Pol Sci 630: Problem Set 11 - Imputation of Missing Data, Regression Diagnostics, and Simulations - Solutions

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Grading Due Date: Friday, November 13th, 12.00 PM (Beginning of Lab)

Insert your comments on the assignment that you are grading above the solution in bold and red text. For example write: "GRADER COMMENT: everything is correct! - 4/4 Points" Also briefly point out which, if any, problems were not solved correctly and what the mistake was.

Use the following scheme to assign points: For problems that were solved correctly in their entirety, assign the full point value of 4. For correctly solved bonus problems, add that value to the total score for a problem but do not go above 4 points per problem. If there are mistakes in any problem, subtract points according to the extent of the mistake. If you subtract points, explain why.

In order to make your text bold and red, you need to insert the following line at the beginning of the document:

```
\usepackage{color}
```

and the following lines above the solution of the specific task:

```
\textbf{\color{red} GRADER COMMENT: everything is correct! - 4/4 Points}
```

R Programming

Problem 1

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/")
library(foreign)
LDC = read.dta("LDC_IO_replication.dta")

LDCs = subset(LDC, ctylabel == "Brazil" | ctylabel == "Jamaica" | ctylabel ==
    "Kuwait" | ctylabel == "HongKong" | ctylabel == "KyrgyzRepublic")
keep = c("ctylabel", "date", "newtar", "fdignp", "gdp_pc_95d", "polityiv_update2")
LDCs = LDCs[, keep]

lm1 = lm(fdignp ~ newtar + gdp_pc_95d + polityiv_update2, data = LDCs)
summary(lm1)

##
## Call:
## lm(formula = fdignp ~ newtar + gdp_pc_95d + polityiv_update2,
##     data = LDCs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9792 -1.0033 -0.1529  0.8006  4.9776
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.9768361  1.6501106   3.016  0.00566 **
## newtar           -0.0710902  0.0312451  -2.275  0.03137 *
## gdp_pc_95d       -0.0002303  0.0003059  -0.753  0.45824
## polityiv_update2 -0.1065855  0.1017928  -1.047  0.30470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.812 on 26 degrees of freedom
```

```
##   (120 observations deleted due to missingness)
## Multiple R-squared:  0.2433,Adjusted R-squared:  0.1559
## F-statistic: 2.786 on 3 and 26 DF,  p-value: 0.0607
```

```
a.out <- amelia(LDCs, m = 5, ts = "date", cs = "ctylabel")
```

```
lm2 <- zelig(fdignp ~ newtar + gdp_pc_95d + polityiv_update2, data = a.out$imputations,
    model = "ls")

##
##
##  How to cite this model in Zelig:
##   Kosuke Imai, Gary King, and Olivia Lau. 2015.
##   "ls: Least Squares Regression for Continuous Dependent Variables"
##   in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,
##   http://gking.harvard.edu/zelig
##

summary(lm2)

##
##   Model: ls
##   Number of multiply imputed data sets: 5
##
## Combined results:
##
## Call:
## lm(formula = formula, weights = weights, model = F, data = data)
##
## Coefficients:
##                          Value   Std. Error     t-stat      p-value
## (Intercept)       4.7825511224 0.9280951222  5.1530829 0.0001891294
## newtar           -0.0692846804 0.0263073710 -2.6336604 0.0280189979
## gdp_pc_95d       -0.0003673317 0.0001984485 -1.8510179 0.1298043418
```

