Probability Theory

- Probability theory is essential to forecasting because we need a confidence bound
- Given CDF, to get inverse CDF (quantile function), make x the subject
  - $F(\alpha) = \frac{\alpha^2}{100}$
  - $F^{-1}(p) = \sqrt{100p}$
- Joint probability only tells association not causality
- Bayesian analysis
  - Conduct forecasting in low data situation
  - Use Bayesian prior to compensate
  - Results in a posterior probability distribution (can see the distribution if we have infinite number of scenarios)
  - Highest posterior density interval is the shortest interval containing a given portion of probability density
- As sample size increase Bayesian tends to frequentist

Hypothesis testing

- Simple returns are log normally distributed
  - $E((r-u)^2) = E[r^2 + u^2 - 2ru] = E[r^2] + E[u^2] - 2E[ru]$
- MEMORIZE
  - $f(r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2}\left(\frac{r-\mu}{\sigma}\right)^2\right)}$
  - $E[r] = \int_{-\infty}^{\infty} r f(r) dr$
  - $V[r] = \int_{-\infty}^{\infty} (r-\mu)^2 f(r) dr$
- Independence is a subset of covariance, if independent covariance = 0
- Correlation is necessary for causation but not sufficient
- $Cov[x, y] = E[(X - \mu_x)(Y - \mu_y)]$
- Estimator
  - Bias – difference between expected value and true parameter, 0 bias is unbiased estimator
  - Consistency – as N increase, expected value becomes closer to true parameter
    - Only applies to biased estimator

Statistical Distribution

- Not everything follows normal distribution
- It should have a pdf or pmf whos equation is known
- Valid PDF is when integrating f(x) across region = 1
- Bernoulli (single coin flip)
  - discrete
  - $\mu = p, \sigma^2 = p(1-p)$
- Binomial distribution
  - $pdf = \binom{N}{n} p^n (1-p)^{N-n}$
  - $\mu = np, \sigma^2 = np(1-p)$
  - If p is unknown, can use $\frac{n}{N}$
  - To show p is unbiased, prove that $E\left[\frac{n}{N}\right] = p$
    - $\int_{n=0}^{N} \frac{n}{N} f(x_i = n) dx = \sum_n^N \binom{N}{n} p^n (1-p)^{N-n}$
- Poisson distribution
  - Discrete
  - $pdf = \frac{\lambda^n}{n!} e^{-\lambda}$
  - $\mu = \lambda, \sigma^2 = \lambda$
  - Relation to binomial, replace p with $\frac{\lambda}{N}$, as sample size increases, binomial tends to poisson
  - Scale poisson with time λt
- Uniform distribution
  - Continuous
  - $\mu = \frac{a+b}{2}, \sigma^2 = \frac{(b-a)^2}{12}$
- Normal distribution
  - Continuous
  - Linear combination of independent normal random variable is also normal
- Chi-squared distribution
  - With k independent standard normal variables, sum of their squares is a chi square distribution
  - As k tends to infinity, chi-squared tends to normal
  - Variance is chi-squared distributed
  - $\mu = k, \sigma^2 = 2k$
- T distribution
  - As k (degree of freedom) approaches infinity, t distribution tends to normal, excess kurtosis decreases
- F distribution
  - Ratio of chi squared distributed variables
  - Order of the 2 degrees of freedom matters
  - As both variables increase to infinity, mean and mode converge to 1
  - Square of a t variable is an F distribution

