4. Forecasting with ARMA models

MSE optimal forecasts

- Let $\hat{Y}_t$ denote a one-step ahead forecast given information $F_t$. For ARMA models, $F_t = Y_t, Y_{t-1}, \ldots$

- The optimal MSE forecast minimizes
  
  $$E(\hat{Y}_{t+1} - Y_{t+1})^2$$
• MSE is a convenient forecast evaluation that has nice theoretical properties.

• MSE is not always a good way to evaluate forecasts from an economic perspective.
  – It is symmetric. Do you care the same about underestimating the time to walk to the bus vs. overestimating?
  – In general, non-linear payoff functions are not optimized by MSE.

The optimal MSE forecast is the conditional expectation

• To see this, let \(g(X_t)\) denote any other forecast.

\[
E(Y_{t+1} - g(X_t))^2 = E[(Y_{t+1} - E(Y_{t+1} | X_t)) + (E(Y_{t+1} | X_t) - g(X_t))^2
\]

\[
= E(Y_{t+1} - E(Y_{t+1} | X_t))^2 + 2E[(Y_{t+1} - E(Y_{t+1} | X_t))(E(Y_{t+1} | X_t) - g(X_t))]
+ E(E(Y_{t+1} | X_t) - g(X_t))^2
\]

• The middle (cross product) term has expectation zero by the Law of Iterated Expectations (LIE) \(E(Y) = E(E(Y | X))\)

\[
E\left[\left\{ E_{t+1} - E(Y_{t+1} | X_t) \right\}(E(Y_{t+1} | X_t) - g(X_t)) \mid X_t \right] = E\left[0\left((E(Y_{t+1} | X_t) - g(X_t))\right)\right] = 0
\]
• So the middle term vanishes and we are left with:

\[ E(Y_{t+1} - g(X_t))^2 = E(Y_{t+1} - E(Y_{t+1} | X_t))^2 + E(E(Y_{t+1} | X_t) - g(X_t))^2 \]

• The smallest the second term can be is zero and it is zero only when \( E(Y_{t+1} | X_t) = g(X_t) \)

• So the optimal MSE forecast is the conditional expectation.

Restricting our attention to linear models, a similar result holds for Least Squares estimators

• Let \( E(Y_{t+1} - X_t \beta)X_t = 0 \) so that \( X_t \beta \) is the linear projection of \( Y_{t+1} \) on \( X_t \), or the least squares solution.

• A nearly identical proof shows that the linear projection provides the MSE optimal forecast in the class of linear predictors.

• Let \( X_t \beta \) denote the least squares solution and let \( X_t g \) denote any other linear combination.
\[ E(Y_{i1} - X, g) = E((Y_{i1} - X, \beta) + (X, \beta - X, g))^2 \]
\[ = E(Y_{i1} - X, \beta)^2 + 2 \cdot E((Y_{i1} - X, \beta)(X, \beta - X, g)) + E(X, \beta - X, g)^2 \]

- The middle term again has expectation zero.

\[ E[(Y_{i1} - X, \beta)(X, \beta - X, g)] = E[(Y_{i1} - X, \beta)X, (\beta - g)] \]
\[ E[(Y_{i1} - X, \beta)X, (\beta - g)] = 0(\beta - g) = 0 \]

- The resulting expression

\[ E(Y_{i1} - g(X_i))^2 = E(Y_{i1} - X, \beta)^2 + E(X, \beta - X, g)^2 \]

is minimized when \( X_i, \beta = X_i g \)

- So if the true conditional expectation is linear then the linear model is the optimal MSE.
- If the true model is non-linear, then the linear model is the best linear predictor.
Forecasting ARMA models

• Consider the ARMA model:
  \[ \phi(L) Y_t = \theta(L) \epsilon_t \]
  
  where \( \phi(L) = (1 - \beta_1 L - \beta_2 L^2 - \ldots - \beta_p L^p) \)
  
  and \( \theta(L) = (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q) \)

• If the model is weakly stationary we get:
  \[ \phi(L)^{-1} \phi(L) Y_t = \phi(L)^{-1} \theta(L) \epsilon_t \]
  \[ Y_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} \]

  where \( \psi_0 = 1 \)

