Advanced Topics in Applied Regression

Day 2: Interactions & Fixed-effects

Constantin Manuel Bosancianu

Wissenschaftszentrum Berlin
Institutions and Political Inequality unit
manuel.bosancianu@wzb.eu

September 30, 2017

Why interactions?

They allow for a much richer set of hypotheses to be put forward and tested.

In my own area of focus (political institutions, economic phenomena, and voter attitudes/behavior), such hypotheses involving moderation are very common.

One prominent example: income inequality's effect on voter turnout at different levels of a person's income (Solt, 2008).

Despite their importance, misunderstandings still persist about how to interpret coefficients/effects in such models.

Basic setup

Why specify interactions?

So far, we've worked with simple models. Think of the example from yesterday, with Boston neighborhood average house prices. Here, I complicated it a bit by also adding a dummy for whether the neighborhood is on the Charles river or not:

$$Prices = a + b_1 Rooms + b_2 River + e \tag{1}$$

Here, the effect of *River* is assumed to be constant, b_2, no matter the level of the other variable in the model.

This is not always the case: effect of SES and union membership on political participation, where b_{union} likely varies.

What if the effect isn't constant?

The riverfront is a desirable real-estate location. Houses with more rooms are certainly more expensive everywhere in Boston, but it's likely that the price difference between $n+1$ and n rooms is higher on the riverfront than elsewhere.

In modelling terms, we might say that the effect of $Rooms$ on $Price$ is different based on the value of the $River$ dummy.

From words to equation (I)

$$
\begin{aligned}
Prices &= a_1 + b_1 Rooms + b_2 River + e \\
b_1 &= a_2 + b_3 River \\
a_1 &= a_3 + b_4 River
\end{aligned}
$$

The second equation gives us how the effect of *Rooms* (b_1) varies depending on *River*.

The third equation makes sure that the intercept varies as well (which usually happens if the slope varies).

From words to equation (II)

$$\begin{aligned} Prices =& a_3 + b_4 River + (a_2 + b_3 River) * Rooms + b_2 River + e \\ =& a_3 + (b_4 + b_2) * River + a_2 Rooms + b_3 River * Rooms + e \\ =& a_3 + (b_4 + b_2) * River + (a_2 + b_3 River) * Rooms + e \qquad (2) \end{aligned}$$

The third row shows most clearly how the effect of *Rooms*, $a_2 + b_3 River$, now varies depending on the precise value of the *River* indicator.

This depends, of course, on the b_3 being statistically significant. If not, then the effect of *Rooms* is always a_2.

Basic interaction model

$$Prices = a_3 + (b_4 + b_2) * River + (a_2 + b_3 River) * Rooms + e \quad (3)$$

If we designate a_3 as γ_1, $b_4 + b_2$ as γ_2, a_2 as γ_3, and b_3 as γ_4, then we get a general form of the interaction:

$$Prices = \gamma_1 + \gamma_2 River + \gamma_3 Rooms + \gamma_4 River * Rooms + e \quad (4)$$

Interaction model (cont.)

When $River = 0$,

$$\begin{aligned} Prices &= \gamma_1 + \gamma_2 0 + \gamma_3 Rooms + \gamma_4 Rooms * 0 + e \\ &= \gamma_1 + \gamma_3 Rooms + e \end{aligned} \tag{5}$$

When $River = 1$,

$$\begin{aligned} Prices &= \gamma_1 + \gamma_2 1 + \gamma_3 Rooms + \gamma_4 Rooms * 1 + e \\ &= \gamma_1 + \gamma_2 + Rooms(\gamma_3 + \gamma_4) + e \end{aligned} \tag{6}$$

The effect of *Rooms* varies depending on the value of *River*.

Symmetry in interpretation

When $Rooms = 0$, then

$$\begin{aligned} Prices =& \gamma_1 + \gamma_2 River + \gamma_3 * 0 + \gamma_4 River * 0 + e \\ =& \gamma_1 + \gamma_2 River + e \end{aligned}$$

When $Rooms = 1$,

$$\begin{aligned} Prices =& \gamma_1 + \gamma_2 River + \gamma_3 * 1 + \gamma_4 River * 1 + e \\ =& \gamma_1 + \gamma_3 + River(\gamma_2 + \gamma_4) + e \end{aligned}$$

The effect of *River* varies depending on the level of *Rooms*.

Interpretation

Wages in 1976

We have information on 526 US workers:

- ✓ `wage`: wage in USD per hour;
- ✓ `educ`: years of education;
- ✓ `gender`: male or female (with 1=female);
- ✓ `exper`: labor force experience (yrs. in labor market);
- ✓ `tenure`: yrs. with current employer.

