# Fixed Effects Regression

*Irfan Kanat*

*August 14, 2017*

One of the assumptions of OLS is that observations are independent of each other. Like coin tosses. The outcome of each toss is independent of the previous toss. We assume that individuals heights are not influenced by other individuals.

There are certain instances where this independence assumption is violated.

Below are some examples:

An individual's height in one year will be highly correlated to his height on previous years. If your dataset tracks multiple observations from same individual, we can not say the observations are independent.

If you track sales of a product over time, the sales performance in each period will be correlated to previous periods' performance.

The ACT scores of different students from one school may be correlated due to quality of education in that school.

The height of trees from one region may be correlated due to weather conditions in that region.

In all these instances we can not claim the observations are independent. There are various ways to address this problem. Here I will touch on a simple solution.

## Idea Behind Fixed Effects

From a linear relation perspective, how would a common factor manifest itself? Let us remember the formula for regression.

$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon$

We assume there is an intercept, and a slope for each variable, and an error term. If there is a common factor that we do not observe in the model, the variance caused by that factor will end up in the error term $\epsilon$.

Let us say we are talking about rent in different neighborhoods. The rent for houses in a neighborhood will be correlated depending on the attractiveness of the neighborhood. A neighborhood with a good school district may command several hundred dollars more per month.

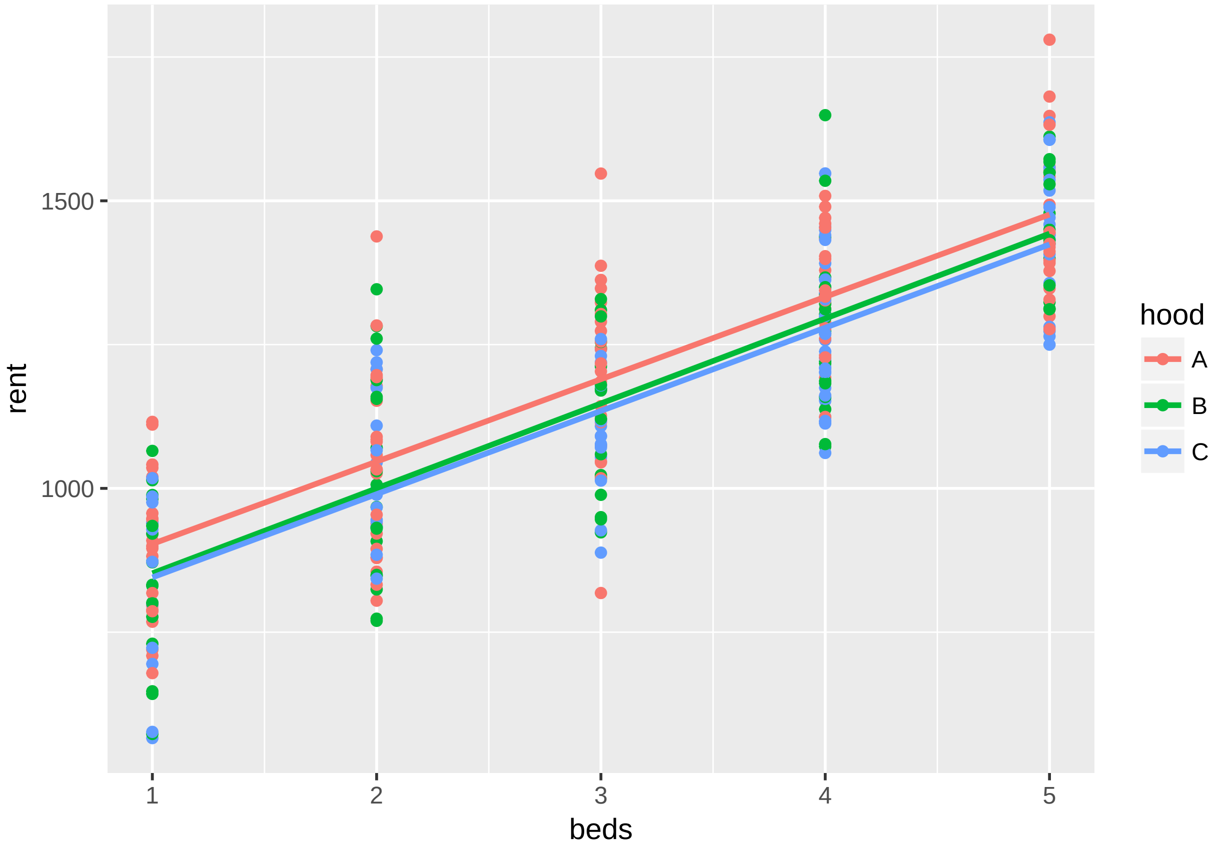
So, can we say that each neighborhood will have a different intercept? Breaking down the error term $\epsilon$ into the intercept for neighborhood ($\alpha_{hood}$) plus some random error ($u$)? $\alpha_{hood}$ will capture that several hundred dollar difference due to the good school, nearby park, etc.

$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \alpha_{hood} + u$

Generally speaking, if your independence assumption is violated it will be harder to get significant results as the error term (and residual variance captured by it) will be inflated.

