Lecture 5

Sample ACF, Diagnostics
Textbook Sections: 1.4, 1.6

Sample Estimates

Given a set of observations $x_1, x_2, \cdots, x_n$, we can obtain estimates of the autocovariance and autocorrelation functions. We estimate $E(X_t)$ by the sample mean

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

We estimate $\gamma(h)$ by the **sample autocovariance function**

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(x_t - \bar{x})(x_{t+h} - \bar{x})$$

for $h = 0, 1, 2, \cdots, n-1$.

We estimate $\rho(h)$ by the **sample autocorrelation function**

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$$

for $h = 0, 1, 2, \cdots, n-1$.

Recall that $\gamma(h) = \gamma(-h)$, so we can use the above estimates for $h = -1, -2, \cdots, -(n-1)$ as well.

Since these statistics are based on a random sample, they are random variables. Later on, we will take a closer look at their distributions, and make confidence intervals for the parameters.

In R, the sample ACVF and ACF values can be obtained with the command `acf()`.

Sample Autocorrelation

If we assume that the data comes from a sequence of normally distributed random variables, then it can be shown that $\hat{\rho}(h)$ is approximately normally distributed with mean $\rho(h)$. In general, the variance has a complicated distribution. However, if $\rho(h) = 0$ (the true autocorrelation value at lag h is 0), then the variance of $\hat{\rho}(h)$ is approximately $1/n$.

We can use this to conduct a hypothesis test with

$$H_0 : \rho(h) = 0, \qquad H_1 : \rho(h) \neq 0.$$

Under H_0, the test statistic $z^* = \sqrt{n}\hat{\rho}(h)$ is distributed $N(0, 1)$.
Using significance level $\alpha = 0.05$, we can reject H_0 if $|z^*| > 1.96$, or if $|\hat{\rho}(h)| > 1.96/\sqrt{n}$.

It is very useful to plot the sample ACF values in order to see if the data exhibits any dependence structure at different lags. It is also common to include horizontal lines at $\pm 1.96/\sqrt{n}$ when making a sample ACF plot. This way, we can quickly see if $\hat{\rho}(h)$ is outside those bounds for any h.

Diagnostics

1. **Unequal Variance.**
   Plot the data against time to check for evidence of unequal variance.

2. **Presence of a Trend.**
   Plot the data against time to check for presence of a trend. There are formal tests for this (such as the rank test), but we won't employ them in this class.

3. **IID Check.**
   Once we have a sequence that seems roughly stationary, it's useful to check whether it is significantly different from i.i.d. data. If we conclude the data is not i.i.d., then we will proceed with fitting a time series model to it.

   - Look at the sample ACF plot with lines at $\pm 1.96/\sqrt{n}$. If for some $h = 1, 2, \cdots$, the value of $\rho(\hat{h})$ is outside those lines, then we have evidence that $\rho(h)$ is significantly different from 0. Thus the sequence has dependence at lag h, and is not i.i.d.
   - For a chosen value of h, we can conduct a hypothesis test with

     $$H_0 : \rho(1) = \cdots = \rho(h) = 0$$
     $$H_1 : \text{at least one of } \rho(1), \cdots, \rho(h) \text{ is nonzero.}$$

     The portmanteau test uses the statistic

     $$Q = n \sum_{j=1}^{h} \hat{\rho}(j)^2,$$

     and the Ljung-Box test uses the statistic

     $$Q_{LB} = n(n+2) \sum_{j=1}^{h} \hat{\rho}(j)^2/(n-j).$$

     In each case, the null distribution of the statistic is approximately Chi-squared with h degrees of freedom, but the approximation is more precise for the Ljung-Box test. This distribution can be used to make conclusions at a specified significance level α. In general, the larger the value of the statistic, the more evidence we have that the sample autocorrelations are too large for the data to have come from an i.i.d. sequence.
   - A few more tests are included in section 1.6 of the textbook.

4. **Normality.**
   We can visually assess this using a histogram or a normal probability plot. We can also use a formal testing procedure, such as the Shapiro-Wilk test.