Session 9

Time Series Regressions

Anthony Tay

This Version: 04 Oct 2025

Agenda

Previously:

- how to describe single time series
  - Deterministic trend, stochastic trend
  - Seasonal dummies (deterministic seasonality)
  - Covariance-Stationary AR(1) for cycles

Agenda

This session:

Time Series Regressions, e.g.,

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

Issues include

- Generally cannot assume data are iid
- Have to account for trend, seasonality, cycles
- Previously assumed ϵ_t iid, perhaps no longer appropriate
- May have to consider dynamic specification

Agenda

We consider regressions with

- Covariance-stationary weakly-dependent time series
- "Trend stationary" time series
- "Difference-stationary time series

Agenda

```
library(tidyverse)  # for data management
library(patchwork)  # for plotting
library(fpp3)       # a suite of packages for forecasting
library(ggfortify)  # for plotting
library(readxl)     # used to read in excel files

ts01 <- read_excel("data\\ts_01.xlsx")
ts01 <- ts01 %>%
  mutate(
    DATE=yearmonth(DATE),
    LN_ELEC_SG = log(ELEC_GEN_SG),
    LN_IP_SG = log(IP_SG),
    LN_IP_SG_1 = lag(LN_IP_SG, 1),
    D_LN_IP_SG = LN_IP_SG - LN_IP_SG_1
    ) %>%
  as_tsibble(index=DATE)
theme0 <- theme_minimal() +
  theme(text=element_text(size=8), axis.text.x = element_text(angle=45, hjust=1))
theme1 <- theme_minimal() + theme(text=element_text(size=6), aspect.ratio=1)
```

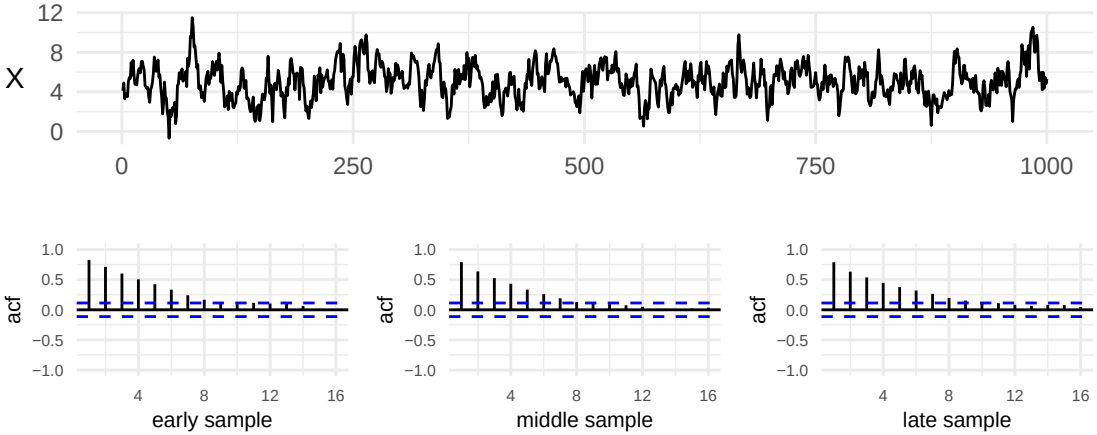
Recap

A **Covariance Stationary** process Y_t is one where

- $E(Y_t)$ is the same finite constant for all t
- $Var(Y_t)$ is the same finite constant for all t
- $Cov(Y_t, Y_{t-k})$ is, for any $k = 1, 2, ...$, the same finite constant for all t
  - May be different for different k
  - but for any given k, same for all t, i.e.,

Weakly-dependent: $ACF(k) \to 0$ as $k \to \infty$

Recap


X
12
8
4
0
0
250
500
750
1000
acf
1.0
0.5
0.0
−0.5
−1.0
4
8
12
16
early sample
middle sample
late sample

Recap

Examples of processes that are covariance-stationary (and weakly-dependent)

- white noise process

$$Y_t = \epsilon_t\,,\ \epsilon_t \sim (0, \sigma^2) \ \text{ where } \ Cov(\epsilon_t, \epsilon_s) = 0 \ \text{ for all } s \neq t\,.$$

- Y_t is covariance-stationary AR(1) if it satisfies:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t \ \text{ with } \ |\beta_1| < 1\,,\ \epsilon_t \overset{iid}{\sim} (0, \sigma^2)$$

Many other types of covariance-stationary processes

Recap

A **trend-stationary** process is one that is:

$$Y_t = f(t) + \text{ cov. stat. process}$$

- Deterministic trend $Y_t = \alpha + \delta t + \epsilon_t\,,\ \epsilon \overset{iid}{\sim} (0, \sigma^2)$
- Deterministic trend + AR(1) errors

$$Y_t = \alpha + \delta t + u_t\,,\ u_t = \rho u_{t-1} + \epsilon_t\,,\ |\rho| < 1\,,\ \epsilon \overset{iid}{\sim} (0, \sigma^2)$$

- Pure seasonal dummies model, seasonal dummies + cov. stationary AR(1) errors are all trend-stationary
- Think of "trend-stationary" as "deterministic + covariance-stationary"
- Trend-stationary processes are *not* covariance-stationary

Recap

Difference-stationary processes:

Recall random-walk: $Y_t = \beta_0 + Y_{t-1} + \epsilon$ which is non-stationary

If we take first-differences we get:

$$\Delta Y_t = Y_t - Y_{t-1} = \beta_0 + \epsilon$$

which is stationary

We say the random walk process is a **difference-stationary process**

- It is a non-stationary process made stationary by taking first differences

Recap

If

- Y_t has "random walk" type behavior and
- $\Delta Y_t \sim$ covariance-stationary AR(1) process

then Y_t is "difference-stationary"

Additional notation:

- cov.-stationary processes sometimes denoted "$I(0)$" ("integrated order zero")
- difference-stationary processes sometimes denoted "$I(1)$" ("integrated order one")

Time Series Regressions (Variance)

We start with regressions where the dependent variable and the regressors are

- covariance-stationary and weakly dependent

$$Y_t = \alpha_0 + \alpha_1 X_t + \epsilon_t \,, \; t = 1, \ldots, T$$

What is different from iid case?

