Modeling and Forecasting Seasonality

Zhenhao Gong

University of Connecticut

This course is designed to be:

1. Introductory
2. Leading by interesting questions and applications
3. Less math, useful, and fun!

Most important:

Feel free to ask any questions!

Enjoy! ☺

Recall

- Unobserved Components: trend, seasonal, cycle, noise.

$$y_t = T_t + S_t + C_t + \varepsilon_t.$$

Or

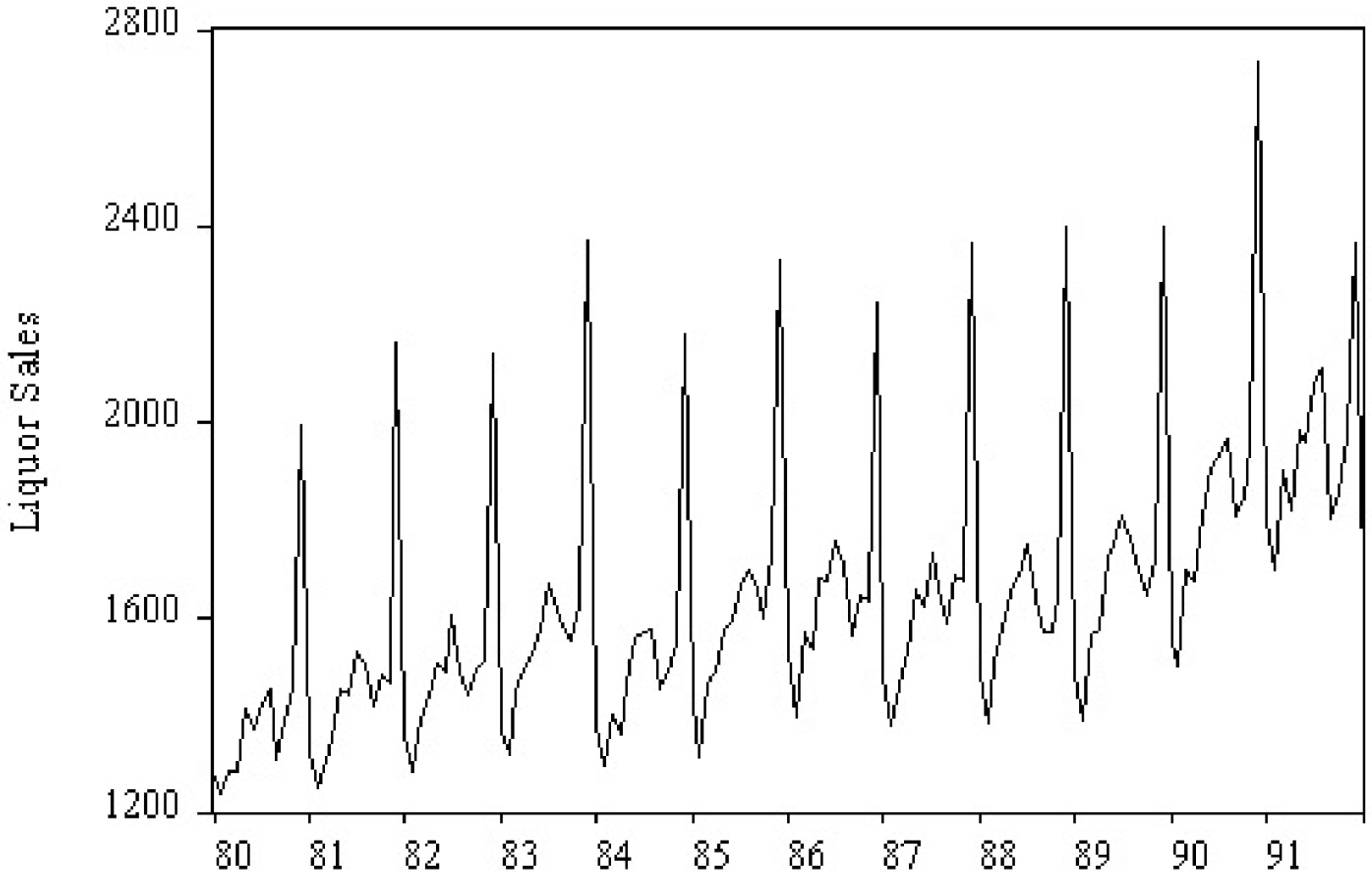
$$y_t = T_t \times S_t \times C_t \times \varepsilon_t.$$

- We focus on seasonal on this lecture.

