# Regression with Categorical Variables

Matthew Seguin

## 2.

```
library(rio);library(plm);library(car)
```

```
## Warning: package 'plm' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
mrd <- import("murder.xls")
```

### a.

Below is the estimated random effects model for $mrdrte_{it} = \beta_0 + \beta_1 exec_{it} + \gamma_1 D90_t + \gamma_2 D93_t + \alpha_i + u_{it}$ where $\alpha_i$ is an unobserved state specific effect.

```
reg1 <- plm(mrdrte ~ exec + d90 + d93, data = mrd[,-2], model = "random")
summary(reg1)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = mrdrte ~ exec + d90 + d93, data = mrd[, -2], model = "random")
##
## Balanced Panel: n = 51, T = 3, N = 153
##
## Effects:
##                  var std.dev share
## idiosyncratic 12.345   3.514 0.145
## individual    73.034   8.546 0.855
## theta: 0.769
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -13.99365  -1.31208  -0.24036   0.52012  26.60731
##
## Coefficients:
##              Estimate Std. Error z-value  Pr(>|z|)
## (Intercept)  7.090413   1.308920  5.4170 6.061e-08 ***
## exec        -0.039526   0.158353 -0.2496   0.80289
## d90          1.387189   0.697829  1.9879   0.04683 *
## d93          1.699039   0.698968  2.4308   0.01507 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Total Sum of Squares:    1926.7
## Residual Sum of Squares: 1843
## R-Squared:      0.043401
## Adj. R-Squared: 0.02414
## Chisq: 6.76007 on 3 DF, p-value: 0.079951
```

## b.

There is not a dummy variable equal to 1 if $t = 1987$ and 0 otherwise because it would introduce perfect multicollinearity which violates the assumptions of the model.
This is because we may write:

$$DY_t = 1 - \sum_{X \neq Y} DX_t$$

If $t = Y$ we get that $DY_t = 1$ and all other dummy variables must be 0 so:

$$DY_t = 1 = 1 - \sum_{X \neq Y} DX_t = 1 - 0 - 0 = 1$$

If $t \neq Y$ we get that $DY_t = 0$ and only one of the other dummy variables must be 1 with the other being 0 so:

$$DY_t = 0 = 1 - \sum_{X \neq Y} DX_t = 1 - 1 - 0 = 0$$

So if the last dummy variable was included there is obviously an issue with perfect multicollinearity, however how it is with only 2 of the dummy variables is fine because you need all three to the above linear combination true.

### c.

The new model is $mrdrte_{it} = \beta_0 + \beta_1 exec_{it} + \beta_2 unemp_{it} + \gamma_1 D90_t + \gamma_2 D93_t + \alpha_i + u_{it}$ where $\alpha_i$ is an unobserved state specific effect. Recall that if there is omitted variable bias then for the original biased $\beta_1$ coefficient we know:

$$plim\ \hat{\beta}_1 = \beta_1 + \beta_2 \frac{cov(exec_{it}, unemp_{it})}{var(exec_{it})}$$

Many murder victims are homeless so we expect an increase in in the murder rate to follow an increase in unemployment (i.e. $\beta_2 > 0$), this is reinforced in the regression. When murder rates are high we expect more executions as more murderers will be caught and the death penalty was still in place. So when unemployment is higher we expect higher murder rates and then higher executions, so we expect for unemployment and executions to be positively correlated. Therefore since we expect $\beta_2 > 0$ and $cov(exec_{it}, unemp_{it}) > 0$ we expect that $plim\ \hat{\beta}_1 > \beta_1$ and hence the estimated coefficient from the first regression was positively biased. Below are the new regression results.

```r
reg2 <- plm(mrdrte ~ exec + unemp + d90 + d93, data = mrd[,-2], model = "random")
summary(reg2)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = mrdrte ~ exec + unemp + d90 + d93, data = mrd[,
##     -2], model = "random")
##
## Balanced Panel: n = 51, T = 3, N = 153
##
## Effects:
##                  var std.dev share
## idiosyncratic 12.400   3.521 0.156
## individual    67.333   8.206 0.844
## theta: 0.7595
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -13.12876  -1.23228  -0.31865   0.61080  26.61983
##
## Coefficients:
##              Estimate Std. Error z-value Pr(>|z|)
## (Intercept)  4.635132   2.179451  2.1267  0.03344 *
## exec        -0.054337   0.159501 -0.3407  0.73335
## unemp        0.394751   0.284813  1.3860  0.16575
## d90          1.732981   0.747856  2.3173  0.02049 *
## d93          1.699913   0.706561  2.4059  0.01613 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1978.6
## Residual Sum of Squares: 1870.9
## R-Squared:      0.054433
## Adj. R-Squared: 0.028877
## Chisq: 8.51977 on 4 DF, p-value: 0.07429
```

```r
reg1$coefficients[2] - reg2$coefficients[2]
```

```
##       exec
## 0.01481142
```

As we can see the estimated coefficient $\hat{\beta}_1$ is larger in the first regression, which we expect since we are saying it is positively biased there and unbiased in the second regression.