- unbiasedness requires $E(\epsilon_t \mid X_1, \ldots, X_T) = 0$
- the usual causes of unbiasedness (e.g., omitted variables) remain
  - any additional issues?
- default variance formula assumes ϵ_t iid, how to amend?

Time Series Regressions (Variance)

Regarding variance formula, data unlikely to be iid, previously formulas unlikely to hold

- Suppose $Y_t = \mu + \epsilon_t\,,\ t = 1, \ldots, 4\,.$
- Want to estimate μ. We have
  - $E(Y_1) = E(Y_2) = E(Y_3) = E(Y_4) = \mu$
  - $Var(Y_t) = E((Y_t - \mu)^2) = E(\epsilon_t^2) = \gamma_0\,,\ t = 1, \ldots, 4$
  - $Cov(Y_t, Y_{t-1}) = E((Y_t - \mu)(Y_{t-1} - \mu)) = E(\epsilon_t \epsilon_{t-1}) = \gamma_1$ for $t = 2, 3, 4$
  - $Cov(Y_t, Y_{t-2}) = E((Y_t - \mu)(Y_{t-2} - \mu)) = E(\epsilon_t \epsilon_{t-2}) = \gamma_2$ for $t = 3, 4$
  - $Cov(Y_t, Y_{t-3}) = E((Y_t - \mu)(Y_{t-3} - \mu)) = E(\epsilon_t \epsilon_{t-3}) = \gamma_3$ for $t = 4$

Time Series Regressions (Variance)

$$\hat{\mu} = \tfrac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$$

$$\begin{aligned} Var(\hat{\mu}) = \frac{1}{16}[\, & Var(Y_1) + Var(Y_2) + Var(Y_3) + Var(Y_4) + 2Cov(Y_1, Y_2) + 2Cov(Y_1, Y_3) + \cdots \\ & 2Cov(Y_1, Y_4) + 2Cov(Y_2, Y_3) + 2Cov(Y_2, Y_4) + 2Cov(Y_3, Y_4)] \end{aligned}$$

$$= \frac{1}{16}[4\gamma_0 + 2 \cdot 3\gamma_1 + 2 \cdot 2\gamma_2 + 2 \cdot 1\gamma_3]$$

$$\neq \frac{\gamma_0}{4} \quad \text{if autocovariances are not zero}$$

For $t = 1, \ldots, T$:

$$Var(\hat{\mu}) = \frac{1}{T^2}(T\gamma_0 + 2(T-1)\gamma_1 + \cdots + 2\gamma_{T-1})$$

Time Series Regressions (Variance)

We seek a consistent estimator $\hat{S}$ for

$$S = T\,Var(\hat{\mu}) = \gamma_0 + \frac{2(T-1)}{T}\gamma_1 + \cdots + \frac{2}{T}\gamma_{T-1}$$

then use the approximation $Var(\hat{\mu}) \approx \hat{S}/T$

Use $\hat{S} = \hat{\gamma}_0 + \frac{2(T-1)}{T}\hat{\gamma}_1 + \cdots + \frac{2}{T}\hat{\gamma}_{T-1}$?

Newey-West

$$\hat{S} = \hat{\gamma}_0 + \sum_{v=1}^{q}\left(1 - \frac{2v}{q+1}\right)\hat{\gamma}_v$$

for some fixed q (NW suggestion: $q = 4(T/100)^{2/9}$)

Time Series Regressions (Variance)

For general k-regressor case

$$Y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_{k-1} X_{t,k-1} + \epsilon_t = X_{t*}\beta + \epsilon_t$$

where X_{t*} is the time-t row vector of regressors (incl. 1 for intercept)

Conditional homoskedasticity and iid sample: $\widehat{Var}(\hat{\beta}) = \widehat{\sigma^2}\left(\sum_{t=1}^{T} X_{t*}^{\mathrm{T}} X_{t*}\right)^{-1} = \widehat{\sigma^2}(X^{\mathrm{T}}X)^{-1}$

Conditional heteroskedasticity and iid sample:

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{t=1}^{T} X_{t*}^{\mathrm{T}} X_{t*}\right)^{-1} \left(\sum_{t=1}^{T} \hat{\epsilon}_t^2 X_{t*}^{\mathrm{T}} X_{t*}\right) \left(\sum_{t=1}^{T} X_{t*}^{\mathrm{T}} X_{t*}\right)^{-1}$$

Time Series Regressions (Variance)

Allowing for heteroskedasticy and autocorrelation in errors

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{t=1}^{T} X_{t*}^{\mathrm{T}} X_{t*}\right)^{-1} \times \left(\sum_{t=1}^{T} \hat{\epsilon}_t^2 X_{t*}^{\mathrm{T}} X_{t*} + \sum_{v=1}^{q} \left(1 - \frac{v}{q+1}\right) (X_{t*}^{\mathrm{T}} X_{t-v,*} + X_{t-v,*}^{\mathrm{T}} X_{t*}) \hat{\epsilon}_t \hat{\epsilon}_{t-v}\right) \times \left(\sum_{t=1}^{T} X_{t*}^{\mathrm{T}} X_{t*}\right)^{-1}$$

