Review of Random Variables, Distributions, and Moments

Discrete Random Variables :

- Discrete probability Dist^{n}s.
- Countable number of values, y_i, $i=1,2,\dots$
- $P(Y_i = y_i) = p_i > 0 \ \forall i$
- $\sum p_i = 1$

E.g: Coin Toss (twice) : Outcome, $O = \{(H,H), (H,T), (T,H), (T,T)\}$

Let $Y =$ # of Heads observed in 2 flips.

$\therefore$ for $y=0$: $f(y) = 0.25$
$y=1$: $f(y) = 0.5$
$y=2$: $f(y) = 0.25$

where $f(y)$ is the probability dist. f^n.

Continuous random variables :

P.D.F $f(y) \geq 0$

$$\int_{-\infty}^{\infty} f(y)\,dy = 1$$

$$\int_{a}^{b} f(y)\,dy = P(a \leq Y \leq b)$$

Moments : Expectations of powers of r.v's. Consider 4:

1.) Mean / Expected value : $E(y) = \sum_i p_i y_i = \mu$

Measure of location, or "central tendency"

2.) Variance : $\sigma^2_y = Var(y) = E(y-\mu)^2$

Measure of dispersion, or scale, of y around its mean

$\sigma = std(y) = \sqrt{E(y-\mu)^2}$

Q: Why prefer σ over σ^2? Same units as 'y'.

3) Skewness : $S = \dfrac{E(y-\mu)^3}{\sigma^3}$ (Cubed deviation, scaled by σ^3 (for technical reasons)

Measure of asymmetry in a distribution.
Large positive value $\Rightarrow$ Long right tail.

4) Kurtosis : $K = \dfrac{E(y-\mu)^4}{\sigma^4}$

Measure of thickness of the tails of a dist.
$K > 3 \Rightarrow$ fat tails
$\Rightarrow$ Leptokurtosis (rel. to Gaussian dist.)
$\Rightarrow$ Extreme events are more likely to occur than under the case of normality.

~~Multivariate Normal Distribution~~

Multivariate Random Variables

$$\text{Cov}(x,y) = E((y-\mu_y)(x-\mu_x)) = \gamma_{x,y}$$

$$\text{corr}(x,y) = \frac{\text{cov}(x,y)}{\sigma_y \sigma_x} = \frac{\gamma_{xy}}{\sigma_x \sigma_y} = \rho_{xy} \in [-1, 1]$$

$\text{Cov}(x,y) > 0 \Rightarrow$ when $y_i > \mu_y$, then x_i tends to be $>$

$\text{Corr}(x,y)$ is unitless ... hence popular.

Statistics

$$\{y_t\}_{t=1}^{T} \sim f(y)$$

f is an unknown population distribution

1.) Sample Mean: $\bar{y} = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} y_t$

Sample variance, $\hat{\sigma}^2 = \dfrac{\sum_1^T (y_t - \bar{y})^2}{T}$

Unbiased estimator of σ^2, $s^2 = \dfrac{\sum_1^T (y_t - \bar{y})^2}{T-1}$

Sample std. deviation, $\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\dfrac{\sum_1^T (y_t - \bar{y})^2}{T}}$

$$s = \sqrt{s^2} = \sqrt{\dfrac{\sum_1^T (y_t - \bar{y})^2}{T-1}}$$

Sample skewness, $\hat{S} = \dfrac{\frac{1}{T}\sum_1^T (y_t - \bar{y})^3}{\hat{\sigma}^3}$

Sample kurtosis, $\hat{K} = \dfrac{\frac{1}{T}\sum_1^T (y_t - \bar{y})^4}{\hat{\sigma}^4}$

If $Y \sim N$, then $Y^2 \sim \chi^2$ dist.

& $\cancel{\chi^2}$

$N/\chi^2 \sim t$ dist.

$\chi^2/\chi^2 \sim F$ dist.

Test of Normality : Jarque-Bera test statistics

$$JB = \frac{T}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{K}-3)^2\right) \underset{\text{Large } T}{\sim} \chi^2(2\,d.f$$

(or $T-k$, for models with k parameters, testing for residuals)

Statistics background for Forecasting

I. Time Series : y_t , $t = 1, 2, \ldots, T$

$$y_1, y_2, \cdots, y_T .$$

Let 'τ' be the forecast lead time,

Then, forecast made at time '$t-\tau$' is denoted by $\hat{y}(t-$

$\therefore$ $\hat{y}(t-\tau)$ is forecast (or predicted) value of 'y_t' that was made at time period '$t-\tau$'

Forecast Error : $e_t(\tau) = y_t - \hat{y}_t(t-\tau)$

'Lead $-\tau$' forecast error.