### d.

Below are the t statistics (though the regression labels them as normal statistics) and associated p-values for $H_0 : \beta_1 = 0$ in each regression (recall $\beta_1$ is the coefficient on $exec_{it}$ in each regression).

```r
reg1_data <- summary(reg1)
reg2_data <- summary(reg2)
t_reg1 <- reg1_data$coefficients[2, "z-value"]
t_reg1
```

```
## [1] -0.2496078
```

```r
p_reg1 <- reg1_data$coefficients[2, "Pr(>|z|)"]
p_reg1
```

```
## [1] 0.8028907
```

```r
t_reg2 <- reg2_data$coefficients[2, "z-value"]
t_reg2
```

```
## [1] -0.340672
```

```r
p_reg2 <- reg2_data$coefficients[2, "Pr(>|z|)"]
p_reg2
```

```
## [1] 0.7333505
```

In the first regression we get $t = -0.2496078$ which has an approximate p-value of 0.8028907 which is very large so we definitely wouldn't want to reject $H_0$.
In the second regression we get $t = -0.340672$ which has an approximate p-value of 0.7333505 which is again very large so we definitely wouldn't want to reject $H_0$.
Therefore in both regressions we can not reject $H_0 : \beta_1 = 0$ which means we can not reject the hypothesis that the predicted effect on the murder rate of the number of executions is zero. This is true for any significance level less than the p-values and so is definitely true at the 10% level.

### e.

Below is the estimated fixed effects model for $mrdrte_{it} = \beta_0 + \beta_1 exec_{it} + \beta_2 unemp_{it} + \gamma_1 D90_t + \gamma_2 D93_t + \alpha_i + u_{it}$ where $\alpha_i$ is an unobserved state specific effect.

```r
reg3 <- plm(mrdrte ~ exec + unemp + d90 + d93, data = mrd[,-2], model = "within")
summary(reg3)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = mrdrte ~ exec + unemp + d90 + d93, data = mrd[,
##     -2], model = "within")
##
## Balanced Panel: n = 51, T = 3, N = 153
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -26.685751  -0.658371  -0.065721   0.674717  13.394112
##
## Coefficients:
##       Estimate Std. Error t-value Pr(>|t|)
## exec  -0.13832    0.17701 -0.7815  0.43642
## unemp  0.22132    0.29638  0.7467  0.45701
## d90    1.55621    0.74533  2.0880  0.03939 *
## d93    1.73324    0.70044  2.4745  0.01506 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1311.5
## Residual Sum of Squares: 1215.2
## R-Squared:      0.073367
## Adj. R-Squared: -0.43723
## F-statistic: 1.93981 on 4 and 98 DF, p-value: 0.10984
```

We weren't asked to find it but I'm curious what the Hausman test says so I'll run it below.

```r
phtest(reg2, reg3)
```

```
##
##  Hausman Test
##
## data:  mrdrte ~ exec + unemp + d90 + d93
## chisq = 5.7757, df = 4, p-value = 0.2165
## alternative hypothesis: one model is inconsistent
```

The p-value is 0.2165 which means we can't reject $H_0 : \alpha_i$ is uncorrelated with the regressors. So it isn't a bad idea to use the regression with random effects because this test tells us it's still consistent.

### f.

If *exec* increases by 30 (i.e. $\Delta exec = 30$) we get the following expected change (keeping all else constant) in murder rate under our model:

$$\Delta mrdrte = mrdrte_1 - mrdrte_0 = (\beta_0 + \beta_1(exec + \Delta exec) + \beta_2 unemp + \gamma_1 D90 + \gamma_2 D93) - (\beta_0 + \beta_1 exec + \beta_2 unemp + \gamma_1 D90 + \gamma_2 D93)$$

$$= \beta_1 \Delta exec = 30\beta_1$$

We can find a 95% confidence interval for $\beta_1$ via the results of our regression as $\left[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)\right]$. The result of which is shown below.

```r
reg3_data <- summary(reg3)
b_1_hat <- as.numeric(reg3_data$coefficients[1, "Estimate"])
b_1_hat
```

```
## [1] -0.1383231
```

```r
se_b_1_hat <- as.numeric(reg3_data$coefficients[1, "Std. Error"])
se_b_1_hat
```

```
## [1] 0.1770059
```

```r
b_1_CI <- c(b_1_hat - 1.96*se_b_1_hat, b_1_hat + 1.96*se_b_1_hat)
b_1_CI
```

```
## [1] -0.4852547  0.2086086
```

