ST4253 Cheatsheet

by Zong Xun

Time Series Regression

Strictly stationary if distribution of $\{x_{t_1}, \cdots, x_{t_k}\}$ is the same as the distribution of $\{x_{t_1+h}, \cdots, x_{t_k+h}\}$.

$$P(x_{t_1} \leq c_1, ..., x_{t_k} \leq c_k) = P(x_{t_1+h} \leq c_1, ..., x_{t_k+h} \leq c_k)$$

for all time points k and time shifts h. In particular,

$$P(x_t \leq c) = P(x_s \leq c), \quad E[x_t] = \mu_t = E[x_s]$$

Weakly stationary if

- The mean value is constant and does not depend on time.
- Autocovariance depends on s and t only through their difference $|s-t|$

Unit Tests

- Aug. Dickey Fuller: H_0 is non-stationary and non-seasonal
- KPSS: H_0 is stationary and non-seasonal

Transformations

Stabilize variance and normalize distribution of data.

- $w_t = \log(y_t)$ and $w_t = y_t^p$
- Box-Cox Transformation

$$f(x) = \begin{cases} \log(y_t) & \text{if } \lambda = 0 \\ (sign(y_t)|y_t|^\lambda - 1)/\lambda & \text{if } \lambda > 0 \end{cases}$$

Time series Decomposition

$$y_t = S_t + T_t + R_t \text{ and } y_t = S_t \times T_t \times R_t$$

Use multiplicative decomposition when the variation in seasonal/trend appears proportional to level of time series.

Recovering the components

1. The estimate of the trend-cycle at time t is obtained by averaging values of the time series within k periods of t.

$$\hat{T}_t = \frac{1}{m}\sum_{j=-k}^{k} y_{t+j} \quad z_t = \frac{1}{2}(T_t + T_{t-1})$$

Use $m = 2k+1$-MA if m is odd and $2 \times m$-MA ow.

2. Calculate the detrended series, $y_t - T_t$
3. Take average of detrended values for each season, and zero-center them to ensure unique decomposition (captures only seasonal deviations).
4. $R_t = y_t - T_t - S_t$

Disadvantages of Classical

1. Trend unavailable for the first and last few observations
2. Over-smoothing of rapid rises and falls
3. Assumes seasonal component repeats YOY
4. Not robust to outliers

Advantages of STL

1. Can estimate trend at endpoints
2. The smoothness can be controlled by user
3. Seasonal component is allowed to change over time, and rate of change can be controlled by the user
4. Robust to outliers

Forecasting Methods

Mean & Naive Method

$$\hat{y}_{T+h} = \frac{1}{T}\sum_{1}^{T} y_t \text{ and } \hat{y}_{T+h} = y_T$$

Seasonal Naive Method

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)}$$

where m is the seasonal period and $k = \lfloor (h-1)/m \rfloor$.

Drift Method

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}(y_T - y_1)$$

Method	h-step Forecast Standard Deviation
Mean	$\hat{\sigma}\sqrt{1+\frac{1}{T}}$
Naïve	$\hat{\sigma}\sqrt{h}$
Seasonal Naïve	$\hat{\sigma}\sqrt{k+1}$
Drift	$\hat{\sigma}\sqrt{h\left(1+\frac{h}{T-1}\right)}$

Measures of forecast accuracy

$$MAE = mean(|e_{T+h}|) \quad RMSE = \sqrt{mean(e_{T+h}^2)}$$

$$MSE = mean(e_{T+h}^2) \quad MAPE = \frac{100\%}{n}\sum_{1}^{h}\left(\frac{|e_{T+h}|}{|y_{T+h}|}\right)$$

For non-seasonal time series, scale errors using naive forecasts:

$$q_{T+h} = \frac{e_{T+h}}{\frac{1}{T-1}\sum_{t=2}^{T}|y_T - y_{t-1}|}$$

For seasonal time series, scale forecast errors using seasonal naive forecasts:

$$q_{T+h} = \frac{e_{T+h}}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_T - y_{t-m}|}$$

Putting them together:

$$MASE = mean(|q_{T+h}|) \quad RMSSE = \sqrt{mean(q_{T+h}^2)}$$

where $T-m$ is not squared.

