1 Checking an Autoregression

Last week, we used the ACF and PACF of y_t to suggest a model. Now we use the ACF and PACF of its **residuals** to check what the model missed.

Step 1: calculate the residuals

For an estimated AR(p),

$$\widehat{y}_t = \widehat{m} + \sum_{j=1}^{p} \widehat{\phi}_j y_{t-j}, \qquad e_t = y_t - \widehat{y}_t.$$

The residual e_t is the actual value minus the fitted value.

Step 2: examine the residual ACF and PACF

ACF at lag k measures the correlation between the current residual e_t and the residual k periods ago, e_{t-k}:

$$\rho_k^e = \text{Corr}(e_t, e_{t-k}).$$

PACF at lag k measures the correlation between those same two residuals after controlling for $e_{t-1}, \ldots, e_{t-k+1}$. For example, at lag 2, we control for e_{t-1}.

A regression-based estimate comes from

$$e_t = a + b_1 e_{t-1} + \cdots + b_k e_{t-k} + u_t.$$

For each k, include the first k lags. The last coefficient, $\widehat{b}_k$, estimates the lag-k PACF.

Step 3: interpret the results

If the model captures the linear time dependence, both measures should be **close to zero** at nonzero lags. Small positive or negative sample values are normal.

Value	Residual ACF at lag k	Residual PACF at lag k
Positive	When e_{t-k} is higher, e_t tends to be higher.	After controlling for the residuals in between, when e_{t-k} is higher, e_t tends to be higher.
Negative	When e_{t-k} is higher, e_t tends to be lower.	After controlling for the residuals in between, when e_{t-k} is higher, e_t tends to be lower.
Close to zero	There is little linear relationship between e_{t-k} and e_t.	After controlling for the residuals in between, e_{t-k} adds little linear information about e_t.

Example: residual ACF and PACF

The table reports sample ACF and PACF values calculated from each fitted model's residuals.

Model	Measure	Lag 1	Lag 2	Lag 3
A	Residual ACF	0.55	0.38	0.26
A	Residual PACF	0.55	0.11	0.02
B	Residual ACF	0.03	−0.04	0.02
B	Residual PACF	0.03	−0.04	0.02

Model A:

- **ACF:** the lag-1 value of 0.55 means a higher residual last period tends to be followed by a higher residual this period. The positive values at lags 2 and 3 show that older residuals are also correlated with the current residual.
- **PACF:** at lag 2, the value falls from an ACF of 0.38 to a PACF of 0.11. After accounting for last period's residual, the residual two periods ago has a much weaker relationship with the current residual. At lag 3, little relationship remains after accounting for the previous two residuals.

The main remaining pattern is at lag 1: past errors still help predict current errors. This suggests that Model A has not fully captured the series' time dependence.

Model B:

- **ACF:** all three values are close to zero. Residuals from one, two, or three periods ago have little linear relationship with the current residual.
- **PACF:** the values are also close to zero. After accounting for the more recent residuals, older residuals add little linear information about the current residual.

There is little evidence of predictable errors at these lags. Model B's residuals therefore appear more consistent with white noise.

The plots below show the same ACF and PACF values. Look for spikes clearly away from zero and patterns across lags. The shaded bands provide an approximate visual reference.

2 Multistep Forecasting

Let T be the last observed date. Consider an AR(1):

$$y_t = m + \phi y_{t-1} + \varepsilon_t.$$

Recursive forecasts: use the same model repeatedly

Estimate the AR(1), then substitute forecasts for future values:

$$\widehat{y}_{T+1|T} = \widehat{m} + \widehat{\phi} y_T,$$
$$\widehat{y}_{T+2|T} = \widehat{m} + \widehat{\phi} \widehat{y}_{T+1|T}.$$

Example

Suppose $\widehat{m} = 1$, $\widehat{\phi} = 0.5$, and the latest observed value is $y_T = 6$.