  - Test if model 2 is better than model 1

Cross Sectional Estimation Frameworks

- Goodness of fit test (compare sum of squared errors with chi squared distribution)
- Always include constant, if not will bias slope coefficient increase t stat and larger slope coefficient
- OLS assumptions
  - Mean of error is 0 => unbiasedness
    - If != 0 can shift the line parallelly to force the mean error to be 0
  - Variance of error is constant σ_e^2 (homoscedastic)
    - Assign equal importance to all data points
    - Weighted least squares, weight is inversely proportional to variance so underweight outliers
    - Tests
      - Goldfeld-Quandt
        - Split sample into 2, run regression and compare variances, the null is both variance is equal (F distribution with larger variance as numerator)
        - Issue: how to split
      - White test
        - Predict variance of error using independent variable, variance of independent variable and pairs of independent variables
        - If all coefficients are 0, means variance cannot be predicted, good
        - Null is error is homoscedastic
        - $N\ obs * R^2 \sim \chi^2(n\ regressor)$
    - Assumption: Positive covariance, estimated standard error < True standard error => absolute value of estimated t-stat > True t-stat => probability of rejecting null > true correct probability of rejecting null => more likely to conclude coefficient is significant when it is noise
    - Mitigation
      - Use log
      - Use Whites heteroscedasticity consistent standard error estimates
      - Dummy variables
  - Error term follows normal with mean 0 and variance σ_e^2
  - No autocorrelation, covariance between error term is 0
    - Error term for one does not help in predicting error term of another
    - Information either in model or error, if error can predict, just increase all values by error term, put the information of the error in the model
    - Positive autocorrelation – cyclical residual plot over time
    - Negative autocorrelation – residuals cross x axis more frequently
    - TEST
      - Durbin-Watson test statistic for the residual of **lag 1** (F distribution)
        - If DW ~ 2, no autocorrelation
          - Close to 0 = positive autocorrelation
          - Close to 4 = negative autocorrelation
      - Breusch-Godfrey test for r order autocorrelation
        - Null is all orders are 0, if any one is not 0, test will fail
        - Estimate linear regression, take the residuals and regress on X plus lag residuals, take the R-squared ($\sim\chi^2$) and check p-value
    - Implication
      - Coefficient estimates still unbiased
      - Standard error inappropriate => wrong inferences
      - R-squared inflated
    - If autocorrelation, use ARIMAX

Estimation Practicum

- Gauss Markov Theorem states that among all linear and unbiased estimators, OLS has minimum variance and are efficient
- $X'y = X'X\beta + X'e, X'e = 0$
- $(X'X)^{-1}X'y = (X'X)^{-1}X'X\beta, making\ \beta\ the\ subject, \beta = (X'X)^{-1}X'y$
- $VAR(\beta) = \sigma_e^2 (X'X)^{-1}, distribution\ of\ variance\ depends\ on\ \sigma_e^2\ aka\ \chi^2\ with\ df\ \boldsymbol{N - K\ independent\ variables, incl. constant}$
- Compare 2 valid OLS
  - 2 models, M1 has all parameters of M2 + additional parameters
    - Check if r-squared between both models is statistically different
    - Use F-test to test all coefficients of a model
      - Restricted vs unrestricted
      - Restricted SSR > unrestricted, harder to explain due to restriction
      - F distribution with df (number of restrictions, n observation – total parameters estimated)
- CML (x-axis = sigma), SML(x-axis = beta)
- If CAPM is correct
  - OLS of excess return against excess market returns would have constant term of risk-free rate
  - OLS against market + other variables, other variables would be statistically insignificant
- Fixed effect
  - Dummy variable for each sector
  - y intercept is the average value of dependent variable
    - y-intercept for sector 1 = $\alpha + \beta_1$
    - y-intercept for sector 2 = $\alpha + \beta_2$
- Interaction variables
  - The unconditional slope coefficient is that of the control variable
  - Slope coefficient – incremental difference between the variable and the control variable
- Jarque-Bera stat
  - Null: regression residuals are normally distributed
- Skew and kurtosis
  - Consider outliers impact
- Handling outliers
  - Winsorize
  - Apply logarithmic to both independent and dependent variables
- Intra cluster correlation (eg. Observations within same industry, violates iid assumption)
  - Overstates t-stat
  - **Doesn't affect R-squared or coefficient or prediction of the model**
  - Only affects statistical inference of the model and confidence of prediction
  - Adjust by clustering
- Multicollinearity
  - Cond. No.
    - Higher, less invertible of the data matrix, caused by high collinearity
  - Non-intuitive estimates of coefficients
  - High R-squared but high standard errors
  - Regression becomes sensitive to small changes
  - Confidence intervals wide
  - Solutions
    - Drop 1 of the collinear variable
    - Transform the highly correlated variables into ratio
    - Collect more data
    - Lasso (can set x to 0) or ridge
- Output of logit model
  - Estimated coefficients are odds-ratio (p(1-p)), p is probability of 1

