# Lecture 13

## Time Series Analysis II

*DSA 8020 Statistical Methods II*

Whitney Huang
Clemson University

Estimating Seasonality
Regression Methods
Seasonal ARIMA Models

# Modeling Trend, Seasonality, and Noise

Recall the trend, seasonality, noise decomposition mentioned last week:

$$Y_t = \mu_t + s_t + \eta_t,$$

where

- $\mu_t$: trend component with $\mathbb{E}(Y_t) = \mu_t$;
- $s_t$: seasonal component with $\mathbb{E}(s_t) = 0$;
- $\eta_t$: random noise with $\mathbb{E}(\eta_t) = 0$

We are going to learn two approaches for estimating $s_t$, the seasonal component

# Seasonal Component Estimation

- Let's consider the situation that a time series consists of seasonal component only (assuming the trend has been estimated/removed), that is,

$$Y_t = s_t + \eta_t,$$

with $\{s_t\}$ having period $d$ (i.e., $s_{t+jd} = s_t$ for all integers $j$ and $t$), $\sum_{t=1}^{d} s_t = 0$ and $\mathbb{E}(\eta_t) = 0$

- Two regression methods to estimate $\{s_t\}$
  - Harmonic regression
  - Seasonal mean model

## Harmonic Regression



Estimating Seasonality

Regression Methods

Seasonal ARIMA Models

- A harmonic regression model has the form

$$s_t = \sum_{j=1}^{k} A_k \cos(2\pi f_j t + \phi_j).$$

For each $j = 1, \cdots, k$:

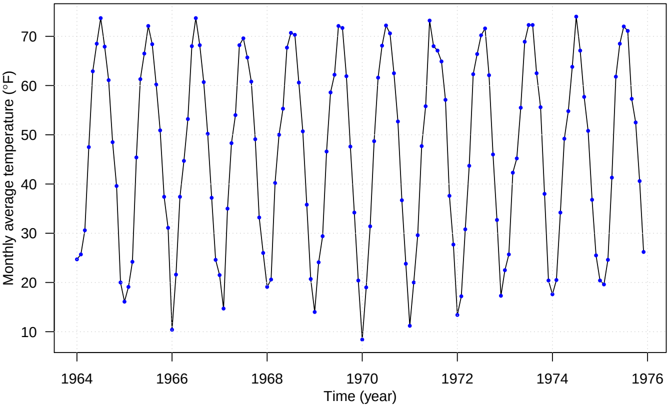
  - $A_j > 0$ is the amplitude of the $j$-th cosine wave
  - $f_j$ controls the the frequency of the $j$-th cosine wave (how often waves repeats)
  - $\phi_j \in [-\pi, \pi]$ is the phase of the $j$-th wave (where it starts)

- The above can be expressed as

$$\sum_{j=1}^{k} \{\beta_{1j} \cos(2\pi f_j t) + \beta_{2j} \sin(2\pi f_j t)\},$$

where $\beta_{1j} = A_j \cos(\phi_j)$ and $\beta_{2j} = -A_j \sin(\phi_j) \Rightarrow$ if $\{f_j\}_{j=1}^{k}$ are known, **we can use regression techniques to estimate the parameters** $\{\beta_{1j}, \beta_{2j}\}_{j=1}^{k}$

# Monthly Temperature in Dubuque, IA [Cryer & Chan, 2008]




Let's assume there is no trend in this time series. Here we want to estimate $s_t$, the seasonal component

# Use a Harmonic Regression to Model Annual Cycles

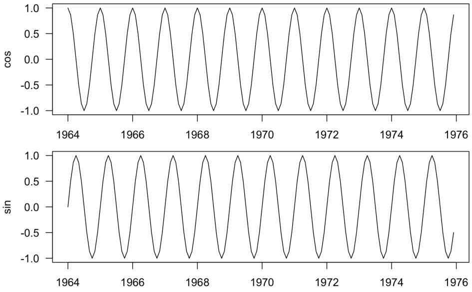


**Model:** $s_t = \beta_0 + \beta_1 \cos(2\pi t) + \beta_2 \sin(2\pi t)$

$\Rightarrow$ annual cycles can be modeled by a linear combination of $\cos$ and $\sin$ with 1-year period.

