Tutorial 7: Dummy Variables and Interactions (II)

Jan Vogler (jan.vogler@duke.edu)

October 9, 2015

Today's Agenda

1. Interaction terms with interval-level variables
2. Graphical representation of interactions
3. Analysis of Variance (ANOVA)
4. Finding variables that represent theoretical concepts
5. Expectations for research outline

Briefly: unit fixed effects

Last time we ran a (pooled) regression and did not include unit-fixed effects. Fixed effects are dummy variables for each unit, i.e. country in our data.

How would we include fixed effects?

```r
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/")
library(foreign)
LDC = read.dta("LDC_IO_replication.dta")
main_fe = lm(newtar ~ l1polity + l1signed + l1office + l1gdp_pc + l1lnpop +
    l1ecris2 + l1bpc1 + l1avnewtar + factor(ctylabel) - 1, data = LDC)
```

How is this different from our old method?

1. Interaction terms with interval-level variables

In today's session we will build upon the last lab and continue to work on interaction effects.

Last time we looked at interactions between a continuous and a binary variable. This time both our variables will be continuous.

For this purpose we will again refer to the dataset by Milner and Kubota.

Let us try to test a hypothesis that combines claims from modernization theory and the literature on FDI.

1. Modernization theory claims that as countries develop economically, societies become more complex and people more educated which leads to a process of democratization. (Lipset 1959, Boix & Stokes 2003)
2. The literature on FDI claims that FDI is an important driver of economic growth. It can increase economic development by allowing for technology transfers from other countries. (Jensen 2003, Damijan et al. 2003)

Following these two claims of the political-economic literature, one might argue that there is a positive interaction of GDP and FDI inflows: because FDI means technology transfers—and potentially educational effects—meaning that societies acquire production processes of higher complexity, the effect of economic development on democratization may be magnified.

Let us test this hypothesis through a simple regression model with an interaction term. We use our data on developing countries as there is diversity in terms of all three variables we are interested in: democracy scores, FDI inflows, and GDP per capita.

Let us first specifically look at the values of the variables that we are interested in.

```
summary(LDC$polityiv_update2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## -10.000  -7.000  -6.000  -2.074   6.000  10.000    2003 
```

```
summary(LDC$l1gdp_pc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       0     442    1266    2888    2999   44160    1823 
```

```
summary(LDC$l1fdi)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -27.2400   0.0269   0.6382   1.7930   1.9900 184.6000     2423 
```

It might be interesting to see what the extreme outliers are in those cases.