Time Series

- $Autocorrelation = \frac{COV(Y_t, Y_{t-k})}{\sqrt{VAR(Y_t)}\sqrt{VAR(Y_{t-k})}}$
- Acf trend, autocorrelation for small lags tend to be large and positive, slowly decrease as lags increase
- Acf can see seasonality, autocorrelations will be larger for seasonal lags
- Slow decay in ACF means not covariance stationary
- Stationarity
  - time series whose moments do not depend on time of observation
  - parameters of models remain constant
  - stationarity is a subset of covariance stationarity
  - test
    - unit root test (KPSS)
    - null: stationary, small p-val => differencing required
- Non-stationary
  - shocks do not die out
  - bias AR coefficients
  - confidence intervals not valid
  - spurious regression
  - induce stationarity
    - differencing
- Removing deterministic effects before modelling
  - calendar adjustments
  - population adjustment
  - inflation adjustment
- Log and sqrt transformation reduces variation (box cox transformation)
- Time series decomposition
  - split time series into individual components, model each before merging back
  - additive decomposition – variation in data doesn't change across time
  - moving average to detect trend, remove trend component, estimate seasonal component (average for each season), adjust to ensure sum to 0, replicate sequence across entire data and compute remainder

- Benchmark models
  - naïve model – last observed value as the forecasted value
  - seasonal naïve model – use last observed value of previous season
  - drift method – last value + average drift
  - average method – average of historical returns
- In sample performance evaluation
  - check if residual still contain time series information (Ljung-Box test, null: no time series information in residuals, $\sim\chi^2$)
    - if have, there will be correlation and non zero mean
- choice of out of sample period
- classical decomposition
  - suffers from end effects
  - trend estimates oversmooths rapid rise and falls
  - assumes seasonal patterns are constant overtime
  - not robust to outliers
  - alternative STL & X11
    - uses rolling window
- STL
  - Handle seasonal period
  - Control rate of change of seasonal component and smoothness
  - Can use on additive only, if data is multiplicative need to log or box cox

ARIMA

- AR – regression of variable against lag values
  - $y_t = c + \phi_1 y_{t-1} + \epsilon_t$
  - AR(1) Expectation = $\frac{c}{1-\phi}$, $|\phi| < 1$ else no finite unconditional mean
  - AR(1) Variance = $\frac{\sigma^2}{1-\phi^2}$, $|\phi| < 1$ else no finite unconditional variance
  - Autocorrelation function – correlation decays as lag grows
    - $Corr(Y_t, Y_{t\pm k}) = \lambda^{|k|}$
  - Hard to distinguish between AR(1) and AR(2) using ACF if coefficients are positive
    - Can use partial autocorrelation – correlation between y and lag after controlling for all intermediate lags
    - ACF exponentially decay, PACF of AR(p) = 0 after lag p
- I – if raw is not stationary, need first differences **before** AR MA
- MA – moving average
  - $y_t = c + \theta_1 \epsilon_{t-1}$
  - Capture short horizon correlation between observations
  - Error term can be correlated
  - Expectation = c
  - Variance = $\sigma^2(1+\theta_1^2)$
  - Autocovariance = $\theta_1\sigma^2$ between lag 1, else 0
  - Autocorrelation = $\frac{\theta_1\sigma^2}{\sigma^2(1+\theta_1^2)} = \frac{\theta_1}{(1+\theta_1^2)}$ between lag 1, else 0
  - ACF drops to 0, indicates MA(p), PACF decays exponentially
- ARMA
  - Stationarity depends solely on AR coefficients
  - ACF and PACF decay exponentially
  - **Slowly decaying ACF need differencing data not stationary**
- White noise
  - Expected value = 0, error cannot be predicted by past lags
  - Variance $= \sigma^2$
  - Cov = 0, behaviour fully describe by y and its lag
- Estimation
  - Information criteria, minimize AIC
- Constant terms meaning
  - C=0, d=0 => long term forecast 0
  - C=0, d=1 => long term forecast non-zero constant
  - C=0, d=2 => long term forecast straight line
  - C!=0, d=0 => long term forecast mean
  - C!=0, d=1 => long term forecast straightline
  - C!=0, d=2 => long term forecast quadratic