Regression residual : $e_t = y_t - \hat{y}_t$

Usually, $e_t < e_t(\tau)$

II. **Graphical displays**

Classical tools of descriptive statistics not very useful if they lose the time dependent features of a series.

E.g: Figure 2.1 & 2.2.

Smoothing Techniques

Moving-averages : of span N

$$M_T = \frac{y_T + y_{T-1} + y_{T-2} + \cdots + y_{T-N+1}}{N}$$

$$= \frac{1}{N} \sum_{t=T-N+1}^{T} y_t$$

$\therefore$ Assigns a weight of $1/N$ to N-most recent observati

~~If~~ Let $Var(y_t) = \sigma^2$

Recall : $Var(a) = 0$

$Var(aX) = a^2 Var(X)$

$$Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) \; *$$

If $X_1, \ldots, X_N$ are such that $Cov(X_i, X_j) = 0, \forall i \neq j$

then, $Var\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} Var(X_i)$

$\therefore$ If we assume that $Var(y_t) = \sigma^2$, AND

Obs. are uncorrelated : $Cov(X_i, X_j) = 0 \; \forall (i \neq j)$, then

$$Var(M_T) = Var\left(\frac{1}{N} \sum_{t=T-N+1}^{N} y_t\right)$$

$$= \frac{1}{N^2} Var\left(\sum_{T-N+1}^{N} y_t\right)$$

$$= \frac{1}{N^2} \sum_{T-N+1}^{N} Var(y_t)$$

$$= \frac{N \cdot \sigma^2}{N^2}$$

$$= \frac{\sigma^2}{N}$$

$$\therefore \boxed{Var(M_T) < Var(y)}$$

Fig. 2.5 & 2.6

Linear filters can also be created using unequal weights

<u>Hanning filter</u> : $M_t^H = 0.25\, y_{t+1} + 0.5\, y_t + 0.25\, y_{t-1}$

This is an example of a "<u>Centered Moving average</u>"

$$M_t = \frac{1}{s+1} \sum_{-s}^{s} y_{t-i} \quad , \text{ where } \underline{\text{span}} = 2s+$$

<u>Disadvantage of linear filters</u> : Emphasizes outliers in the span duration.

Use <u>moving medians</u> instead!

$$m_t^{[N]} = \text{median}(y_{t-u}, \cdots, y, \cdots, y_{t+u}), \text{ wher}$$

$$\text{span} = N = 2u+1$$

$$\therefore \quad m_t^{[3]} = \text{med}\,(y_{t-1}, y_t, y_{t+1})$$

III. Numerical Descriptions

<u>Stationary time series</u> : If

$$F_y(y_t, y_{t+1}, \cdots, y_{t+n}) = F_y(y_{t+k}, y_{t+k+1}, \cdots, y_{t+k+}$$

1) $\mu_y = E(y) = \int_{-\infty}^{\infty} y \cdot f(y)\, dy \Rightarrow$ Constant mean

2.) $\sigma_y^2 = Var(y) = \int_{-\infty}^{\infty} (y - \mu_y)^2 \cdot f(y)\, dy \Rightarrow$ Constant Variance

3.) $E\left[\left(y_{t+j} - E(y_{t+j})\right)\left(y_{t+k} - E(y_{t+k})\right)\right]$ only depends on $(j-k)$

Covariance stationarity!

y_t is <u>Covariance stationary</u> iff :

1.) $E(y_t) = E(y_{t-s}) = \mu$

2.) $E[(y_t - \mu)^2] = E[(y_{t-s} - \mu)^2] = \sigma_y^2$

$var(y_t) = var(y_{t-s}) = \sigma_y^2$

3.) $E[(y_t - \mu)(y_{t-s} - \mu)] = E[(y_{t-j} - \mu)(y_{t-j-s} - \mu)] = \gamma$

$cov(y_t, y_{t-s}) = cov(y_{t-j}, y_{t-j-s}) = \gamma_s$

This is called <u>weak stationarity</u>, second-order stat
or wide-sense stationary proce

strongly stationary processes are not required to hav
a finite mean &/or variance.

Simply put, a time series is covariance stationary if its mean and all autocovariances are unaffected by a change of time origin

<u>Sample</u> :

$$\bar{y} = \hat{\mu}_y = \frac{1}{T} \sum_{t=1}^{T} y_t$$

$$s^2 = \hat{\sigma}_y^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \bar{y})^2$$

Autocovariance and Autocorrelation

Useful tip: plot y_t vs. y_{t+1}.