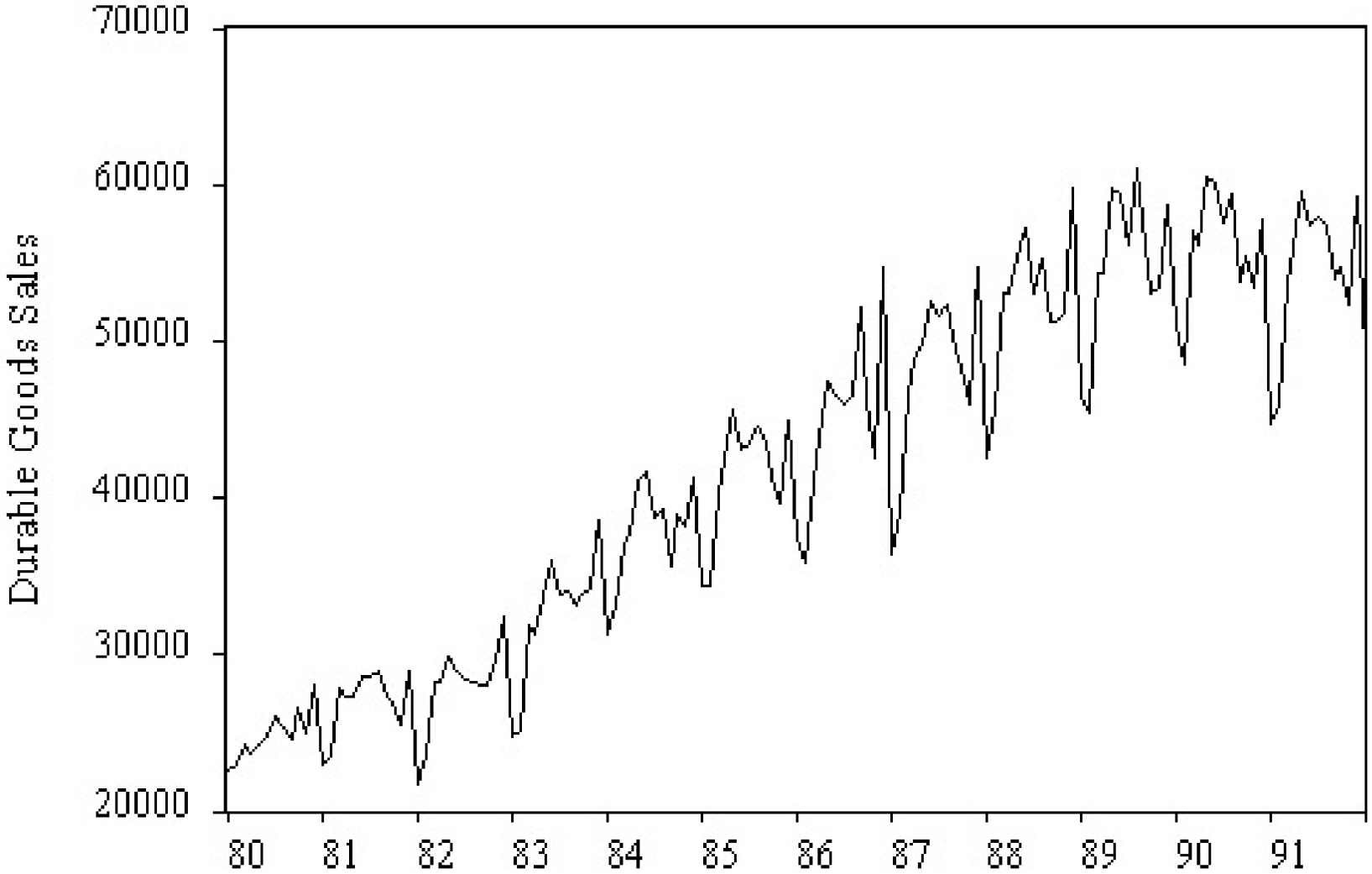
A seasonal pattern is one that repeats itself every year. It arises from links of technologies, preferences and institutions to the calendar.

We focus on the deterministic seasonality in which the annual repetition can be **exact**.

