**Lab Assignment week 13**

**I. Objectives**: The primary objective is to explore interactions among variables in linearregressionmodels. Interpretation of regression coefficients is complicated by the fact that each variable’s marginal effects will be conditional on the value of any variable it is interacted with, and by the fact that each coefficient can only be interpreted as the marginal effect of a variable when any variable it is interacted with equals zero.

**II. Dataset**: *alexseev.dta*

**III. Packages**: *compactr*, *foreign*, *ggplot*

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear any data in memory by typing rm(list=ls())

3) Download dataset *alexseev.dta* and place it in a working directory

4) Download R script “Lab 13 2021.R” and place it in a working directory.

5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

6) Run lines 2−4 in the R script to load packages that you can use to make more attractive figures and the package that you will need to open the dataset.

**V. Instructions for Lab 13**

You will use data on ethnocentric (anti-immigrant) voting during the 2003 Russian State Duma elections. This data was collected by Mikhail Alexseev and for his 2006 *Political Behavior* article, “Ballot-Box Vigilantism: Ethnic Population Shifts and Xenophobic Voting in Post-Soviet Russia.” William Berry, Matt Golder, and Daniel Milton re-analyzed the data on pages 665–668 of their 2012 *Journal of Politics* article, “Improving Tests of Theories Positing Interactions.” I posted both articles in Blackboard.

Load the dataset by running lines 5–6, changing the directory if needed. The dataset that can be downloaded from the internet has two additional lines of NA values, which create some trouble in trying to find means and other summary statistics. You will work with a data frame named alexseev.

The main dependent variable, *e03ld*, is the percent of votes cast for the extreme nationalist party, Zhirinovsky Bloc, in 2003. The main party in this bloc, LDPR, persistently ran on a campaign platform of “Russia for Ethnic Russians.”

Run lines 7–8 to examine the normality of the dependent variable (*e03ld*) visually. In line 7, you generate the normal quantile plot of *e03ld* against the normal distribution; any deviation from a straight line indicates a non-normal distribution. In line 8, you generate a histogram of *e03ld*; any deviation from a bell curve would indicate a non-normal distribution.

> qqnorm(alexseev$e03ld)

> hist(alexseev$e03ld, freq = FALSE, breaks=19)

One reason for concern about the distribution of the dependent variable is that it is truncated below at 0.

The two main independent variables are *slav89*, which is the percent of the population that belonged to the dominant ethnic group (Slavs) in 1989, and *nonslav8*, which the change in the percentage of the population accounted for by ethnic minorities between 1989 and 2002.

Alexseev’s preferred theory, named the “Defended Nationhood” hypothesis, posits that a greater Slavic share of the population should be positively associated with the Zhirinovsky Bloc’s vote share and that this effect should be enhanced by a large influx of ethnic minorities, hence the inclusion of a multiplicative interaction term.

Alexseev also argues that greater increases in the percentage of ethnic minorities should increase support for anti-immigrant parties regardless of Slavic population.[[1]](#footnote-1)

Alexseev also includes six control variables:

* *inc9903* is a measure of change in average personal income from 1999 to 2003;
* *eduhi02* is the percent of the population with a college education in 2002;
* *unemp02* is the percent of the working age population claiming unemployment benefits in 2002;
* *apt9200* is the percent of apartments or houses converted to private ownership from the start of privatization in 1992 to 2000;
* *vsall03* is a region’s population size; and
* *brdcont* is a dummy variable identifying regions located along Russia’s borders over which it had a dispute with a neighboring state at any time between 1991 and 2003.

Run lines 10–11 to estimate Alexseev’s main model, which he labels as Test 1 in his Table 2 (2006: 225), and to inspect the results. (You may spend some time inspecting the constant and the control variables’ effects, but our focus in the remainder of lab will be on the *slav89* and *nonslav8* variables, and their interaction.) The remaining columns in Table 2 focus on specific non-slavic ethnic groups.