Let us check this with the following commands:

```
which(LDC$l1gdp_pc > 20000)
```

```
##  [1] 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1655 1656 1657 1658
## [15] 1659 1660 1661 1662 1663 1664 1665 1666 1667 1672 1895 1896 1897 1898
## [29] 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 2034 2035 2036 2037
## [43] 2038 2039 2040 2365 2366 2367 2368 2369 2370
```

```
LDC[1472, ]
```

```
##      country ctylabel date gatt_wto_new aclpn bpc1 dopen_wacz2 ecris2
## 1472     443   Kuwait 1971            1     0   NA          NA     NA
##      fdignp gdp_pc_95d l1aclpn l1bpc1 l1ecris2 newtar polityiv_update2
## 1472     NA   32146.39       0     NA       NA     NA               -8
##      signed yrsoffic     usheg   l1usheg l1fiveop l1gdp_pc      avsw
## 1472      0       NA 0.2678981 0.2756129     10.2 31940.31 0.1505376
##      avnewtar    l1avsw l1avnewtar    lnpop l1lnpop l1office
## 1472        0 0.1505376          0 13.58728 13.5198       NA
##      l1partyage2000 l1fdi l1polity l2polity l3polity l1signed milit2 sp2
## 1472             NA    NA       -9       NA       NA        0      0   0
##      pers2 l1milit2 l1sp2 dictator1 l1dictator1 yr70 yr80 l1ssch closedyr
## 1472     0        0     0         2           2    1    0  1.232       NA
##      _spline1 _spline2 _spline3 l1gatt_wto_new
## 1472       NA       NA       NA              1
```

```r
# It's Kuwait

which(LDC$l1fdi > 50)
```

```
## [1]  984 2847 2848
```

```r
LDC[2847, ]
```

```
##      country         ctylabel date gatt_wto_new aclpn bpc1 dopen_wacz2
## 2847     642 EquatorialGuinea 1996            0     0    1          NA
##      ecris2   fdignp gdp_pc_95d l1aclpn l1bpc1 l1ecris2 newtar
## 2847      0 184.5647    517.211       0      1        0     NA
##      polityiv_update2 signed yrsoffic   usheg   l1usheg l1fiveop l1gdp_pc
## 2847               -5      0       17 0.26817 0.2627195     13.2 410.9776
##           avsw avnewtar    l1avsw l1avnewtar    lnpop  l1lnpop l1office
## 2847 0.6344086 15.04722 0.5913978   16.13231 12.92255 12.89672       16
##      l1partyage2000    l1fdi l1polity l2polity l3polity l1signed milit2
## 2847            9.5 82.97678       -5       -5       -5        0      0
##      sp2 pers2 l1milit2 l1sp2 dictator1 l1dictator1 yr70 yr80 l1ssch
## 2847   0     0        0     0         2           2    0    0     NA
##      closedyr _spline1 _spline2 _spline3 l1gatt_wto_new
## 2847       NA       NA       NA       NA              0
```

```r
# It's Equatorial Guinea
```

Now let us estimate a model with an interaction effect.

```r
# Second: estimate the model
intmodel = lm(polityiv_update2 ~ l1gdp_pc + l1fdi + l1gdp_pc * l1fdi, data = LDC)
summary(intmodel)
```

```
## 
## Call:
## lm(formula = polityiv_update2 ~ l1gdp_pc + l1fdi + l1gdp_pc *
##     l1fdi, data = LDC)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -21.687  -4.959  -2.689   7.206  12.278 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -2.724e+00  1.951e-01 -13.957  < 2e-16 ***
## l1gdp_pc        1.006e-03  1.042e-04   9.653  < 2e-16 ***
## l1fdi          -5.063e-02  3.696e-02  -1.370    0.171    
## l1gdp_pc:l1fdi  1.726e-04  3.475e-05   4.967 7.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.755 on 2372 degrees of freedom
##   (2994 observations deleted due to missingness)
## Multiple R-squared:  0.08849,    Adjusted R-squared:  0.08734 
## F-statistic: 76.76 on 3 and 2372 DF,  p-value: < 2.2e-16
```

How would we interpret the interaction of GDP per capita and FDI? Is it problematic that one of the terms is not significant?

In-class exercise: interaction terms with two interval-level variables

Assume that you have the following linear model

Y = 10 + (5) * X1 + (2) * X2 + (1) * X1 * X2 + epsilon

1. Calculate the derivative of Y with respect to X1 and X2.
2. Use R to plot the marginal effect of X1 at different levels of X2. Assume that X2 is an integer that varies between -10 and 10.

Hint: To plot this in R you need to create a vector with values for X2 for your x-axis and a vector with the respective marginal effect of X1 on your y-axis.

How does this relate to the above problem?

How would we plot the marginal effect of GDP per capita at different values of FDI inflows?

