Regression III: Lab 1

Dave Armstrong
University of Western Ontario
Department of Political Science
Department of Statistics and Actuarial Science (by courtesy)

e: dave.armstrong@uwo.ca
w: www.quantoid.net/teachicpsr/regression3/

1 Question 1

Do the following in R:

```
library(foreign)
dat <- read.dta("http://www.quantoid.net/files/reg3/lab1_nes.dta")
```

These are data from the American National Election Study. The variables included are:

pid	7-point party ID variable (Strong Dem=1, Strong Rep = 7)
pid3	3-point party ID variable (1=Dem, 2=Ind, 3=Rep)
demtherm	Democratic feeling thermometer
reptherm	Republican feeling thermometer
diffthermm	Difference between democratic and republican thermometers
age	Survey respondent's age
income	23-category income variable
race	Un-recoded race variable
racerec	Recoded race variable
libcon	Liberal-conservative ideology (smaller = more liberal)

1. Is the interaction necessary?

```
mod <- lm(demtherm ~ racerec + pid3 + libcon*age, data=dat)
summary(mod)
```

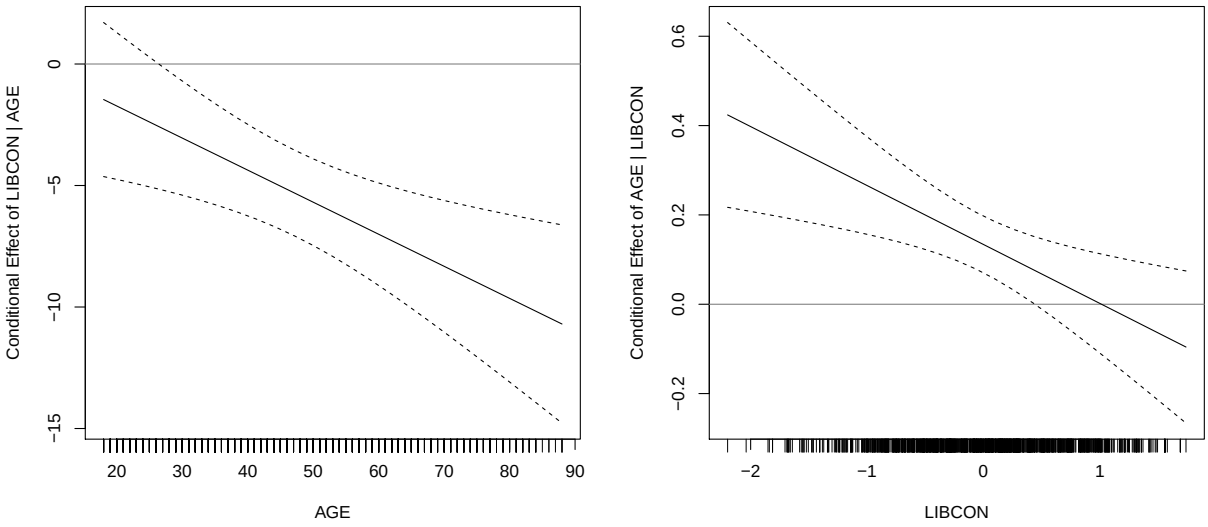
```
##
## Call:
## lm(formula = demtherm ~ racerec + pid3 + libcon * age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.534 -10.129   0.356  11.963  60.344
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              70.44821    2.08037  33.863  < 2e-16 ***
## racerecasian             -1.89569    3.51634  -0.539  0.58992
## racerecnative american   -1.13055    4.70462  -0.240  0.81014
## racerechispanic          -1.26132    2.46313  -0.512  0.60870
## racerecwhite             -7.78947    1.57034  -4.960 8.11e-07 ***
## pid3Independent         -16.14028    1.99211  -8.102 1.39e-15 ***
## pid3Republican          -26.77403    1.32084 -20.270  < 2e-16 ***
## libcon                    0.91250    2.34797   0.389  0.69762
## age                       0.13385    0.03231   4.143 3.68e-05 ***
## libcon:age               -0.13196    0.04594  -2.872  0.00415 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 17.94 on 1134 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.455,Adjusted R-squared:  0.4507
## F-statistic: 105.2 on 9 and 1134 DF,  p-value: < 2.2e-16
```

Yes, because the interaction term is significant

2. Plot the conditional effects of both `age` and `libcon`

```
library(DAMisc)
DAintfun2(mod, c("libcon","age"))
```



3. How do you interpret each of the plots?

   The left-hand plot tell us that as people get older (higher values on age), the effect of libcon gets increasingly negative and goes from insignificant for the youngest people to signifcant and strong for the oldest people. On the other hand, the effect of age seems to decrease with liberal-conservatism. Thus, the oldest liberals are farther away from the oldest conservatives on the dependent variable than the youngest liberals are away from the youngest conservatives, on average, holding constant the other model variables.