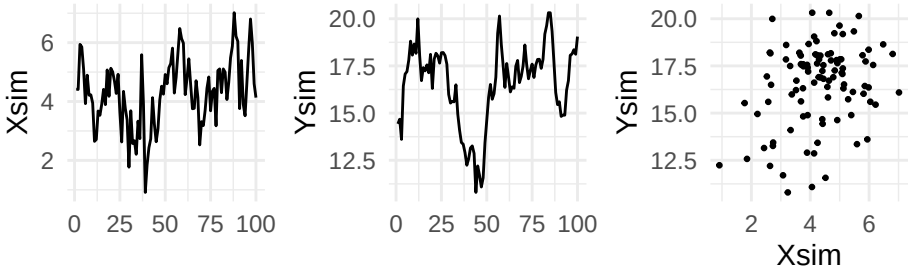
- "Heteroskedasticity and Autocorrelation Consistent" or (HAC) var-cov matrix estimators
- several kinds, the above is "Newey-West"

Time Series Regressions (Variance)

Simulation Example: $\{X_t, Y_t\}_{t=1}^{100}$ where

$$X_t = 0.8 + 0.8X_{t-1} + \epsilon_t \, , \, \epsilon_t \overset{iid}{\sim} N(0,1)$$

$$Y_t = 0.8 + 0X_t + u_t \, , \, u_t = 0.95u_{t-1} + v_t \, , \, v_t \overset{iid}{\sim} N(0,1)$$



Time Series Regressions (Variance)

```
mdlsim <- lm(Ysim~Xsim, data=df)
cat("OLS with Default Standard Errors\n")
mdlsim %>% lmtest::coeftest()
```

```
OLS with Default Standard Errors

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 14.24286    0.79751 17.8592 < 2.2e-16 ***
Xsim         0.53330    0.17928  2.9748  0.003692 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS Properties

```
mdlsim <- lm(Ysim~Xsim, data=df)
cat("OLS with Heteroskedasticity-Robust S.E.\n")
lmtest::coeftest(mdlsim, vcov=sandwich::vcovHC(mdlsim, type="HC2"))
```

```
OLS with Heteroskedasticity-Robust S.E.

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.24286    0.85027 16.7509  < 2e-16 ***
Xsim         0.53330    0.18137  2.9404  0.00409 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Not much difference is s.e., since errors are not heteroskedastic

Time Series Regressions (Variance)

```
mdlsim <- lm(Ysim~Xsim, data=df)
cat("OLS with Heteroskedasticity and Autocorrelation (HAC) Robust S.E.\n")
lmtest::coeftest(mdlsim, vcov=sandwich::NeweyWest)
```

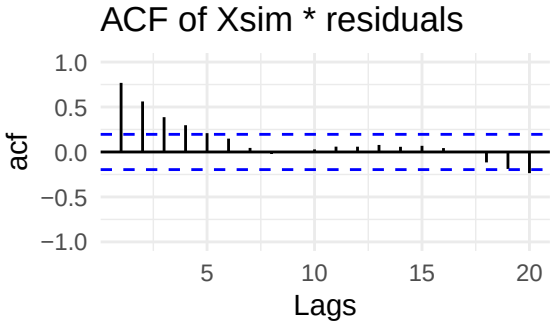
```
OLS with Heteroskedasticity and Autocorrelation (HAC) Robust S.E.

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 14.24286    2.06874  6.8848 5.551e-10 ***
Xsim         0.53330    0.34783  1.5332    0.1284
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time Series Regressions (Variance)