Trend and Seasonality in ACF plots

- Trend $\rightarrow$ ACF for small lags tend to be large and positive, slowly decreasing as lags increase.
- Seasonal $\rightarrow$ ACF will be larger for km lags

Linear Models

Assumptions: Reasonable approximation and normal errors that are uncorrelated and unrelated to the predictor variables. Uses **least square estimation** to estimate the coefficients.

$$\sum_{t=1}^{T}\epsilon_t^2 = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2$$

- Use $m-1$ dummies for m categories ow the regression will be singular and inestimable.

Goodness of fit

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = 1 - \frac{\sum_{t=1}^{T}e_t^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

- Never decreases, can cause overfitting.
- Proportion of variation explained by model.

$$R_{adj} = 1 - \frac{SSE/T-k-1}{TSS/T-1} = 1 - (1-R^2)\frac{T-1}{T-k-1}$$

- Accounts for the number of estimated parameters of the model, does not always increase.
- $SSE/T-k-1$ is a measure of fit.

$$AIC = Tlog(SSE/T) + 2(k+2)$$
$$BIC = Tlog(SSE/T) + 2(k+2)log(T)$$

Prediction Interval

For simple regression,

$$\hat{y} \pm 1.96\sigma\sqrt{1 + \frac{1}{T} + \frac{(x-\bar{x})^2}{(T-1)S_x^2}}$$

where $S_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})^2$

Notes

- BIC penalizes more heavily.
- Do not drop predictors based on scatterplot.
- Do not use p-values alone to select. Can be misleading when two or more predictors are correlated with each other.
- Select highest adj. R^2, lowest CV, AIC, BIC
- Note that log transformation requires all values to be positive, so use $\log(x+1)$.
- **Residuals**: No pattern against predictors (nonlinearity).
- **Residuals**: No pattern against fitted values (heteroskedasticity).

Matrix formulation

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t.$$

Let $\boldsymbol{y} = (y_1, \ldots, y_T)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \ldots & x_{k,T} \end{bmatrix}.$$

Then

$$\boldsymbol{y} = \underset{n \times k}{\boldsymbol{X}}\overset{k \times 1}{\boldsymbol{\beta}} + \overset{n \times 1}{\boldsymbol{\varepsilon}}.$$

ε has mean 0 and varaince $\sigma^2 \boldsymbol{I}$.

is a theoretical construct, capturing all the randomness or "noise" in the relationship between xi and yi

Least Square Estimation

- Minimize: $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$
- Differentiate wrt $\boldsymbol{\beta}$ gives

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- This equation is called the normal equation.
- Requires $(\boldsymbol{X}'\boldsymbol{X})$ to be invertible. If you fall for the dummy variable trap, $(\boldsymbol{X}'\boldsymbol{X})$ is a singular matrix.
- The residual variance is estimated using

residual refers to observed difference between the data and the fitted model.

$$\hat{\sigma}^2 = \frac{1}{\underset{\text{number of parameters}}{T-k-1}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

Likelihood

If the errors are iid and normally distributed, then

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^T(2\pi)^{T/2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})\right)$$

which is maximized when $(\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$ is minimized. So **MLE = OLS**.

Fitted values

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y} \qquad \text{where } \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.\end{aligned}$$

LOO cross-validation MSE

$$\mathrm{CV} = \frac{1}{T}\sum_{t=1}^{T}[e_t/(1-h_t)]^2$$

- e_t = residual at time t (from fitting model to all data)
- $h_1, \ldots, h_T$ are the diagonals of $\boldsymbol{H}$.

Multiple Regression Forecasts

Optimal forecasts

$$\hat{y}^* = \mathrm{E}(y^*|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{x}^*) = \boldsymbol{x}^*\hat{\boldsymbol{\beta}} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\boldsymbol{x}^*$ is a row vector containing the values of the predictors for the forecasts (in the same format as $\boldsymbol{X}$).

Forecast variance

$$\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*) = \sigma^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right]$$

- This ignores any errors in $\boldsymbol{x}^*$.
- 95% prediction intervals assuming normal errors:

$$\hat{y}^* \pm 1.96\sqrt{\mathrm{Var}(y^*|\boldsymbol{X}, \boldsymbol{x}^*)}.$$

Piecewise Linear Models

$$x_1 = x$$

$$x_2 = (x - c_1)_+ = \begin{cases} 0, & x < c_1 \\ x - c, & x \geq c_1 \end{cases}$$

where c_1 is the first knot and $c_i > c_j$ for $i > j$.