4. What would you do differently if, instead, the interaction were between `pid3` and `age` with `libcon` as the control (i.e., not in the interaction)?

```
mod <- lm(demtherm ~ racerec + pid3*age + libcon, data=dat)
summary(mod)


##
## Call:
## lm(formula = demtherm ~ racerec + pid3 * age + libcon, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.386 -10.266  -0.004  11.883  61.058
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              66.50723    2.46072  27.028  < 2e-16 ***
## racerecasian             -2.01148    3.51291  -0.573  0.56703
## racerecnative american   -1.18985    4.70080  -0.253  0.80022
## racerechispanic          -1.18281    2.46068  -0.481  0.63084
```

```
## racerecwhite            -7.79753    1.56871  -4.971 7.70e-07 ***
## pid3Independent        -17.78472    5.88375  -3.023  0.00256 **
## pid3Republican         -16.49752    3.40090  -4.851 1.40e-06 ***
## age                      0.21517    0.04319   4.981 7.29e-07 ***
## libcon                  -5.27791    0.89853  -5.874 5.59e-09 ***
## pid3Independent:age      0.03338    0.11905   0.280  0.77921
## pid3Republican:age      -0.21825    0.06655  -3.279  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.92 on 1133 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.4567,Adjusted R-squared:  0.4519
## F-statistic: 95.25 on 10 and 1133 DF,  p-value: < 2.2e-16


Anova(mod)


## Anova Table (Type II tests)
##
## Response: demtherm
##           Sum Sq   Df  F value    Pr(>F)
## racerec    10413    4   8.1091 1.905e-06 ***
## pid3      133940    2 208.6073 < 2.2e-16 ***
## age         5676    1  17.6801 2.819e-05 ***
## libcon     11077    1  34.5035 5.587e-09 ***
## pid3:age    3809    2   5.9329  0.002734 **
## Residuals 363732 1133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

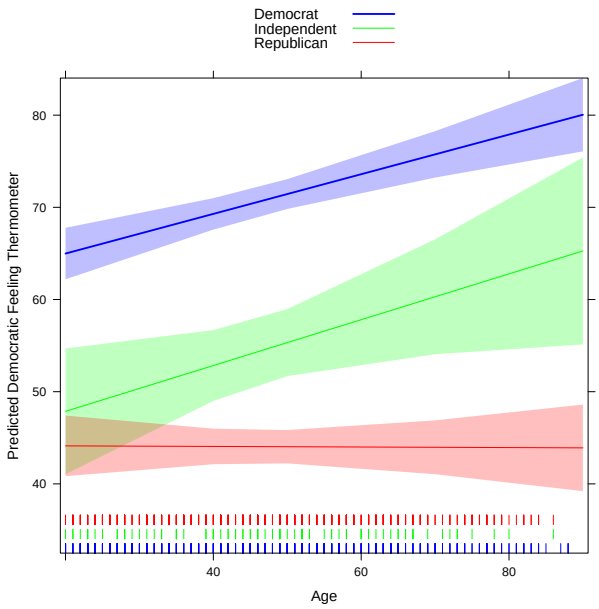
The interaction is significant. Now, we need to figure out what the interaction looks like.

```
intQualQuant(mod, c("pid3", "age"), type="slopes", plot=F)