Therefore using this result we have that the 95% confidence interval for $\Delta mrdrte$ as a result of an increase in *exec* by 30 is given by $30\, CI(\beta_1)$ where $CI(\beta_1)$ is the confidence interval for $\beta_1$ (found above). The result of which is shown below.

```r
chg_mrdrte_CI <- 30*b_1_CI
chg_mrdrte_CI
```

```
## [1] -14.557641   6.258258
```

## g.

Below is the test of the hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$ in the regression $mrdrte_{it} = \beta_0 + \beta_1 exec_{it} + \beta_2 unemp_{it} + \gamma_1 D90_t + \gamma_2 D93_t + \alpha_i + u_{it}$ where $\alpha_i$ is an unobserved state specific effect.

```r
linearHypothesis(reg3, c("d90=0","d93=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## d90 = 0
## d93 = 0
##
## Model 1: restricted model
## Model 2: mrdrte ~ exec + unemp + d90 + d93
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    100
## 2     98  2 7.2641    0.02646 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
reg3_data$coefficients[3, "Estimate"]
```

```
## [1] 1.556215
```

```r
reg3_data$coefficients[4, "Estimate"]
```

```
## [1] 1.733242
```

For this test we get the p-value 0.0265 so we can't reject $H_0$ at the 1% level but we can reject $H_0$ at the 5% level (Note that R is using a Chi-squared stat instead of an F stat to calculate this, this doesn't matter because $F_{(q,\infty)} = \frac{X_q^2}{q}$). The 5% level is standard so we can say that the true coefficients on $d90$ and $d93$ are nonzero, that is there is some change in the murder rate relative to 1987 in the years 1990 and 1993. Since the coefficients estimated on each of the year variables are positive (shown above) we take this as evidence that the murder rate increased relative to 1987 in both the years 1990 and 1993.

### h.

We want to test $H_0 : \gamma_1 = \gamma_2$ (i.e $\gamma_1 - \gamma_2 = 0$). We can use a simple t test for this. We know:

$$Var(\hat{\gamma}_1 - \hat{\gamma}_2) = Var(\hat{\gamma}_1) + Var(\hat{\gamma}_2) - 2Cov(\hat{\gamma}_1, \hat{\gamma}_2)$$

We also know under $H_0$ that $\gamma_1 - \gamma_2 = 0$, so the t-statistic will be:

$$t = \frac{(\hat{\gamma}_1 - \hat{\gamma}_2) - 0}{\sqrt{Var(\hat{\gamma}_1) + Var(\hat{\gamma}_2) - 2Cov(\hat{\gamma}_1, \hat{\gamma}_2)}} \approx \frac{\hat{\gamma}_1 - \hat{\gamma}_2}{\sqrt{(SE(\hat{\gamma}_1))^2 + (SE(\hat{\gamma}_1))^2 - 2Cov(\hat{\gamma}_1, \hat{\gamma}_2)}}$$

The result of the test is shown below (note we are given $Cov(\hat{\gamma}_1, \hat{\gamma}_2) = 0.24$).

```
g_1_hat <- as.numeric(reg3_data$coefficients[3, "Estimate"])
g_1_hat
```

```
## [1] 1.556215
```

```
se_g_1_hat <- as.numeric(reg3_data$coefficients[3, "Std. Error"])
se_g_1_hat
```

```
## [1] 0.7453273
```

```
g_2_hat <- as.numeric(reg3_data$coefficients[4, "Estimate"])
g_2_hat
```

```
## [1] 1.733242
```

```
se_g_2_hat <- as.numeric(reg3_data$coefficients[4, "Std. Error"])
se_g_2_hat
```

```
## [1] 0.7004381
```

```
cov_g1_g2 <- 0.24
se_diff_hat <- sqrt((se_g_1_hat)^2 + (se_g_2_hat)^2 - 2*cov_g1_g2)
t <- (g_1_hat - g_2_hat)/se_diff_hat
t
```

```
## [1] -0.2352793
```

```
p_val <- 2*pnorm(-abs(t))
p_val
```

```
## [1] 0.813992
```

So out t-statistic is -0.235 which gives a p-value of 0.814 which is very large so we definitely don't want to reject $H_0$. Therefore we can say that there is no significant difference between $\gamma_1$ and $\gamma_2$ which means that the increase in the murder rate for 1990 and 1993 was the same relative to 1987.