ETS Models

Simple Exponential Smoothing

$$\begin{aligned}\hat{y}_{t+1|t} &= \alpha y_t + (1-\alpha)\hat{y}_{t|t-1} \\ &= \sum_{j=0}^{t-1}\alpha(1-\alpha)^j y_{T-j} + (1-\alpha)^t \ell_0\end{aligned}$$

- Short term forecasting, no trend, no seasonality.
- Minimize $\sum_{i=1}^{T} y_i - \hat{y}_{i|i-1}$ (SSE), no closed form.

Holt's Linear (Damped) Trend

- Two smoothing parameters $0 \leq \alpha, \beta^* \leq 1, 0 < \phi < 1$
- Choose ℓ_0, b_0 to minimise SSE.
- As $h \rightarrow \infty$, $\hat{y}_{t+h|t} \rightarrow \ell_t + \frac{\phi}{(1-\phi)}b_T$.
- If $\phi = 1$, $\sum_{i=1}^{h}\phi^i$ is identical to no damping.

Holt and Winter's Method

- Smoothing parameter for seasonality, $0 \leq \gamma \leq 1-\alpha$
- The seasonal component is expressed as

$$s_t = \gamma^*(y_t - \ell_t) + (1-\gamma^*)s_{t-m}$$

where $\gamma^*(1-\alpha) = \gamma$ (only for additive errors)

Innovation State Space Models

Let $x_t = (\ell_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})$ and $\epsilon_t \sim^{iid} N(0, \sigma^2)$

$$y_t = h(x_{t-1}) + k(x_{t-1})\epsilon_t$$
$$x_t = f(x_{t-1}) + g(x_{t-1})\epsilon_t$$

For additive errors,

$$k(x_{t-1}) = 1 \quad y_t = \mu_t + \epsilon_t$$

For multiplicative errors,

$$k(x_{t-1}) = \mu_t \quad y_t = \mu_t(1+\epsilon_t)$$

where $\epsilon_t = (y_t - \mu_t)/\mu_t$ is relative error.

Estimation

- Maximise the likelihood function.
- Likelihood represents the probability of data given model
- Estimate the smoothing parameters and initial states.
- For additive errors, equivalent to minimising SSE.
- For multiplicative errors, **not** equivalent to minimising SSE.

Model Selection

$$AIC = -2log(L) + 2k \quad AIC_c = AIC + \frac{2k(k+1)}{T-k-1}$$
$$BIC = AIC + k[log(T) - 2]$$

where L is the likelihood and k is the number of parameters & initial states estimated.

Notes

- $ETS(A,*,*)$ and $ETS(M,*,*)$ have the same point forecasts but different prediction intervals.
- Wider for multiplicative errors.
- **For all point forecasts, set** $\epsilon_t = 0$ **for** $t > T$.
- Multiplicative models require all values to be strictly positive. Zero or negative values will break the model.
- We set $\beta = \alpha\beta^*$ in ECF
- $E(y_{t+h|x_t}$ is point forecast only if seasonality is additive.

ARIMA Models

Time series composed of two components, a nonstationary trend component, μ_t and a zero-mean stationary component, x_t. If μ_t is a polynomial of the k-th order, then the differenced series $(1-B)^k y_t$ is stationary.

AR

- Complex roots of $1 - \phi_1 z + \cdots + \phi_p z^p$ lie outside of unit circle on the complex plane.
- Stationarity assumption implies that $\forall i \in 1..t, E[y_i] = \mu$
- Can represent all stationary $AR(p)$ models as MA(∞).
- When $h > p$, the regression of y_{t+h} on $\{y_{t+1} \ldots y_{t+h-1}\}$ is

$$\hat{y}_{t+h} = \sum_{j=1}^{p} \phi_j y_{t+h-j}$$

and $\phi_{hh} = corr(y_{t+h} - \hat{y}_{t+h}, y_t - \hat{y}) = corr(\epsilon_{t+h}, y_t - \hat{y}) = 0$.