Let us first check out the coefficients of our model.

```r
intmodel$coefficients
```

```
##    (Intercept)       l1gdp_pc          l1fdi l1gdp_pc:l1fdi 
##  -2.7235846254   0.0010063348  -0.0506297183   0.0001726174 
```

Now we need to look at possible values of FDI inflows:

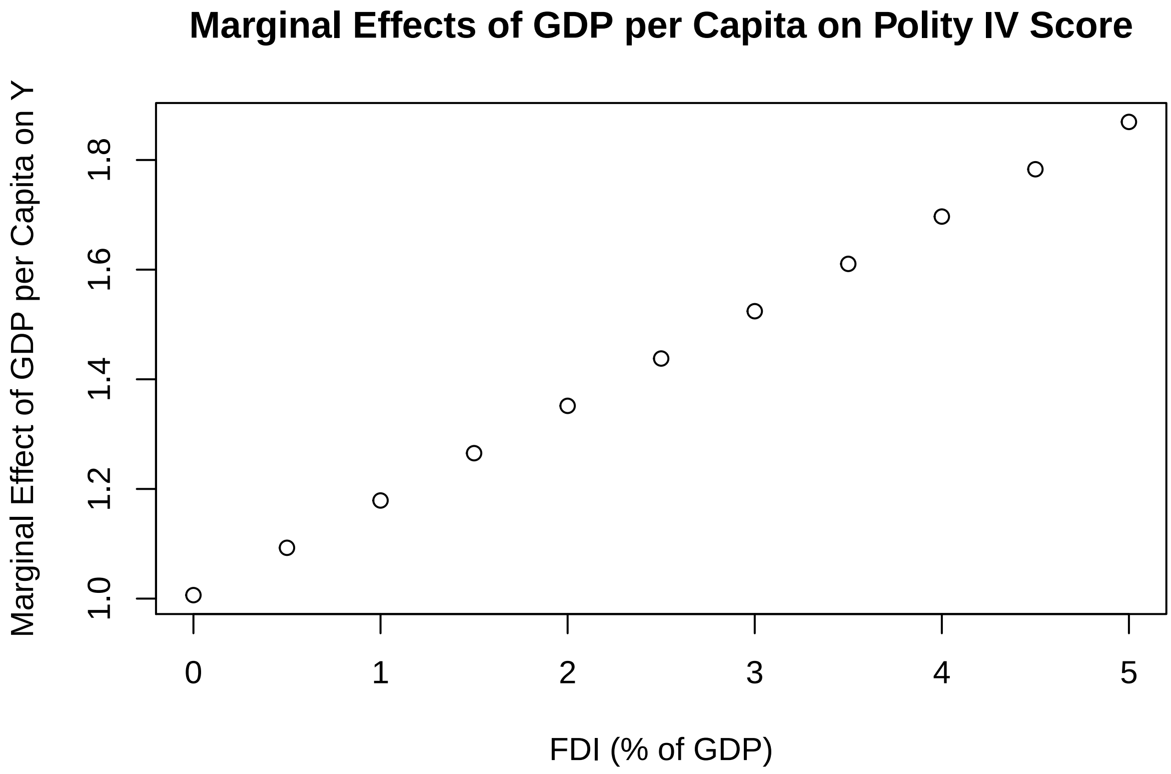
```r
summary(LDC$fdignp)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -27.2400   0.0361   0.6644   1.8960   2.0830 184.6000     2294 
```

```r
quantile(LDC$fdignp, probs = c(0.1, 0.9), na.rm = T)
```

```
##      10%      90% 
## 0.000000 5.202046 
```

```r
fdi_values = seq(0, 5, by = 0.5)

### Note the marginal effect is for an increase of GDP per Capita by 1 USD In
### order to make interpretation easier, we will look at increases by 1000 USD

marginal = rep(0.0010063348 * 1000, 11) + 0.0001726174 * fdi_values * 1000
plot(fdi_values, marginal, type = "p", main = "Marginal Effects of GDP per Capita on Polity IV Score",
    xlab = "FDI (% of GDP)", ylab = "Marginal Effect of GDP per Capita on Y")
```



2. Graphical representation of interactions

In order to graphically represent the effect that GDP per capita has at different levels of FDI, we create several new dataframes.

```
quantile(LDC$l1fdi, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
```

```
##       25%       50%       75% 
## 0.0269332 0.6382053 1.9903931
```