If you have run the model correctly, then you should see the following regarding his main hypothesis:

* *slav89* has a positive regression coefficient, which is consistent with Alexseev’s theory, but it is statistically insignificant;
* *nonslav8* has a negative regression coefficient, which is inconsistent with Alexseev’s theory, and it is statistically significant;
* the interaction of *slav89* and *nonslav8* has a positive regression coefficient, which is consistent with Alexseev’s theory, but it is statistically insignificant.

If you have kept up with the readings, then by now you should be able to answer the following question: Why are these three bulleted findings virtually meaningless – especially the first two?

Recall that when you are examining regression coefficients for variables that are in a multiplicative interaction, the coefficient only tells you the impact when the other variable in the interaction equals zero.

That is, when including an interaction of *x* and *z*, the marginal effect of *x* on *y* equals *βx* + *βxz*⋅*z*; this implies that the marginal effect equals only when *z* = 0. So, it is more useful to examine the marginal effect of each independent variable at their *typical* values instead. Moreover, multiplicative interactions have a tendency to inflate standard errors, so statements about statistical significance are dubious.

Run lines 13–15 to examine the distribution of the interaction term, *sl89nsl8*, and its two constituent variables, *slav89* and *nonslav8*; save the mean values of the constituent terms, named *s.bar* and *ns.bar*:

> s.bar = summary(alexseev$slav89)[4]

> ns.bar = summary(alexseev$nonslav8)[4]

Run line 16 to calculate the interaction of the mean values (which is not equal to the mean of the interaction term). Next, run line 18 to find the marginal effect of *slav89* when *nonslav8* is set at its mean value:

> model$coef[2] + model$coef[4]\*ns.bar

Then, run line 19 to find the marginal effect of *nonslav8* when *slav89* is set at its mean value:

> model$coef[3] + model$coef[4]\*s.bar

Later in this lab, we will come back to whether these marginal effects are statistically significant.

To get more insight into the marginal effects, let us explore how the predicted values of vote for the Zhirinovsky Bloc changes when the values of the *slav89* and *nonslav8* variables change from their 25th percentile to their median, and from their median to the 75th percentile.

Begin by running line 21 to define the mean value of all the control variables, and then run line 22 to create a baseline value of *e03ld* by adding the intercept to the products found by multiplying each coefficient times its respective control variable’s mean value. If you were to inspect the summary of *e03ld*, then you might discover that the average is about 12.5 percent; our baseline value is a bit low because it implicitly holds the *slav89* and *nonslav8* variables, as well as the interaction term, *sl89nsl8*, equal to zero. You can run line 23 to check an equality that we found back in week 2: the fitted value of the dependent variable holding all independent variables at their averages equals the average value of the dependent variable.

When calculating a bunch of predicted values in order to inspect marginal effects, it is often extremely useful to create a new dataset of inputs. The approach we are about to take is not the most elegant, but it is practical. We are going to have three variables (*slav89*, *nonslav8*, and their interaction) set at three different values (first quartile, median, and third quartile), and then calculate predicted values by multiplying each model coefficient times a particular value of the corresponding variable, and adding the baseline value.

If you run lines 25–28, you will create the new data frame. The *slav89* variable is set at first quartile, median, and third quartile – in that order – three times. The *nonslav8* variable is set at its first quartile three times, at its median three times, and at its third quartile three times. The interaction is created from the product of *slav89* and *nonslav8*. (Therefore, the fifth observation out of nine has all three variables at their median value.)

Run line 29 to create a column of predicted values in the data frame, which is named *new.data*. If you simply type *new.data* in the **Console** window, then RStudio will display the data frame for you to view. The first three columns show you the values that were set for the explanatory variables, and the fourth column shows you the predicted value of the dependent variable.

If you focus on the fourth, fifth and sixth predicted values, this shows you how the predicted value of *e03ld* changes in response to increases in *slav89*, holding *nonslav8* at its median. You should observe that increases in *slav89* correspond to increasing support for the Zhirinovsky Bloc, which is consistent with Alexseev’s theory.