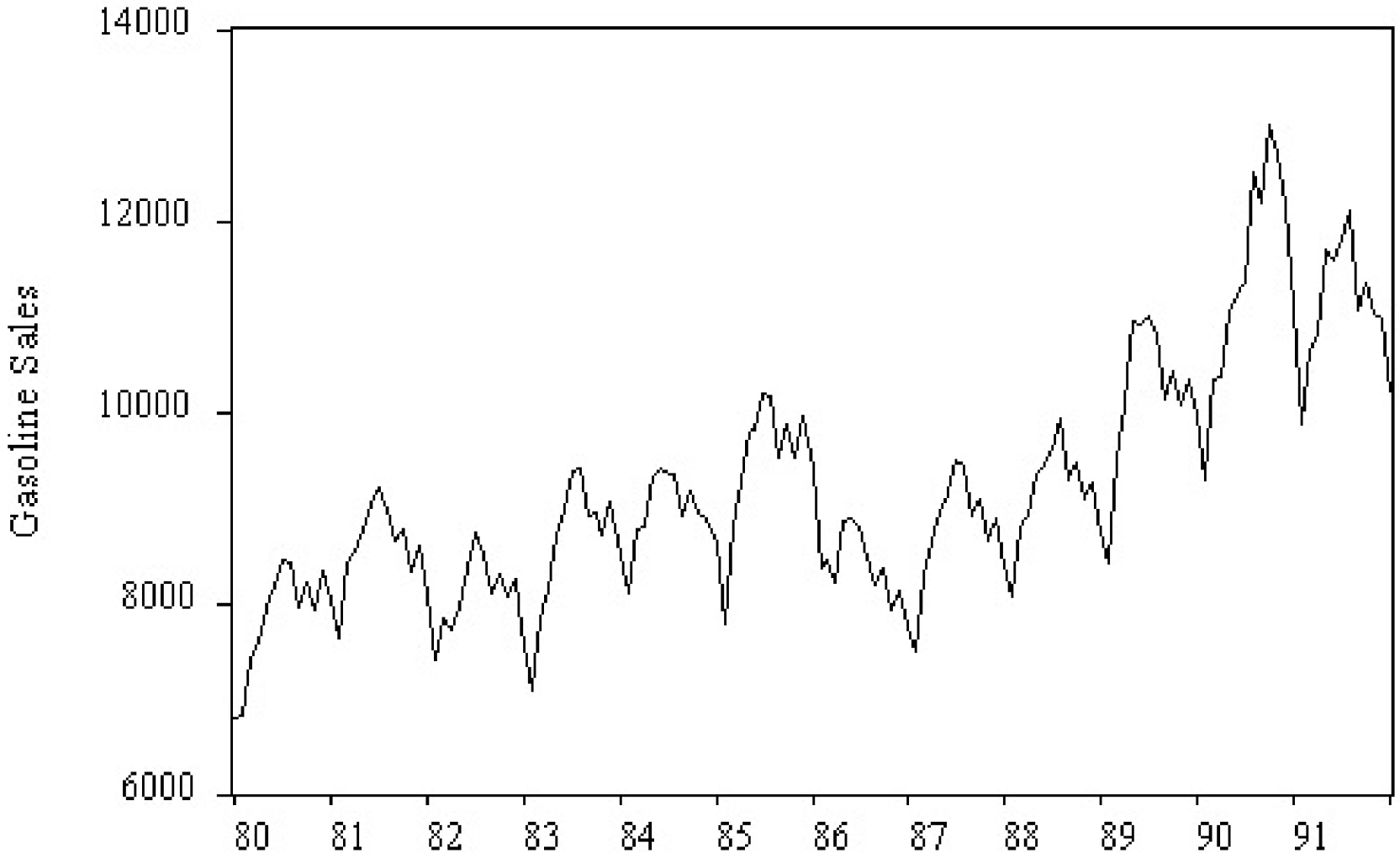
- Example: monthly U.S. current-dollar liquor sales 1980.01 - 1992.01: very high in Nov. and Dec.



- Example: Monthly U.S. current-dollar durable goods sales, 1980.01 - 1992.01: fall in Dec.



- Example: Monthly U.S. current-dollar gasoline sales, 1980.01 - 1992.01: higher in summertime



Seasonally adjusted

One way to deal with seasonality in a series is simply to remove it, and then to model and forecast the **seasonally adjusted series**.

Useful when interest centers explicitly on forecasting nonseasonal fluctuations.

Often inappropriate in business forecasting situations in which all the variation in a series are interested.

A key technique for modeling seasonality is regression on **seasonal dummies**.

Let **s** be the number of seasons in year. We can think it as the number of observations on a series in each year.

- $s = 4$ if we have quarterly data
- $s = 12$ if we have monthly data
- $s = 52$ if we have weekly data

Regression analysis can also be used when the regressor is **binary**, that is, when it takes on only two values, 0 or 1:

- $X = 1$ if small class size, $= 0$ if not
- $X = 1$ if female, $= 0$ if male
- $X = 1$ if treated (experimental drug), $= 0$ if not

Binary regressors are sometimes called "dummy" variables.

Interpreting regressions with a binary regressor

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \quad i = 1, 2, \cdots, n$$

where X is binary ($X_i = 0$ or 1):

- When $X_i = 0$, $Y_i = \beta_0 + u_i$, the conditional mean of Y_i given $X_i = 0$ is β_0. That is , $E(Y_i|X_i = 0) = \beta_0$.
- When $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$, the conditional mean of Y_i given $X_i = 1$ is $\beta_0 + \beta_1$. That is , $E(Y_i|X_i = 1) = \beta_0 + \beta_1$.

So:

$$\begin{aligned}\beta_1 &= E(Y_i|X_i = 1) - E(Y_i|X_i = 0) \\ &= \text{population difference in group means.}\end{aligned}$$

Construct seasonal dummy variables:

- Use four seasons as an example,

$$
\begin{aligned}
D_1 &= (1,0,0,0,1,0,0,0,1,0,0,0,\cdots)\\
D_2 &= (0,1,0,0,0,1,0,0,0,1,0,0,\cdots)\\
D_3 &= (0,0,1,0,0,0,1,0,0,0,1,0,\cdots)\\
D_4 &= (0,0,0,1,0,0,0,1,0,0,0,1,\cdots).
\end{aligned}
$$

- D_1 indicates whether we're in the first quarter and so on.
- At any given time, we can be in only one of the four quarters, so one seasonal dummy is 1, and all others are zero.

Pure seasonal dummy model

The pure seasonal dummy model is

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t.$$

- We're just regressing on an intercept, but allow for a different intercept in each season.
- Those different intercepts are called the **seasonal factors**; they summarize the seasonal pattern over the year.
- In the absence of seasonality, we can drop all the seasonal dummies and instead simply include an intercept in the usual way.