The goal is to predict wages.[1]

[1] In fact, we'll be predicting $log(wage)$, as wages tend to be right skewed, which causes problems with the normality of errors.

Interpreting coefficients

	DV: Log hourly wage (USD)
(Intercept)	1.762***
	(0.025)
Female	−0.311***
	(0.037)
Yrs. education	0.089***
	(0.007)
Yrs. experience	0.005**
	(0.002)
Yrs. tenure	0.021***
	(0.003)
Female * Tenure	−0.013*
	(0.006)
R^2	0.399
Adj. R^2	0.393
Num. obs.	526
RMSE	0.414

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Continuous variables were demeaned.

Specification with interaction: Female * Tenure

Interpreting coefficients (cont.)

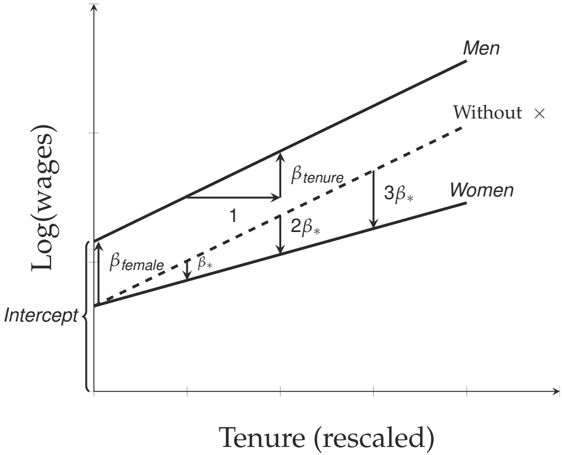
How do you interpret $\beta_{female} = -0.311$?

Important: after demeaning, the "0" for variable X refers to the mean of X, $\overline{X}$.

How do you interpret $\beta_{tenure} = 0.021$?

How do you interpret $\beta_{female*tenure} = -0.013$? How is the effect of *tenure* different for men, compared to women?

Graphical depiction



Example with wages (graph adapted from Brambor et al., 2005). β_* means $\beta_{female*tenure}$.

Difference between *coefficients* and *effects*

For linear models without interactions, $coefficient = effect$. A $\beta_X = 2$ means the *effect* of 1-unit increase in X on Y is 2.

For linear models with (significant) interactions, $coefficient \neq effect$. Rather, the effect of an interacted variable is a function of 2 coefficients.

$$
\begin{aligned}
Wage =& 1.762 - 0.311 * Fem. + 0.021 * Tnr. - 0.013 * Fem. * Tnr. + \dots \\
=& 1.762 + 0.021 * Tnr. + \underbrace{(-0.311 - 0.013 * Tnr.)}_{\text{effect}} * Fem. + \dots
\end{aligned}
$$

2nd example: differences in salaries

	DV: Salary in company
(Intercept)	14180.85***
	(333.93)
Experience	452.66***
	(60.18)
Management	7172.32***
	(506.82)
Exper.*Managem.	222.74*
	(104.09)
R^2	0.88
Adj. R^2	0.87
Num. obs.	46
RMSE	1701.23

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$. Experience has been centered by subtracting 7.5 from each value.

Experience measured in years, management is dichotomous indicator (1=manager)

3rd example: Boston house prices

	Model 1	Model 2
(Intercept)	22.251***	22.250***
	(0.302)	(0.301)
Average num. rooms	9.024***	8.967***
	(0.440)	(0.416)
Charles river	4.194***	4.083***
	(1.186)	(1.151)
Charles*Rooms	−0.536	
	(1.355)	
R^2	0.496	0.496
Adj. R^2	0.493	0.494
Num. obs.	506	506
RMSE	6.547	6.541

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Number of rooms has been demeaned.

Predicting house price in neighborhood

Interactions – other measurement scales

The interpretations carry over perfectly, e.g. when both are continuous (we will practice more during the lab).

$$Y = a + b_1X_1 + b_2X_2 + b_3(X_1 * X_2) + e \tag{7}$$

b_2 is the effect of X_2 on Y when X_1 is 0.

The converse interpretation, for b_1, is also identical.