If you focus on the second, fifth and eighth predicted values, this shows you how the predicted value of *e03ld* changes in response to increases in *nonslav8*, holding *slav89* at its median. You should observe that increases in *nonslav8* actually correspond to slight decreases in support for the Zhirinovsky Bloc, which is inconsistent with Alexseev’s theory.

Demonstrating that predicted support for the Zhirinovsky Bloc increased with increases in *slav89* or decreased with increases in *nonslav8* does not prove that either variable has a statistically significant effect on the value of the dependent variable. We will postpone a discussion of statistical significance until after we perform two more tasks and then use R to calculate standard errors. Also, we have not directly addressed the main question of whether increases in *nonslav8* tend to heighten the effect of changes in *slav89*. We can view this indirectly by seeing the following:

* when the change in percent minority population is low (*nonslav8* = 0.1568), *yhat* increases from 12.96 to 13.45 as *slav89* increases from 82.14 to 97.16, a difference of \_\_\_\_\_
* when the change in percent minority population is high (*nonslav8* = 2.2550), *yhat* increases from 12.59 to 13.34 as *slav89* increases from 82.14 to 97.16, a difference of \_\_\_\_\_

Three major tasks remain: to put R’s power to work by graphing predicted values, to put the power of R to work by graphing marginal effects, and to perform tests of statistical significance on marginal effects.

You are about to create a new variable with many values of *slav89*, and then use the basic equation for predicted values, i.e., , to inspect how the value of *e03ld* is expected to vary as *slav89* increases, at the same three values of *nonslav8* identified above.

Run the code in line 31 to create a variable with 101 values, evenly spaced between the minimum and the maximum values of *slav89*. Next, run lines 32–34 to create predicted vote shares for the Zhirinovsky Bloc for all values in the domain of *slav89*, using the baseline value we created earlier (line 22), and setting *nonslav8* equal to its first quartile (line 32), its median (line 33), and its third quartile (line 34). I have included the code from line 33 for your inspection:

> yhat.q2 <- baseline + model$coef[2]\*index + model$coef[3]\*sumns[3] + model$coef[4]\*index\*sumns[3]

Thus, you are adding the (a) baseline value, plus (b) the product of the coefficient for *slav89* and a value of *slav89*, plus (c) the product of the coefficient for *nonslav8* and the median value of *nonslav8*, plus (d) the product of the interaction term and the median value for *nonslav8* and a value of *slav89*.

Run the code in lines 35−36 to plot the three lines generated when *nonslav8*’s value is set at its first quartile (shown in red), at its median (shown in black), and at its third quartile (shown in blue). You should have expected to see all three of these lines being upward sloping.

If the posited interaction effect for *nonslav8* is correct, then the steepest line should be the blue line (more rapid increase in minority population) and the shallowest line should be the red line (slower increase in minority population). Are these expectations met?

If the posited its main effect for *nonslav8* is correct, the blue line’s value should diverge above the black line as *slav89*’s value increases, and the red line’s value should diverge below the black line as *slav89*’s value increases. Are these expectations met?

The next step is to investigate the interaction effect at all values in the domain of *nonslav8*. To foreshadow, you are about to create a new variable with many values of *nonslav8*, and then use the basic equation for marginal effects for *x*, i.e., , to inspect how the value of the partial derivative with respect to *slav89* (∂ *e03ld* /∂ *slav89*) changes as the value of *nonslav8* increases.

Run the code in line 38 to create a variable with 101 values, evenly spaced between the minimum and the maximum values of *nonslav8*. Next, run line 39 to create marginal effects for *slav89* at all values in the domain of *nonslav8*. I have included the code from line 39 for your inspection:

> dydsl89 <- model$coef[2]+model$coef[4]\*smooth

You can run lines 40–45 or type the following code to plot marginal effects over the domain of *nonslav8*:

> plot(smooth, dydsl89, cex=.25, xlim=c(min(smooth),max(smooth)))

Take the time to examine closely the equations for creating this plot. Students who have learned calculus will have an easier time understanding that the marginal effect of *x* is the first derivative with respect to *x* of the function *y* = *f*(*x,z*) = *b*0 + *b*1*x + b*2*z + b*3*xz*; using the rules for derivatives yields ∂*y*/∂*x* = *b*1 + *b*3*z*.