```
df <- df %>%
  mutate("res"=resid(mdlsim), "Xres"=Xsim*res)
ACF(df, Xres) %>%
  autoplot() + theme_minimal() + ylim(-1,1) +
  xlab("Lags") + ggtitle("ACF of Xsim * residuals")
```



- this ACF suggests that the estimator standard errors in third regression are appropriate
- standard errors in first two regressions too small since they ignore correlations in residuals

Time Series Regressions (Specification)

Possibility of dynamic specifications

- Static Regression: $Y_t = \alpha_0 + \alpha_1 X_t + \epsilon_t$
- Distributed Lag Models: $Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + ...\beta_q X_{t-q} + \epsilon_t$
- Autoregressions: $Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \epsilon_t\,,\ |\alpha_1| < 1$
- Autoregressive Distributed Lag (ARDL) models

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_q X_{t-q} + \epsilon_t$$

How should we interpret the parameters of such models?

Time Series Regressions (Dynamic Specifications)

Effect of one-unit one-period "impulse" in X_t

$$
\begin{aligned}
Y_t &= \alpha_0 + \boxed{\beta_0} X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_q X_{t-q} + \epsilon_t \\
Y_{t+1} &= \alpha_0 + \beta_0 X_{t+1} + \boxed{\beta_1} X_t + \beta_2 X_{t-1} + ... + \beta_q X_{t-q+1} + \epsilon_{t+1} \\
Y_{t+2} &= \alpha_0 + \beta_0 X_{t+2} + \beta_1 X_{t+1} + \boxed{\beta_2} X_t + ... + \beta_q X_{t-q+2} + \epsilon_{t+2} \\
&\vdots \\
Y_{t+q} &= \alpha_0 + \beta_0 X_{t+q} + \beta_1 X_{t+q-1} + \beta_2 X_{t+q-2} + ... + \boxed{\beta_q} X_t + \epsilon_{t+q} \\
Y_{t+q+1} &= \alpha_0 + \beta_0 X_{t+q+1} + \beta_1 X_{t+q} + \beta_2 X_{t+q-1} + ... + \beta_q X_{t+1} + \epsilon_{t+q+1}
\end{aligned}
$$

Coefficients are called "dynamic multipliers"

Time Series Regressions (Dynamic Specifications)

Effect of a permanent shift in X_t

$$
\begin{aligned}
Y_t &= \alpha_0 + \boxed{\beta_0} X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_q X_{t-q} + \epsilon_t \\
Y_{t+1} &= \alpha_0 + \boxed{\beta_0} X_{t+1} + \boxed{\beta_1} X_t + \beta_2 X_{t-1} + ... + \beta_q X_{t-q+1} + \epsilon_{t+1} \\
Y_{t+2} &= \alpha_0 + \boxed{\beta_0} X_{t+2} + \boxed{\beta_1} X_{t+1} + \boxed{\beta_2} X_t + ... + \beta_q X_{t-q+2} + \epsilon_{t+2} \\
&\vdots \\
Y_{t+q} &= \alpha_0 + \boxed{\beta_0} X_{t+q} + \boxed{\beta_1} X_{t+q-1} + \boxed{\beta_2} X_{t+q-2} + ... + \boxed{\beta_q} X_t + \epsilon_{t+q} \\
Y_{t+q+1} &= \alpha_0 + \boxed{\beta_0} X_{t+q+1} + \boxed{\beta_1} X_{t+q} + \boxed{\beta_2} X_{t+q-1} + ... + \boxed{\beta_q} X_{t+1} + \epsilon_{t+q+1}
\end{aligned}
$$

We refer to $\beta_0 + \beta_1 + \cdots + \beta_q$ as "long-run cumulative dynamic multiplier"

Time Series Regressions (Dynamic Specifications)

Interpretation of AR(1)?

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \epsilon_t\,,\ |\alpha_1| < 1$$

- A tool for describing "stable cycles"
- β_1 is the autocorrelation of Y_t at lag one
- Can be viewed as "reduced form" expression of cyclical behavior implied by economic interactions
- Member of the ARMA class of models (not covered in this course)
- A useful forecasting tool

Time Series Regressions (Dynamic Specifications)

Consider ARDL(1,1)

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \beta_0 X_t + \beta_1 X_{t-1} + \epsilon_t$$

This implies an "Infinite Distributed Lag Structure". Assume $|\alpha_1| < 1$.

- Lag: $Y_{t-1} = \alpha_0 + \alpha_1 Y_{t-2} + \beta_0 X_{t-1} + \beta_1 X_{t-2} + \epsilon_{t-1}$
- Substitute in ARDL

$$Y_t = \alpha_0(1+\alpha_1) + \alpha_1^2 Y_{t-2} + \beta_0 X_t + (\beta_1 + \alpha_1\beta_0) X_{t-1} + \alpha_1\beta_1 X_{t-2} + \epsilon_t + \alpha_1\epsilon_{t-1}$$

- Repeat with Y_{t-2}, then Y_{t-3}, and so on
- Y_t depends on X_t and infinite number of lags of X_t

Key Assumption for Consistency

Consider simple linear regression

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Key assumption for unbiasedness is $E(\epsilon_i \mid X_1, X_2, \dots, X_n) = 0$

In time series context, this assumption becomes

$$E(\epsilon_t \mid X_T, X_{T-1}, \dots, X_1) = 0$$

which turns out often to be too strong

Key Assumption for Consistency

E.g. 1: Regressions with lagged dependent variable, such as the AR(1)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t \,,\ \epsilon_t \overset{iid}{\sim} (0, \sigma^2) \,,\ t = 2, 3, ..., T \,.$$

- Assumption $E(\epsilon_t \mid X_T, X_{T-1}, ..., X_1) = 0$ is $E(\epsilon_t \mid Y_T, Y_{T-1}, ..., Y_1) = 0$
- but this is impossible since ϵ_t must be correlated with Y_t

E.g. 2: Regressions $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$ where noise term may contain information that can predict future X, implies correlation between ϵ_t and future X_s, $s > t$

Key Assumption for Consistency

The weaker assumption

$$E(\epsilon_t \mid X_t) = 0 \qquad \text{"contemporaneous exogeneity"}$$

is much more likely to hold

E.g., for AR(1)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

this assumption becomes $E(\epsilon_t \mid Y_{t-1}) = 0$, which is possible

Key Assumption for Consistency

If variables are covariance-stationary and weakly dependent, and if contemporaneous exogeneity holds, then

- although OLS estimator is still biased, it will nonetheless be consistent

$$\hat{\beta}_1^{ols} = \beta_1 + \frac{(1/T)\sum_{t=1}^{T}(X_t - \overline{X})\epsilon_t}{(1/T)\sum_{t=1}^{T}(X_t - \overline{X})^2}$$

As long as $cov(X_t, \epsilon_t) = 0$ and variables are cov. stationary weakly dependent, a CLT guarantees that numerator of second term converges to zero

Time Series Regressions, Autocorrelations in noise terms

We continue with simple linear regression case

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

- Assume that noise term satisfies contemporaneous exogeneity
- Assume (for the moment) that $X_t \neq Y_{t-1}$

With time series data, the default and Heteroskedasticity-robust standard errors formulas are often not appropriate

- in numerator of second term of RHS in previous slide, variance of sum is not sum of variance

Time Series Regressions with AR Errors

We now consider the special case of time series regressions with a particular kind of autocorrelation in the error term

"**Regression with (zero-mean) AR(1) errors**"

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \,,\; \epsilon_t = \rho \epsilon_{t-1} + u_t \,,\; |\rho| < 1 \,,\; u_t \overset{iid}{\sim} (0, \sigma^2)$$

How to estimate? A few options:

- Continue with OLS, use HAC variance estimators to get s.e. (ok, but not efficient)
- "Cochrane-Orcutt procedure" (a kind of "Generalized Least Squares)
- Transformation into "Dynamically Complete" ARDL

Time Series Regressions with AR Errors

Cochrane-Orcutt and ARDL aprroaches use the following transformation:

$$
\begin{aligned}
Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \,,\, \epsilon_t = \rho\epsilon_{t-1} + u_t \,,\, |\rho| < 1 \,,\, u_t &\overset{iid}{\sim} (0, \sigma^2) \\
\Rightarrow \quad \rho Y_{t-1} &= \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho\epsilon_{t-1} \\
\Rightarrow \quad Y_t - \rho Y_{t-1} &= (1-\rho)\beta_0 + \beta_1(X_t - \rho X_{t-1}) + \epsilon_t - \rho\epsilon_{t-1} \\
\Rightarrow \quad Y_t^* &= \beta_0^* + \beta_1 X_t^* + u_t
\end{aligned}
$$

where $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$

- Transformed regression is regression without autocorrelation or heteroskedasticity
- If $E(\epsilon_t \mid X_T, X_{T-1}, \ldots, X_1) = 0$ then all requirements for Gauss-Markov Theorem are met in the *transformed regression*, and OLS estimator from transformed regression will be best linear unbiased

Time Series Regressions with AR Errors

But ρ is unknown, and must be estimated

"Cochrane-Orcutt" suggestion:

- Estimate $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$, get $\hat{\epsilon}_t$
- Run regression $\hat{\epsilon}_t = \rho \hat{\epsilon}_{t-1} + u_t$, get $\hat{\rho}$
- Compute $Y_t^* = Y_t - \hat{\rho} Y_{t-1}$ and $X_t^* = X_t - \hat{\rho} X_{t-1}$
- Estimate regression $Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t$ using OLS

$$\hat{\beta}_1^{gls} = \frac{\sum_{t=2}^{T} (X_t^* - \overline{X^*})(Y_t^* - \overline{Y^*})}{\sum_{t=2}^{T} (X_t^* - \overline{X^*})^2}$$

Time Series Regressions with AR Errors

Alternative: since

$$Y_t - \rho Y_{t-1} = (1-\rho)\beta_0 + \beta_1(X_t - \rho X_{t-1}) + u_t$$

Estimate ARDL version

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 X_t + \alpha_3 X_{t-1} + u_t$$

although this is not exactly the same as the original

- original has 3 parameters
- ARDL version has 4 parameters
- To make them exactly the same, have to restrict $\alpha_1\alpha_2 + \alpha_3 = 0$

Time Series Regressions with AR Errors

Both approaches can be extended to multiple regressors, and also higher-ordered AR processes, e.g.,

$$Y_t = \beta_0 + \beta_1 X_{1t} + ... \beta_k X_{kt} + \epsilon_t \,, \;\; \epsilon_t = \rho_1 \epsilon_{t-1} + \cdots + \rho_p \epsilon_{t-p} + u_t$$

We omit discuss of higher-ordered AR processes in this course

Most researchers nowadays will either

- use OLS with HAC standard errors
- use ARDL approach, adding lags of Y_t and regressors until residuals do not indicate autocorrelations

Testing for Autocorrelation

To check for autocorrelation in noise terms

- check sample a.c.f. of regresson residuals

A more formal approach:

- regress

$$\hat{\epsilon}_t \text{ on } \hat{\epsilon}_{t-1}, \ldots, \hat{\epsilon}_{t-p}, X_{1t}, \ldots, X_{kt}$$

  where $X_{1t}, \ldots, X_{kt}$ are the regressors in X_t

- test for significance of the coefficients on the lagged residuals

Regression with Non-Stationary Series

Regression with Non-Stationary Series

- Regressions on trending and seasonal series
- Regressions on persistent series (containing random walk characteristics)

Regression with Deterministic Trend

If trend (and seasonality) are deterministic, they should be included in the regression

- Otherwise you will almost always get a significant regression, regardless of variables (basically an omitted variable situation)

E.g., Let Y_t is *log IP_SG* and X_t is *POULTRY_US* and consider the regressions

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$
$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + \beta_3 t^2 + \epsilon_t$$
$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + \beta_3 t^2 + \text{seasonal dummies} + \epsilon_t$$

Regression with Deterministic Trend

