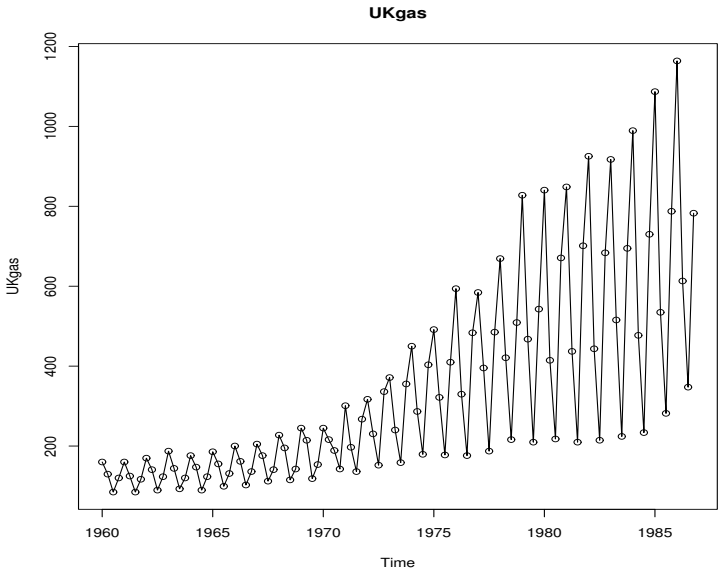
Eco311: Predict Quarterly UK Gas Consumption

(Jing Li, Miami University)

1. This note shows how to build step by step time series model for Quarterly UK gas consumption in millions of therms from 1960Q1 to 1986Q4

2. A highly recommended practice for time series analysis is visualizing the data before conducting any analysis, since the specification of model is based on the patterns shown in the graph. The time series plot for original data is



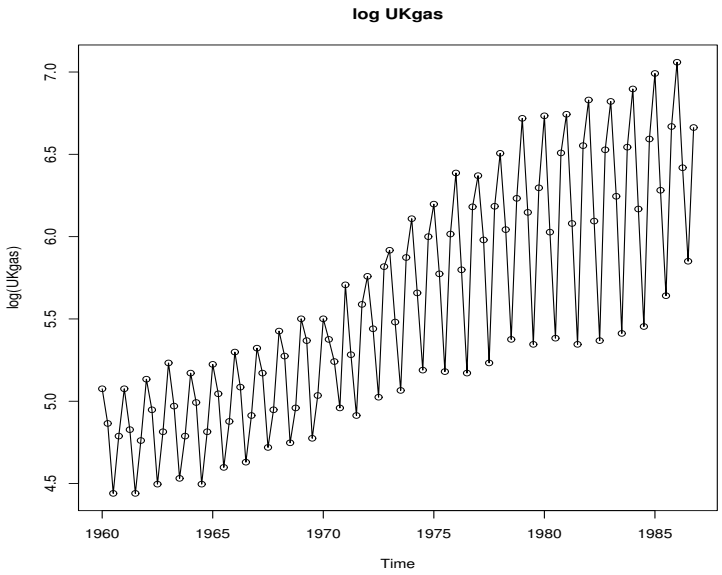
   There are three noticeable patterns:

   (a) There is an upward trend that appears to be exponential, partly because of growing population or rising living standard

   (b) There is seasonality—in a given year, gas consumption is at the highest level in Q1, while lowest in Q3. This is expected because 80% of UK homes use gas for space heating and hot water

   (c) The variation within a year increases over time (heteroskedasticity)

3. The finding of exponential trend and heteroskedasticity suggests taking log of the series



After taking log we see the variation is more uniform (homoskedasticity) than before. Moreover, the trend now becomes approximately linear

4. So the first model we try is log trend model

$$\log(UKgas) = \beta_0 + \beta_1 t + u \Rightarrow 100 \times \beta_1 = \frac{100 \times d\log(UKgas)}{dt} = percent\ growth\ rate \tag{1}$$

```
> data("UKgas")
> ts.plot(UKgas,type="o",main="UKgas")
> ts.plot(log(UKgas),type="o",main="log UKgas")
> y = log(UKgas)
> n = length(UKgas)
> trend = 1:n
> m1 = lm(y~trend)
> summary(m1)$coef
              Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 4.60416665 0.077901440 59.10246 5.722180e-83
trend       0.01788662 0.001240732 14.41618 9.826612e-27
> summary(m1)$r.squared
[1] 0.6622336
```

The slope coefficient 0.01788662 multiplied by 100 indicates that the gas consumption

tends to grow by 1.78 percent every quarter (or grow by $1.78 \times 4 = 7.12$ percent every year). The trend alone can explain 66 percent variation of log gas consumption. The R squared is far from one, so there is room for improvement by accounting for other patterns such as seasonality.