Collinearity

High correlations in interactions

```
require(MASS)
out <- mvrnorm(300, # number of observations
               mu = c(5,5), # means of the variables
               # correlation matrix
               Sigma = matrix(c(1,0.35,0.35,1), ncol = 2),
               empirical = TRUE)
colnames(out) <- c("x1","x2")
out <- as.data.frame(out)
cor(out$x1, out$x2) # So, that's the correlation

[1] 0.35

out$inter <- out$x1*out$x2 # Construct the interaction term
cor(out$x1, out$inter) # Correlation

[1] 0.8179821

cor(out$x2, out$inter) # Correlation

[1] 0.8137691
```

In these situations, the VIF becomes very large, making the sampling variance for coefficients large as well.

High correlations – "solution"

Essentially, it's justified that we have large SEs—the software is telling us it doesn't have enough *unique* information to estimate the effect precisely.

The "solution": center the variable, i.e. subtract the mean/median from all observations on the variable.

$$X_i^* = X_i - \overline{X} \tag{8}$$

High correlations – “solution”

```
out$x1mod <- out$x1 - mean(out$x1)
out$x2mod <- out$x2 - mean(out$x2)
cor(out$x1mod, out$x2mod) # cor(X1,X2) is the same

[1] 0.35

out$intermod <- out$x1mod*out$x2mod
cor(out$x1mod, out$intermod) # Correlation

[1] 0.02988741

cor(out$x2mod, out$intermod) # Correlation

[1] -0.005910737
```

Not so much a solution; more of a *re-specification* of the original model (Kam & Franzese Jr., 2007, pp. 93–99).

Centering will produce different bs, a and SEs, simply because these refer to different quantities.

Presentation

Significance testing in interactions

With interactions, significance tests also take on a different interpretation (Braumoeller, 2004).

$$Y_i = a + b_1 X1_i + b_2 X2_i + b_3 (X1_i * X2_i) + e_i \qquad (9)$$

The significance test on b_1 is only valid for instance when $b_2 = 0$.

At other levels of b_2, this significance test might no longer produce a positive result.

Sampling variance

$$Y_i = a + b_1 X1_i + b_2 X2_i + b_3 (X1_i * X2_i) + e_i \quad (10)$$

Since it's an interaction, b_1 is the coefficient of $X1$, and eff_{X1} is the effect of $X1$ on Y. If b_3 is significant, $b_1 \neq eff_{X1}$

$$V(eff_{X1}) = V(b_1) + X_2^2 V(b_3) + 2X_2 Cov(b_1, b_3) \quad (11)$$

This makes it clear that the variance varies depending on X_2 as well.

Presenting results

There is little need to use the formula in Equation 11 to compute things by hand.[2]

The best way to do present results from a specification with interactions is by plotting both the effect and its associated uncertainty.

An easy way to do this is with the `effects` package in R (but also check out Thomas Leeper's `margins` package).

[2] An example that shows you how to do this can be found in today's script.

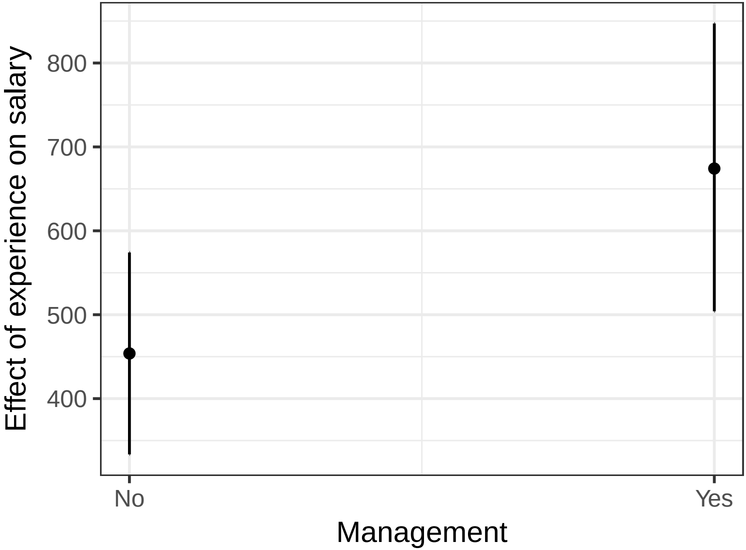
Predicting salaries

	DV: Salary in company
(Intercept)	14180.85***
	(333.93)
Experience	452.66***
	(60.18)
Management	7172.32***
	(506.82)
Exper.*Managem.	222.74*
	(104.09)
R^2	0.88
Adj. R^2	0.87
Num. obs.	46
RMSE	1701.23