Multi-step ahead forecast

• From \( Y_{t+k} = \sum_{j=0}^{\infty} \psi_j \epsilon_{t+k-j} \), the multi-step ahead forecast is easily constructed.

\[
E(Y_{t+k} \mid F_t) = Y_{t+k}^k = \left( \sum_{j=0}^{\infty} \psi_j \epsilon_{t+k-j} \mid F_t \right)
\]

\[
= E\left( \sum_{j=0}^{k-1} \psi_j \epsilon_{t+k-j} \mid F_t \right) + \sum_{j=k}^{\infty} \psi_j \epsilon_{t+k-j} = \sum_{j=k}^{\infty} \psi_j \epsilon_{t+k-j}
\]
What is the k-step ahead forecast error?

• The forecast error is defined as the difference between the outcome and it’s forecast is denoted by: \( e^k_t = Y_{t+k} - Y^k_t \)

• Notice:
  \[
  Y_{t+k} = \sum_{j=0}^{\infty} \psi_j e_{t+k-j} = \sum_{j=0}^{k-1} \psi_j e_{t+k-j} + E(Y_{t+k} \mid F_t)
  \]

• So \( e^k_t = Y_{t+k} - E(Y_{t+k} \mid F_t) = \sum_{j=0}^{k-1} \psi_j e_{t+k-j} \)

• The k-step ahead forecast error variance is then given by:
  \[
  Var(e^k_t) = Var\left(\sum_{j=0}^{k-1} \psi_j e_{t+k-j}\right) = \sigma^2 \sum_{j=0}^{k-1} \psi_j^2
  \]

• If the errors are Gaussian then the 95% prediction interval is given by:
  \[
  Y^t + 2\sigma \sqrt{\sum_{j=0}^{k-1} \psi_j^2}
  \]
Example

• Consider the AR(1) model

\[(1 - \beta L) Y_t = \beta_0 + \epsilon_t\]

• Pre-multiply both sides of \(Y_{t+k}\) by:

\[\left(1 + \beta L + \beta^2 L^2 + \ldots + \beta^{k-1} L^{k-1}\right)\]

yielding

\[Y_{t+k} = \beta_0 \sum_{j=0}^{k-1} \beta^j + \sum_{j=0}^{k-1} \beta^j \epsilon_{t+k-j}\]

or

\[Y_{t+k} = \beta^k Y_t + \beta_0 \sum_{j=0}^{k-1} \beta^j + \sum_{j=0}^{k-1} \beta^j \epsilon_{t+k-j}\]

• The k-step ahead forecast is:

\[E(Y_{t+k} \mid F_t) = \beta^k Y_t + E\left(\sum_{j=0}^{k-1} \beta^j \epsilon_{t+k-j}\right) = \beta_0 \sum_{j=0}^{k-1} \beta^j + \beta^k Y_t\]

• The forecast error is \(\sum_{j=0}^{k-1} \beta^j \epsilon_{t+k-j}\)

• The forecast error variance is:

\[Var\left(\sum_{j=0}^{k-1} \beta^j \epsilon_{t-j}\right) = \sigma^2 \sum_{j=0}^{k-1} \beta^{2j}\]
• Notice that we could pre-multiply by $(1 + \beta L + \beta^2 L^2 + \ldots + \beta^{k-1} L^{k-1})$ regardless of whether $\beta$ is less than one (we didn’t actually use the inverse operator).

• If $|\beta|<1$ the point forecast is

$$E(Y_{t+k} | F_t) = \beta_0 \sum_{j=0}^{k-1} \beta^j Y_t + \beta^k Y_t = \beta_0 \frac{1 - \beta^k}{1 - \beta} + \beta^k Y_t$$

$$= (1 - \beta^k) \mu + \beta^k Y_t$$

and the forecast error variance is:

$$Var(e_t^k) = Var\left( \sum_{j=0}^{k-1} \beta^j e_{t-j} \right) = \sigma^2 \sum_{j=0}^{k-1} \beta^{2j} = \sigma^2 \frac{1 - \beta^{2k}}{1 - \beta^2}$$
If $\beta = 1$, the point forecast is given by:

$$E(Y_{t+k} \mid F_t) = \beta_0 \sum_{j=0}^{k-1} \beta^j + \beta^k Y_t = \beta_0 k + Y_t$$