In `R`, we can easily create these harmonics using the `harmonic` function in the `TSA` package

```
harmonics <- harmonic(tempdub, 1)
```



# R Code & Output



```
```{r}
harReg <- lm(tempdub ~ harmonics)
summary(harReg)
```

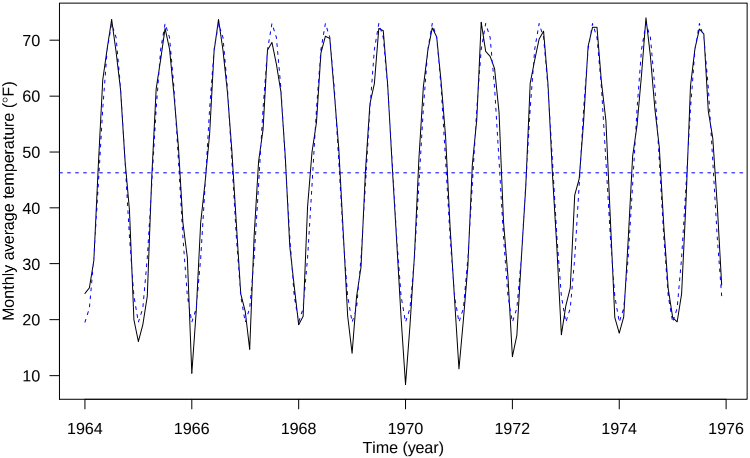
```
Call:
lm(formula = tempdub ~ harmonics)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1580  -2.2756  -0.1457   2.3754  11.2671

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           46.2660     0.3088 149.816  < 2e-16 ***
harmonicscos(2*pi*t) -26.7079     0.4367 -61.154  < 2e-16 ***
harmonicssin(2*pi*t)  -2.1697     0.4367  -4.968 1.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
```

# The Harmonic Regression Model Fit





# Seasonal Means Model



- **Harmonics regression** assumes the seasonal pattern has a regular shape, i.e., the height of the peaks is the same as the depth of the troughs

- A less restrictive approach is to model $\{s_t\}$ as

$$s_t = \begin{cases} \beta_1 & \text{for } t = 1, 1+d, 1+2d, \cdots \quad ; \\ \beta_2 & \text{for } t = 2, 2+d, 2+2d, \cdots \quad ; \\ \vdots & \vdots \quad ; \\ \beta_d & \text{for } t = d, 2d, 3d, \cdots \quad . \end{cases}$$

- This is the seasonal means model, the parameters $(\beta_1, \beta_2, \cdots, \beta_d)^T$ can be estimated under the linear model framework (think about ANOVA)

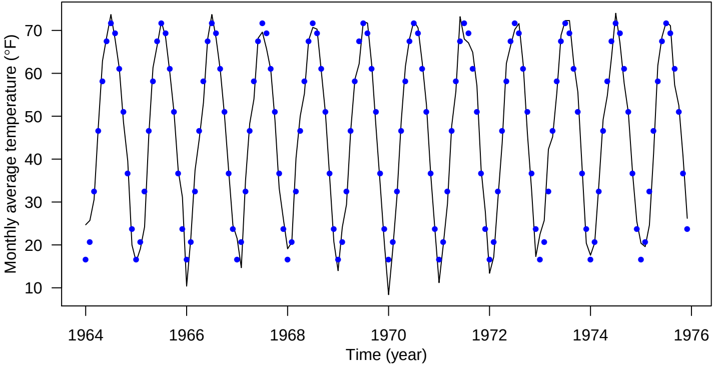
# R Output

```
Call:
lm(formula = tempdub ~ month - 1)

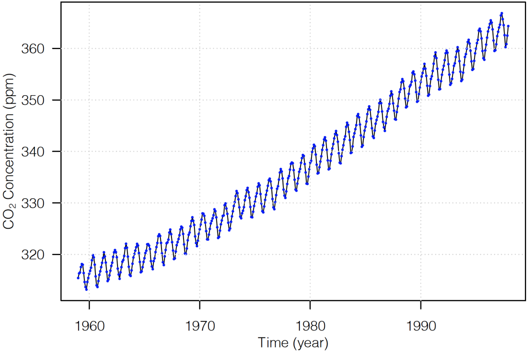
Residuals:
    Min      1Q  Median      3Q     Max
-8.2750 -2.2479  0.1125  1.8896  9.8250

