Lecture 9

Time Series Regression

Arnab Hazra

Classical regression

- Suppose $\{X_t, t = 1, \dots, \}$ is being influenced by a collection of possible inputs or independent series, say, $Z_{t1}, Z_{t2}, \dots, Z_{tq}$.

- We first regard the inputs as fixed and known, i.e., say the realizations are $z_{t1}, z_{t2}, \dots, z_{tq}$ and we build the model conditionally.

- We express this relation through the linear regression model

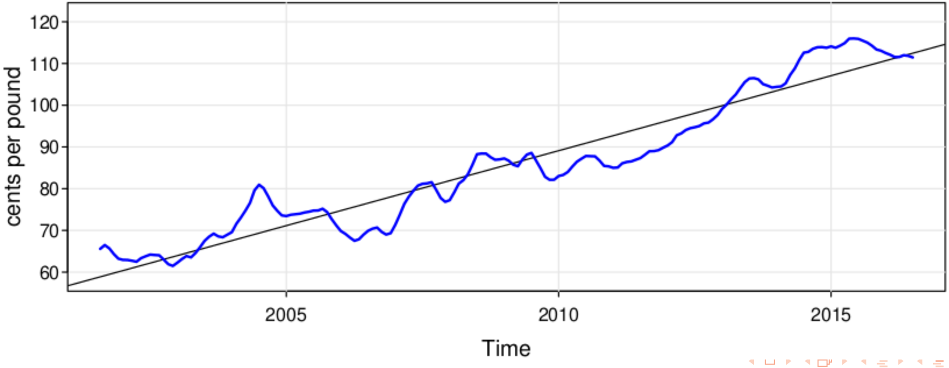
$$X_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_q z_{tq} + W_t,$$

where we assume $W_t \overset{IID}{\sim} \mathrm{Normal}(0, \sigma_W^2)$.

Example

- Consider the monthly price (per pound) of a chicken in the US from mid-2001 to mid-2016 (180 months).
- We might fit the model

$$X_t = \beta_0 + \beta_1 z_t + W_t, \quad z_t = 2001\frac{7}{12}, 2001\frac{8}{12}, \ldots, 2016\frac{6}{12}, \; W_t \overset{IID}{\sim} \mathrm{Normal}(0, \sigma_W^2)$$



Example: Inference

- Suppose the realizations are $x_1, \ldots, x_T$.
- In OLS, we minimize the error sum of squares

$$Q = \sum_{t=1}^{T} (x_t - \beta_0 - \beta_1 z_t)^2$$

- The estimators are

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{T} (x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^{T} (z_t - \bar{z})^2}, \quad \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z}.$$

- Using R, the estimated slope coefficient is $\hat{\beta}_1 = 3.59$ (with a standard error of 0.08) yielding a significant estimated increase of about 3.59 cents per year.

Classical regression: generic notations

- We can rewrite the linear regression model as

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \ldots + \beta_q z_{tq} + w_t = \boldsymbol{z}_t' \boldsymbol{\beta} + w_t,$$

where $\boldsymbol{z}_t = (1, z_{t1}, \ldots, z_{tq})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$.

- In OLS, we minimize the error sum of squares $Q = \sum_{t=1}^{T} (x_t - \boldsymbol{z}_t' \boldsymbol{\beta})^2$.
- The normal equation is given by

$$\left[\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' \right] \hat{\boldsymbol{\beta}} = \sum_{t=1}^{T} \boldsymbol{z}_t x_t.$$

- If $\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t'$ is nonsingular, $\hat{\boldsymbol{\beta}} = \left[\sum_{t=1}^{T} \boldsymbol{z}_t \boldsymbol{z}_t' \right]^{-1} \left[\sum_{t=1}^{T} \boldsymbol{z}_t x_t \right]$.

Classical regression: generic notations

- $SSE = ?$
- $\mathrm{Cov}(\hat{\beta}) = ?$
- What is an unbiased estimator for the variance σ_W^2?
- What is the test statistic for checking the significance of β_i?