Recall, autocovariance at lag 'k':

$$\gamma_k = Cov(y_t, y_{t+k}) = E([y_t-\mu][y_{t+k}-\mu])$$

Autocovariance function: $\{\gamma_k\} = \{\gamma_0, \gamma_1, \gamma_2, \dots\}$

Recall, autocorrelation at lag 'k' (for a stationary series)

$$\rho_k = \frac{E[(y_t-\mu)(y_{t+k}-\mu)]}{\sqrt{E[(y_t-\mu)^2]\cdot E[(y_{t+k}-\mu)^2]}}$$

$$= \frac{cov(y_t, y_{t+k})}{Var(y_t)} = \frac{\gamma_k}{\sigma^2} = \frac{\gamma_k}{\gamma_0}$$

Autocorrelation f^n (ACF) : $\{\rho_k\}_{k=0,1,2,\dots}$

Properties: (1.) $\rho_0 = 1$

(2.) Dimensionless / Independent of scale

(3.) Symmetric around zero.

$\Rightarrow \rho_k = \rho_{-k}$.

Sample: $c_k = \hat{\gamma}_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t-\bar{y})(y_{t+k}-\bar{y})$; $k=0,1,2,\dots$

ACF : $r_k = \hat{\rho}_k = \frac{c_k}{c_0}$, $k=0,1,2,\dots,K$.

Should have $T \geq 50$; K upto $T/4$.

Variogram : $G_k = \dfrac{Var\,(y_{t+k} - y_t)}{Var\,(y_{t+1} - y_t)}$; $k = 1, 2, \ldots$

$\Rightarrow$ Variance of differences b/w observations that are 'k' lag's apart and diff variance of the differences that are one time unit apart.

If y_t is stationary, $G_k = \dfrac{1-\rho_k}{1-\rho_1}$.

Stationary Series $\Rightarrow$ Variogram converges to $1/_{1-\rho_1}$ asymptotically

Non-Stationary Series $\Rightarrow$ Variogram increases monotonically!

Let $d_t^k = y_{t+k} - y_t$

$$\therefore \overline{d_t^k} = \frac{1}{T-k} \sum_{t=1}^{T-k} d_t^k$$

$$\therefore s_k^2 = \hat{\sigma}^2_{d_t^k} = \sum_{t=1}^{T-k} \frac{(d_t^k - \bar{d}^k)^2}{T-k-1}$$

$$\therefore \hat{G}_k = \frac{s_k^2}{s_1^2} \quad ; \quad k = 1, 2, \ldots$$

Data Transformations

- Used to stabilize the variance of data.
- Non Constant variance common in time series data.

Power transforms of data

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, & \lambda \neq 0 ; \\ \dot{y} \ln y, & \lambda = 0 \end{cases}$$

where $\dot{y} = \exp\left[\frac{1}{T} \sum_{t=1}^{T} \ln y \right]$ = geometric mean of

$\lambda = 0 \Rightarrow$ log transform

$\lambda = 0.5 \Rightarrow$ Square-root transform

$\lambda = -0.5 \Rightarrow$ Reciprocal square-root transform

$\lambda = -1 \Rightarrow$ Inverse transform.

Also, $\lim_{\lambda \to 0} \frac{y^{\lambda} - 1}{\lambda} = \ln y.$

Example: Log transform.

$\{y_t\}_{t=1}^{T}$. Let $x_t = \frac{y_t - y_{t-1}}{y_{t-1}} \times 100$

$= \%$ change in y_t

$$100[\ln(y_t) - \ln(y_{t-1})] = 100 \ln\left[\frac{y_t}{y_{t-1}}\right]$$

$$= 100 \ln\left[\frac{y_{t-1} + y_t - y_{t-1}}{y_{t-1}}\right]$$

$$= 100 \ln\left[1 + \frac{y_t - y_{t-1}}{y_{t-1}}\right]$$

$$= 100 \ln\left[1 + \frac{x_t}{100}\right]$$

$\left[\ln(1+z) \underset{z_t \to 0}{\cong} z_t\right]$ $\cong x_t$

Detrending data

Linear trend : $E(y_t) = \beta_0 + \beta_1 t$
Quadratic trend : $E(y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$
Exponential trend : $E(y_t) = \beta_0 e^{\beta_1 t}$

Differencing data

Let $x_t = y_t - y_{t-1} = \nabla y_t$

Let Backshift Operator, B be defined as:

$$B y_t \equiv y_{t-1} \quad / \quad (1-B) = \nabla$$

$$\therefore \; x_t = (1-B) y_t = \nabla y_t = y_t - y_{t-1}$$

Second difference: $x_t = \nabla^2 y_t = \nabla(\nabla y_t)$

$$= (1-B)^2 y_t$$
$$= (1-2B+B^2) y_t$$
$$= y_t - 2By_t + B^2 y_t$$
$$= y_t - 2y_{t-1} + y_{t-2}$$

$$\therefore \; x_t = \nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2}.$$

Defn:
$$B^d y_t = y_{t-d}$$
$$\nabla^d = (1-B)^d.$$

Why Second difference?