AR(1)

$$\gamma(h) = \frac{\sigma^2\phi^h}{1-\phi^2}(ACV) \quad \rho(h) = \phi^h(ACF)$$

- Require $|\phi| < 1$ for stationarity and convergence.

$$y_t = e_t + \sum_{j=1}^{\infty}\phi^j e_{t-j} \quad MA(\infty) \quad \phi^j = \psi_j \text{ (causal)}$$

- Mean is equal to $\mu = \frac{c}{1-\phi}$

AR(2)

- $|\phi_2| < 1$ and $\phi_2 + \phi_1 < 1$ and $\phi_2 - \phi_1 < 1$
- Autocovariance

$$\gamma(0) = \frac{\sigma^2(1-\phi_2)}{1-\phi_2(\phi_1^2 - \phi_1^2 - \phi_2 + \phi_2^2)} \quad \gamma(1) = \gamma(0)\frac{\phi_1}{1-\phi_2}$$

- Autocorrelation

$$\rho(0) = 1 \quad \rho(1) = \frac{\phi_1}{1-\phi_2}$$
$$\rho(h) = \phi_1\rho(h-1) + \phi_2\rho(h-2), h \geq 1$$

- When z_1 and z_2 are real and distinct, then

$$\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h}$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.
- When $z_1 = z_2 = z_0$ are real and equal, then

$$\rho(h) = z_0^{-h}(c_1 + c_2 h)$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.
- When $z_1 = \bar{z_2}$ are a complex conjugate pair, then $c_2 = \bar{c_1}$ (because $\rho(h)$ is real), and

$$\rho(h) = a|z_1|^{-h}cos(h\theta + b)$$

where a and b are determined by initial conditions. $\rho(h)$ dampens to zero exponentially fast as $h \to \infty$, but it does so in a sinusoidal fashion.

MA(q)

$$\gamma(h) = \sigma^2\sum_{j=0}^{q-h}\theta_j\theta_{j+h} \quad 0 \leq h \leq q$$
$$\rho(h) = \frac{\sum_{j=0}^{q-h}\theta_j\theta_{j+h}}{1+\theta_1^2+\cdots+\theta_q^2} \quad 1 \leq h \leq q$$

- MA(q) models are not unique.
- **Any finite-order MA is weakly stationary**
- Any MA(q) process can be written as AR(∞) if stationary
- Complex roots of $1 + \theta_1 z + \theta_2 z^2 + \ldots + \theta_q z^q$ lie outside of the unit circle on the complex plane.
- PACF will not cut off.

MA(1)

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma^2, & h = 0 \\ \theta\sigma^2, & h = 1 \\ 0, & h > 1 \end{cases}$$

$$\phi_{hh} \text{ (PACF)} = -\frac{(-\theta)^h(1-\theta^2)}{1-\theta^{2(h+1)}}, \quad h \geq 1$$

$$\pi_i = -\theta^i \text{ (invertibility)}$$

- Require $|\theta| < 1$ for stationarity and invertability.
- $|\theta_2| < 1$ and $\theta_2 + \theta_1 > -1$ and $\theta_1 - \theta_2 < 1$

ARMA(p, q)

If y_t has a nonzero mean μ, we set $c = \mu(1 - \phi_1 - \cdots - \phi_p)$ and write the model as

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + \cdots + \theta_q\epsilon_{t-q}$$
$$\phi(B)y_t = \theta(B)\epsilon_t$$

- ARMA models are stationary
- An ARMA model is causal if $\phi(z) = 0 \implies |z| > 1$

$$y_t = \sum_{j=0}^{\infty}\psi_i\epsilon_{t-j} \quad \sum_{j=0}^{\infty}|\psi_j| < \infty \quad \psi_0 = 1$$

- An ARMA model is invertible if $\theta(z) = 0 \implies |z| > 1$

$$\epsilon_t = \sum_{j=0}^{\infty}\pi_j y_{t-j} \quad \sum_{j=0}^{\infty}|\pi_j| < \infty \quad \pi_0 = 1$$

- ACF is a combination of AR and MA, hard to decipher which ARMA model it belongs.