Dummy variable trap

Instead of including a full set of **s** seasonal dummies, we can include any **(s-1)** seasonal dummies and an intercept.

- the constant term is the intercept for the omitted season.
- the coefficients on the seasonal dummies give the seasonal increase or decrease relative to the omitted season.

In no case, however, should we include s seasonal dummies and an intercept.

- cause the issue of perfect multicollinearity (dummy variable trap) if we do so.

Trend and seasonal model

Trend may be included as well, in which case the model is

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t.$$

- Generalize what we did in modeling trend.
- The idea of seasonality may be extended to allow for more general calendar effects, such as holiday variation and trading-day variation.

Holiday variation

Holiday variation refers to the fact that some holidays' dates change over time.

- Arrive at approximately the same time each year, the exact dates differ. (Easter)
- The behavior of many series depends in part on the timing of such holidays.
- As with seasonality, holiday effects may be handled with dummy variables.

Example: In a monthly model, in addition to a full set of seasonal dummies, we might include an "Easter dummy," which is 1 if the month contains Easter and 0 otherwise.

Trading-day variation

Trading-day variation refers to the fact that different months contain different numbers of trading days or business days.

Example: In a monthly forecasting model of volume traded on the London Stock Exchange, in addition to a full set of seasonal dummies, we might include a trading day variable, whose value each month is the number of trading days that month.

Complete model

Allowing for the possibility of holiday or trading day variation gives the complete model

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{v_1} \delta_i^{HD} \mathrm{HDV}_{it} + \sum_{i=1}^{v_2} \delta_i^{TD} \mathrm{TDV}_{it} + \varepsilon_t.$$

- The HDVs are the relevant holiday variables (there are v_1 of them).
- The TDVs are the relevant trading day variables (there are v_2 of them).
- This is just a standard regression equation and can be estimated by ordinary least squares.

We consider constructing an h-step-ahead point forecast, $y_{T+h,T}$, at time T:

$$\mathrm{y_{T+h,T}} = \beta_1 \mathrm{TIME_{T+h}} + \sum_{\mathrm{i=1}}^{\mathrm{s}} \gamma_{\mathrm{i}} \mathrm{D_{i,T+h}} + \sum_{\mathrm{i=1}}^{\mathrm{v_1}} \delta_{\mathrm{i}}^{\mathrm{HD}} \mathrm{HDV_{it}} + \sum_{\mathrm{i=1}}^{\mathrm{v_2}} \delta_{\mathrm{i}}^{\mathrm{TD}} \mathrm{TDV_{it}}.$$

Interval forecast: $\mathrm{y_{T+h,T}} \pm 1.96\sigma$, assuming $\varepsilon_t \sim N(0, \sigma^2)$.
Density forecast: $N(\mathrm{y_{T+h,T}}, \sigma^2)$.

To make the point forecast operational, we can replace the unknown parameters with estimates:

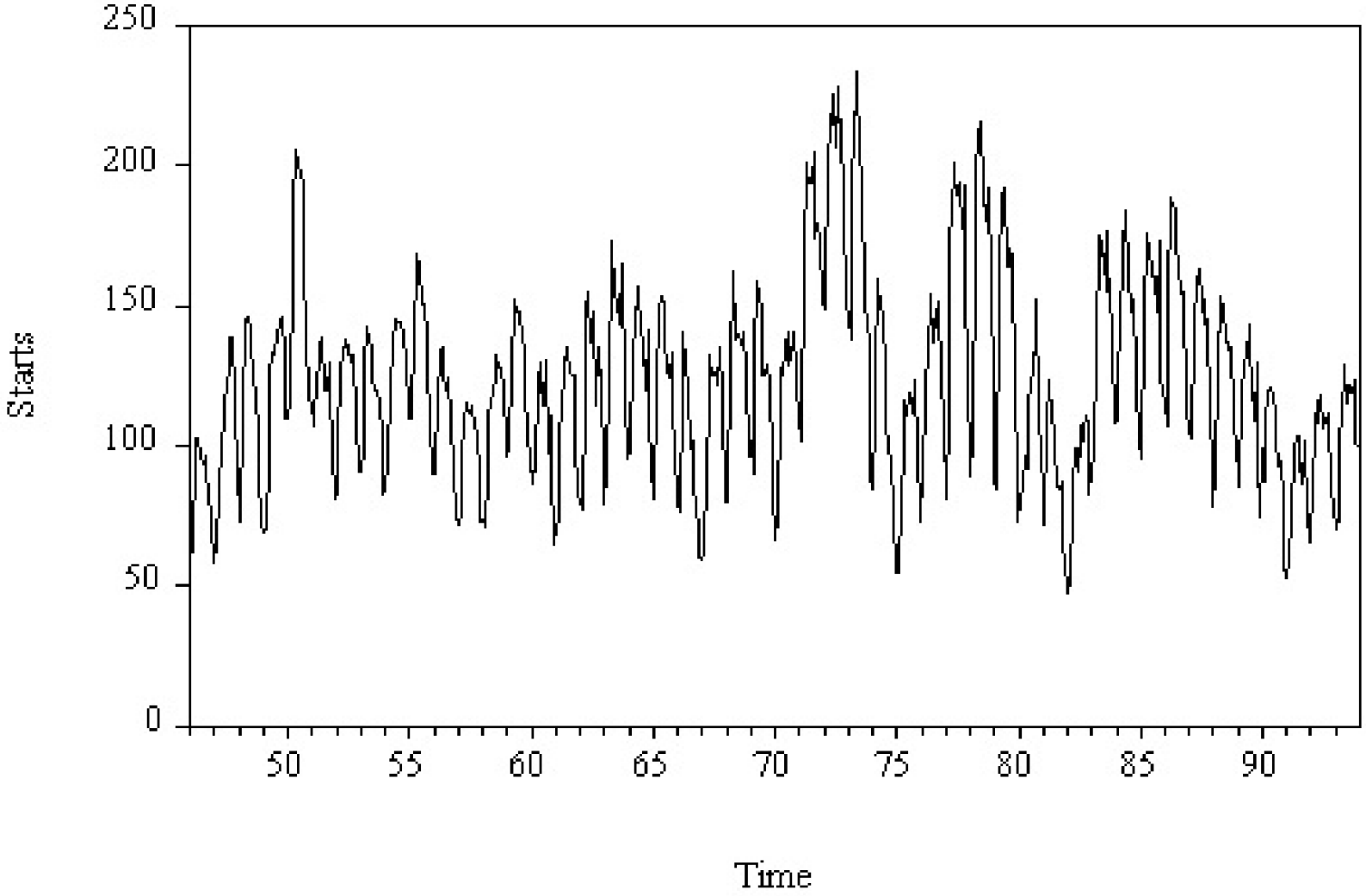
$$\hat{y}_{T+h,T} = \hat{\beta}_1 \text{TIME}_{T+h} + \sum_{i=1}^{s} \hat{\gamma}_i D_{i,T+h} + \sum_{i=1}^{v_1} \hat{\delta}_i^{HD} \text{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \hat{\delta}_i^{TD} \text{TDV}_{i,T+h}.$$