First difference: Accounts for trend (which changes the mean)

Second difference: Accounts for change in slo

Seasonal differencing:

$$\nabla_d y_t = (1-B^d) y_t = y_t - y_{t-d}$$

For monthly data: $\nabla_{12} y_t = (1-B^{12}) y_t$
$= y_t - y_{t-12}$

Steps in Time Series modeling & forecasting

I. Plot data, determine basic features:
- Trend
- Seasonality
- Outliers
- Change over time in the above.

II. Eliminate trend, eliminate seasonality, transform ..
Produce "Stationary Residuals"

III. Develop a forecasting model
- Use historical data to determine model fit.

IV. Validate model performance

V. Compare values of actual y_t and forecast values of untransformed data

VI. Construct prediction intervals

VII. Monitor forecasts : evaluate stream of forecast err

Evaluating forecast model

One-step ahead forecast errors: $e_t(1) = y_t - \hat{y}_t(t-1)$

Let $e_t(1)$, $t = 1, 2, \dots, n$ = `n' 1-step ahead forecast errors for n forecasts, then

Mean Error, $ME = \frac{1}{n} \sum_{t=1}^{n} e_t(1)$;

Mean absolute deviation (error), $MAD = \frac{1}{n} \sum_{t=1}^{n} |e_t(1)|$;

Mean squared error, $MSE = \frac{1}{n} \sum_{t=1}^{n} (e_t(1))^2$

- we want forecasts to be unbiased $\Rightarrow (ME) \approx 0$
- MAD and MSE measure the variability in forecast error
  In fact $MSE = \cancel{\sigma^2_{e(1)}}\ \hat{\sigma}^2_{e(1)}$. Why?

To get a unitless forecast error measurement, use

Relative forecast error (percent), or percent forecast error, $re_t(1) = \left(\frac{e_t(1)}{y_t}\right) * 100$; $y_t \neq 0$

Mean percent forecast error, $MPE = \frac{1}{n} \sum_{t=1}^{n} re_t(1)$.

Mean absolute percent forecast error, $MAPE = \frac{1}{n} \sum_{t=1}^{n} |re_t(1)|$

Finally, do a "normal probability plot"

If e_t is white noise, then sample autocorrelation coefficient at lag k (for large samples):

$$r_k \sim N(0, \frac{1}{T})$$

Calculate Z-stat i.e. $\frac{y - \bar{y}}{s.e.}$

$$Z_0 = \frac{r_k}{\sqrt{1/T}} = r_k \sqrt{T}$$

Check to see if $|Z\text{-stat}| > Z_{\alpha/2}$, say $Z_{0.025} = 1.96$
[Individual autocorrelation coeff] $Z_{0.005} = 2.58$.

If we want to evaluate a "set of autocorrelations" jointly to determine if they are white noise;

Let $Q_{BP} = T \sum_{k=1}^{K} r_k^2$ = Box-Pierce Statistic

(Recall, if $x \sim N$, then $x^2 \sim \chi^2$).

$\therefore Q_{BP} \sim \chi^2(K)$; H_0 : r is white noise

In small samples, use Ljung-Box goodness-of-fit stat

$$Q_{LB} = T(T+2) \sum_{k=1}^{K} \frac{r_k^2}{T-k} \sim \chi^2(k)$$

Model Selection

- Avoid overfitting ; prefer parsimony.
- Use data splitting to produce "out-of-sample" forecast errors. ~~Standard~~
- Minimise ~~wta~~ MSE of out-of-sample forecast errors
- This is called "cross-validation".

$$\text{MSE of residuals}, \; s^2 = \frac{\sum_1^T e_t^2}{T-p}$$

T = periods of data used to fit the model

p = # parameters

e_t = residual from model fitting in period t.

$$R^2 = 1 - \sum_{t=1}^{T} e_t^2 \Big/ \sum_{t=1}^{T} (y_t - \bar{y})^2 = 1 - \frac{RSS}{TSS}$$