```
## polityiv_update2 -0.0139006929 0.0502203761 -0.2767939 0.7852298382
##
## For combined results from datasets i to j, use summary(x, subset = i:j).
## For separate results, use print(summary(x), subset = i:j).
```

The results that every student will get can differ significantly. However, it is most likely that the second regression will yield different results than the first one.

Are these results from the first or the second regression more accurate (i.e. closer to the truth)? This is impossible to tell. Although multiple imputation potentially is a powerful tool in the sense that it allows us to make inferences about missing data points,these guesses have some uncertainty attached and we can never be fully certain about their accuracy.

Before we ran any regression we had a fairly small number of observations and many missing values for several of our variables. Does this mean that the imputation of data has improved our results? Considering that the imputation itself is also a process of statistical inference that relies on our data and considering that imputation is more precise the more data we have, the results of our imputation here are questionable in their accuracy.

It could be the case that the results from the second regression are more accurate than from the first one. The reverse could also be true. In short, we cannot say with certainty which of the two is better. Given that observations were missing in a systematic fashion, we should not have too much confidence in the results.

For grader: If you grade another students homework, then assign full points only if the student arrives at the conclusion that we cannot be certain which results are closer to the truth.

Problem 2

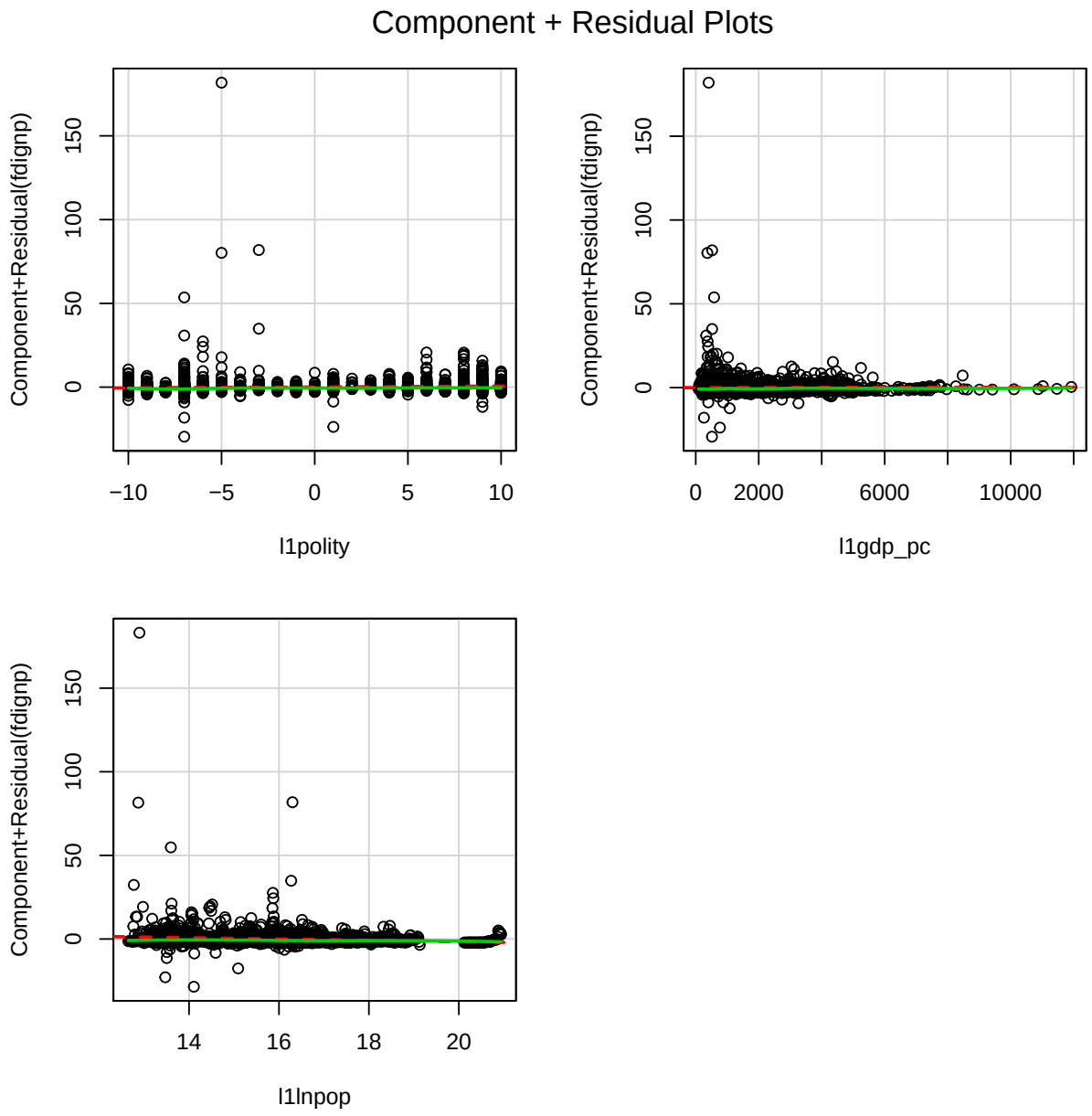
```
lm3 = lm(fdignp ~ l1polity + l1gdp_pc + l1lnpop, data = LDC)
summary(lm3)