```
fit_chicken <- ts01 %>%
  model(
    m1 = TSLM(LN_IP_SG ~ I(POULTRY_US/1000000)),
    m2 = TSLM(LN_IP_SG ~ I(POULTRY_US/1000000) + season()),
    m3 = TSLM(LN_IP_SG ~ I(POULTRY_US/1000000) + trend() + I(trend()^2/1000)),
    m4 = TSLM(LN_IP_SG ~ I(POULTRY_US/1000000) + trend() + I(trend()^2/1000) + season())
  )
fit_chicken %>% select(m1) %>% coefficients()
```

```
# A tibble: 2 x 6
  .model term                estimate std.error statistic   p.value
  <chr>  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
1 m1     (Intercept)             1.06    0.0616      17.3 7.64e- 51
2 m1     I(POULTRY_US/1e+06)     4.45    0.0975      45.6 2.01e-164
```

Regression with Deterministic Trend

```
fit_chicken %>% select(m2) %>% coefficients()
```

```
# A tibble: 13 x 6
   .model term                estimate std.error statistic   p.value
   <chr>  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
 1 m2     (Intercept)           0.939     0.0716    13.1    4.74e- 33
 2 m2     I(POULTRY_US/1e+06)   4.57      0.0931    49.0    6.72e-173
 3 m2     season()year2         0.126     0.0606     2.07   3.87e-  2
 4 m2     season()year3         0.0889    0.0605     1.47   1.42e-  1
 5 m2     season()year4         0.0482    0.0605     0.797  4.26e-  1
 6 m2     season()year5        -0.0729    0.0605    -1.20   2.29e-  1
 7 m2     season()year6        -0.00449   0.0605    -0.0743 9.41e-  1
 8 m2     season()year7         0.00729   0.0605     0.121  9.04e-  1
 9 m2     season()year8        -0.0996    0.0606    -1.64   1.01e-  1
10 m2     season()year9         0.122     0.0605     2.01   4.49e-  2
11 m2     season()year10       -0.0353    0.0605    -0.583  5.60e-  1
12 m2     season()year11        0.233     0.0606     3.86   1.34e-  4
13 m2     season()year12        0.219     0.0605     3.63   3.22e-  4
```

Regression with Deterministic Trend

