Dummies in regressions

(Not regressions for dummies)

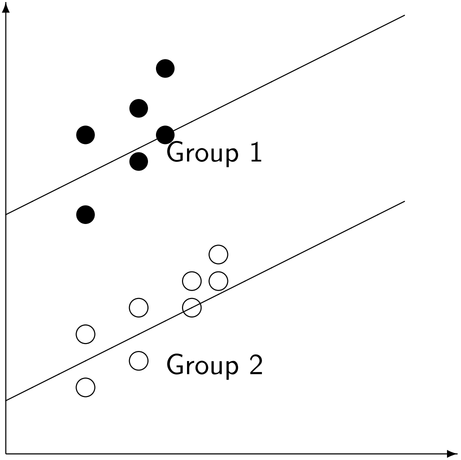
> *Let us remember the unfortunate econometrician who, in one of the major functions of his system, had to use a proxy for risk and a dummy for sex.*

Fritz Machlup, Journal of Political Economy, 1974.

Some uses for dummy variables

1. Allowing for differences in the intercept term
2. Allowing for differences in slopes
3. Test for stability of regression coefficients
4. Ameliating outliers
5. Panel data (fixed effects)

Dummy variables for changes in the intercept term



Correct for seasonalities.
If the behaviour of a variable varies across quarters, say, we can introduce quarterly dummies.

$$D_1 = \begin{cases} 1 & \text{if observation is from first quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{if observation is from second quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{if observation is from third quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$D_4 = \begin{cases} 1 & \text{if observation is from fourth quarter} \\ 0 & \text{otherwise} \end{cases}$$

Exercise

In finance one has identified various "calendar anomalies", that stock returns depend on calendar time in surprising ways. One of these is the "January effect," that stock returns seem to be higher in January.

Using returns for the S&P 500 in the period after 1950, test the hypothesis that the returns in January is different from other months.

In implementing this use indicator variables in a regression framework, where January is the only explanatory variable.

Implement the tests in R.

Solution

We ask whether returns in january are fundamentally different from the rest. Regression to run

$$r_m = E[r_m] + \beta D_{january} + e$$

If january is different, $\beta \neq 0$.

Reading in data and generating

```
library(zoo)
library(xts)
INSP500d <- read.zoo("../data/sp500_daily.csv",
                     format="%Y-%m-%d",sep=",",header=TRUE)
sp500d <- as.xts(INSP500d[,6])
sp500m <-sp500d[endpoints(sp500d,on="months")]
Rsp500m <- diff(log(sp500m))
```

Now, using this return series:

```
> Rm <- Rsp500m;
> dates  <-  as.POSIXlt(index(Rm))
> jan <- as.numeric(dates$mon==0)
> reg <- lm(Rm~jan)

Residuals:
      Min         1Q    Median        3Q       Max
-0.250944 -0.023895  0.003518  0.028917  0.145527

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.005516   0.001603   3.442  0.00061 ***
jan         0.004578   0.005551   0.825  0.40982

Residual standard error: 0.04219 on 754 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared: 0.0009012,Adjusted R-squared: -0.000423
F-statistic: 0.6801 on 1 and 754 DF,  p-value: 0.4098
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0055	0.0016	3.44	0.0006
jan	0.0046	0.0056	0.82	0.4098

There is an economically larger return in january, but not statistically significant.

Exercise

In finance one has identified various "calendar anomalies", that stock returns depend on calendar time in surprising ways. One of these is the "Day of the week effect," that stock returns seem to be different across days of the week.

Using returns for the S&P 500, test the hypothesis that the expected return is different across days of the week.

In implementing this use indicator variables in a regression framework.

Implement the analysis in R.

Solution

Preliminary, reading the data

```
> library(xtable)
> library(car)
> source("read.R")
> Rm <- Rsp500d;
> dates  <-  as.POSIXlt(index(Rm))
> Rm <- as.matrix(Rm)
> mon <- as.numeric(dates$wday==1)
> tue <- as.numeric(dates$wday==2)
> wed <- as.numeric(dates$wday==3)
> thu <- as.numeric(dates$wday==4)
> fri <- as.numeric(dates$wday==5)
```

First estimate dummy for each day, with no constant term.

```
> reg1 <- lm(Rm~0+mon+tue+wed+thu+fri)
> summary(reg1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.228293 -0.004455  0.000164  0.004691  0.110276

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
mon -0.0007038  0.0001772  -3.971 7.19e-05 ***
tue  0.0003388  0.0001723   1.967  0.04919 *
wed  0.0007367  0.0001721   4.279 1.89e-05 ***
thu  0.0003510  0.0001733   2.026  0.04279 *
fri  0.0006426  0.0001739   3.695  0.00022 ***