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
monthJanuary     16.608      0.987   16.83   <2e-16 ***
monthFebruary    20.650      0.987   20.92   <2e-16 ***
monthMarch       32.475      0.987   32.90   <2e-16 ***
monthApril       46.525      0.987   47.14   <2e-16 ***
monthMay         58.092      0.987   58.86   <2e-16 ***
monthJune        67.500      0.987   68.39   <2e-16 ***
monthJuly        71.717      0.987   72.66   <2e-16 ***
monthAugust      69.333      0.987   70.25   <2e-16 ***
monthSeptember   61.025      0.987   61.83   <2e-16 ***
monthOctober     50.975      0.987   51.65   <2e-16 ***
monthNovember    36.650      0.987   37.13   <2e-16 ***
monthDecember    23.642      0.987   23.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# The Seasonal Means Model Fit



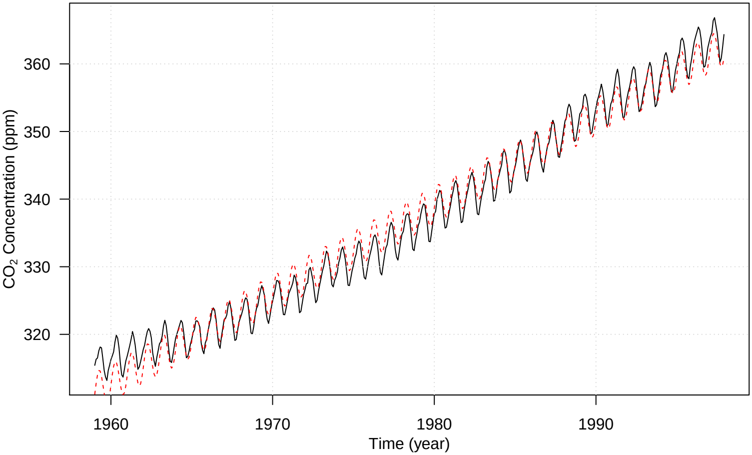
# Estimating the Trend and Seasonal Components Together




Let's perform a regression analysis to model both $\mu_t$ (assuming a linear time trend) and $s_t$ (using $\cos$ and $\sin$)

```{r}
time <- as.numeric(time(co2))
harmonics <- harmonic(co2, 1)

lm_trendSeason <- lm(co2 ~ time + harmonics)
summary(lm_trendSeason)
```

# The Regression Fit



# Differencing

- We define the first order difference operator $\nabla$ as

$$\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t,$$

 where $B$ is the **backshift operator** and is defined as $BY_t = Y_{t-1}$.

- Similarly the general order difference operator $\nabla^q Y_t$ is **defined recursively** as $\nabla[\nabla^{q-1} Y_t]$

- The backshift operator of power $q$ is defined as $B^q Y_t = Y_{t-q}$

- A seasonal difference is the difference between an observation and the previous observation from the same season:

$$Y_t - Y_{t-s} = Y_t - B^s Y_t = (1 - B^s) Y_t$$

# The Seasonal ARIMA (SARIMA) Model



Let $d$ and $D$ be non-negative integers. Then $\{X_t\}$ is a seasonal ARIMA$(p, d, q) \times (P, D, Q)$ process with period $s$ if

$$Y_t = \nabla^d \nabla_s^D X_t = (1 - B)^d (1 - B^s)^D X_t,$$

is a casual ARMA process define by

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t,$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$.

$\{Y_t\}$ is causal if $\phi(z) \neq 0$ and $\Phi(z) \neq 0$, for $|z| \leq 1$, where

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p;$$
$$\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P.$$

# An Illustration of Seasonal Model



Consider a monthly time series $\{X_t\}$ with both a trend, and a seasonal component of period $s = 12$.

- Suppose we know the values of $d$ and $D$ such that $Y_t = (1 - B)^d (1 - B^{12})^D X_t$ is stationary
- We can arrange the data this way:

| | Month 1 | Month 2 | $\cdots$ | Month 12 |
|---|---|---|---|---|
| Year 1 | $Y_1$ | $Y_2$ | $\cdots$ | $Y_{12}$ |
| Year 2 | $Y_{13}$ | $Y_{14}$ | $\cdots$ | $Y_{24}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\cdots$ | $\vdots$ |
| Year $r$ | $Y_{1+12(r-1)}$ | $Y_{2+12(r-1)}$ | $\cdots$ | $Y_{12+12(r-1)}$ |

# The Inter-annual Model

Here we view each column (month) of the data table from the previous slide as a separate time series

