{y}-\mathbf{X}\beta) \\ &= (\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{X}\beta)^\top((\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top)^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{X}\beta).\end{aligned}$$

Recall that since $\mathbf{X}$ is full rank, therefore from the property of pseudo-inverse, we known its left pseudo-inverse exists therefore the following transformation is well defined,

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{\mathbf{-1}}\mathbf{X}^\top\mathbf{Y}.$$

Let's take $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top = \mathbf{X}^+$. Thus $\mathbf{A}^+ = \mathbf{X}$. Therefore we have

$$(\mathbf{A}^\top\mathbf{A})^{-\mathbf{1}}\mathbf{A}^\top = \mathbf{X}$$

So, plug it back we have

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) &= (\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\beta)^\top((\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top)^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top(\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\beta)\\ &= (\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\beta)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{A}\mathbf{y} - \mathbf{A}\mathbf{X}\beta).\end{aligned}$$

Since $\mathbf{X}$ is full rank (in column), therefore it is trivial to show that A is also full rank (in column), therefore we have

$$d^p\mathbf{A}\mathbf{y} = \sqrt{\det(\mathbf{A}^\top\mathbf{A})}d^n\mathbf{y}.$$

$$d^n\mathbf{y} = \frac{d^p\mathbf{A}\mathbf{y}}{\sqrt{\det(\mathbf{A}^\top\mathbf{A})}}$$

The reason for the above equality is not trivial. My understand is that ne should get a feeling about the geometrical meaning of SVD decomposition of A, which is *rotation-scaling-rotation*, where only the *scaling* step can change the mapped surface area. So one can imagine that the product of scaling components, i.e., the product of singular values corresponds to the area changing.

Further, consider economic SVD $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, one can easily show that

$$\det(\mathbf{A}^\top\mathbf{A}) = \det(\mathbf{U}\mathbf{\Sigma}^{-\mathbf{2}}\mathbf{U}^\top) = \frac{1}{\det(\mathbf{X}^\top\mathbf{X})}$$

Therefore, consider the transformation on $\mathbf{y}$ we have

$$\begin{aligned}f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{\sqrt{(2\pi)^n}\sigma^n}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)^\top(\mathbf{y}-\mathbf{X}\beta)} \qquad (A.2)\\ &= \frac{1}{\sqrt{(2\pi)^n}\sigma^n}e^{-\frac{1}{2\sigma^2}(\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{X}\beta)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{A}\mathbf{y}-\mathbf{A}\mathbf{X}\beta)} \qquad (A.3)\\ &= \frac{1}{\sqrt{(2\pi)^n}\sigma^n}e^{-\frac{1}{2\sigma^2}(\mathbf{u}-\beta)^\top\mathbf{X}^\top\mathbf{X}(\mathbf{u}-\beta)} \qquad (A.4)\\ \to f_{\mathbf{A}\mathbf{Y}}(\mathbf{u}) &\sim e^{-\frac{1}{2}(\mathbf{u}-\beta)^\top(\sigma^{\mathbf{2}}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}})^{-1}(\mathbf{u}-\beta)} \qquad (A.5)\\ \to \hat{\beta} &\sim \mathcal{N}(\beta, \sigma^2\mathbf{X}^\top\mathbf{X}^{-1}). \qquad (A.6)\end{aligned}$$

□

Further, we can easily see that the result above is simply the result of Bayesian posteriori of a uniform prior. So usually it is called *Gaussian linear models* which implies the likelihood is a *Gaussian* and inside the exponent function the parameters *linear*. Therefore, it is consistent with the Bayesian framework and can lead to many interesting variants.

A.2 Unbiased Estimation for $\sigma_\epsilon^2 = SSR/(n-p-1)$

Note that for the least-square residual, $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta} = (\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^\top)\mathbf{y} = \mathbf{Q}\mathbf{y}$. Therefore it must be a multi-variate Normal distribution.