5. Next, we create a categorical variable "quarter" that equals one for Q1, two for Q2, three for Q3, and four for Q4. It is categorical (not numerical) because we can not say $1 + 2 = 3$. In other words, if we mistakenly treat quarter as numerical variable, we implicitly assume the change in gas consumption between Q2 and Q1 is the same as between Q3 and Q2, and between Q4 and Q3. This restriction of constant marginal effect is invalid for this problem.

6. The proper way is to use a set of dummy variables to represent each quarter, and run DVR

```
> quarter = rep(1:4,27)
> D2 = as.numeric(quarter==2)
> D3 = as.numeric(quarter==3)
> D4 = as.numeric(quarter==4)
> summary(lm(y~trend+D2+D3+D4))$coef
              Estimate   Std. Error    t value      Pr(>|t|)
(Intercept)  5.0304422 0.0472188997 106.534506 3.341535e-107
trend        0.0180918 0.0005772298  31.342455  1.757800e-54
D2          -0.4205932 0.0508694131  -8.268097  5.041798e-13
D3          -0.9858963 0.0508792371 -19.377183  3.929468e-36
D4          -0.3433413 0.0508956063  -6.745991  9.085014e-10
>
> m2 = lm(y~trend+factor(quarter))
> summary(m2)$coef
                   Estimate   Std. Error    t value      Pr(>|t|)
(Intercept)       5.0304422 0.0472188997 106.534506 3.341535e-107
trend             0.0180918 0.0005772298  31.342455  1.757800e-54
factor(quarter)2 -0.4205932 0.0508694131  -8.268097  5.041798e-13
factor(quarter)3 -0.9858963 0.0508792371 -19.377183  3.929468e-36
factor(quarter)4 -0.3433413 0.0508956063  -6.745991  9.085014e-10
> summary(m2)$r.squared
```

```
[1] 0.9290536
```

(a) We do not include the dummy for Q1 to avoid dummy variable trap. Q1 (the base group) is represented by intercept 5.0304422.

(b) The coefficient of D2 -0.4205932 is the difference between Q2 and Q1, and is significant (t value is -8.268097)

(c) The coefficients of D2, D3, and D4 confirm seasonality—Q1 sees highest consumption, while Q3 sees lowest consumption.

(d) Instead of creating dummy variables, we can use factor function as a shortcut and obtain the same results

(e) Trend combined with quarterly dummy variables explain 92.9 percent of variation of log gas consumption. Because the R squared is close to one, the room for further improvement is small

7. The sample ends at 1986Q4. Next we can predict log gas consumption at 1987Q1 and 1987Q2

```
> predict(m2,data.frame(trend=(n+1),quarter=1))
7.002448
> predict(m2,data.frame(trend=(n+2),quarter=2))
6.599947
> co = coef(m2)
> co[1]+co[2]*(n+1)
7.002448
> co[1]+co[2]*(n+2) + co[3]
6.599947
```

8. Due to the dependent nature of time series, next we include the first lag of log consumption as additional regressor

```
> ylag1 = c(NA, y[1:(n-1)])
> m3 = lm(y~trend+factor(quarter)+ylag1)
> summary(m3)$coef
                   Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)       4.5254617 0.464427891   9.744164 3.286156e-16
```

```
trend             0.0161578 0.001882485   8.583226 1.165582e-13
factor(quarter)2 -0.4566200 0.061781536  -7.390882 4.332818e-11
factor(quarter)3 -0.9766131 0.052102328 -18.744136 1.173512e-34
factor(quarter)4 -0.2731534 0.081775588  -3.340280 1.173685e-03
ylag1             0.1077659 0.098920299   1.089422 2.785604e-01
> summary(m3)$r.squared
[1] 0.929538
```

The R squared barely changes, and the t value of lag y 1.089422 is less than 1.96. Both indicate that it is unnecessary to include the new regressor. Anyway, for teaching purpose, the forecasts for 1987Q1 and 1987Q2 based on the new regression are

```
> yhat = predict(m3,data.frame(trend=(n+1),quarter=1,ylag1=y[n]))
> yhat
7.004693
> yhat = predict(m3,data.frame(trend=(n+2),quarter=2,ylag1=yhat))
> yhat
6.601066
```