Classical regression: Model selection

- Various competing models are often of interest to isolate or select the best subset of independent variables.
- Suppose a proposed model specifies that only a subset $r < q$ covariates, say, $Z_{t,1:r} = \{Z_{t1}, Z_{t2}, \ldots, Z_{tr}\}$ is influencing the response X_t. The reduced model is

$$X_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \ldots + \beta_r z_{tr} + W_t$$

- Whether the remaining variables are important predictors or not can be determined by

$$H_0 : \beta_{r+1} = \ldots = \beta_q = 0.$$

- We can do that using the F-test as

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(T-q-1)} = \frac{MSR}{MSE},$$

where SSE_r is the error sum of squares under the reduced model.

ANOVA table

Table 2.1. Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square	F
$z_{t,r+1:q}$	$q-r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q-r)$	$F = \frac{MSR}{MSE}$
Error	$\top - (q+1)$	SSE	$MSE = SSE/(\top - q - 1)$	

Your paragraph text

Model selection: Information criteria

- Suppose we consider a normal regression model with k coefficients and denote the maximum likelihood estimator for the variance as

$$\hat{\sigma}_k^2 = \frac{SSE(k)}{T}$$

  where $SSE(k)$ denotes the residual sum of squares under the model with k regression coefficients.

- Akaike's Information Criterion (AIC): $AIC = \log(\hat{\sigma}_k^2) + \frac{T+2k}{T}$
- Bias Corrected AIC (AICc): $AICc = \log(\hat{\sigma}_k^2) + \frac{T+k}{T-k-2}$
- Bayesian Information Criterion (BIC): $AIC = \log(\hat{\sigma}_k^2) + \frac{k\log(T)}{T}$

Pollution, Temperature, and Mortality

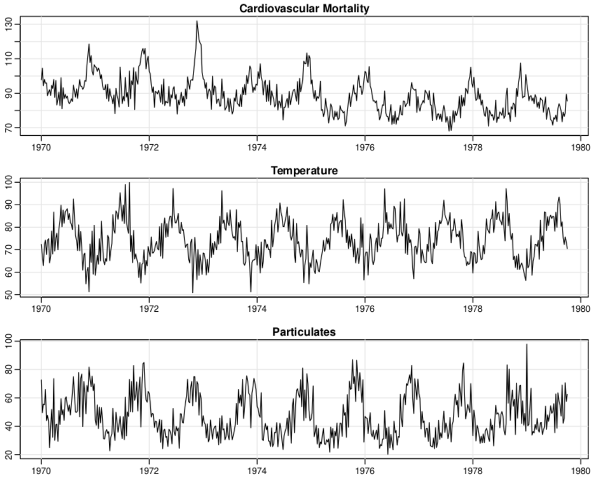


Fig. 2.2. *Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.*

Pollution, Temperature, and Mortality

- M_t denotes cardiovascular mortality, T_t denotes temperature and P_t denotes the particulate levels.

- Four possible models are
  - $M_t = \beta_0 + \beta_1 t + W_t$
  - $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T_\cdot) + W_t$
  - $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T_\cdot) + \beta_3 (T_t - T_\cdot)^2 + W_t$
  - $M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T_\cdot) + \beta_3 (T_t - T_\cdot)^2 + \beta_4 P_t + W_t$

Summary statistics

k	SSE	df	MSE	R^2	AIC	BIC
2	40,020	506	79.0	.21	5.38	5.40
3	31,413	505	62.2	.38	5.14	5.17
4	27,985	504	55.5	.45	5.03	5.07
5	20,508	503	40.8	.60	4.72	4.77

- A model with only trend could be compared to the full model, $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, using $q = 4, r = 1, n = 508$, and we have

$$F_{3,503} = \frac{(40,020 - 20,508)/3}{20,508/503} = 160$$

which exceeds $F_{3,503}(.001) = 5.51$.

Regression With Lagged Variables

- We have seen Southern Oscillation Index (SOI) measured at time $t - 6$ months is associated with the Recruitment series at time t.
- Consider the following regression,

$$R_t = \beta_0 + \beta_1 S_{t-6} + W_t$$

- The fitted model is

$$\hat{R}_t = 65.79 - 44.28_{(2.78)} S_{t-6}$$

with $\hat{\sigma}_W = 22.5$ on 445 degrees of freedom.

Thank you!