So, you must add the coefficient for *slav89* and the coefficient for the interaction times a given value of *nonslav8*.

Nine lines of code (47–51 plus 54–57) provide one way to plot marginal effects and confidence intervals around those marginal effects. Recall that in a linear model, the marginal effect equals the regression coefficient, and there is a single confidence interval around that effect. For a model with an interaction, however, the marginal effect will depend on the value of *z* (as you saw above), but the confidence interval also depends on the value of *z*. You will start by saving the variance-covariance matrix of the estimators:

> vce=vcov(model)

Next you will save three entries: the variance of the main effect coefficient for *slav89*, the variance of the interaction effect coefficient, and the covariance of the aforementioned main effect coefficient and the interaction effect coefficient:

> varsl89<-vce[2,2]

> varsl89nsl8<-vce[4,4]

> cova<-vce[2,4]

Next you will combine these three terms using the equation for the standard error for a marginal effect, i.e., , at all 1000 values of *nonslav8* that you generated earlier:

> sedydsl89<-sqrt(varsl89 + (smooth^2)\*varsl89nsl8 + 2\*smooth\*cova)

If you wish to see how the standard error varies with the value of *nonslav8*, then you can run line 52.

To generate the confidence intervals around the marginal effect, you will need to multiply a *t* critical value times the standard error. Many of the scholarship about interaction terms (Kam and Franzese; Matt Golder and his co-authors) advise using a 90% confidence interval. Line 54 provides code for finding the *t* critical value associated with 5% in the upper tail of the *t* distribution, for degrees of freedom equal to *n*–*k*–1 from the regression model run earlier.

> t.star = qt(.95,model$df.residual)

Run lines 55–56 to create the upper and lower bounds for the confidence interval for the slope; you are adding and subtracting 1.96 times the standard error (computed in line 69) to/from the slope.

> upperci <- dydsl89 + t.star\*sedydsl89

> lowerci <- dydsl89 - t.star\*sedydsl89

Finally, run line 57 to plot the marginal effect and its confidence interval, with a horizontal line for zero. The curve shown by running line 57 is not the predicted value of *e03ld* given the value of *slav89*; rather the curve shows the slope of the function relating share of the population belonging to the ethnic majority to support for the Zhirinovsky Bloc.

When the marginal effect is above zero, it indicates that in this range, increases in *slav89* have a positive effect on support for the Zhirinovsky Bloc. When the marginal effect plot slopes upward, it indicates that the impact of ethnic majority population on support for the Zhirnovsky Bloc increases as the change in the percent of ethnic minorities increases.

You might be wondering whether someone has automated this process. The answer is yes; the *interplot* package is one way of doing this. Run line 59 to install the package and run line 60 to activate it.

Run line 61 of to re-estimate the model with an interaction rather than the interaction variable; take a moment to inspect the difference between model (which you ran in line 10) and model2.

You can run lines 62–69 to show a basic plot (most of the code has to do with labels), or alternatively, you can run lines 71–78 to include a histogram of values of *nonslav8* (the rest of the code is identical). Note that the default is to use a 95% confidence level; this can be changed easily by typing ci = 0.90, inside the first set of parentheses.

Pretty cool, right?

To clear the **Environment**, type rm(list=ls()) or click on the broom icon.

To clear the **Console** window, type Ctrl-*l*

1. Alexseev’s theory is based on studies of anti-migrant vigilantism in American cities and anti-minority hate crimes in New York’s 51 boroughs, so this is an interesting application of American scholarship to comparative politics. [↑](#footnote-ref-1)