The forecast error variance is given by:

$$\text{Var}(e_t^k) = \text{Var}\left(\sum_{j=0}^{k-1} \beta^j e_{t+k-j}\right) = \sigma^2 \sum_{j=0}^{k-1} \beta^{2j} = \sigma^2 k$$

A simple algorithm for multi-step ahead forecasts

1 step:

$$E(Y_{t+1} \mid F_t) = \sum_{j=1}^{k} \beta_j Y_{t+1-j} + \sum_{i=1}^{k} \theta_i e_{t+1-j}$$

2 step

$$E(Y_{t+2} \mid F_t) = \beta_1 E(Y_{t+1} \mid F_t) + \sum_{j=2}^{k} \beta_j Y_{t+2-j} + \theta_1 E(e_{t+1} \mid F_t) + \sum_{j=1}^{k} \theta_j e_{t+2-j}$$

= $\beta_1 E(Y_{t+1} \mid F_t) + \sum_{j=2}^{k} \beta_j Y_{t+2-j} + \sum_{i=1}^{k} \theta_i e_{t+2-j}$

3 step

$$E(Y_{t+3} \mid F_t) = \beta_1 E(Y_{t+2} \mid F_t) + \beta_2 E(Y_{t+1} \mid F_t) + \sum_{j=3}^{k} \beta_j Y_{t+3-j} + \sum_{j=1}^{k} \theta_j e_{t+3-j}$$
• So once you have the one-step you can get the two step, once you have the two-step, you can get the three-step and so on.
• Plug in conditional expected values for future values of \( Y \) and zero for the expectation of future values of \( \varepsilon \). Plug in know past values for values of \( Y \) and \( \varepsilon \) at or before time \( t \).

5. Maximum Likelihood Estimation

• Let’s start with a simple example.
• Suppose that we take an iid sample of size 10 of whether a voter will vote for candidate A.
• Suppose the data look like this:
  \[
  1 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 1
  \]
  where a 1 denotes “favor candidate A”.
• Let \( p \) denote the true fraction of voters that favor candidate A. \( p \) is the parameter we want to estimate.
• If we knew p, how likely would it be to observe this sequence of data?
• Well, they are all independent and identically distributed.
  1 0 1 1 1 0 0 1 0 1
  The joint probability that \( x_1=1, x_2=0, x_3=1 \ldots \) is given by:
  \[
  p(1-p)ppp(1-p)(1-p)p(1-p)p=p^6(1-p)^4
  \]

• More generally, for a sample of size n we can write the probability of observing the sample as
  \[
  p^n (1 – p)^{n_0}
  \]
• This is simply the joint probability of n outcomes.
For a given value of $p$, $L(data | p) = p^n (1 - p)^{n_0}$ tells us how likely it is to observe this particular data set. We therefore call this the likelihood function.

The goal of maximum likelihood is to find the parameter value $p$ that maximizes the likelihood function.

In doing so we are finding the parameter value that makes it most likely that we observe our given sample.

The log function is a monotonically increasing function. This means that the value of $p$ that maximizes the likelihood function $L$ is the same as the value of $p$ that maximizes the log of the likelihood function.

$$L(data | p) = \ln(L) = n_1 \ln(p) + n_0 \ln(1 - p)$$
• We can maximize the likelihood by taking the derivative and setting it to zero.
• The p that maximizes the log likelihood function will be our estimate of p so we denote it by \( \hat{p} \).

Taking the derivative and setting to zero yields:

\[
\frac{dL(p \mid data)}{dp} = \frac{n_1}{p} - \frac{n_0}{(1 - p)} = 0
\]

• The solution is:

\[
\hat{p} = \frac{n_1}{(n_0 + n_1)} = \frac{n_1}{n}
\]
• For continuous random variables the density function describes probabilities.
• Let $f(\cdot | \theta)$ denote a density function that depends on a possible vector of parameters (i.e. for the normal it depends on the mean and variance).