1. **Forecast one period ahead.** Use the observed value $y_T = 6$:

$$\widehat{y}_{T+1|T} = 1 + 0.5(6) = 4.$$

2. **Forecast two periods ahead.** We do not yet observe y_{T+1}, so replace it with our forecast of 4:

$$\widehat{y}_{T+2|T} = 1 + 0.5(4) = 3.$$

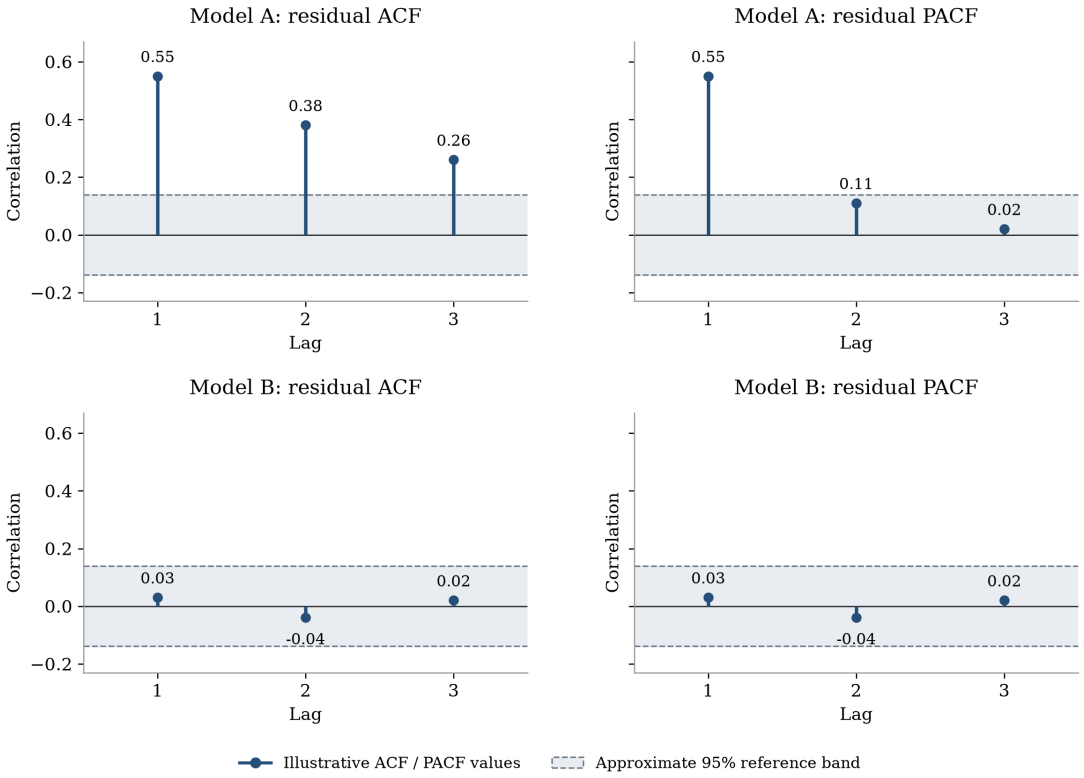


Figure 1: Residual ACF and PACF for Models A and B. Model A shows remaining correlation, especially at lag 1. Model B's values are close to zero at all three lags.

Both forecasts are made at time T. The second step uses a forecast, not a new observation.

Direct forecasts: estimate a model for each horizon

For horizon h, estimate

$$y_{t+h} = \alpha_h + \beta_h y_t + u_{t+h}.$$

Then forecast directly from the observed y_T:

$$\widehat{y}^D_{T+h|T} = \widehat{\alpha}_h + \widehat{\beta}_h y_T.$$

Each horizon has its own coefficients. Only use estimation observations with $t + h \leq T$.

Example

Suppose the latest observed value is again $y_T = 6$.

1. **Forecast one period ahead.** Regress y_{t+1} on a constant and y_t. Suppose the estimates are $\widehat{\alpha}_1 = 1$ and $\widehat{\beta}_1 = 0.5$:
$$\widehat{y}^D_{T+1|T} = 1 + 0.5(6) = 4.$$

2. **Forecast two periods ahead.** Run a separate regression of y_{t+2} on a constant and y_t. Suppose the estimates are $\widehat{\alpha}_2 = 1.2$ and $\widehat{\beta}_2 = 0.3$:
$$\widehat{y}^D_{T+2|T} = 1.2 + 0.3(6) = 3.$$

Both forecasts use the observed value $y_T = 6$. The two-step forecast does not use the one-step forecast of 4. The coefficients above happen to give the same forecasts as the recursive example; this need not occur in practice.

Neither method is always more accurate. Compare out-of-sample MSPE using the same targets and horizons.

Mean reversion: how quickly does a forecast approach the mean?