- For each month $m$, we assume the same ARMA$(P, Q)$ model. We have

$$
\begin{aligned}
Y_{m+12s} - \sum_{i=1}^{P} \Phi_i Y_{m+12(s-i)} \\
= U_{m+12s} + \sum_{j=1}^{Q} \Phi_j U_{m+12(s-j)},
\end{aligned}
$$

  for each $s = 0, \cdots, r-1$, where $\{U_{m+12s:s=0,\cdots,r-1}\} \sim \mathrm{WN}(0, \sigma_U^2)$ for each $m$
- We can write this as

$$
\Phi(B^{12})Y_t = \Theta(B^{12})U_t,
$$

  and this defines the inter-annual model

# The Intra-Annual Model



We induce correlation between the months by letting the process $\{U_t\}$ follow an ARMA$(p, q)$ model,

$$\phi(B)U_t = \theta(B)Z_t,$$

where $Z_t \sim \mathrm{WN}(0, \sigma^2)$

- This is the intra-annual model
- The combination of the inter-annual and intra-annual models for the differenced stationary series,

  $$Y_t = (1 - B)^d(1 - B^{12})^D X_t,$$

  yields a SARIMA model for $\{X_t\}$

## Steps for Modeling SARIMA Processes

1. Transform data if necessary

2. Find $d$ and $D$ so that

$$Y_t = (1 - B)^d (1 - B^s)^D X_t$$

is stationary

3. Examine the sample ACF/PACF of $\{Y_t\}$ at lags that are multiples of $s$ for plausible values for $P$ and $Q$

4. Examine the sample ACF/PACF at lags $\{1, 2, \cdots, s - 1\}$, to identify possible values for $p$ and $q$

## Modeling SARIMA Processes (Cont'd)



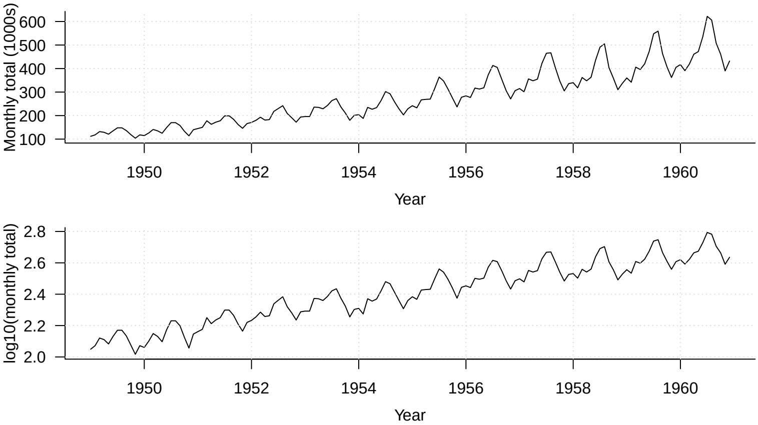
5. Use maximum likelihood method to fit the models

6. Use model summaries, diagnostics, AIC (AICC) to determine the best SARIMA model

7. Conduct forecast

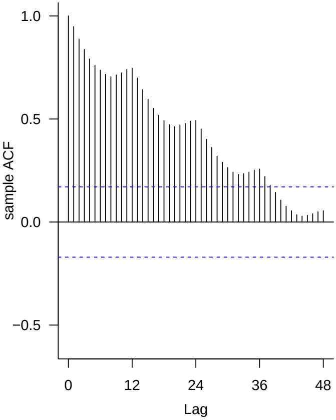
# Airline Passengers Example

We consider the data set `airpassengers`, which are the monthly totals of international airline passengers from 1949 to 1960, taken from Box and Jenkins [1970]