```
fit_chicken %>% select(m3) %>% coefficients()
```

```
# A tibble: 4 x 6
  .model term                estimate std.error statistic   p.value
  <chr>  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
1 m3     (Intercept)           2.39     0.0419       56.9  1.86e-198
2 m3     I(POULTRY_US/1e+06)   0.450    0.134         3.37 8.29e-  4
3 m3     trend()               0.00657  0.000393     16.7  2.41e- 48
4 m3     I(trend()^2/1000)    -0.00402  0.000659     -6.10 2.39e-  9
```

Regression with Deterministic Trend

```
fit_chicken %>% select(m4) %>% coefficients()
```

```
# A tibble: 15 x 6
   .model term                 estimate std.error statistic   p.value
   <chr>  <chr>                   <dbl>     <dbl>     <dbl>     <dbl>
 1 m4     (Intercept)           2.48      0.0508     48.9    6.11e-172
 2 m4     I(POULTRY_US/1e+06)   0.0224    0.158       0.142  8.87e-  1
 3 m4     trend()               0.00771   0.000446   17.3    1.53e- 50
 4 m4     I(trend()^2/1000)    -0.00574   0.000718   -7.99   1.38e- 14
 5 m4     season()year2        -0.100     0.0230     -4.36   1.62e-  5
 6 m4     season()year3         0.113     0.0216      5.23   2.72e-  7
 7 m4     season()year4        -0.000401  0.0217     -0.0185 9.85e-  1
 8 m4     season()year5        -0.00324   0.0217     -0.149  8.82e-  1
 9 m4     season()year6         0.0427    0.0217      1.97   4.93e-  2
10 m4     season()year7         0.0345    0.0216      1.60   1.11e-  1
11 m4     season()year8         0.0445    0.0222      2.00   4.59e-  2
12 m4     season()year9         0.0831    0.0216      3.84   1.41e-  4
13 m4     season()year10        0.0591    0.0219      2.70   7.22e-  3
14 m4     season()year11        0.00804   0.0229      0.351  7.26e-  1
15 m4     season()year12        0.0825    0.0220      3.74   2.08e-  4
```

Regression with Deterministic Trend

E.g., LN_ELEC_SG vs LN_IP_SG

```
fit_elec <- ts01 %>%
  model(
    elec1 = TSLM(LN_ELEC_SG ~ LN_IP_SG),
    elec2 = TSLM(LN_ELEC_SG ~ LN_IP_SG + trend() + I(trend()^2/1000) + season()),
    elec3 = TSLM(LN_ELEC_SG ~ LN_IP_SG + lag(LN_ELEC_SG)+ trend() + I(trend()^2/1000) + season()),
    elec4 = ARIMA(LN_ELEC_SG ~ LN_IP_SG + trend() + I(trend()^2/1000) + season() +
                               pdq(3,0,0) + PDQ(0,0,0))
  )
fit_elec %>% select(elec1) %>% coefficients()
```

```
# A tibble: 2 x 6
  .model term        estimate std.error statistic   p.value
  <chr>  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 elec1  (Intercept)    4.51    0.0290       156. 0
2 elec1  LN_IP_SG       0.836   0.00749      112. 3.68e-313
```

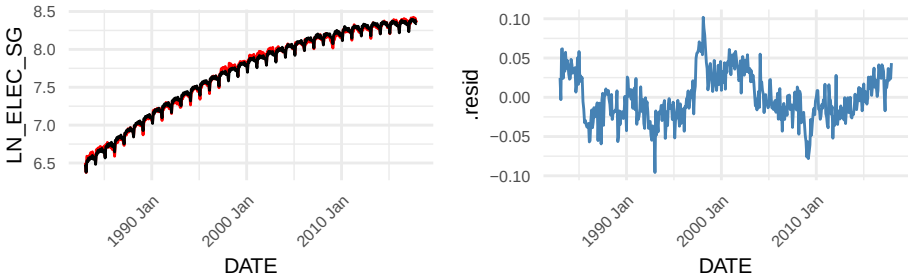
Regression with Deterministic Trend

```
fit_elec %>% select(elec2) %>% coefficients()
```

```
# A tibble: 15 x 6
   .model term                estimate std.error statistic   p.value
   <chr>  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
 1 elec2  (Intercept)          6.23      0.0429     145.    0
 2 elec2  LN_IP_SG             0.0888    0.0170       5.22  2.85e-  7
 3 elec2  trend()              0.00775   0.000141    54.8   1.49e-189
 4 elec2  I(trend()^2/1000)  -0.00890   0.000152   -58.7   4.63e-200
 5 elec2  season()year2      -0.0880    0.00759    -11.6   4.96e- 27
 6 elec2  season()year3       0.0330    0.00764      4.32  1.96e-  5
 7 elec2  season()year4       0.0240    0.00739      3.24  1.28e-  3
 8 elec2  season()year5       0.0640    0.00739      8.65  1.19e- 16
 9 elec2  season()year6       0.0313    0.00743      4.21  3.19e-  5
10 elec2  season()year7       0.0552    0.00742      7.44  6.04e- 13
11 elec2  season()year8       0.0457    0.00743      6.15  1.82e-  9
12 elec2  season()year9       0.0146    0.00753      1.93  5.39e-  2
13 elec2  season()year10      0.0458    0.00746      6.14  1.95e-  9
14 elec2  season()year11     -0.000946  0.00740     -0.128 8.98e-  1
15 elec2  season()year12     -0.00347   0.00753     -0.461 6.45e-  1
```

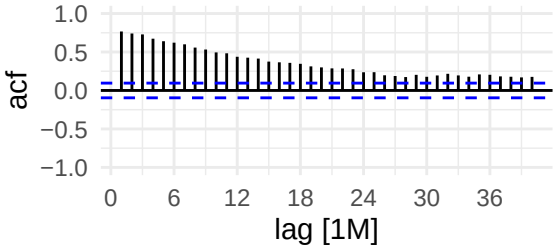
Regression with Deterministic Trend