## Conditional effects of  age :
##             B      SE(B) t-stat Pr(>|t|)
## Democrat    0.215  0.043 4.981  0.000
## Independent 0.249  0.111 2.229  0.026
## Republican  -0.003 0.052 -0.060 0.953
## $out
##                       eff         se       tstat       pvalue
## Democrat     0.215171137 0.04319451  4.98144590 7.292862e-07
## Independent  0.248553726 0.11148770  2.22942725 2.598092e-02
## Republican  -0.003078053 0.05170858 -0.05952693 9.525429e-01
##
## $varcor
##              [,1]         [,2]         [,3]
## [1,] 1.865766e-03 6.178038e-05 5.501350e-05
## [2,] 6.178038e-05 1.242951e-02 5.482559e-05
## [3,] 5.501350e-05 5.482559e-05 2.673777e-03
```

The slopes of age for Democrats and Independents are positive and significant, meaning that with age, people who identify with the Democrats or no party at all tend to see increased feeling thermometer scores toward the democrats. For Republicans, age appears to not soften their stance towards Democrats. This can also be seen in the following plot.

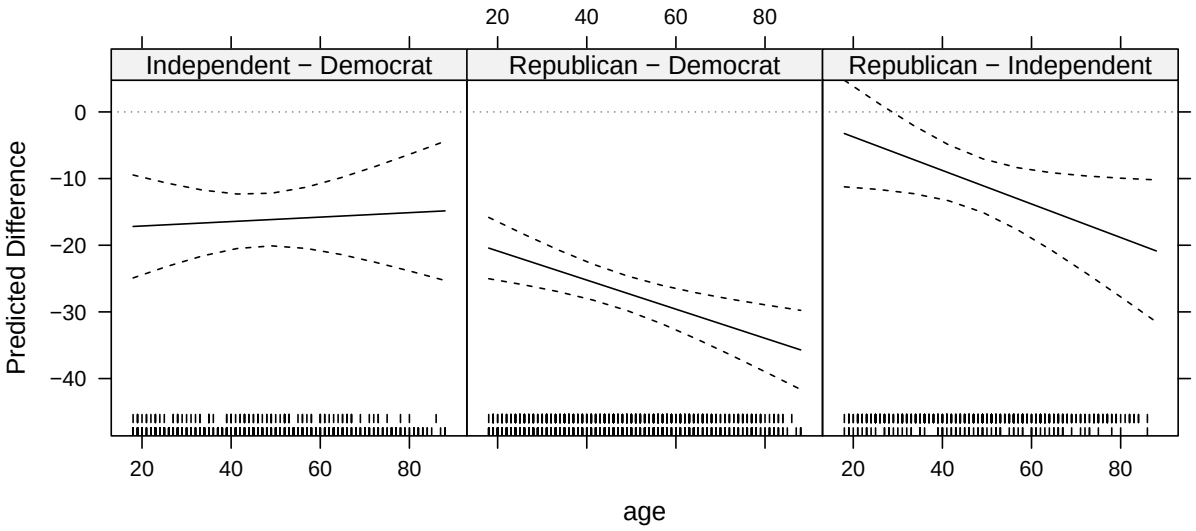
```
library(lattice)
trellis.par.set(superpose.line=list(col=c("blue", "green", "red")),
   superpose.polygon = list(col=c("blue", "green", "red")))
p1 <- intQualQuant(mod, c("pid3", "age"), type="slopes", plot=T)
update(p1, xlab="Age", ylab="Predicted Democratic Feeling Thermometer")
```



We can also look at the other side of the interaction, too.

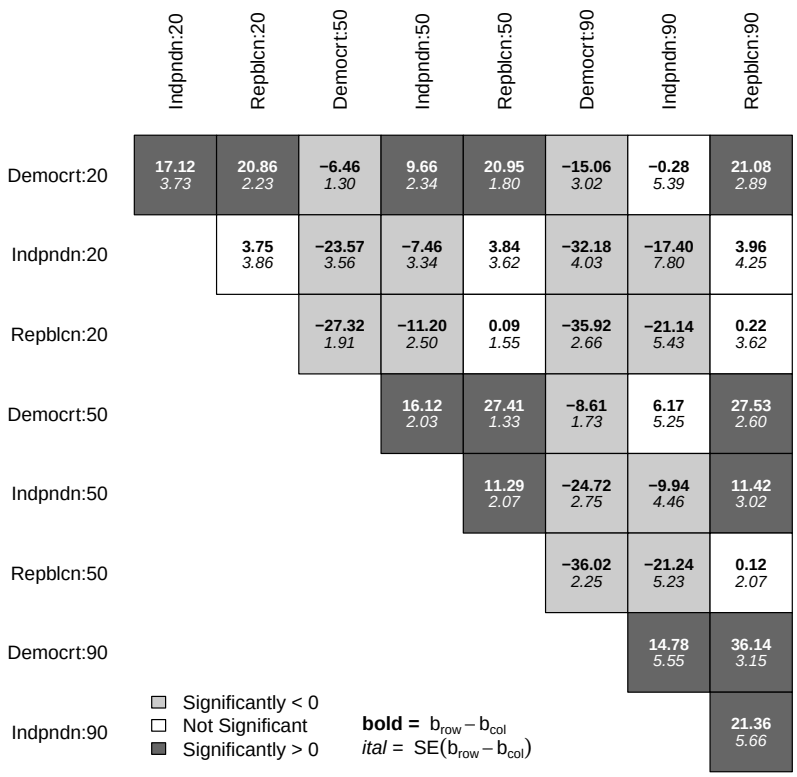
```
p2 <- intQualQuant(mod, c("pid3", "age"), type="facs", plot=T)
```

```
update(p2, layout=c(3,1), aspect=1)
```



You could also use the `effects` package along with factorplot:

```
library(factorplot)
library(effects)
eff <- effect("pid3*age", mod, xlevels=3)
plot(factorplot(eff))
```



2 Question 2

The data you will use below come from the World Bank. Do the following to read in the data:

```
dat <- read.dta("http://www.quantoid.net/files/reg3/lab1_wb.dta")
```

If you want to see the variable descriptions, you could do the following:

```
library(DAMisc)
searchVarLabels(dat, "")
```

In the exercises below, I want you to use life expectancy (`life_exp`) as the dependent variable.

1. Estimate a model of `life_exp` on civilization codes (`civ2`), percentage of the total population living in urban areas (`urban_pct_total`) and the interaction of the natural logarithm of GDP/capita (PPP) (`loggdp_pc_ppp`) and primary school expenditures as a percentage of GDP (`expend_prim`).
   (a) Is there a significant interaction?
   (b) If so, what is the nature of that interaction? What sort of inferences make sense from the data?
   (c) How would you present the results of the `civ2` variable and how would you talk about them?
2. Estimate a model of `life_exp` on primary school expenditures as a percentage of GDP (`expend_prim`), the percentage of the total population living in urban areas (`urban_pct_total`) and the interaction of the natural logarithm of GDP/capita (PPP) (`loggdp_pc_ppp`) and civilization codes (`civ2`).
   (a) Estimate a model of `life_exp` on civilization codes (`civ2`), percentage of the total population living in urban areas (`urban_pct_total`) and the interaction of the natural logarithm of GDP/capita (PPP) (`loggdp_pc_ppp`) and primary school expenditures as a percentage of GDP (`expend_prim`).

i. Is there a significant interaction?

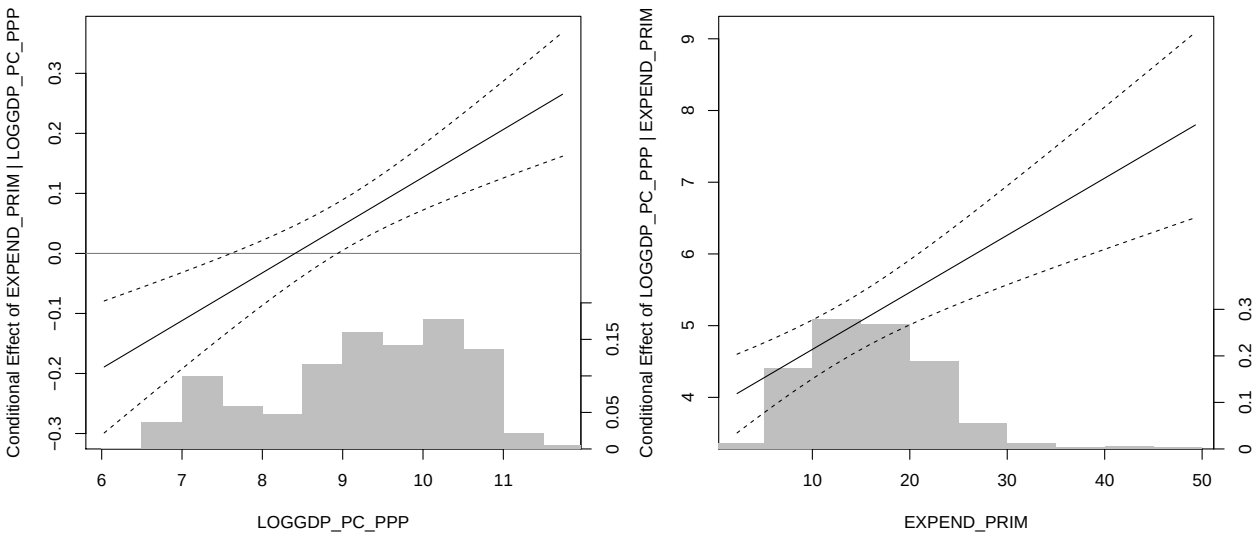
```
mod <- lm(life_exp ~ civ2 + urban_pct_total + loggdp_pc_ppp*expend_prim, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = life_exp ~ civ2 + urban_pct_total + loggdp_pc_ppp *
##     expend_prim, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5401  -1.7452   0.3833   1.9324  13.1614
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                32.95924    2.63303  12.518  < 2e-16 ***
## civ2African                -9.99139    0.46901 -21.303  < 2e-16 ***
## civ2Islamic                -2.21763    0.47175  -4.701 2.97e-06 ***
## civ2Latin American          1.69984    0.49184   3.456 0.000572 ***
## civ2Orthodox                0.29821    0.71505   0.417 0.676735
## civ2Western                -1.06781    0.43605  -2.449 0.014509 *
## urban_pct_total             0.03303    0.01024   3.225 0.001303 **
## loggdp_pc_ppp               3.87248    0.30878  12.541  < 2e-16 ***
## expend_prim                -0.66920    0.15868  -4.217 2.71e-05 ***
## loggdp_pc_ppp:expend_prim   0.07961    0.01749   4.552 5.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.847 on 958 degrees of freedom
##   (8806 observations deleted due to missingness)
## Multiple R-squared:  0.8448,Adjusted R-squared:  0.8434
## F-statistic: 579.6 on 9 and 958 DF,  p-value: < 2.2e-16
```

Since the coefficient on the interaction regressor is significant, there is a significant interaction.

ii. If so, what is the nature of that interaction? What sort of inferences make sense from the data?

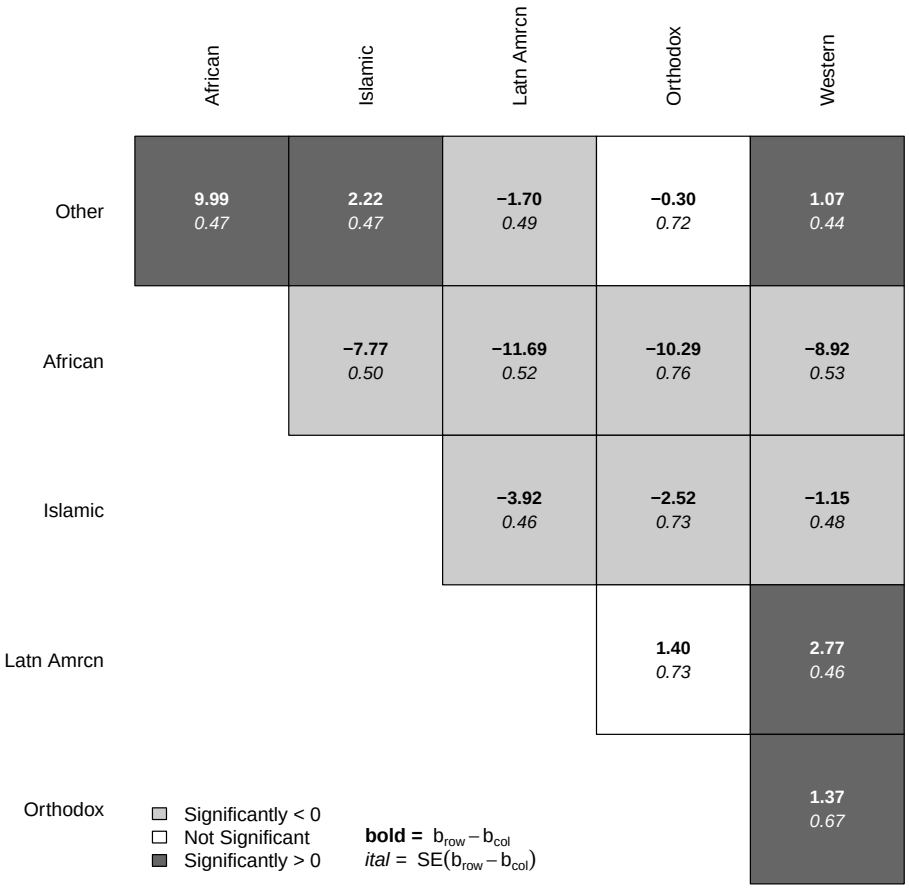
```
DAintfun2(mod, c("loggdp_pc_ppp", "expend_prim"), hist=T, rug=F, scale.hist=.3, plot.type="pdf")
```



Here, the effect of primary school expenditure is actually negative for low GDP/capita countries, but it looks like it becomes not significant around 7.5. Incidentally, only about 14% of the observations fall below 7.5 on `loggdp_pc_ppp`. After about 9 on `loggdp_pc_ppp` (around the 35^{th} percentile), the effect of primary school spending is significant and positive on life expectancy. The data are reasonably well distributed over the values of `loggdp_pc_ppp`, so most inferences we would be interested in are valid. The effect of GDP/capita on life expectancy is always positive (right panel), but it gets bigger with primary school expenditures. The data for primary school expenditures is between 5 and 30 for the most part

iii. How would you present the results of the `civ2` variable and how would you talk about them?