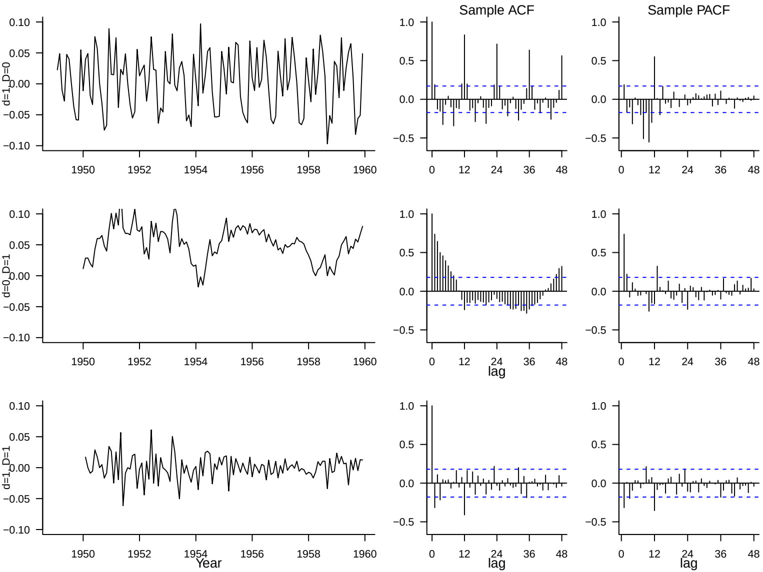

Here we stabilize the variance with a $\log_{10}$ transformation

# Sample ACF/PACF Plots



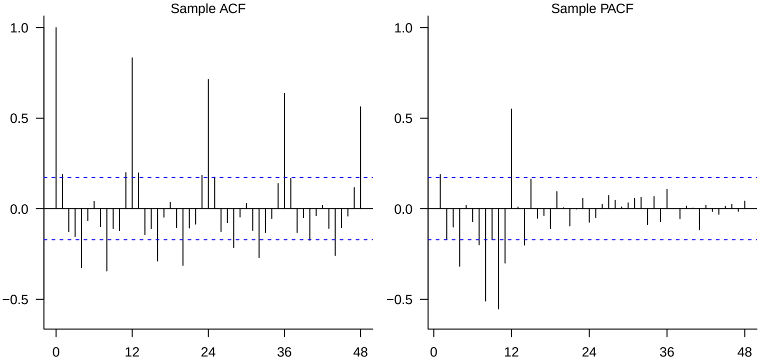

- The sample ACF decays slowly with a wave structure ⇒ seasonality
- The lag one PACF is close to one, indicating that differencing the data would be reasonable

# Trying Different Orders of Differencing



# Choosing Candidate SARIMA Models

We choose a SARIMA$(p, 1, q) \times (P, 0, Q)$ model. Next we examine the sample ACF/PACF of the process $Y_t = (1 - B)X_t$

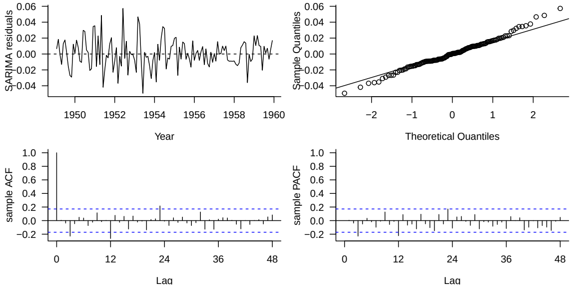



Now we need to choose $P$ , $Q$, $p$, and $q$

# Fitting a SARIMA$(1,1,0) \times (1,0,0)$ model

```
> fit1 <- arima(diff.1.0, order = c(1, 0, 0), seasonal = list(order = c(1, 0, 0), period = 12))
> fit1

Call:
arima(x = diff.1.0, order = c(1, 0, 0), seasonal = list(order = c(1, 0, 0),
    period = 12))

Coefficients:
          ar1    sar1  intercept
      -0.2667  0.9291     0.0039
s.e.   0.0865  0.0235     0.0096

sigma^2 estimated as 0.0003298:  log likelihood = 327.27,  aic = -646.54
```

```
> Box.test(fit1$residuals, lag = 48, type = "Ljung-Box")

        Box-Ljung test

data:  fit1$residuals
X-squared = 55.372, df = 48, p-value = 0.2164
```



## A Discussion of the Model Fit

- The spread of the residuals is larger in 1949-1955 compared to the later years and the residual distribution has heavy tails
- The Ljung-Box test result indicates the fitted SARIMA $(1,1,0)\times(1,0,0)$ has sufficiently account for the temporal dependence
- 95% CI for $\phi_1$ and $\Phi_1$ do not contain zero $\Rightarrow$ no need to go with simpler model

