Indicator Variables

In class we called this a "nonparametric" seasonality method. We can pursue it that way, by simply taking the average of each $y_t, y_{t+d}, y_{t+2d}, ...$ for some period d. But it will be simpler for us to use indicator functions. For logical condition c,

$$I(c) = \begin{cases} 1 & \text{if } c \text{ is true} \\ 0 & \text{if } c \text{ is false} \end{cases}.$$

For example, if we wanted to indicate which month is January,

$$I(January) = \begin{cases} 1 & \text{if the month is January} \\ 0 & \text{if } c \text{ is false} \end{cases}.$$

Here we'll look at problem 2.1 in the book, slightly adjusted. For the Johnson & Johnson data (`JohnsonJohnson` in `R`), time t is in quarters $(1960.00, 1960.25, \ldots)$ so one unit of time is a year. Here we'll look at log-transformed earnings, i.e., $\log(Y_t)$.

1. Fit the regression model

$$\log(Y_t) = \beta t + \alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t) + W_t.$$

   where βt is the linear trend, and $Q_i(t) = 1$ if time t corresponds to quarter i = 1, 2, 3, 4, and zero otherwise. The $Q_i(t)$'s are indicator variables. We will assume for now that w_t is a Gaussian white noise sequence.

2. If the model is correct, what is the estimated average annual increase in the logged earnings per share?

3. If the model is correct, does the average logged earnings rate increase or decrease from the third quarter to the fourth quarter?

4. What happens if you do or do not include an intercept term in the model in part 1?

5. Graph the data and superimpose the fitted values on the graph. Examine the residuals and state your conclusions. Does it appear that the model fits the data well (do the residuals look white)?

Review of Sinusoids

Because seasonal trends repeat themselves after some period d, trying to model them via polynomial regression doesn't make sense. Instead, we can try modelling seasonal trends as a linear combination waves - or sinusoids - of period d.

We define the set of sinusoid functions as

$$\{g(t) = R\cos(2\pi ft + \Phi) \ : \ R \in R_+, \ f \in R_+, \ \Phi \in [0, 2\pi/f)\},$$

where

- R is called the **amplitude**
- f is called the **frequency**
- Φ is called the **phase**
- $1/f = d$ is called the **period**

Note: Replacing the cosine with sine in the above equation will also give sinusoids too, but with different parameter of the phase shift.

6. In R, plot a sinusoid with $R = 1, f = 1, \Phi = 0$.

7. Try modifying R, f, Φ one at a time. What happens to the sinusoid when you change each parameter?

8. In lecture, we discussed that the sinusoid above is difficult to use in linear models (linear models do great at estimating coefficients like the amplitude, but not values inside the cosine like the frequency and phase). Frequency domain methods will help us choose the frequency, and we can rewrite the sinusoid to remove the phase shift by adding another linear coefficient. Let $A = R\cos(\Phi)$ and $B = -R\sin(\Phi)$. Show that the above expression for a sinusoid is equivalent to

$$\{g(t) = A\cos(2\pi ft) + B\sin(2\pi ft) \ : \ A, B \in R, \ f \in R_+\}.$$

9. We're pathologists trying to model lung disease deaths over time, so we can strategically allocate resources and staff to critical periods in the future. We know that lung disease deaths follow yearly trend, so we want to fit some combination of sinusoids of the appropriate period.

   For this exercise, we'll be working the `ldeaths` dataset in R, which contains the number of deaths from lung disease per month in the UK. The goal is to fit a linear combination of sinusoids to the data. Here are some things to consider:

   - What is the period of the `ldeaths` time series? We want to fit sinusoids of the matching period.
   - We must construct the sinusoids before passing them into the `lm` function.
   - What parameters should we vary when constructing the sinusoids? Which ones matter, and which ones don't?