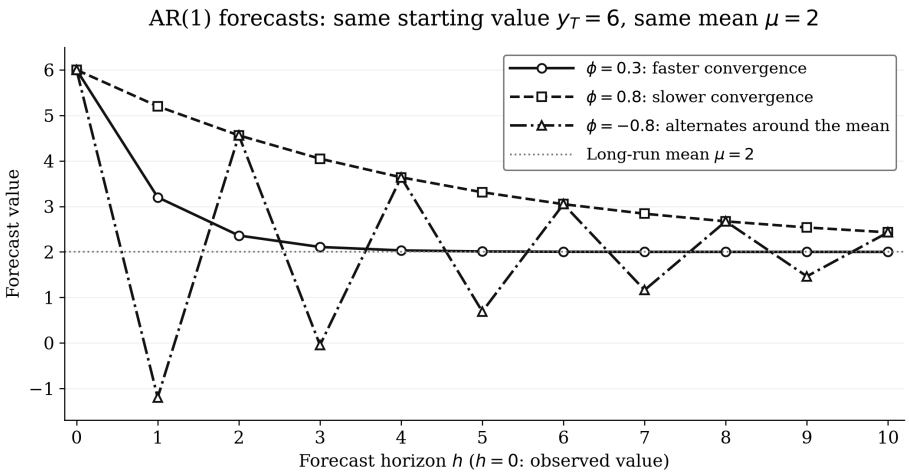


Figure 2: AR(1) forecasts starting from $y_T = 6$ with long-run mean $\mu = 2$. Smaller $|\phi|$ gives faster convergence. A negative ϕ makes forecasts alternate around the mean.

For a stationary AR(1), with known parameters and future shocks having conditional mean zero,

$$\mu = \frac{m}{1-\phi}, \qquad \widehat{y}_{T+h|T} - \mu = \phi^h (y_T - \mu).$$

- Smaller $|\phi|$: faster convergence to the mean.
- Negative ϕ: forecasts alternate around the mean.
- For AR(p), the coefficients jointly determine the speed; the order p alone does not.

3 Outliers and Time Alignment

An AR(1) uses last month's value to explain this month's value:

$$y_t = m + \phi y_{t-1} + \varepsilon_t.$$

Dropping regression rows does not change the dates

Suppose March's value of 40 is an outlier. First construct the lag column using the original dates. Then exclude any regression row where March's value appears as either the outcome or the regressor:

Month	Outcome: y_t	Regressor: y_{t-1}	Use in regression?
February	11	10	Keep
March	**40**	11	Drop: outlier as outcome
April	12	**40**	Drop: outlier as regressor
May	13	12	Keep

After excluding March and April, the remaining rows are:

Month	Outcome: y_t	Regressor: y_{t-1}
February	11	10
May	13	12

May still uses April's value of 12, **not** February's value of 11. We drop April's regression row, not its original value.

Rule: construct lags using the original dates, then exclude the chosen regression rows. Do not recalculate lags from the shortened list of rows.

Exercises

Exercise 1: Checking residuals

Two AR models are fitted to the same series. Their residual ACF and PACF are:

Model	Measure	Lag 1	Lag 2	Lag 3
A	ACF	0.46	0.31	0.20
A	PACF	0.46	0.13	0.05
B	ACF	0.04	−0.06	0.03
B	PACF	0.04	−0.06	0.04

For this exercise, use approximate pointwise reference bands of $[-0.15, 0.15]$.

1. Which model leaves clearer evidence of time dependence in its residuals? Explain using the table.

2. A student says: "Model B's residual correlations are all inside the bands, so Model B must be the true model." Is this conclusion justified? Explain briefly.

Exercise 2: Multistep forecasts

An estimated AR(1) has $\widehat{m} = 2$ and $\widehat{\phi} = 0.5$. The latest observation is $y_T = 12$.

A separately estimated two-step direct regression is

$$y_{t+2} = \alpha_2 + \beta_2 y_t + u_{t+2}, \qquad \widehat{\alpha}_2 = 2.5, \quad \widehat{\beta}_2 = 0.3.$$

1. Calculate the one-step and two-step recursive forecasts, and the two-step direct forecast.

2. Calculate the fitted AR(1)'s long-run mean. Do its recursive forecasts move toward that mean? Can the forecast values alone tell us whether the direct method is more accurate? Explain briefly.

Exercise 3: Keeping lags aligned

Consider an AR(2) fitted to the following monthly data:

Month	Jan	Feb	Mar	Apr	May	Jun
y_t	8	9	35	10	11	12

March is flagged as an outlier. Consider only regression rows from March through June.