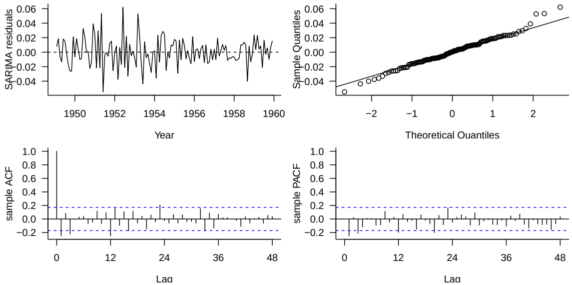
Our estimated model is

$$(1+0.2667B)(1-0.9291B^{12})(X_t-0.0039)=Z_t,$$

where $\{Z_t\} \overset{i.i.d.}{\sim} \mathrm{N}(0,\sigma^2)$ with $\hat{\sigma}^2=0.00033$

# Comparing with a SARIMA$(0, 1, 0) \times (1, 0, 0)$ Model

```
> (fit2 <- arima(diff.1.0, seasonal = list(order = c(1, 0, 0), period = 12)))

Call:
arima(x = diff.1.0, seasonal = list(order = c(1, 0, 0), period = 12))

Coefficients:
        sar1  intercept
      0.9081     0.0040
s.e.  0.0278     0.0108

sigma^2 estimated as 0.0003616:  log likelihood = 322.75,  aic = -639.51
> Box.test(fit2$residuals, lag = 48, type = "Ljung-Box")

        Box-Ljung test

data:  fit2$residuals
X-squared = 80.641, df = 48, p-value = 0.002209
```



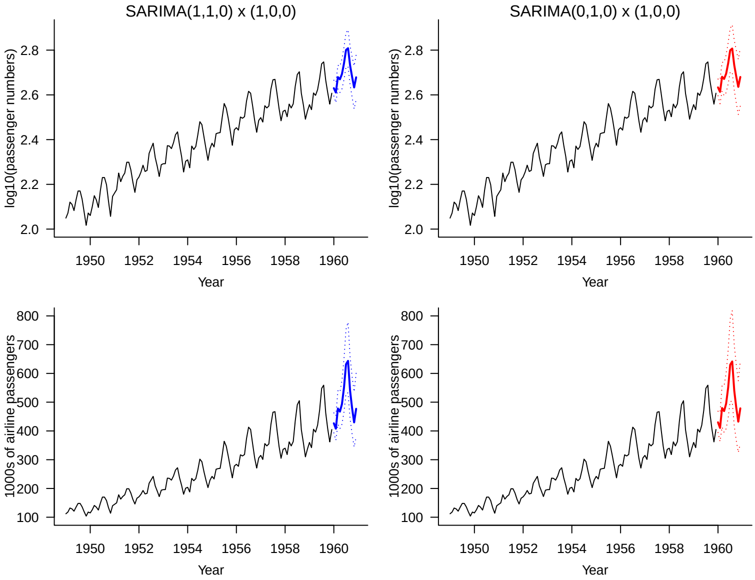
## A Discussion of Model Fit2

Here we drop the AR(1) term

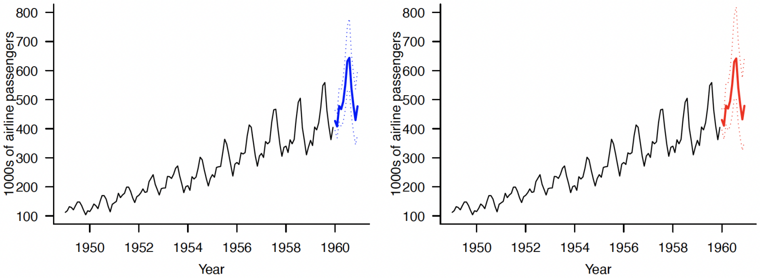
- The residual plots looks quite similar to before: The spread of the residuals is larger in 1949-1955 compared to the later years and the residual distribution has heavy tails
- Both $\hat{\sigma}^2$ and AIC increase (compared with model fit1)
- The lag 1 of ACF and PACF now lies outside the IID noise bounds. The Ljung-Box P-value of 0.0022, leads us to reject the IID residual assumption

In conclusion, the SARIMA$(1,1,0) \times (1,0,0)$ model fits better than SARIMA$(0,1,0) \times (1,0,0)$

# Forecasting the 1960 Data



# Evaluating Forecast Performance




| Metrics | Model Fit1 | Model Fit2 |
|---|---|---|
| Root Mean Square Error | 30.36 | 31.32 |
| Mean Relative Error | 0.057 | 0.060 |
| Empirical Coverage | 0.917 | 1.000 |

## Summary

This slides cover two methods for estimating seasonality:

- Harmonic regression models
- Seasonal ARIMA Models
- Ways to evaluate forecasting performance