Interval forecast: $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}$.
Density forecast: $N(\hat{y}_{T+h,T}, \hat{\sigma}^2)$.

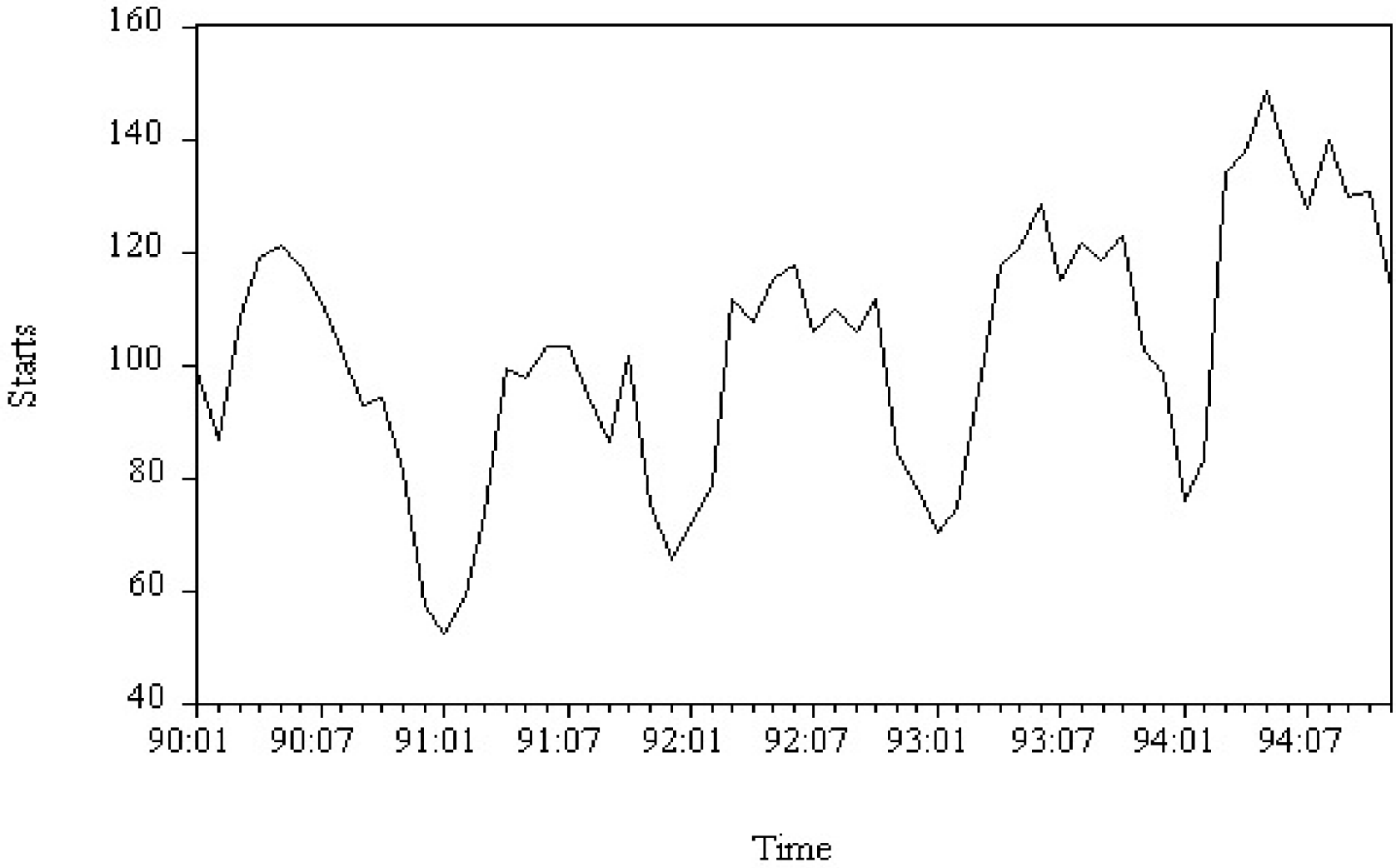
Housing starts are **seasonal** because it's usually preferable to start houses in the spring, so that they're completed before winter arrives.