1. Complete the lag columns using the original dates.

Month	y_t	y_{t-1}	y_{t-2}
Mar	35		
Apr	10		
May	11		
Jun	12		

2. Which regression rows must be excluded under each rule?

   - Exclude March's value only as an outcome.
   - Exclude March's value as an outcome or as either lagged regressor.

Answers

Exercise 1: Checking residuals

Two AR models are fitted to the same series. Their residual ACF and PACF are:

Model	Measure	Lag 1	Lag 2	Lag 3
A	ACF	0.46	0.31	0.20
A	PACF	0.46	0.13	0.05
B	ACF	0.04	−0.06	0.03
B	PACF	0.04	−0.06	0.04

For this exercise, use approximate pointwise reference bands of $[-0.15, 0.15]$.

1. Which model leaves clearer evidence of time dependence in its residuals? Explain using the table.

   Model A. Its residual ACF values at lags 1, 2, and 3 all exceed the upper reference bound of 0.15. Its residual PACF at lag 1 also exceeds the bound.

   The smaller PACF values at lags 2 and 3 suggest that older residuals add relatively little information after controlling for the more recent residuals. However, the strong lag-1 relationship remains.

   Model B's residual ACF and PACF values are all close to zero and inside the reference bands.

2. A student says: "Model B's residual correlations are all inside the bands, so Model B must be the true model." Is this conclusion justified? Explain briefly.

   No. The results show little evidence of remaining linear time dependence at the three lags examined. They do not prove that Model B is the true model. Relationships at other lags or other forms of dependence may still be present.

Exercise 2: Multistep forecasts

An estimated AR(1) has $\widehat{m} = 2$ and $\widehat{\phi} = 0.5$. The latest observation is $y_T = 12$.

A separately estimated two-step direct regression is

$$y_{t+2} = \alpha_2 + \beta_2 y_t + u_{t+2}, \qquad \widehat{\alpha}_2 = 2.5, \quad \widehat{\beta}_2 = 0.3.$$

1. Calculate the one-step and two-step recursive forecasts, and the two-step direct forecast.

   The one-step recursive forecast uses $y_T = 12$:

   $$\widehat{y}_{T+1|T} = 2 + 0.5(12) = \boxed{8}.$$

   The two-step recursive forecast substitutes the one-step forecast of 8:

   $$\widehat{y}_{T+2|T} = 2 + 0.5(8) = \boxed{6}.$$

   The two-step direct forecast uses $y_T = 12$ and the separately estimated coefficients:

   $$\widehat{y}^{D}_{T+2|T} = 2.5 + 0.3(12) = \boxed{6.1}.$$

2. Calculate the fitted AR(1)'s long-run mean. Do its recursive forecasts move toward that mean? Can the forecast values alone tell us whether the direct method is more accurate? Explain briefly.

   The fitted long-run mean is

   $$\widehat{\mu} = \frac{\widehat{m}}{1-\widehat{\phi}} = \frac{2}{1-0.5} = \boxed{4}.$$

   Yes. Starting from 12, the recursive forecasts are 8 and 6, moving toward 4. Their distances from the mean shrink:

   $$12-4=8 \quad \longrightarrow \quad 8-4=4 \quad \longrightarrow \quad 6-4=2.$$

   The forecast values alone do not tell us which method is more accurate. We need actual outcomes to calculate forecast errors. To compare performance, evaluate multiple forecasts of the same dates and horizons, using only data available when each forecast was made.

Exercise 3: Keeping lags aligned

Consider an AR(2) fitted to the following monthly data:

Month	Jan	Feb	Mar	Apr	May	Jun
y_t	8	9	35	10	11	12

March is flagged as an outlier. Consider only regression rows from March through June.

1. Complete the lag columns using the original dates.

Month	y_t	y_{t-1}	y_{t-2}
Mar	35	9	8
Apr	10	35	9
May	11	10	35
Jun	12	11	10

2. Which regression rows must be excluded under each rule?

   - Exclude March's value only as an outcome.
   - Exclude March's value as an outcome or as either lagged regressor.

   Exclude only as an outcome: drop March. Keep April, May, and June.

   Exclude as an outcome or either regressor: drop March, April, and May:

   - March contains 35 as the outcome.
   - April contains 35 as the first lag.
   - May contains 35 as the second lag.

   Keep June. Its regression row is

   $$y_t = 12, \qquad y_{t-1} = 11, \qquad y_{t-2} = 10.$$

   June still uses May's and April's original values, even though their regression rows are excluded. Dropping rows does not change the calendar lags.