```
library(factorplot)
fp <- factorplot(mod, factor.variable="civ2")
plot(fp)
```



(b) Estimate a model of `life_exp` on primary school expenditures as a percentage of GDP (`expend_prim`), the percentage of the total population living in urban areas (`urban_pct_total`) and the interaction of the natural logarithm of GDP/capita (PPP) (`loggdp_pc_ppp`) and civilization codes (`civ2`).

i. Is there a significant interaction?

```
mod2 <- lm(life_exp ~ expend_prim + urban_pct_total + loggdp_pc_ppp*civ2, data=dat)
summary(mod2)
```

```
##
## Call:
## lm(formula = life_exp ~ expend_prim + urban_pct_total + loggdp_pc_ppp *
##     civ2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.441  -1.684   0.184   1.678  13.662
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     10.85791    2.76586   3.926 9.27e-05 ***
## expend_prim                      0.01429    0.02167   0.659 0.509894
## urban_pct_total                  0.02413    0.01010   2.390 0.017054 *
## loggdp_pc_ppp                    6.46423    0.33819  19.114  < 2e-16 ***
## civ2African                     18.71639    3.60511   5.192 2.55e-07 ***
## civ2Islamic                     11.72582    3.32554   3.526 0.000442 ***
## civ2Latin American              -1.73823    7.11281  -0.244 0.806989
## civ2Orthodox                    20.30653    8.39144   2.420 0.015710 *
## civ2Western                     -1.66552    4.21284  -0.395 0.692678
```

```
## loggdp_pc_ppp:civ2African        -3.45449    0.42541  -8.120 1.43e-15 ***
## loggdp_pc_ppp:civ2Islamic        -1.59723    0.37124  -4.302 1.86e-05 ***
## loggdp_pc_ppp:civ2Latin American  0.34428    0.77996   0.441 0.659019
## loggdp_pc_ppp:civ2Orthodox       -2.15750    0.88501  -2.438 0.014958 *
## loggdp_pc_ppp:civ2Western        -0.05148    0.42947  -0.120 0.904606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.729 on 954 degrees of freedom
##   (8806 observations deleted due to missingness)
## Multiple R-squared:  0.8549,Adjusted R-squared:  0.8529
## F-statistic: 432.2 on 13 and 954 DF,  p-value: < 2.2e-16
```

```
Anova(mod2)
```

```
## Anova Table (Type II tests)
##
## Response: life_exp
##                      Sum Sq  Df  F value  Pr(>F)
## expend_prim             6.0   1   0.4346 0.50989
## urban_pct_total        79.4   1   5.7109 0.01705 *
## loggdp_pc_ppp        9033.3   1 649.5803 < 2e-16 ***
## civ2                 9288.4   5 133.5859 < 2e-16 ***
## loggdp_pc_ppp:civ2   1221.4   5  17.5663 < 2e-16 ***
## Residuals           13266.6 954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction according to the ANOVA results.

ii. If so, what is the nature of that interaction? What sort of inferences make sense from the data?