$$\text{(general) } \gamma(h) = \sum_{j=1}^{p}\phi_j\gamma(h-j) + \sigma^2\sum_{j=h}^{q}\theta_j\psi_{j-h}$$
$$\text{(causal) } \gamma(h) = \sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+h}$$

Use general if $0 \leq h < max(p, q+1)$

ARMA(1, 1)

$$\gamma(0) = \phi\gamma(1) + \sigma^2(1+\theta\phi) \quad \gamma(1) = \phi\gamma(0) + \sigma^2\theta$$
$$\gamma(0) = \frac{\sigma^2(1+2\theta\phi+\theta^2)}{1-\phi^2} \quad \gamma(1) = \frac{\sigma^2(1+\theta\phi)(\phi+\theta)}{1-\phi^2}$$
$$\gamma(h) = \gamma(1)\phi^{h-1}$$
$$\phi^j = \psi_j \text{ (causal)} \quad \pi_i = \theta^i \text{ (invertibility)}$$

Partial ACF

Do regression on $y_{t+1} \ldots y_{h-1}$ and minimize

$$E[(y_{t+h} - \sum_{j=1}^{h-1}\beta_j y_{t+j})^2]$$

by taking the derivative w.r.t β.

$$\phi_{hh} = Cov(y_{t+h} - \hat{y}_{t+h}, y_t - \hat{y}_t)$$

where $\phi_{11} = \rho(1)$ by definition.

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Durbin Levinson Algorithm

Let $\phi_{00} = 0, \phi_{11} = \rho(1)$. For $n \geq 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1}\phi_{n-1,k}\rho(n-k)}{1 - \sum_{k=1}^{n-1}\phi_{n-1,k}\rho(k)}, n \geq 1$$

$\phi_{nk} = \phi_{n-1,k} - \phi_{nn}\phi_{n-1,n-k}, n \geq 2$ and $k = 1 \ldots n-1$

Derive the relevant autocorrelation functions and start from the base cases.

Model Selection

$$AIC = -2log(L) + 2(p+q+k+1)$$
$$AIC_c = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}$$
$$BIC = AIC + [log(T) - 2](p+q+k+1)$$

where L is the likelihood and $k = 1$ if $c \neq 0$ and $k = 0$ otherwise.

Notes

- MoM: Replace $\gamma(h)$ with $\gamma(\hat{h})$ and solve the equations.
- For point forecasts, replace all y_t with y_{t+h}, replace all future observations with their forecasts, **future errors with zero**, and past errors with corresponding residuals.

ARIMA(0,0,q)

$$y_{t+h} = \epsilon_{t+h} + \sum_{i=1}^{q}\theta_i\epsilon_{t+h-i} \quad \hat{y}_{t+h} = \sum_{j=h}^{q}\theta_j\epsilon_{t-j+h}$$
$$\hat{\sigma}_h^2 = \sigma^2(1 + \sum_{i=1}^{h-1}\hat{\theta}^2) \text{ for } h = 2, 3, \ldots$$

Use this equivalence to obtain prediction intervals for AR(1). For stationary models, intervals will converge and become the same. Intervals are narrower as only variations in errors are accounted for.

Misc

- A series can have more than one seasonal pattern.
- Autocovariance

$$\gamma_x(s,t) = E[(x_s - \mu_s)(x_t - \mu_t)] = E[st] - E[s]E[t]$$

- Autocorrelation (time series and its own lagged values)

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\sum_{t=K+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

- Residuals $\sim N(0, \sigma^2)$ are uncorrelated and unrelated to predictor variables/fitted values. ow: information left that should be included or variance is not constant.

$$\hat{\sigma}_e = \sqrt{\frac{1}{T-K-1-M}\sum_{t=1}^{T}e_t^2}$$

where $K+1$ (#parameters) and M (#missing data).
- Geometric Series

$$\sum_{i=0}^{\infty}ar^i = a(\frac{1}{1-r}) \quad \sum_{0}^{n-1}ar^i = \frac{a(1-r^n)}{1-r}$$

- Arithmetic Series

$$a + (a+d) + \cdots + (a+(n-1)\cdot d) = \frac{n}{2}(2a + (n-1)\cdot d)$$

- A moving average models of errors has the same ACF function as m.a. of simple linear model.