```
fitted_elec2 <- fit_elec %>% select(elec2) %>% augment()
p1 <- fitted_elec2 %>% ggplot(aes(x=DATE)) +
  geom_line(aes(y = LN_ELEC_SG), color="red") +
  geom_line(aes(y = .fitted), color="black", linewidth=0.5) + theme0
p2 <- fitted_elec2 %>% ggplot(aes(x=DATE)) +
  geom_line(aes(y = .resid), color="steelblue") + theme0
p1 | p2
```



Regression with Deterministic Trend

```
fitted_elec2 %>% ACF(.resid, lag_max=40) %>% autoplot() + theme_minimal() + ylim(-1,1)
```



Model `elec2` is dynamically incomplete

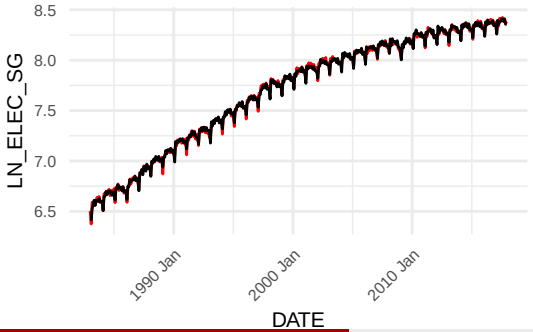
Regression with Deterministic Trend

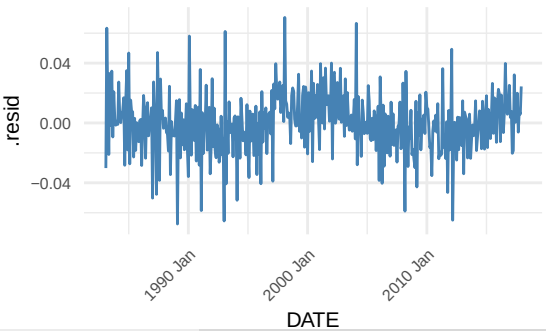
```
fit_elec %>% select(elec3) %>% coefficients()
```

```
# A tibble: 16 x 6
   .model term                estimate std.error statistic  p.value
   <chr>  <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
 1 elec3  (Intercept)          1.65     0.210        7.85  3.67e-14
 2 elec3  LN_IP_SG             0.0636   0.0116       5.50  6.68e- 8
 3 elec3  lag(LN_ELEC_SG)      0.721    0.0326      22.1   1.51e-71
 4 elec3  trend()              0.00185  0.000284     6.51  2.27e-10
 5 elec3  I(trend()^2/1000)  -0.00226  0.000318    -7.09  6.16e-12
 6 elec3  season()year2      -0.0879   0.00514    -17.1   1.00e-49
 7 elec3  season()year3       0.109    0.00619     17.5   1.72e-51
 8 elec3  season()year4      -0.00439  0.00518     -0.847 3.97e- 1
 9 elec3  season()year5       0.0494   0.00506      9.75  2.60e-20
10 elec3  season()year6      -0.0109   0.00541     -2.02  4.42e- 2
11 elec3  season()year7       0.0335   0.00514      6.52  2.15e-10
12 elec3  season()year8       0.00756  0.00535      1.41  1.58e- 1
13 elec3  season()year9      -0.0165   0.00531     -3.11  1.99e- 3
14 elec3  season()year10      0.0342   0.00510      6.71  6.56e-11
15 elec3  season()year11     -0.0350   0.00526     -6.65  9.56e-11
16 elec3  season()year12      0.00152  0.00512      0.297 7.67e- 1
```

Regression with Deterministic Trend

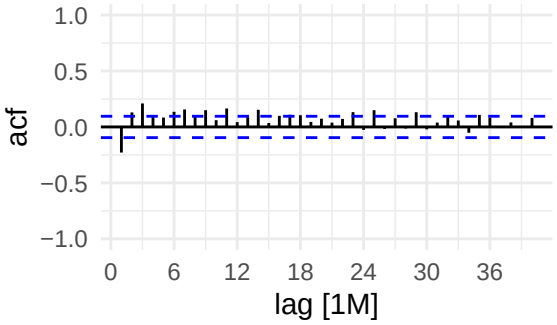
```
fitted_elec3 <- fit_elec %>% select(elec3) %>% augment()
p1 <- fitted_elec3 %>% ggplot(aes(x=DATE)) + geom_line(aes(y = LN_ELEC_SG), color="red") +
  geom_line(aes(y = .fitted), color="black", linewidth=0.5) + theme0
p2 <- fitted_elec3 %>% ggplot(aes(x=DATE)) + geom_line(aes(y = .resid), color="steelblue") + theme0
p1 | p2
```





Regression with Deterministic Trend

```
fitted_elec3 %>% ACF(.resid, lag_max=40) %>% autoplot() + theme_minimal() + ylim(-1,1)
```



Better!

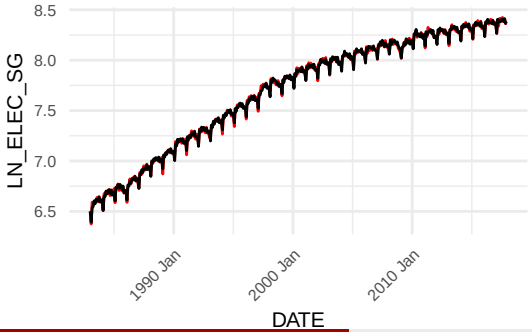
Regression with Deterministic Trend

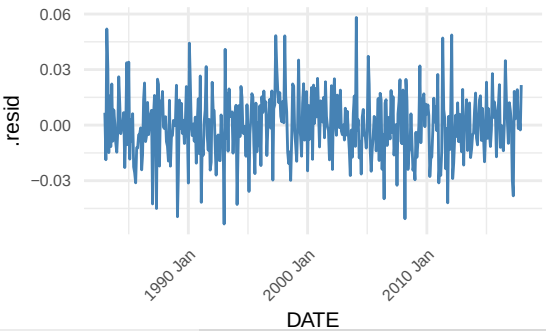
```
fit_elec %>% select(elec4) %>% coefficients()
```