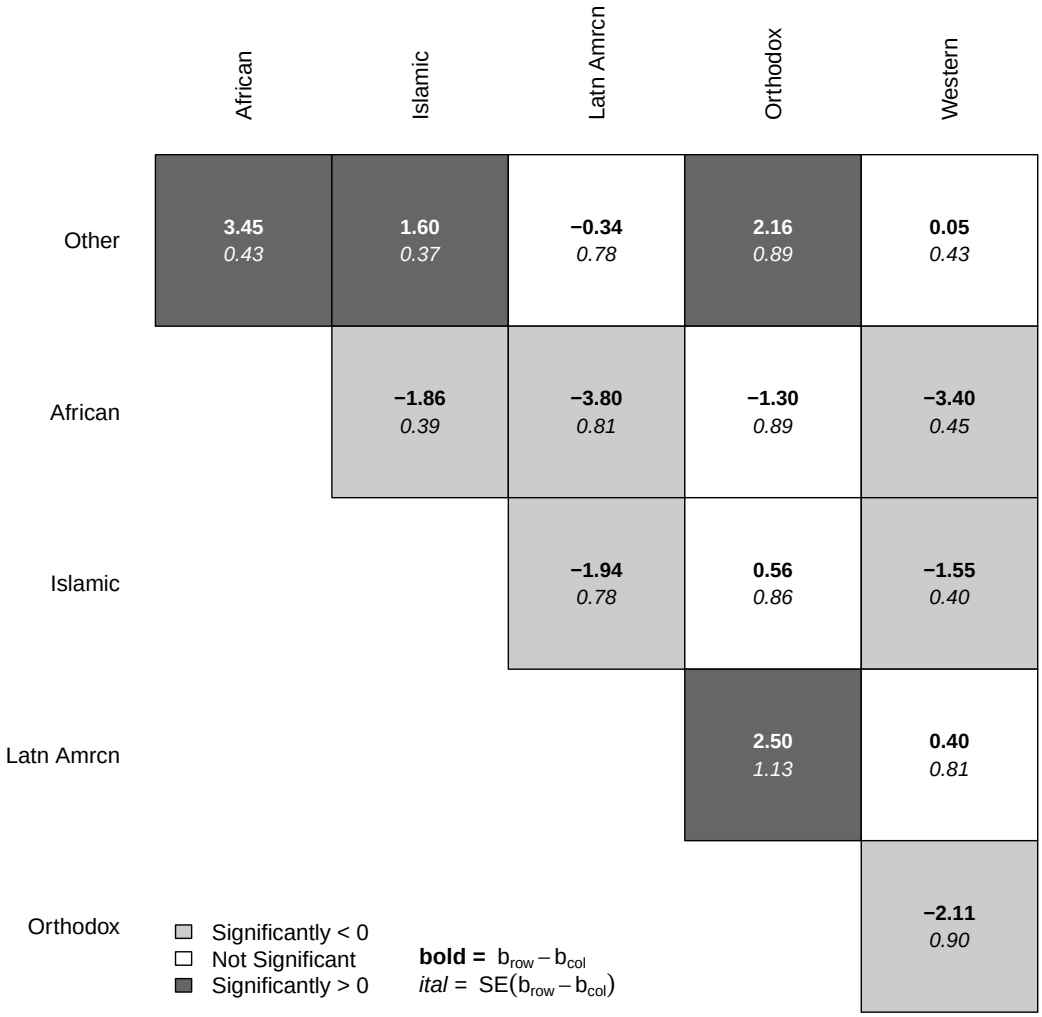
Considering the effect of GDP/capita first, we could do:

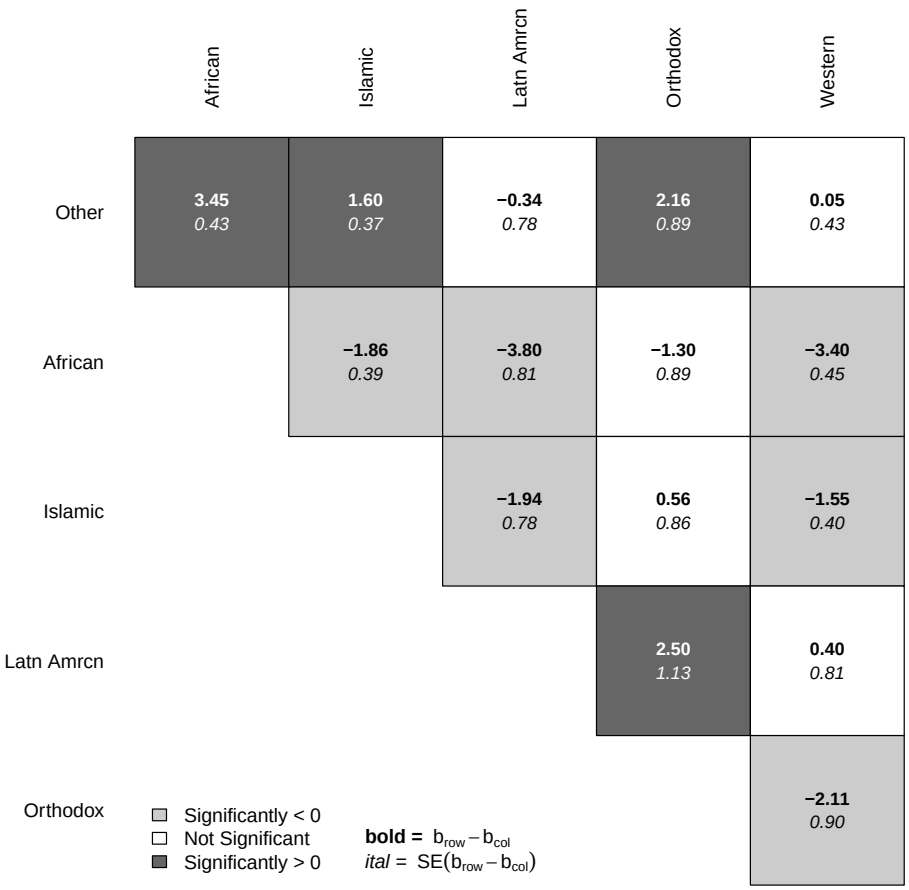
```
iqq <- intQualQuant(mod2, c("loggdp_pc_ppp", "civ2"), type="slopes", plot=F)
```

```
## Conditional effects of  loggdp_pc_ppp :
##                B     SE(B) t-stat Pr(>|t|)
## Other          6.464 0.338 19.114 0.000
## African        3.010 0.325 9.268  0.000
## Islamic        4.867 0.274 17.737 0.000
## Latin American 6.809 0.771 8.825  0.000
## Orthodox       4.307 0.835 5.157  0.000
## Western        6.413 0.349 18.396 0.000
```