• If we take an iid sample of a continuous random variable then the likelihood of observing the sample is given by the product of the marginal densities.

$$f(y_1, y_2, \ldots, y_n | \theta) = f(y_1 | \theta) f(y_2 | \theta) \cdots f(y_n | \theta)$$
• It is convenient to take the log of the joint density function in which case we get:

\[
\ln f(y_1, y_2, \ldots, y_n | \theta) = \sum_{i=1}^{n} \ln f(y_i | \theta)
\]

• As before, the value of \( \theta \) that maximizes the function \( f \) will also maximize \( \ln(f) \) and is called the Maximum Likelihood Estimate \( \hat{\theta} \).

• Example: iid Normal likelihood function. (on white board)
How do we set up the likelihood for dependent time series data?

- The factorization of the likelihood into the product of marginal densities relied on an iid sample of data which is clearly not true for dependent time series data.

- Fact: without any loss of generality, the joint density function $f(y_T, y_{T-1}, \cdots, y_1 | \theta)$ can always be expressed as:

$$f(y_T, y_{T-1}, \cdots, y_1 | \theta) = f(y_T | y_{T-1}, y_{T-2}, \cdots, y_1; \theta) f(y_{T-1} | y_{T-2}, y_{T-3}, \cdots, y_1; \theta)$$

$$f(y_{T-2} | y_{T-3}, y_{T-4}, \cdots, y_1; \theta) \cdots f(y_2 | y_1; \theta) f(y_1; \theta)$$

- Taking logs of the joint density then gives:

$$\ln f(y_1, y_2, \cdots, y_n | \theta) = \sum_{i=1}^{n} \ln f(y_i | y_{i-1}, y_{i-2}, \cdots, y_1; \theta)$$

- The value of $\theta$ that maximizes the likelihood is called the Maximum Likelihood Estimate (MLE) of $\theta$. 
• All the models that considered in class are different models for this conditional distribution.
• The result says that if we can write the conditional density of \( f(y_{t+1} | F_t) \) we can construct the likelihood.
• But this is exactly how we specified the models discussed so far. We defined the dynamics conditional of \( F_t \).

• AR(1) model likelihood under Gaussian errors.
  (on white board)
Variance of maximum likelihood parameter estimates

• The value of $\theta$ that maximizes the likelihood is our estimate, but how accurate is the estimate?
• If the assumed distribution is correct then the variance covariance matrix of the parameter estimates is given by inverse of the information matrix denoted by $I^{-1}$.

Constructing the information matrix

• Write the log likelihood function as: $L = \sum_{t=1}^{T} l_t$ where $l_t = \ln f(y_t | y_{t-1}, y_{t-2}, ..., y_1; \theta)$
• There are two ways to estimate the information matrix:

1) $\hat{I}_{2d} = -\frac{1}{T} \frac{d^2 L}{d\theta d\theta'}|_{\hat{\theta}} = -\frac{1}{T} \sum_{t=1}^{T} \frac{d^2 l_t}{d\theta d\theta'}|_{\hat{\theta}}$

2) $\hat{I}_{op} = \frac{1}{T} \sum_{t=1}^{T} \frac{d}{d\theta} \frac{d l_t}{d\theta'}|_{\hat{\theta}}$
• The if $\theta_0$ is the true parameter value, the distribution of the parameter estimates is given by:

$$\hat{\theta} \sim N\left(\theta_0, \frac{1}{T} \hat{I}^{-1}\right)$$

• Where $\hat{I}$ is one of the two estimates $\hat{I}_{2d}$ or $\hat{I}_{op}$.

• When the model is correctly specified the two estimate of the information matrix are the same in large samples. In small samples they will differ a little although neither is preferred.

• There are several ways the model can be wrong.
  – First, we could misspecify the dynamics (i.e. have the wrong AR model)
  – Second, we could use the wrong shape for the conditional distribution (i.e. use a Normal when we should have used a t-distribution)
Getting the standard errors right

• Interestingly, if we have the dynamics right but falsely assume Normality, the parameter estimates are still consistent under very general conditions.
• The standard errors will be wrong, however.
• The good news is that we know how to fix them.

White Robust Standard Errors

• If the dynamics are correctly specified, but the assumed distribution is wrong the estimates are consistent and the variance covariance matrix can be estimated by:

\[ \hat{\theta} \sim N\left(\theta_0, \frac{1}{T} \hat{I}_d^{-1} \hat{I}_{op} \hat{I}_d^{-1}\right) \]