The forecasts are almost the same as those without using ylag1

9. But wait! Maybe Q1 this year is related to Q1 last year more than Q4 last year. How about we include the fourth lag?

```
> ylag4 = c(NA, NA, NA, NA, y[1:(n-4)])
> m4 = lm(y~trend+factor(quarter)+ylag4)
> summary(m4)$coef
                    Estimate Std. Error   t value     Pr(>|t|)
(Intercept)       0.95376237 0.27076689  3.522448 6.510965e-04
trend             0.00358263 0.00104261  3.436214 8.667869e-04
factor(quarter)2 -0.09119084 0.03619562 -2.519389 1.337289e-02
factor(quarter)3 -0.19898501 0.06025074 -3.302615 1.337744e-03
factor(quarter)4 -0.06641827 0.03392157 -1.957995 5.307420e-02
ylag4             0.81935384 0.05441864 15.056493 3.082867e-27
> summary(m4)$r.squared
[1] 0.9785069
> predict(m4,data.frame(trend=(n+1),quarter=1,ylag4=y[n-3]))
```

```
7.128523
> predict(m4,data.frame(trend=(n+2),quarter=2,ylag4=y[n-2]))
6.515706
```

The fourth lag is statistically significant, and R squared rises to 0.9785069. The two predicted values change accordingly. This exercise highlights the importance of using common sense in data analysis.

Optional Advanced Stuff

10. How to predict y based on a model for $\log(y)$? Mathematically, the prediction is based on mean value

$$E(y) = E\left(e^{\log(y)}\right) = e^{\widehat{\log(y)}+\frac{\sigma^2}{2}} \tag{2}$$

where σ^2 is the variance of residual in the model for $\log(y)$. Thus, the predicted gas consumption based on the last model is

```
> exp(7.128523+summary(m4)$sigma^2/2)
[1] 1253.605
> exp(6.515706+summary(m4)$sigma^2/2)
[1] 679.2315
```

They are greater than the incorrect forecasts that fail to account for $\frac{\sigma^2}{2}$.

11. Other than using quarterly dummies to capture seasonality, we can try seasonal AR(1) model

```
> dy4 = diff(y,4)
> dy4lag = c(NA,dy4[1:(length(dy4)-1)])
> summary(lm(dy4~dy4lag))$coef
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  0.07810227 0.01224579  6.377887 5.461458e-09
dy4lag      -0.17908370 0.09793213 -1.828651 7.040303e-02
```

where the seasonal (quarter-to-quarter) difference is $\Delta_4 y_t \equiv y_t - y_{t-4}$, and AR(1) model is fitted for the seasonal difference $\Delta_4 y_t = \beta_0 + \beta_1 \Delta_4 y_{t-1}$. The estimated model is

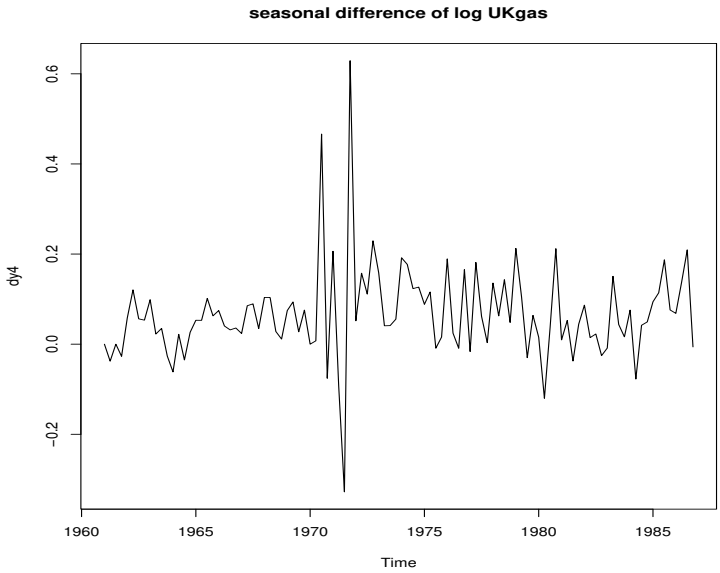
$$(y_t - y_{t-4}) = 0.07810227 - 0.17908370(y_{t-1} - y_{t-5}) \tag{3}$$

or equivalently

$$y_t = 0.07810227 - 0.17908370y_{t-1} + y_{t-4} + 0.17908370y_{t-5} \tag{4}$$

The forecasts can be computed based on (4)

12. Finally, plotting the seasonal difference reveals something interesting called structural break



The volatility in early 1970s is remarkable, and below is explanation from Chatgpt— "In the mid-1960s, substantial natural gas reserves were discovered in the North Sea. The UK saw this as a major energy opportunity, and a nationwide conversion program began. From 1967 to 1977, the UK carried out an ambitious plan to switch households and industry from town gas (made from coal) to natural gas. By the early 1970s, a significant portion of this conversion was complete, and natural gas demand surged."

13. The existence of structural break suggests that, to avoid apple-to-orange (town gas to natural gas) comparison, we may focus on the subsample after 1975 and ignore the subsample before 1975. We may redo our analysis using only the post-1975 subsample.