I created a version of rent dataset from earlier that is suitable for our purposes here.

```
qplot(data = rent2, x = beds, y = rent, col = hood) +
  geom_smooth(method = "lm", se = F)
```



Each neigborhood follows a similar tracectory (slopes are the same) but the starting points (intercept) are different.

## How Can We Fit This Model?

If the basic idea is that there is a separate intercept for each grouping (neighborhood, person, company…) then we can use dummy variables.

Let us say you have three neighborhoods: A, B, and C. To capture the information in three (n) levels in binary variables you would need two (n-1) binary variables

A = 0 0

B = 1 0

C = 0 1

Very often in practice we call the Level captured as 0 0 the reference level and label the binary variables after the remaining levels.

So the data:

```
head(rent2, n = 5)
```

```
##     beds baths sqft hood     rent
## 14     1     2  463    C 565.8082
## 87     1     1  545    B 572.9239
## 58     1     1  624    C 576.5924
## 188    1     2  687    A 719.3036
## 151    1     2  739    A 709.0037
```

Becomes:

```
data.frame(beds = rent2$beds[1:5], baths = rent2$baths[1:5],
           sqft = rent2$sqft[1:5], b = c(0, 1, 0, 0, 0), c = c(1, 0, 1, 0, 0),
           rent = rent2$rent[1:5])
```

```
##   beds baths sqft b c     rent
## 1    1     2  463 0 1 565.8082
## 2    1     1  545 1 0 572.9239
## 3    1     1  624 0 1 576.5924
## 4    1     2  687 0 0 719.3036
## 5    1     2  739 0 0 709.0037
```

You won't need to do this by hand, I am just demonstrating the effects of dummy coding a categorical variable.

If you fit a model with these dummy variables, it is no different than any other model with binary variables. Remember the example of gender from Interactions learning activity.

Let us fit a model with dummies for neigborhood.

```
 # I use factor to declare categorical variables
rent2_lm_0 <- lm(rent ~ beds + baths + sqft + factor(hood), data = rent2)
summary(rent2_lm_0)
```

```
##
## Call:
## lm(formula = rent ~ beds + baths + sqft + factor(hood), data = rent2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -214.907  -52.053   -0.166   51.423  237.035
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   136.32570   29.14009   4.678 4.42e-06 ***
## beds          147.66913    3.27807  45.048  < 2e-16 ***
## baths          58.90626    8.99555   6.548 2.59e-10 ***
## sqft            0.51964    0.02214  23.466  < 2e-16 ***
## factor(hood)B -30.37572   10.90531  -2.785  0.00569 **
## factor(hood)C -49.15798   10.79172  -4.555 7.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.26 on 294 degrees of freedom
## Multiple R-squared:  0.8977, Adjusted R-squared:  0.896
## F-statistic:   516 on 5 and 294 DF,  p-value: < 2.2e-16
```

Looking at the output you will notice that neighborhood A is not shown in results. That is because it is used as reference level.

All other dummies will be in relation to reference level.

So we will say, houses in neighborhood B are -30.38 cheaper than those in neigborhood A.

Houses in neighborhood C are -49.16 cheaper than those in neighborhood A.

The intercept for neighborhood B will then be $\beta_0 + \alpha_B = 105.95$.

Here is a simple exercise for you, calculate the intercept for neighborhoods A and C.

### More Efficiency

Very often you would not care about the actual value of these fixed effects. All you want is to remove the variance caused by these fixed effects from the model and reduce type 2 error (rejecting a significant relationship).

If that is all you care about, then you can simply remove these intercepts by "demeaning". Substracting the neighborhood average from each observation would essentially remove the effect of neighborhood.

Luckily you won't need to do this by hand. The excellent plm package handles fixed effects models nicely. Good thing, that you DO KNOW how to install and activate packages.

```r
rent2_plmFE_0 <- plm(rent ~ beds + baths + sqft, data = rent2,
                     # model="within" for fixed effects
                     # index = 'hood' for clustering variable
                     model = "within", index = "hood")
summary(rent2_plmFE_0)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = rent ~ beds + baths + sqft, data = rent2, model = "within",
##     index = "hood")
##
## Unbalanced Panel: n=3, T=92-111, N=300
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -215.000  -52.100   -0.166   51.400  237.000
##
## Coefficients :
##         Estimate Std. Error t-value  Pr(>|t|)
## beds  147.669130   3.278068 45.0476 < 2.2e-16 ***
## baths  58.906263   8.995549  6.5484 2.591e-10 ***
## sqft    0.519638   0.022144 23.4663 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    17018000
## Residual Sum of Squares: 1754700
## R-Squared:      0.89689
## Adj. R-Squared: 0.89514
## F-statistic: 852.474 on 3 and 294 DF, p-value: < 2.22e-16
```

If you compare the coefficients, you will see that the demeaned model is identical to dummy variables model.

There are many other ways of handling the violation of "Independence of Observations" assumption, basically anything more would turn into a full course rather than a learning activity. If you are interested, look into random effects, mixed effects, and Generalized Method of Moments models.

## Solutions to Exercises

1 - Calculate the intercept for neighborhoods A and C.

Intercept for Neighborhood A would be $\beta_0 = 136.33$.

The intercept for neighborhood C will then be $\beta_0 + \alpha_C = 87.17$.