```
# A tibble: 18 x 6
   .model term                estimate std.error statistic   p.value
   <chr>  <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
 1 elec4  ar1                   0.406    0.0478       8.50  3.31e- 16
 2 elec4  ar2                   0.228    0.0511       4.46  1.07e-  5
 3 elec4  ar3                   0.260    0.0477       5.44  9.17e-  8
 4 elec4  LN_IP_SG              0.118    0.0139       8.47  4.06e- 16
 5 elec4  trend()               0.00741  0.000255    29.1   9.99e-103
 6 elec4  I(trend()^2/1000)    -0.00846  0.000520   -16.3   1.60e- 46
 7 elec4  season()year2        -0.0852   0.00370    -23.0   1.86e- 76
 8 elec4  season()year3         0.0297   0.00393      7.56  2.48e- 13
 9 elec4  season()year4         0.0239   0.00353      6.77  4.49e- 11
10 elec4  season()year5         0.0639   0.00387     16.5   1.53e- 47
11 elec4  season()year6         0.0298   0.00399      7.47  4.56e- 13
12 elec4  season()year7         0.0540   0.00395     13.7   1.74e- 35
13 elec4  season()year8         0.0442   0.00400     11.0   4.84e- 25
14 elec4  season()year9         0.0119   0.00405      2.93  3.52e-  3
15 elec4  season()year10        0.0438   0.00365     12.0   9.48e- 29
16 elec4  season()year11       -0.00158  0.00362     -0.437 6.62e-  1
```

Regression with Deterministic Trend

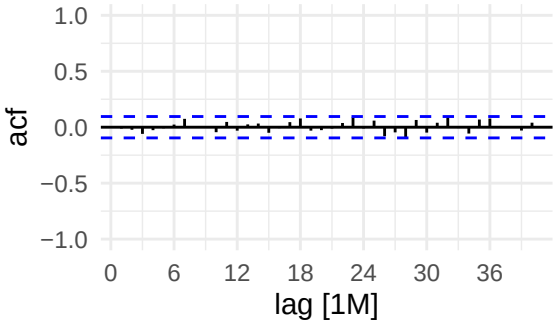
```
fitted_elec4 <- fit_elec %>% select(elec4) %>% augment()
p1 <- fitted_elec4 %>% ggplot(aes(x=DATE)) + geom_line(aes(y = LN_ELEC_SG), color="red") +
  geom_line(aes(y = .fitted), color="black", linewidth=0.5) + theme0
p2 <- fitted_elec4 %>% ggplot(aes(x=DATE)) + geom_line(aes(y = .resid), color="steelblue") + theme0
p1 | p2
```





Regression with Deterministic Trend

```
fitted_elec4 %>% ACF(.resid, lag_max=40) %>% autoplot() + theme_minimal() + ylim(-1,1)
```



Regression with Persistent Series (A Warning)

We end with a remark on regression with persistent series

Simulate 200 pairs of random walks $\{X_t^{(r)}, Y_t^{(r)}\}_{t=1}^{100}$, $r = 1, 2, ..., 200$:

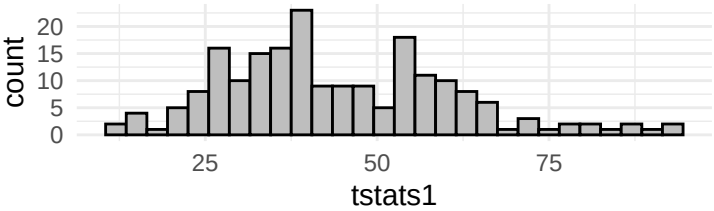
$$X_t^{(r)} = \alpha_X + X_{t-1}^{(r)} + u_t^{(r)}$$
$$Y_t^{(r)} = \alpha_Y + Y_{t-1}^{(r)} + v_t^{(r)}$$

where $u_t^{(r)}$ and $v_t^{(r)}$ are independent Normal$(0, 1)$ noise terms, $\alpha_X = 0.5$ and $\alpha_Y = 0.8$.

For each replication r

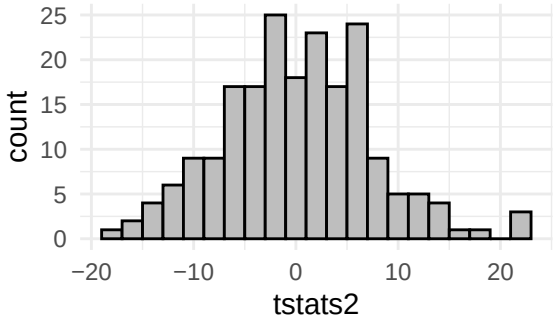
- regress $Y_t^{(r)}$ on $X_t^{(r)}$, with intercept
- collect the t-statistic on the coefficient of $X_t^{(r)}$

Regression with Persistent Series (A Warning)



Regression with Persistent Series (A Warning)

Repeat with $\alpha_X = \alpha_Y = 0$ (i.e., no drift)



Regression with Persistent Series

- Regressions with persistent series not always spurious, can give very good results
- Simple (maybe incomplete) solution for persistent series is to take first differences (i.e., transform to stationarity)
- Stochastic vs Deterministic Trend?
- Will have to leave further details to "next course"