

Econometrics: ECON-3740
Creation of Quarterly Dummy Variables for Seasonality Test &
A Worked Example of an F-test in R
D Prescott, Department of Economics and Finance

Objective: Create 4 Quarterly Dummy Variables

The objective is to test the hypothesis that house prices have no seasonal pattern against the alternative that house prices vary from quarter to quarter. The test is conducted in the context of a linear model of house prices. These notes explain how quarterly dummy variables can be created in R from the date-variable “month” which records the month of the sale. The variable “month” has a range 1-72 which spans the 6 year period, 1983 to 1988. The first step is to create a variable that identifies the month of the year in which the sale took place. This variable takes integer values 1-12. Next, quarterly dummy variables are created: Q1-Q4 which take the values of 0, 1 to indicate which quarter of the year the sale took place.

The following command shows summary statistics for the variable “month”

```
> summary(month)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00  18.00   31.00   36.29  58.00   72.00
```

To convert “month” into a variable “mo1” that takes integer values 1-12 where 1 implies “January” and 12 implies “December”, calculate the remainder of the division: month/12. The possible remainders are 0-11 where 0 implies “December” and 1 implies “January”. Next, swap the value 12 for 0 so that 12 implies “December.” Two commands in R will complete this task:

```
> mo <- month - (month %/% 12)*12
> summary(mo)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   5.000   5.565   8.000  11.000
> mo1 <- mo + 12*(mo < 1)
> summary(mo1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   6.000   6.181   9.000  12.000
```

Note that the statement: **(month %/% 12)** calculates the integer part of the division of “month” by 12. Hence, **month - (month %/% 12)*12** is the remainder of the division of “month” by 12.

The statement `mo1 <- mo + 12*(mo < 1)` creates “mo1” which differs from “mo” only in that 12 replaces 0.

The next 4 commands create the quarterly dummy variables Q1-Q4:

```
q1 <- 1*(mo1 < 4)
q2 <- 1*(mo1 > 3 & mo1 < 7)
q3 <- 1*(mo1 > 6 & mo1 < 10)
q4 <- 1*(mo1 > 9)
```

To check that the dummy variables do indeed represent the quarter in which a sale took place, check observation 18.

```
> month[18]
[1] 34
> mo1[18]
[1] 10
> q4[18]
[1] 1
> q1[18]
[1] 0
> q2[18]
[1] 0
> q3[18]
[1] 0
```

The 18th observation on “month” shows the sale took place on month 34, which is October 1985. The variable “mo1” takes the value 10, which is the 10th month i.e. October. Finally, the values of Q1-Q4 show that in position 18, $Q4[18] = 1$ while the other 3 dummy variables are all zero.

The means of Q1-Q4 represent the share of house sales by quarter. For example, consider the means of Q3 and Q4:

```
> mean(q3)
[1] 0.2622585
> mean(q4)
[1] 0.2013373
```

A Sample F-Test Procedure in R

To illustrate the calculation of an F-statistic in R consider the following linear model in which the square of lot size is included along with lot size itself. In this example, the unrestricted model is:

```
> modf <- lm(price~size+lsize+ls2+dagar+sager+firep+bathp+month+m2+m3)
```

Call:

```
lm(formula = price ~ size + lsize + ls2 + dagar + sager + firep +  
    bathp + month + m2 + m3)
```

Residuals:

Min	1Q	Median	3Q	Max
-57574	-8764	-125	8740	131194

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.220e+04	2.565e+03	-4.756	2.19e-06	***
size	3.541e+01	1.417e+00	24.994	< 2e-16	***
lsize	1.564e+00	1.965e-01	7.963	3.58e-15	***
ls2	-1.733e-05	2.899e-06	-5.976	2.94e-09	***
dagar	1.790e+04	1.736e+03	10.313	< 2e-16	***
sager	9.495e+03	1.064e+03	8.928	< 2e-16	***
firep	7.523e+03	1.042e+03	7.220	8.71e-13	***
bathp	2.637e+03	2.909e+02	9.067	< 2e-16	***
month	-9.096e+02	2.086e+02	-4.360	1.40e-05	***
m2	5.185e+01	6.756e+00	7.674	3.20e-14	***
m3	-3.221e-01	6.226e-02	-5.173	2.65e-07	***

The unrestricted model is referred to as “modf”. The null hypothesis is that the true coefficients on “lsize” and its square “ls2” are both zero. Examination of the results above shows that individually, the two coefficients are both statistically significantly different from zero, so it seems very likely that the F-test for the hypothesis that both true coefficients are simultaneously zero will be rejected. Nevertheless, this hypothesis will be tested using an appropriate F-statistic.

The relevant F-statistic is:

$$F(r, n - K) = (SSRR - SSRU / SSR) * (n - K / r)$$

SSRR and *SSRU* are the sum of squared residuals from the restricted and unrestricted models. There are two degrees of freedom: *r* is the number of restrictions in the null hypothesis and $(n - K)$ is the number of observations minus the number of parameters in the unrestricted model. The information that must be stored from the unrestricted regression is: *SSRU* and $(n - K)$. The following commands store this information:

```
> SSRU <- sum(resid(modf)^2)  
> SSRU
```

```
[1] 342625938910
> df2 <- length(price) - length(coef(modf))
> df2
[1] 1335
```

Next, define and estimate the restricted model which imposes the restrictions in the null hypothesis, namely that the coefficients on lsize and ls2 are both zero. The model “modfr” (the restricted model) excludes the two variables lsize and ls2.

```
> modfr <- lm(price~size+dagar+sager+firep+bathp+month+m2+m3)
> summary(modfr)
```

Call:

```
lm(formula = price ~ size + dagar + sager + firep + bathp + month +
    m2 + m3)
```

Residuals:

```
    Min      1Q  Median      3Q     Max
-51158  -9197   -467   9079 133925
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-7.090e+03	2.545e+03	-2.786	0.00541	**
size	3.690e+01	1.438e+00	25.665	< 2e-16	***
dagar	1.980e+04	1.761e+03	11.242	< 2e-16	***
sager	1.038e+04	1.083e+03	9.582	< 2e-16	***
firep	7.841e+03	1.066e+03	7.356	3.29e-13	***
bathp	2.791e+03	2.967e+02	9.408	< 2e-16	***
month	-8.553e+02	2.133e+02	-4.010	6.41e-05	***
m2	4.961e+01	6.901e+00	7.189	1.08e-12	***
m3	-3.022e-01	6.356e-02	-4.754	2.21e-06	***

```
> SSRR <- sum(resid(modfr)^2)
> SSRR
[1] 359636385635
> r <- 2
> fstat <- ((SSRR-SSRU)/SSRU)*(df2/r)
> fstat
[1] 33.13956
```

The calculated F-statistic is 33.14 and is compared to the relevant critical value in order to decide if the null is rejected or not. The critical values for tests at the 5% and 1% levels of significance are reported by R:

```
> qf(.95,r,df2)
[1] 3.002465
> qf(.99,r,df2)
[1] 4.621093
```

The computed F-statistic is well above both critical values so the hypothesis is rejected at the 1% level of significance.