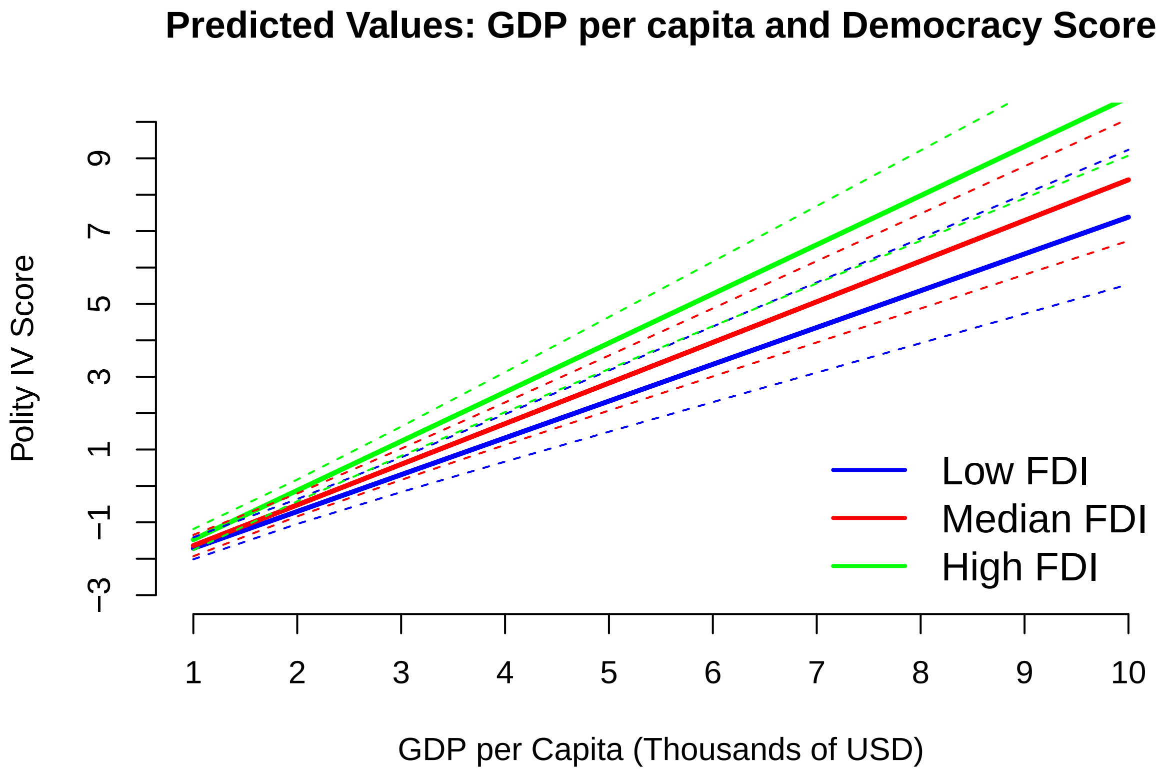
```
nd1 = data.frame(l1gdp_pc = seq(1000, 10000, by = 1000), l1fdi = rep(0.0269332,
    10))
nd2 = data.frame(l1gdp_pc = seq(1000, 10000, by = 1000), l1fdi = rep(0.6382053,
    10))
nd3 = data.frame(l1gdp_pc = seq(1000, 10000, by = 1000), l1fdi = rep(1.9903931,
    10))
```

Next we use the model we estimated to predict values based on this new dataframe.

```
pred.p1 = predict(intmodel, type = "response", se.fit = TRUE, newdata = nd1)
pred.p2 = predict(intmodel, type = "response", se.fit = TRUE, newdata = nd2)
pred.p3 = predict(intmodel, type = "response", se.fit = TRUE, newdata = nd3)
pred.table1 = cbind(pred.p1$fit, pred.p1$se.fit)
pred.table2 = cbind(pred.p2$fit, pred.p2$se.fit)
pred.table3 = cbind(pred.p3$fit, pred.p3$se.fit)
```