ARMIA-X

- All inputs should be stationary
  - Remove spurious regression
  - Else biasness in estimation of coefficients
  - If inputs are stationary then estimate a ARMA model not ARIMA
- 1st stage errors don't have to be stationary, 2nd stage errors must pass Ljung Box test

SARIMA

- (p, d, q) SEASONAL(P, D, Q)
  - Imagine seasonal lags without bars in between for <u>PDQ</u>
  - Look at each year by itself for <u>pdq</u>

Vector Autoregression and Vector error correction models

- Model selection
  - For K variables and p lags, number of coefficients = $K + pK^2$
  - Use BIC for evaluation
    - AIC tends to pick models with larger number of lags due to lower penalty factor
  - Modelling seasonally adjusted components with ETS, ARIMA, VAR
  - Use Ljung-Box test on residuals (remb null is no time series so need to fail to reject null)
  - Evaluate goodness of fit using out of sample
- Spurious regression – when y and t are non-stationary and have a time trend and both receive a positive shock at the same time
- Cointegration
  - Non-stationary time series but move together over time
  - Linear combination of the series will be stationary
  - Test
    - Engle-Granger
    - Check order of integration for both variables are the same using KPSS
      - If different order, cannot have cointegration
    - If cointegration exist, then $y_t = \beta_0 + \beta_1 x_{1,t} + \epsilon_t$
    - Residual is stationary, use ADF test, null = unit root in residuals, no cointegration
- Scenarios
  - If 2 variables stationary then use OLS if assumptions satisfied
  - If not stationary and not cointegrated => spurious regression, use first difference before modelling
  - If not stationary but cointegrated, OLS estimator is super consistent
- Error correction model for cointegrated variables
  - Model nonstationary directly without differencing
  - $\Delta y_t = \gamma + \beta_1 \Delta X_t + \alpha(Y_{t-1} - \beta_0 - \beta_1 X_{t-1}) + v_t$
    - Note $(Y_{t-1} - \beta_0 - \beta_1 X_{t-1})$ is the error term from prior period forecast with expectation of 0 hence a large value should revert back to 0 over time
- Building VECM
  - Estimate VAR(p) model for variables
  - Determine number of cointegrating vectors using maximum likelihood
    - Johansen procedure (null: there is n cointegrated variable, large t stat => no cointegration, if fail to reject null then there is cointegrations)
  - If 0 cointegrated vectors then no VECM go back to VAR
  - Make variables stationary
  - Run VARSelect for starting order of model
  - Estimate model based on VARSelect, run LjungBox test
  - If pass then finish, if fail keep increasing order of model until pass
  - If still fail consider if need to difference an additional time

Workflow

- Given new dataset, try OLS, ARIMA, ARIMA-X, VAR, VECM
- From each class, choose the best representive using their own metrics (R-squared, AICC, BIC)
- Evaluate out of sample RSME, MAPE

GARCH

- EWMA
  - Follows MA model with m lag
  - More towards volatility clustering
  - Unable to account for mean reversion
- GARCH(1,1)
  - ARMA(1,1) with constant OR ARIMAX with long run mean as constant
  - Unconditional stationary
    - Long term finite mean and variance
    - Unconditional mean is long run variance rate
  - Use Ljung Box test to make sure no more time series information else need more lags

EXAM

- AR(1) process unconditional mean
  - $\frac{c}{1-\phi}$
- Auto correlation
  - $\lambda^{|k|}$
- autoplot, monthdays, BoxCox, nsdiffs, diff (as many times as is necessary), ndiffs, diff (as many times as is necessary), tsdisplay, Arima, checkresiduals
- high kurtosis more observations in tails than normal, skew is on one side of tail, means outliers, how to control use log, hyperbolic tangent