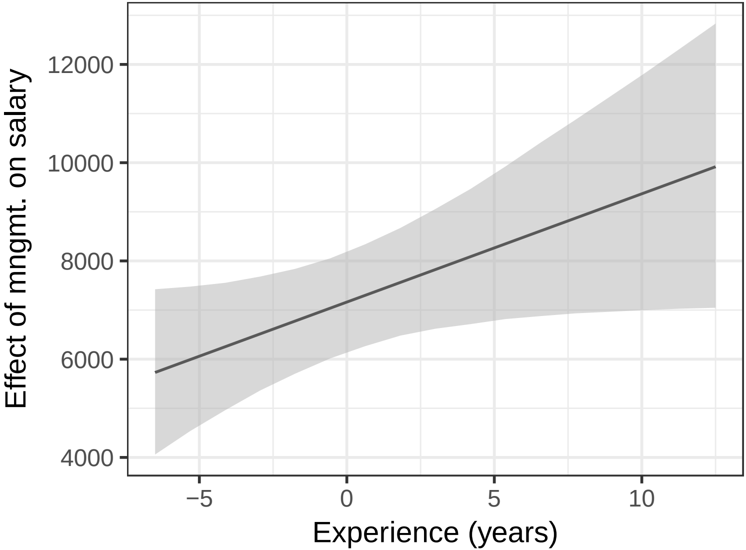
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Experience has been centered by subtracting 7.5 from each value.

Experience measured in years, management is dichotomous indicator (1=manager)

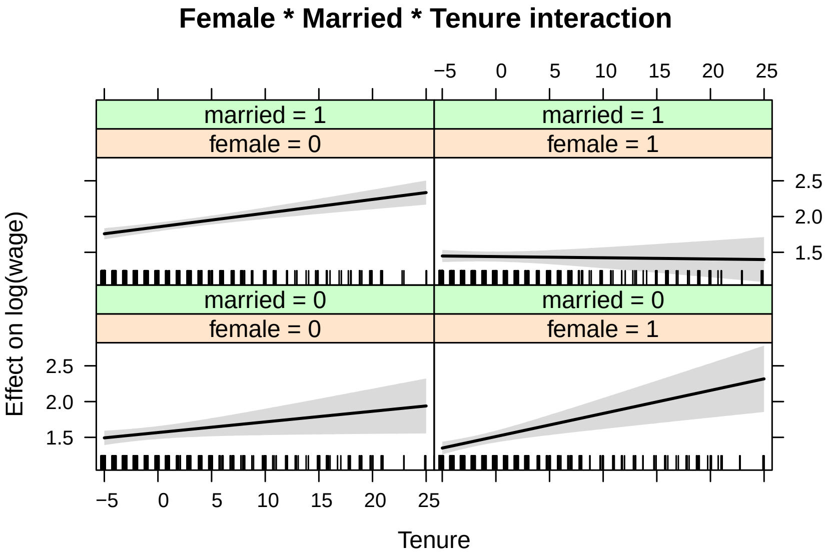
Predicting salaries – effect of experience



Predicting salaries – effect of management



Predicting hourly wage – 3-way interaction



Fixed effects

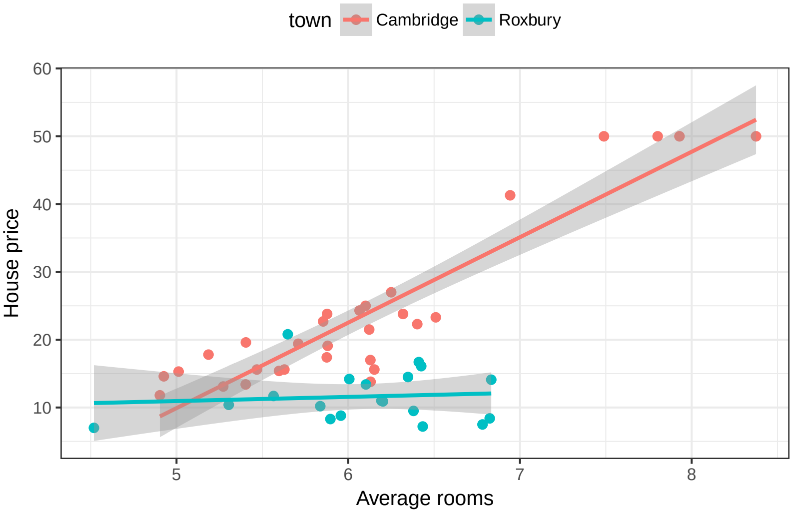
Why fixed effects?

	DV: House price (ave.)
(Intercept)	-42.757^{***}
	(9.620)
Average num. rooms	10.139^{***}
	(1.568)
R^2	0.471
Adj. R^2	0.460
Num. obs.	49
RMSE	8.277

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Predicting house price using number of rooms

Why fixed effects?