Here, all of the coefficients are significant and positive, though some with higher magnitudes than others. If you wanted to do all of the pairwise tests, you could use the `factorplot` command.

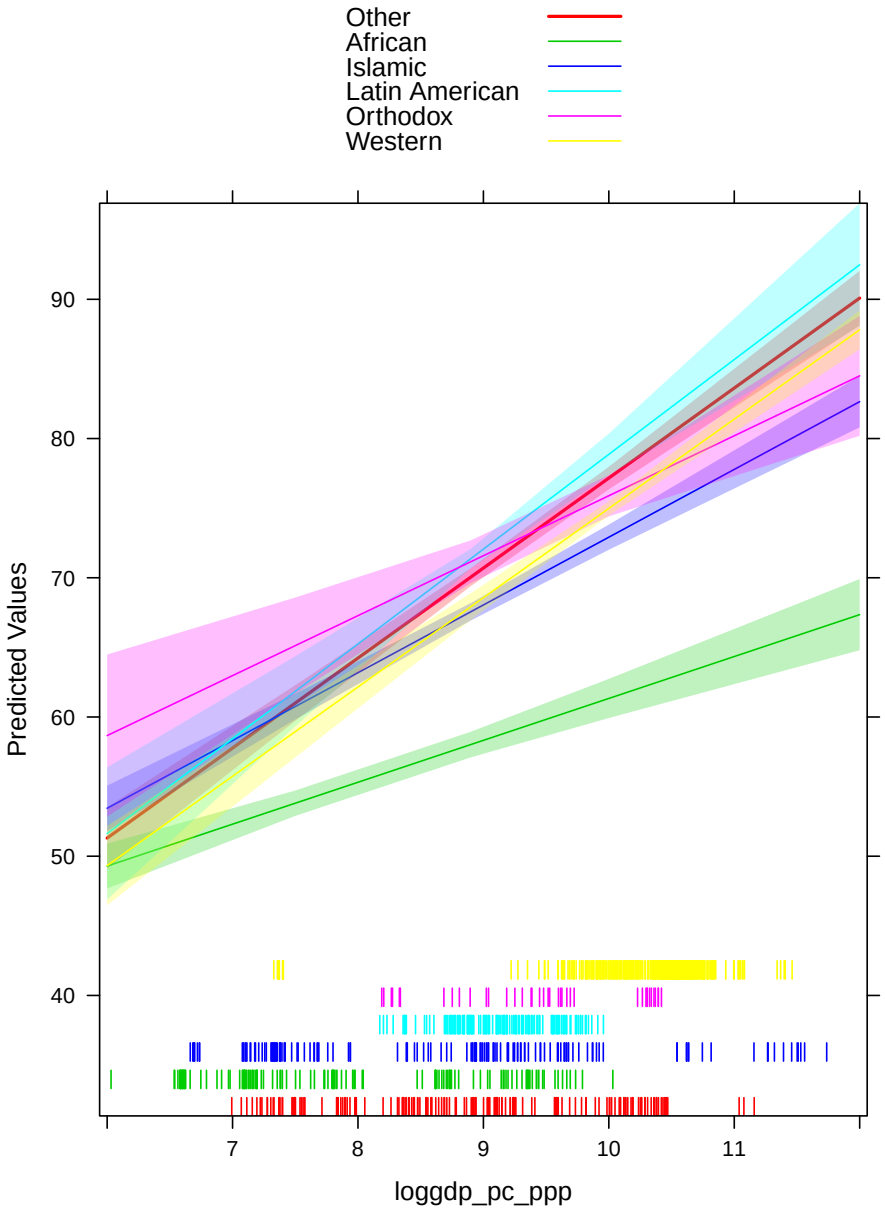
```
eff <- iqq$out$eff
names(eff) <- rownames(iqq$out)
fp <- factorplot(eff, var=iqq$varcor, resdf=mod2$df.residual)
plot(fp)
```


African
Islamic
Latn Amrcn
Orthodox
Western
Other
3.45
0.43
1.60
0.37
−0.34
0.78
2.16
0.89
0.05
0.43
African
−1.86
0.39
−3.80
0.81
−1.30
0.89
−3.40
0.45
Islamic
−1.94
0.78
0.56
0.86
−1.55
0.40
Latn Amrcn
2.50
1.13
0.40
0.81
Orthodox
−2.11
0.90
Significantly < 0
Not Significant
Significantly > 0
bold = b_row − b_col
ital = SE(b_row − b_col)