##
## Call:
## lm(formula = fdignp ~ l1polity + l1gdp_pc + l1lnpop, data = LDC)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -29.260  -1.389  -0.716   0.276 181.953
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.999e+00  1.131e+00   7.074 1.97e-12 ***
## l1polity     5.425e-02  1.617e-02   3.354 0.000808 ***
## l1gdp_pc     1.541e-05  7.461e-05   0.207 0.836337
## l1lnpop     -3.972e-01  7.045e-02  -5.638 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.372 on 2392 degrees of freedom
##   (2974 observations deleted due to missingness)
## Multiple R-squared:  0.01755,Adjusted R-squared:  0.01632
## F-statistic: 14.24 on 3 and 2392 DF,  p-value: 3.354e-09
```

```
library(car)
```

```
##
## Attaching package:  'car'
##
## The following object is masked from 'package:boot':
##
##    logit
```

```
crPlots(lm3)
```



The component plus residual plots do not indicate any non-linear relationships for any of the variables utilized. There is a small amount of outliers for each of the three variables. However, those outliers do change the impression that there are generally linear trends.

```
library(lmtest)

## Loading required package:  zoo
##
## Attaching package:  'zoo'
```

```
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
resettest(lm3, power = 2:3, type = c("regressor"), data = LDC)
```

The Ramsey RESET test yields a significant test result with $p < 0.001$, indicating that our model has not been specified correctly. We have to reject the null hypothesis that our model correctly captures the relationship. Therefore, including squared or cubic terms of our independent variables has the potential to improve the fit of our model.

Note that we should nonetheless be critical about including additional terms. The component plus residual plots have not revealed any strongly visible non-linear relationships. Like in other situations, there might be a trade off between overfitting the data and gaining higher R-squared values. (This second paragraph is not required from students.)

Problem 3

```
LDC = LDC[with(LDC, complete.cases(newtar, l1polity, l1signed, l1office, l1gdp_pc,
    l1lnpop, l1ecris2, l1bpc1)), ]

lm4 = lm(newtar ~ l1polity + l1signed + l1office + l1gdp_pc + l1lnpop + l1ecris2 +
    l1bpc1, data = LDC)
summary(lm4)
```