Finally, we create the plot. Let's start by adding the first values.

```
plot(pred.p1$fit, type = "l", ylim = c(-3, 10), main = "Predicted Values: GDP per capita and Democracy S
    xlab = "GDP per Capita (Thousands of USD)", ylab = "Polity IV Score", axes = FALSE,
    col = "blue", lwd = 2.5)
axis(1, at = seq(1, 10), labels = seq(1, 10))
axis(2, at = seq(-3, 10), labels = seq(-3, 10))

### Next: we add lines

lines(pred.p2$fit, col = "red", lwd = 2.5)
lines(pred.p3$fit, col = "green", lwd = 2.5)

### Let us add a legend to our plot, so it's more obivous what we did here.

legend("bottomright", c("Low FDI", "Median FDI", "High FDI"), lty = 1, lwd = 2,
    col = c("blue", "red", "green"), bty = "n", cex = 1.25)

### We can also add confidence intervals to our plot, though this will make it
### a little less clear.

fit1 = pred.p1$fit
low1 = pred.p1$fit - 2 * pred.p1$se.fit
high1 = pred.p1$fit + 2 * pred.p1$se.fit
cis1 = cbind(fit1, low1, high1)

fit2 = pred.p2$fit
low2 = pred.p2$fit - 2 * pred.p2$se.fit
high2 = pred.p2$fit + 2 * pred.p2$se.fit
cis2 = cbind(fit2, low2, high2)

fit3 = pred.p3$fit
low3 = pred.p3$fit - 2 * pred.p3$se.fit
high3 = pred.p3$fit + 2 * pred.p3$se.fit
cis3 = cbind(fit3, low3, high3)

matlines(cis1[, c(2, 3)], lty = 2, col = "blue")
matlines(cis2[, c(2, 3)], lty = 2, col = "red")
matlines(cis3[, c(2, 3)], lty = 2, col = "green")
```



Be very cautious about any causal interpretation here. We probably have several problems associated with causal inference, including omitted variable bias and endogeneity.

Let's naively assume that we don't deal with those problems here. Would the predicted values support our initial hypothesis?

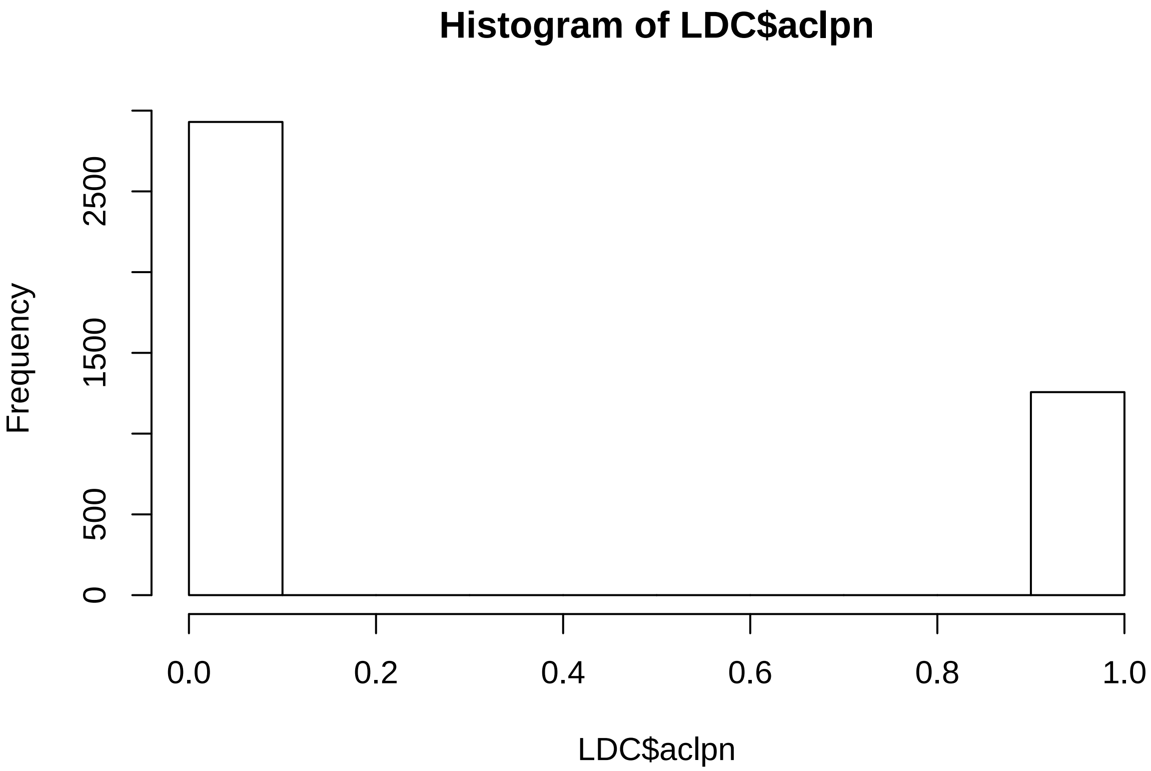
3. Analysis of Variance (ANOVA)