Residual standard error: 0.009784 on 15852 degrees of freed
  (1 observation deleted due to missingness)
Multiple R-squared: 0.003502,Adjusted R-squared: 0.003188
F-statistic: 11.14 on 5 and 15852 DF,  p-value: 9.737e-11
```

	Estimate	Std. Error	t value	Pr(>|t|)
mon	-0.0007	0.0002	-3.97	0.0001
tue	0.0003	0.0002	1.97	0.0492
wed	0.0007	0.0002	4.28	0.0000
thu	0.0004	0.0002	2.03	0.0428
fri	0.0006	0.0002	3.70	0.0002

in this setting need to construct hypothesis tests for equality

```
> C <- c(c(1, -1, 0, 0, 0), c(0, 1, -1, 0, 0 ), c(0, 0 ,1,
> C <- matrix(C,nrow=4,ncol=5,byrow=TRUE)
> r <- c(0, 0, 0, 0)
> linearHypothesis(reg1,hypothesis.matrix=C,rhs=r)
Linear hypothesis test

Hypothesis:
mon - tue = 0
tue - wed = 0
wed - thu = 0
thu - fri = 0

Model 1: restricted model
Model 2: Rm ~ 0 + mon + tue + wed + thu + fri

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1  15856 1.5214
2  15852 1.5174  4 0.0040662 10.62 1.362e-08 ***
```

Estimate regression with constant term, leaving out one observation (constant = monday)

```
> reg2 <- lm(Rm~tue+wed+thu+fri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.228293 -0.004455  0.000164  0.004691  0.110276

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0007038  0.0001772  -3.971 7.19e-05 ***
tue          0.0010427  0.0002472   4.219 2.47e-05 ***
wed          0.0014405  0.0002471   5.830 5.65e-09 ***
thu          0.0010549  0.0002479   4.256 2.10e-05 ***
fri          0.0013464  0.0002483   5.423 5.96e-08 ***

Residual standard error: 0.009784 on 15852 degrees of freed
  (1 observation deleted due to missingness)
Multiple R-squared: 0.002673,Adjusted R-squared: 0.002421
F-statistic: 10.62 on 4 and 15852 DF,  p-value: 1.362e-08
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0007	0.0002	-3.97	0.0001
tue	0.0010	0.0002	4.22	0.0000
wed	0.0014	0.0002	5.83	0.0000
thu	0.0011	0.0002	4.26	0.0000
fri	0.0013	0.0002	5.42	0.0000