Further note that $\mathbf{Q}$ kills $\mathbf{X}$. $\mathbf{Q}\mathbf{y} = \mathbf{Q}(\mathbf{X}\beta + \epsilon) = \mathbf{Q}\epsilon$. It implies that $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is not full rank. So it means the residual must also be rank-deficient after this transformation.

Let consider the matrix $\mathbf{Q}$. So we can observe that $\mathbf{Q}^\top = \mathbf{Q} = \mathbf{Q}^2$. So it is Herimitian and it is a projection matrix. Therefore the minimal polynomial of Q must divide $z(1-z)$, so it only has eigenvalue either 0 or 1 or both.

Next, consider the trace of $\mathbf{Q}$, using *cyclic relation,*

$$\mathrm{tr}(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{\mathbf{-1}}\mathbf{X}^\top) = n - p - 1. \tag{A.7}$$

Note that $\mathbf{Q}$ is Herimitian, so there exists a real *unitary* matrix $\mathbf{V}$ that transform $\mathbf{Q}$ into real diagonal matrix, which contains diagonal elements as its eigenvalues which is $n-p-1$ times 1 and $p+1$ times 0.

$$\mathbf{\Lambda} = \mathbf{V}^\top\mathbf{Q}\mathbf{V} \tag{A.8}$$

Note that we are interested in the variance scalar of the true residual ϵ, σ_ϵ^2.

$\because \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \because \hat{\epsilon} = \mathbf{Q}\epsilon, \therefore \hat{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}\sigma_\epsilon^2\mathbf{Q}^\top)$. Consider $\mathbf{V}^\top\hat{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^{\mathbf{2}}\mathbf{V}^\top\mathbf{Q}\mathbf{Q}^\top\mathbf{V}) = \mathcal{N}(0, \sigma_\epsilon^2\mathbf{\Lambda}^2) = \mathcal{N}(0, \sigma_\epsilon^2\mathbf{\Lambda})$.

Based on the above, we have the following statistic,

$$SSE = \text{sum of squares of residuals over } \sigma_\epsilon^2 = \|\hat{\epsilon}/\sigma_\epsilon\|_2^2 = \|\mathbf{V}^\top\hat{\epsilon}/\sigma_\epsilon\|_2^2 \sim \chi^2(n-p-1). \tag{A.9}$$

Therefore $\frac{\mathbb{E}(SSE)}{n-p-1} = \sigma_\epsilon^2$. So according to the definition, a unbiased estimator for $\sigma_\epsilon^2 = SSE/(n-p-1)$.

A.3 t-test for significance on single parameter

Recall that for constructing a t-distribution, we need

- standard Normal distribution
- χ^2 with whatever degree of freedom

Note that

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma_\epsilon^2(X^\top X)^{-1}).$$

The diagonal element of the covariance matrix has its statistical meaning just as the name. So the $j+1$ diagonal element is meaning the variance of $\hat{\beta}_j$. Therefore, we obtain our standard Normal distribution as

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}} \sim \mathcal{N}(0,1) \tag{A.10}$$

Let's find the χ^2 distribution. Note that the $j+1$ diagonal element is nothing but $\nu_j\sigma_\epsilon^2$. It is a constant, but we don't know the noise level. Fortunately, note that we have a statistic $\hat{\sigma_\epsilon} = \sqrt{\frac{SSE}{n-p-1}}$, which is a random variable. It is because SSE is a random variable and it follows that $SSE/\sigma_\epsilon^2 \sim \chi^2(n-p-1)$. Thus, we just find a χ^2 distribution.

Then, the final step is to construct the t distribution

$$t_j = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}}{\sqrt{\frac{SSE/\sigma_\epsilon^2}{n-p-1}}} \sim t_{n-p-1} \tag{A.11}$$

Note that

$$\frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j)}}}{\sqrt{\frac{SSE/\sigma_\epsilon^2}{n-p-1}}} = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\nu_j\sigma_\epsilon^2}}}{\sqrt{\frac{SSE/\sigma_\epsilon^2}{n-p-1}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\nu_j}\sqrt{\frac{SSE}{n-p-1}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\nu_j}\hat{\sigma}_\epsilon}.$$