Let us conduct an analysis of variance. We are interested in the level of FDI inflows and we look at two different binary variables:

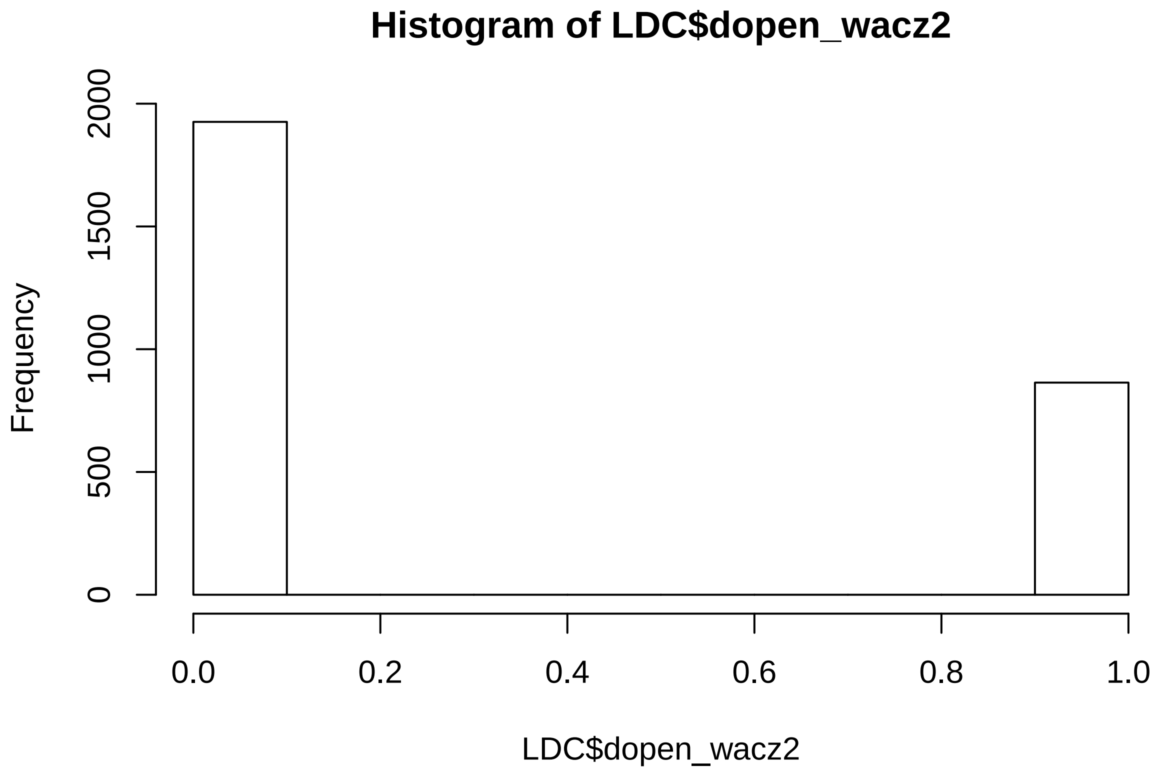
1. Democracy - coded as 0 and 1
2. Open Economy - coded as 0 and 1