Another alternative, only friday

```
> reg3 <- lm(Rm~fri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.229190 -0.004432  0.000190  0.004675  0.109379

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.928e-04  8.694e-05   2.218   0.0266 *
fri         4.497e-04  1.946e-04   2.311   0.0208 *

Residual standard error: 0.009794 on 15855 degrees of freed
  (1 observation deleted due to missingness)
Multiple R-squared: 0.0003368,Adjusted R-squared: 0.0002738
F-statistic: 5.342 on 1 and 15855 DF,  p-value: 0.02082
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0002	0.0001	2.22	0.0266
fri	0.0004	0.0002	2.31	0.0208

Find support to think friday is different.

Dummy variables for changes in slope coefficients

Dummy variables for parameter stability testing

You are investigating the market model

$$r_{it} = a + br_{mt} + e_{it}$$

in the Norwegian Market, and apply it to the company Norsk Hydro (NHY). Collect monthly returns for NHY for the period 1980-2006, and monthly returns for a value weighted market index for the same period.
After having estimated the model you worry that the NHY beta (The parameter b) has changed over time. You therefore split the sample into two periods, 1980–1989 and 1990–2006.
Test whether there are reasons to believe the b parameter has changed different in the two periods.

Consider the following regression.

$$r_{it} = a + b_1 r_{mt} + b_2 D r_{mt} + e_{it}$$

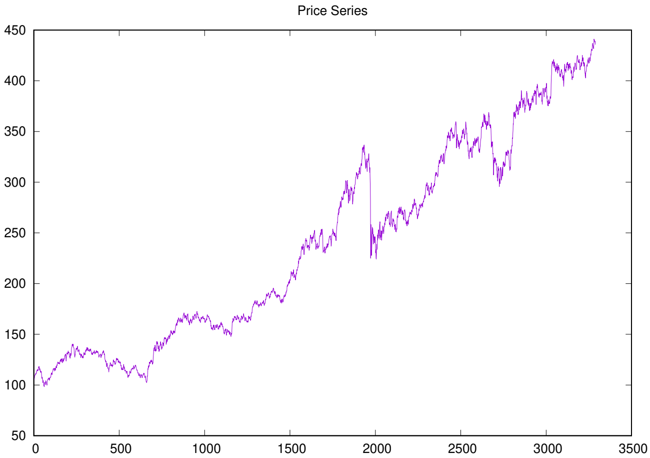
This achieve the desired test, by testing wheter $b_2 = 0$ we test the null of no change in beta.

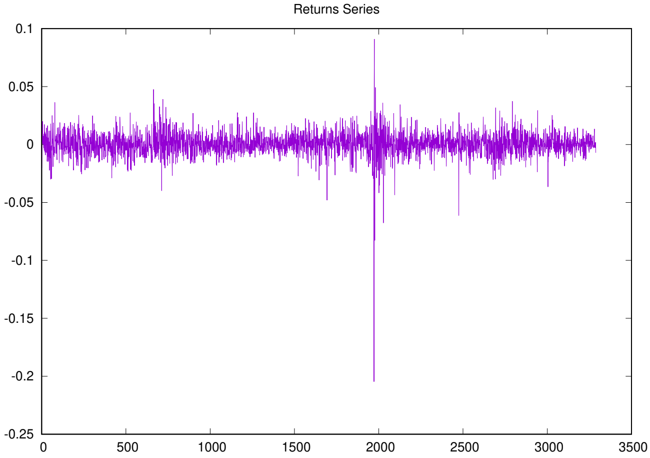
Running the regression

Variable	coeff	serr	t-val	p-val(t)
Constant	-0.00814	0.00318	-2.56	0.011
Rm	1.19274	0.06825	17.48	0.000
D	-0.14959	0.09054	-1.65	0.099
R^2	0.631	F	274.40	
$Adj\bar{R}^2$	0.629	pval F	0.0000	
DW	1.88			

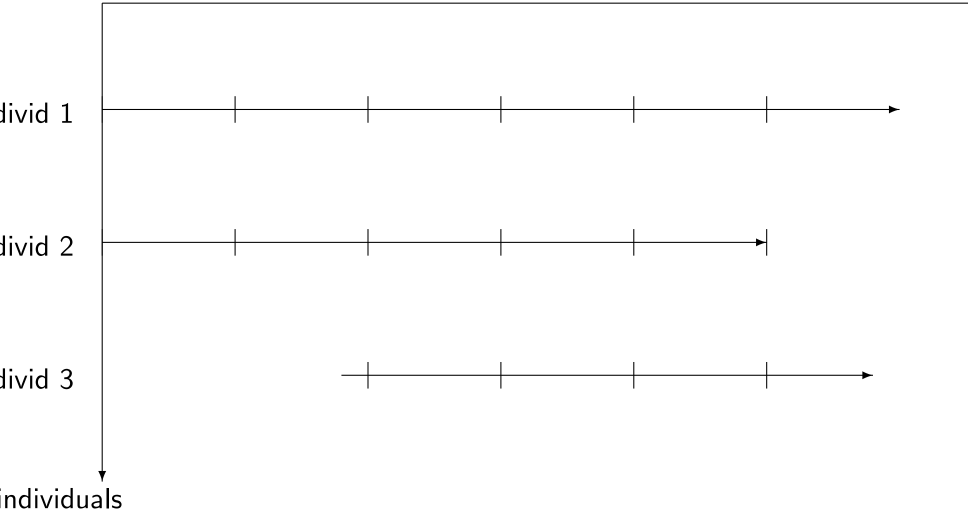
The p-value on the D is not significant at the 5% level. Do therefore not reject a null of no change.

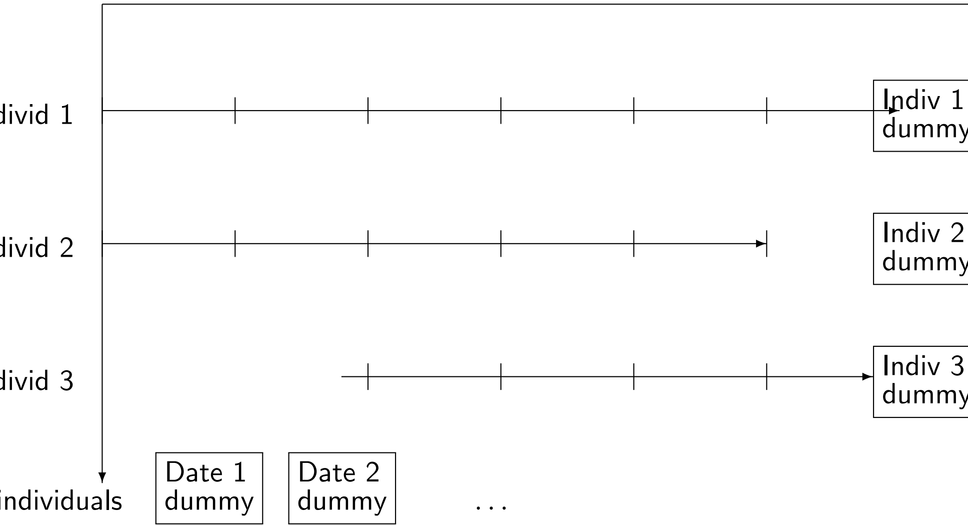
Using dummies to ameliate outliers




Returns Series
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
0
500
1000
1500
2000
2500
3000
3500

Dummy variables in panel data





divid 1
divid 2
divid 3
individuals
Indiv 1
dummy
Indiv 2
dummy
Indiv 3
dummy
Date 1
dummy
Date 2
dummy
...