Date:
Monthly data on U.S. housing starts 1946.01 - 1994.11;
Use the 1946.01 - 1993.12 period for estimation;
1994.01 - 1994.11 period for out-of-sample forecasting.

- Housing Starts, 1946.01 - 1994.11



- Housing Starts, 1990.01 - 1994.11



The figures reveal that there is no trend, so we'll work with the pure seasonal model,

$$\mathbf{y_t} = \sum_{\mathbf{i}=1}^{\mathbf{s}} \gamma_{\mathbf{i}} \mathbf{D}_{\mathrm{it}} + \boldsymbol{\varepsilon}_{\mathbf{t}}.$$

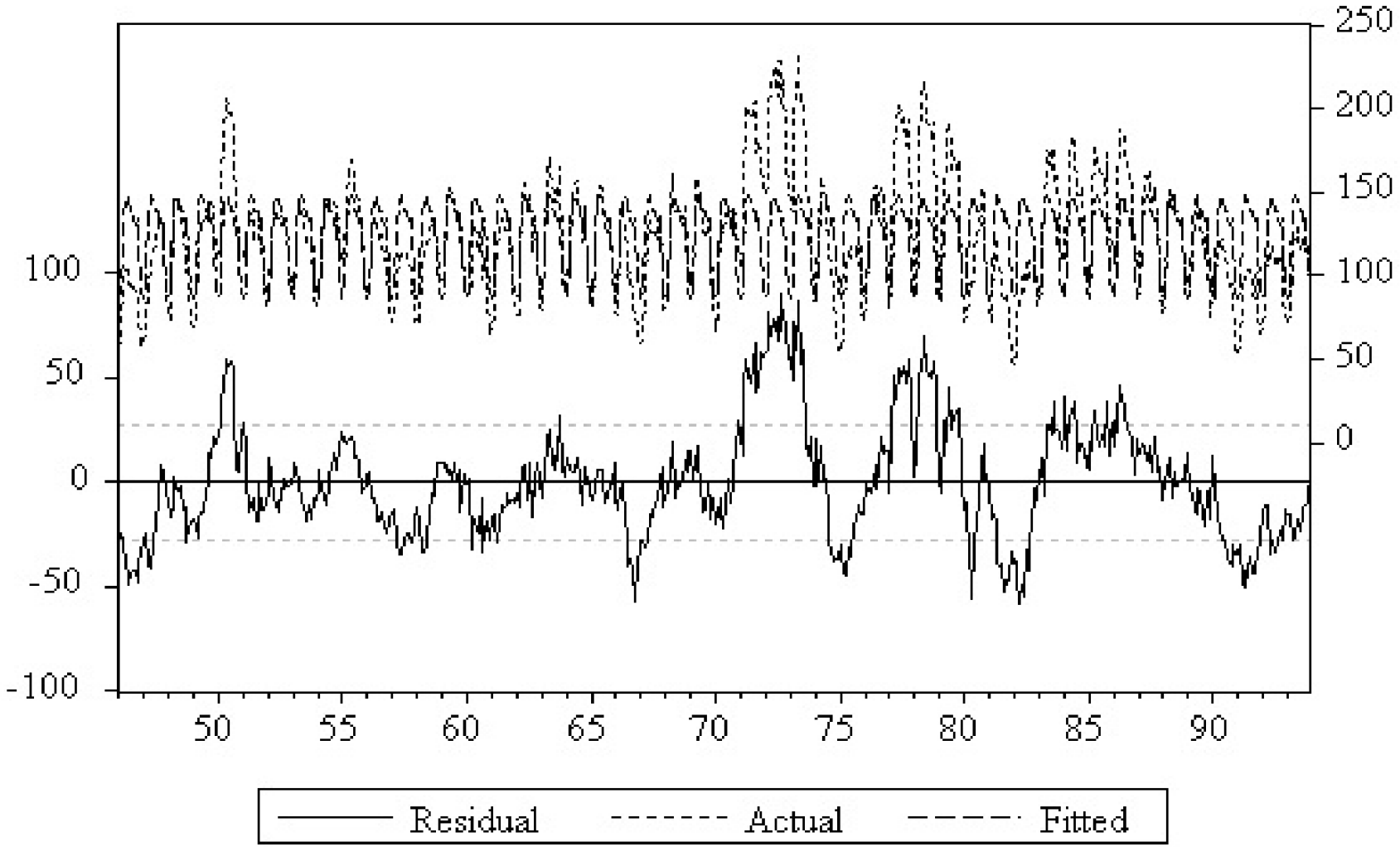
Here we let $s = 12$, which means we have twelve seasonal dummies in the forecasting model.