How would we expect these variables to influence the level of FDI inflows?

```
hist(LDC$aclpn)
```



```
hist(LDC$dopen_wacz2)
```



Let us run look at a one-way ANOVA table.

```
anovamodel1 = lm(fdignp ~ aclpn, data = LDC)
summary(anovamodel1)
```

```
##
## Call:
## lm(formula = fdignp ~ aclpn, data = LDC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.736  -1.501  -1.114   0.219 183.064
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5007     0.1189  12.623  < 2e-16 ***
## aclpn         1.1995     0.2023   5.928 3.41e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.293 on 3025 degrees of freedom
##   (2343 observations deleted due to missingness)
## Multiple R-squared:  0.01148,    Adjusted R-squared:  0.01116
## F-statistic: 35.14 on 1 and 3025 DF,  p-value: 3.408e-09
```

```
anova(anovamodel1)
```

```
## Analysis of Variance Table
##
## Response: fdignp
##             Df Sum Sq Mean Sq F value    Pr(>F)
## aclpn        1    984  984.48  35.144 3.408e-09 ***
## Residuals 3025  84739   28.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How would we interpret these results? Compare to the results of the OLS regression.

Let us look at a two-way ANOVA table.

```
anovamodel2 = lm(fdignp ~ aclpn + dopen_wacz2 + aclpn * dopen_wacz2, data = LDC)
summary(anovamodel2)
```

```
##
## Call:
## lm(formula = fdignp ~ aclpn + dopen_wacz2 + aclpn * dopen_wacz2,
##     data = LDC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.294  -1.059  -0.710   0.313  82.253
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.05857    0.09506  11.136  < 2e-16 ***
## aclpn              0.04466    0.20076   0.222 0.823982
## dopen_wacz2        0.40981    0.20384   2.010 0.044501 *
## aclpn:dopen_wacz2  1.06032    0.32010   3.312 0.000939 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 2360 degrees of freedom
##   (3006 observations deleted due to missingness)
## Multiple R-squared:  0.02566,    Adjusted R-squared:  0.02442
## F-statistic: 20.72 on 3 and 2360 DF,  p-value: 3.008e-13
```

```
anova(anovamodel2)
```

```
## Analysis of Variance Table
##
## Response: fdignp
##                     Df  Sum Sq Mean Sq F value    Pr(>F)
## aclpn                1   258.3  258.27  22.630 2.083e-06 ***
## dopen_wacz2          1   325.9  325.85  28.551 1.000e-07 ***
## aclpn:dopen_wacz2    1   125.2  125.23  10.972 0.0009388 ***
## Residuals         2360 26934.8   11.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How would we interpret these results? Compare to the results of the OLS regression.

4. Finding variables that represent theoretical concepts

We are interested in finding a measurement for veto players. Veto players can be described as the number of institutions in a political system whose approval is required when changes to the status quo are attempted to be made.

How can we measure this theoretical concept? Which variables might be most useful?

In class-exercise: turning a concept into a measurement

1. Find the raw data and the codebook of the Polity IV Score.
2. Download both.
3. Answer the following questions:

a. Which component of the Polity IV Score best represents the theoretical concept of the number and importance of veto players?
b. What are the values that the Polity IV Score can take?
c. Find Afghanistan's Polity IV Score in 1993. What is the meaning of this score?

5. Expectations for research outline

Due on Thursday after fall break.

We won't have a problem set over fall break.

1. What is your theory?
2. How can you turn this theory into a testable hypothesis?
3. Which data is out there that would allow you to test your hypothesis?

Have a great fall break!