African
Islamic
Latn Amrcn
Orthodox
Western
Other
3.45
0.43
1.60
0.37
−0.34
0.78
2.16
0.89
0.05
0.43
African
−1.86
0.39
−3.80
0.81
−1.30
0.89
−3.40
0.45
Islamic
−1.94
0.78
0.56
0.86
−1.55
0.40
Latn Amrcn
2.50
1.13
0.40
0.81
Orthodox
−2.11
0.90
Significantly < 0
Not Significant
Significantly > 0
bold = b_row − b_col
ital = SE(b_row − b_col)

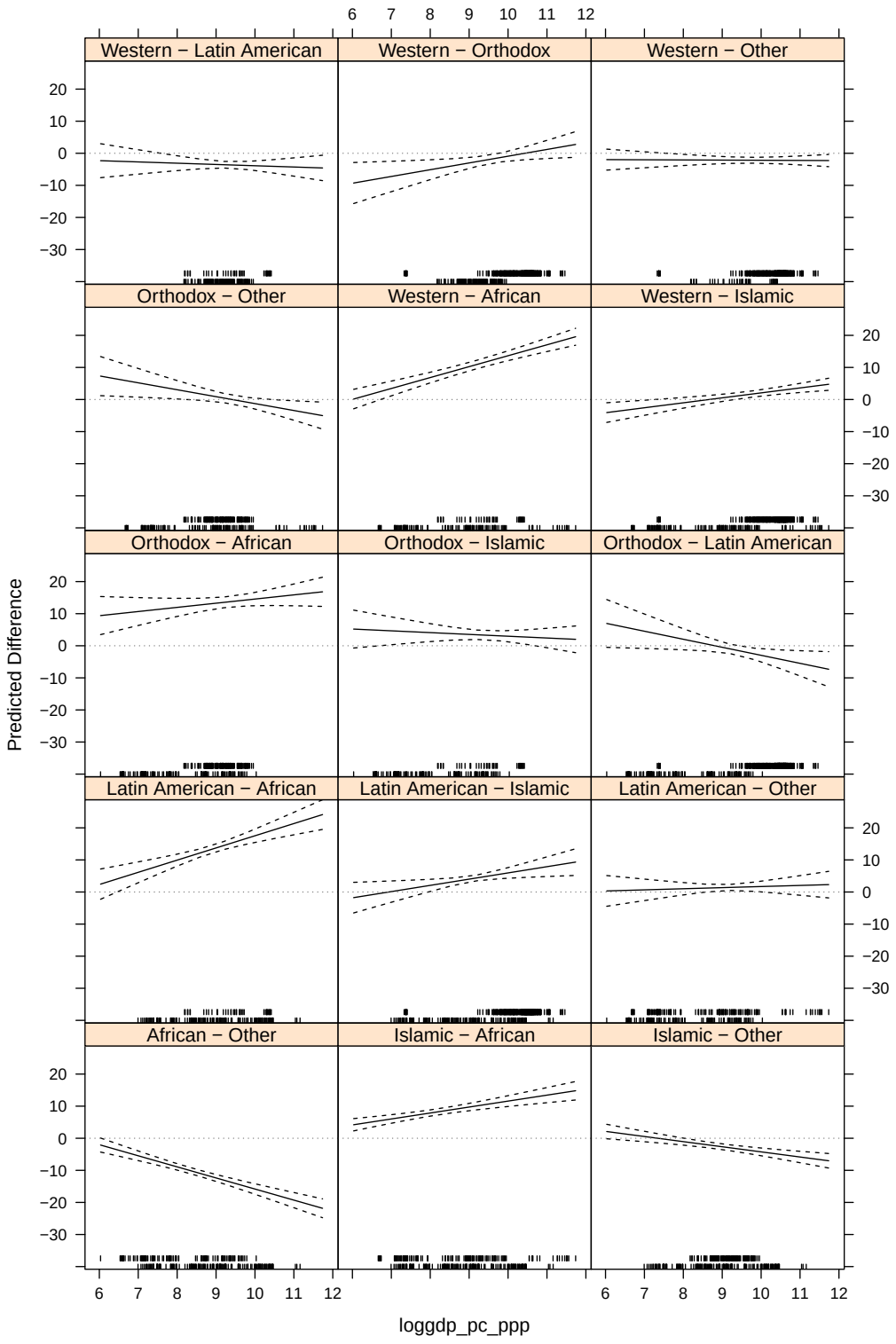
Finally, you could plot the lines:

```
trellis.par.set(
   superpose.line = list(col=palette()[-1]),
   superpose.polygon = list(col=palette()[-1]))
intQualQuant(mod2, c("loggdp_pc_ppp", "civ2"), type="slopes", plot=T, rug=T)
```



On the other side of the interaction, we want to know what the effect is of all of the pairwise changes between civilizations.

```
iq <- intQualQuant(mod2, c("loggdp_pc_ppp", "civ2"), type="facs",
    plot=T, rug=T,layout=c(3,5))
```


Western − Latin American
Western − Orthodox
Western − Other
Orthodox − Other
Western − African
Western − Islamic
Orthodox − African
Orthodox − Islamic
Orthodox − Latin American
Latin American − African
Latin American − Islamic
Latin American − Other
African − Other
Islamic − African
Islamic − Other
Predicted Difference
loggdp_pc_ppp