Why fixed effects?

1. As a solution to the issue of heteroskedasticity, when the problem is caused by different trends in each of the groups.
2. As a solution to the issue of omitted variable bias, on the road to a better causal estimate of the effect of X on Y.

These two issues are related, inasmuch as the trends in the groups are caused by variables which our model specification does not include.

Classic example

We have 172 children assessed with a test at 3 points in time.

The goal is to understand what predicts their test scores, and whether extra courses helps.

Measurements at multiple points in time are great for boosting sample size, and lowering SEs, but they add complications to the analysis: clustering.

Classic example

	DV: Test score
(Intercept)	48.613^{***}
	(0.661)
Female	1.647^{*}
	(0.764)
SES index	-1.712^{**}
	(0.531)
AP courses	4.812^{***}
	(0.447)
R^2	0.196
Adj. R^2	0.191
Num. obs.	516
RMSE	8.605

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Predicting test scores

What if other factors, e.g. genetic or psychological, are at play both for AP courses and test scores?

Standard model

$$Score_i = a + b_1 X1_i + \cdots + b_k Xk_i + e_i \tag{12}$$

In the standard model, one of the assumptions is that es are distributed $\mathcal{N}(0, \sigma_e^2)$.

This is no longer the case is there are omitted predictors Z, which were not included in the model.[3]

[3] The bigger implication here is also the fact that the effects of $X1$, $\ldots$, Xk are likely biased in this case.

The error term

$$Score_{it} = b_1 X1 + \cdots + b_k Xk + \underbrace{\alpha_i + e_{it}}_{e} \qquad (13)$$

Now the error is decomposed into an individual-specific term, α_i, and an observation-specific one, e_{it}.[4]

If any time-invariant factors not in the model have an effect on test score, this means estimates for some Xs are biased.

[4]This observation can be understood as a "individual i at time t" case.

Within- and between-

2 sources of variance: between-individuals and within-individuals (over time).

Suppose that over time we have a good model. However, the between-individual variance is the source of problems, as it may include variables we cannot observe in the data: drive to succeed, or genetic factors.

The solution adopted by FE is to do away with the problematic variance, as either way our interest is in the time-varying factor: number of AP courses.

FE strategy: demeaning

If we average the values over time for each student, $\bar{Y}$, $\bar{X1}, \ldots, \bar{Xk}$, and then subtract observations over time from these averages, we get

$$Score_{it} - \overline{Score_i} = (X1_i - \bar{X1})\beta_1 + \cdots + (Xk_i - \bar{Xk})\beta_k + e_{it} - \bar{e}_i \quad (14)$$

This takes care of the problematic between-variance, as all that remains is within-variance.

	Raw			Demeaned		
	t_1	t_2	t_3	t_1	t_2	t_3
Individual 1	10	20	30	-10	0	10
Individual 2	60	70	80	-10	0	10

FE "cousins": LSDV

Least Squares Dummy Variable (LSDV) regression.

Add a set of $i - 1$ dummy indicators[5] for persons, which capture *all* the between-person variation—the problematic one.

$$Score_{it} = a + b_1 X1 + \cdots + b_k Xk + \underbrace{P_1 + \cdots + P_{i-1}}_{i-1 \text{ terms}} + e_{it} \tag{15}$$

These allow for the causal effect to be estimated only based on within-variance.

LSDV and FE will be *identical*.

[5] That's because we still want to estimate an intercept.

FE "cousins": first differences (FD)

Particularly valuable for cases where auto-correlation of measurements proximate in time might be an issue.

Instead of trying to explain raw scores, this approach focuses on score differences between adjacent time points.

$$\Delta Y_t = \Delta X1_t \beta_1 + \cdots + \Delta Xk_t \beta_k + \Delta e_{it} \tag{16}$$

where $\Delta Y_t = Y_{t+1} - Y_t$.

FE and FD will be identical *only* in instances with 2 time points.

Thank you for the kind attention!

References

Brambor, T., Clark, W. R., & Golder, M. (2005). Understanding Interaction Models: Improving Empirical Analyses. *Political Analysis*, *14*(1), 63–82.

Braumoeller, B. F. (2004). Hypothesis Testing and Multiplicative Interaction Terms. *International Organization*, *58*(4), 807–820.

Kam, C. D., & Franzese Jr., R. J. (2007). *Modeling and Interpreting Interactive Hypotheses in Regression Analysis*. Ann Arbor, MI: University of Michigan Press.

Solt, F. (2008). Economic Inequality and Democratic Political Engagement. *American Journal of Political Science*, *52*(1), 48–60.