Note: A large value of R^2 does NOT ensure that out-of-sample one-step-ahead forecast errors will be small.

$$R^2_{adj} = 1 - \frac{RSS/T-p}{TSS/T-1} = 1 - \frac{\sum_1^T e_t^2 / T-p}{\sum_1^T (y_t-\bar{y})^2 / T-1} = 1 - \frac{s^2}{TSS/}$$

Information Criteria

$$AIC := \ln\left(\frac{\sum_1^T e_t^2}{T}\right) + \frac{2p}{T}$$

$$BIC/SBC := \ln\left(\frac{\sum_{t=1}^T e_t^2}{T}\right) + \frac{p \cdot \ln(T)}{T}$$

Both penalize the model for additional parameters

Consistent ~~parameter~~ criterion :

i.) If true model is present among those being considered, then it selects the true model with prob. $\to 1$ as $T \to \infty$

ii.) If true model is not among those under consideration, then it selects the best approximation with prob. $\to 1$ as $T \to \infty$.

R^2, R^2_{adj}, AIC are Inconsistent !

SBC / BIC is consistent ! Heavier "size" adjustment penalty.

For asymptotically efficient criterion, use AIC.

$$AIC_c = \ln\left(\frac{\sum_1^T e_t^2}{T}\right) + \frac{2T(p+1)}{T-p-2}$$

Monitoring a forecasting model

a.) Shewhart Control Charts :

a) Center line : zero(for unbiased forecast) or ME $\left(= \frac{\sum_1^n e_t(1)}{n} \right)$

b) Control limits : $\pm$ 3 s.d's of the center line

Define <u>moving range</u> as the absolute value of the difference b/w any two successive one-step ahead forecast errors ; i.e,

$$MR = \sum_{t=2}^{n} \left| e_t(1) - e_{t-1}(1) \right|$$

$$\hat{\sigma}_{e(1)} = \frac{0.8865 \, MR}{n-1} = \frac{0.8865 \sum_{t=2}^{n} \left| e_t(1) - e_{t-1}(1) \right|}{n-1}$$

$$\therefore \quad \hat{\sigma}_{e(1)} = 0.8865 \, \overline{MR}$$

Use this to construct Upper control limit and lower control limit.

b) CUSUM : Cumulative Sum Control Chart.

- effective at monitoring "level" changes in the monitored variable.

- Set target value, T_0 (either zero, or ME)
- Upper ~~(control)~~ CUSUM statistic :

$$C_t^+ = \max[0, e_t(1) - (T_0 + K) + C_{t-1}^+]$$

- Lower CUSUM statistic :

$$C_t^- = \min[0, e_t(1) - (T_0 - K) + C_{t-1}^-]$$

Here, $K = 0.5\,\sigma_{e(1)}$

If $C^+ > 5\,\sigma_{e(1)}$ or $C^- < -5\sigma_{e(1)}$

Note : Can use the MR method to calculate $\hat{\sigma}_{e(1)}$

c) EWMA: Exponentially Weighted Moving Average

$$\bar{e}_t(1) = \lambda e_t(1) + (1-\lambda)\bar{e}_{t-1}(1) \quad ; 0.05 < \lambda < 0$$

λ = smoothing constant

Note that this is a differential equation!

$\therefore$ If $\bar{e}_0(1) = 0$ (or = ME), then

$$\bar{e}_t(1) = \lambda e_t(1) + (1-\lambda)\bar{e}_{t-1}(1)$$
$$= \lambda e_t(1) + (1-\lambda)[\lambda e_{t-1}(1) + (1-\lambda)\bar{e}_{t-2}(1)]$$

Substituting recursively ...

$$= \lambda e_t(1) + \lambda(1-\lambda)e_{t-1}(1) + (1-\lambda)^2\bar{e}_{t-2}(1)$$
$$\vdots$$

For n forecast series,

$$\bar{e}_n(1) = \lambda \sum_{j=0}^{n-1}(1-\lambda)^j e_{T-j}(1) + (1-\lambda)^n \bar{e}_0(1)$$

Also, $\sigma_{\bar{e}_t(1)} = \sigma_{e(1)}\sqrt{\dfrac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2t}\right]}$

$\therefore$ for EWMA,

a.) Center line $= T$

b.) UCL $= T + 3\sigma_{\bar{e}_t(1)}$

c.) LCL $= T - 3\sigma_{\bar{e}_t(1)}$.

Also see Cumulative Error tracking signal (CETS)
Smoothed Error tracking signal (SETS).