LS // Dependent Variable is STARTS
Sample: 1946:01 1993:12
Included observations: 576

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D1	86.50417	4.029055	21.47009	0.0000
D2	89.50417	4.029055	22.21468	0.0000
D3	122.8833	4.029055	30.49929	0.0000
D4	142.1687	4.029055	35.28588	0.0000
D5	147.5000	4.029055	36.60908	0.0000
D6	145.9979	4.029055	36.23627	0.0000
D7	139.1125	4.029055	34.52733	0.0000
D8	138.4167	4.029055	34.35462	0.0000
D9	130.5625	4.029055	32.40524	0.0000
D10	134.0917	4.029055	33.28117	0.0000
D11	111.8333	4.029055	27.75671	0.0000
D12	92.15833	4.029055	22.87344	0.0000

R-squared	0.383780	Mean dependent var	123.3944
Adjusted R-squared	0.371762	S.D. dependent var	35.21775
S.E. of regression	27.91411	Akaike info criterion	6.678878
Sum squared resid	439467.5	Schwarz criterion	6.769630
Log likelihood	-2728.825	F-statistic	31.93250
Durbin-Watson stat	0.154140	Prob(F-statistic)	0.000000

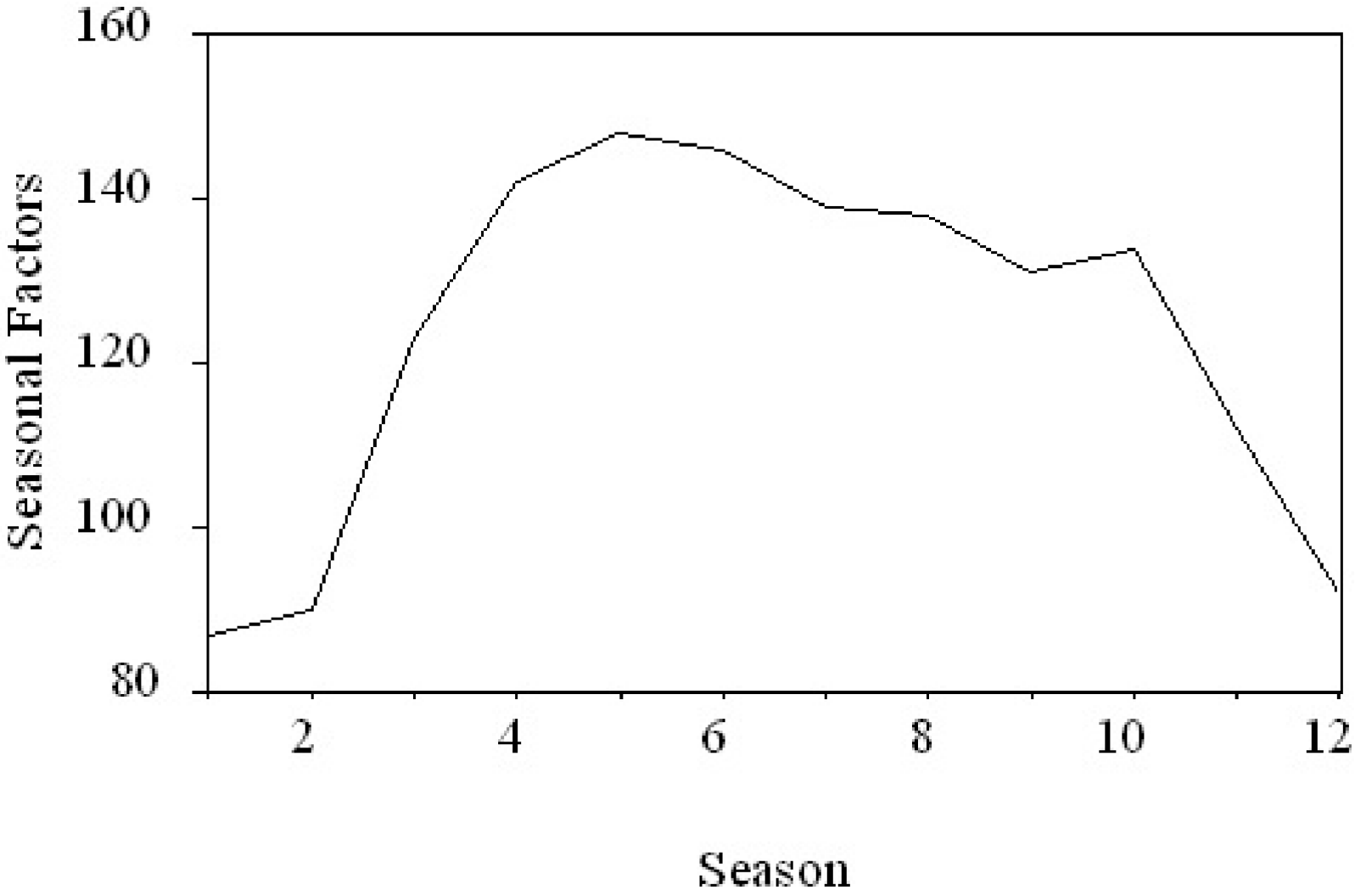
Residual plot



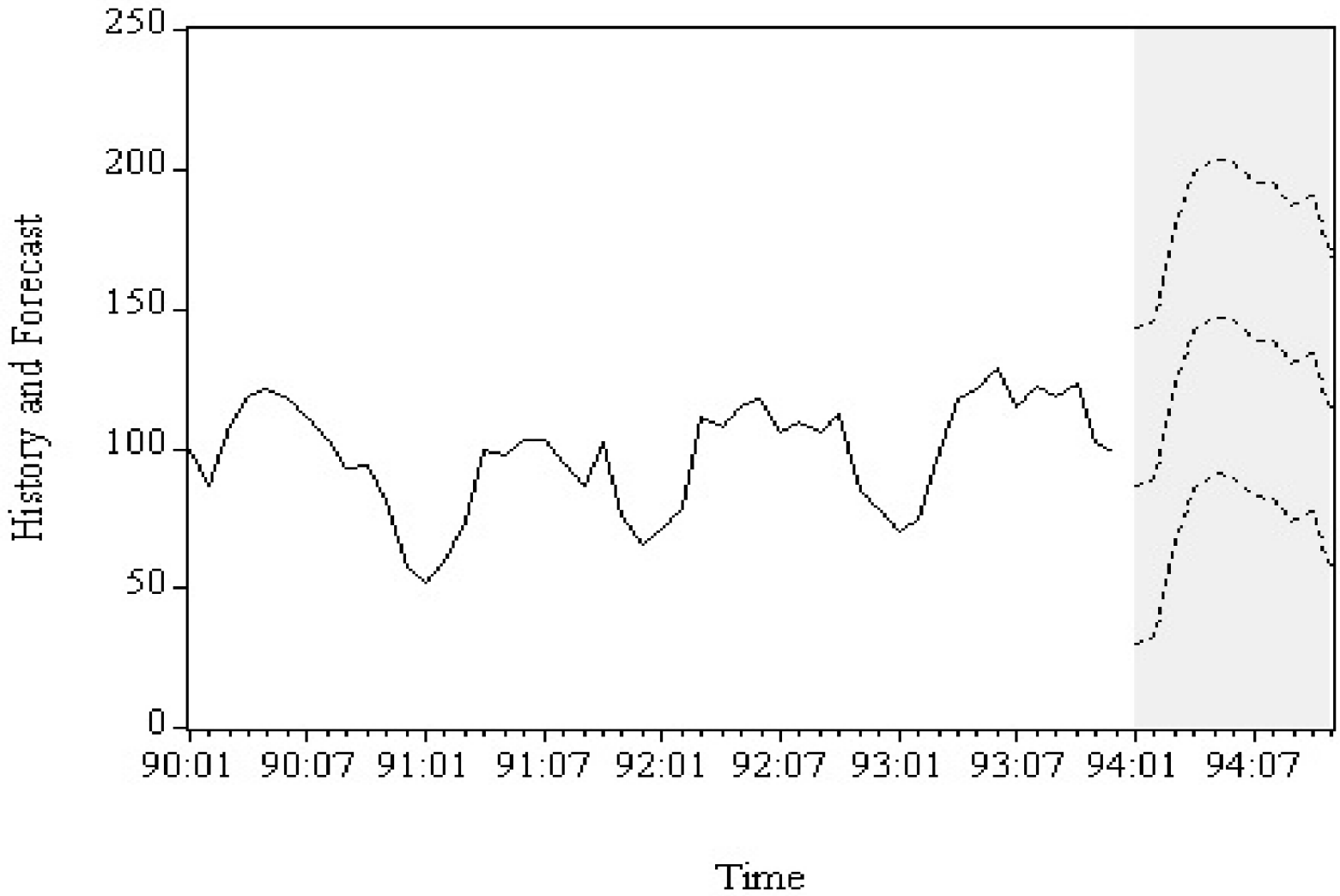
In summary

- The twelve seasonal dummies account for more than a third of the variation in housing starts, as adjusted $R^2 = .371$.
- At least some of the remaining variation is cyclical, which the model is not designed to capture. (Very low Durbin-Watson statistic.)
- Rigid seasonal pattern (there's nothing in the model other than deterministic seasonal dummies) picks up a lot of the variation in housing starts.

- Estimated Seasonal Factors for Housing Starts
  (Twelve estimated coefficients)



- Housing Starts
  History, 1990.01-1993.12; Forecast, 1994.01-1994.11



- Housing Starts
  History, 1990.01-1993.12; Forecast, 1994.01-1994.11;
  Realization, 1994.01-1994.11;