```
##
## Call:
## lm(formula = newtar ~ l1polity + l1signed + l1office + l1gdp_pc +
##     l1lnpop + l1ecris2 + l1bpc1, data = LDC)
```

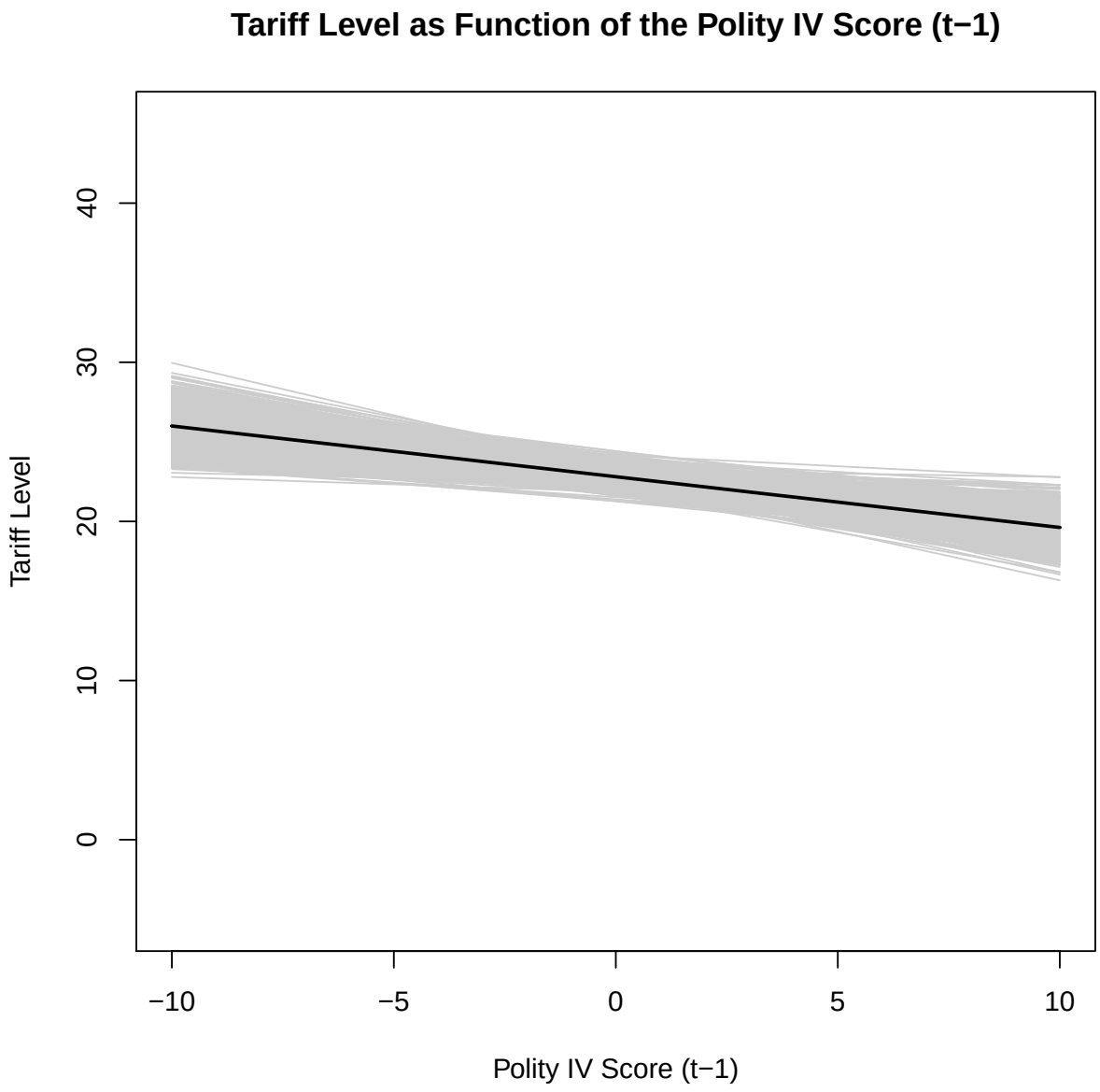
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.354  -8.404  -2.452   5.225  65.782
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.499e+01  5.958e+00  -5.873 6.68e-09 ***
## l1polity    -3.188e-01  8.649e-02  -3.686 0.000246 ***
## l1signed     1.195e+00  1.151e+00   1.038 0.299548
## l1office    -1.870e-01  7.337e-02  -2.549 0.011014 *
## l1gdp_pc    -1.237e-03  1.569e-04  -7.880 1.28e-14 ***
## l1lnpop      3.786e+00  3.376e-01  11.215  < 2e-16 ***
## l1ecris2    -8.246e+00  1.562e+00  -5.278 1.75e-07 ***
## l1bpc1       6.834e-01  1.016e+00   0.673 0.501424
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.77 on 686 degrees of freedom
## Multiple R-squared:  0.3134,Adjusted R-squared:  0.3064
## F-statistic: 44.73 on 7 and 686 DF,  p-value: < 2.2e-16
```

```
library(arm)
```

```
## Loading required package:  Matrix
## Loading required package:  lme4
##
## arm (Version 1.8-6, built:  2015-7-7)
##
## Working directory is C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/W11
##
##
## Attaching package:  'arm'
##
## The following object is masked from 'package:car':
```

```
##
##    logit
##
## The following object is masked from 'package:Zelig':
##
##    sim
##
## The following object is masked from 'package:boot':
##
##    logit
```

```
set.seed(2015)
model.sims = sim(lm4, n.sims = 1000)

curve(coef(lm4)[1] + coef(lm4)[2] * x + coef(lm4)[3] * mean(LDC$l1signed, na.rm = T) +
    coef(lm4)[4] * mean(LDC$l1office, na.rm = T) + coef(lm4)[5] * mean(LDC$l1gdp_pc,
    na.rm = T) + coef(lm4)[6] * mean(LDC$l1lnpop, na.rm = T) + coef(lm4)[7] *
    mean(LDC$l1ecris2, na.rm = T) + coef(lm4)[8] * mean(LDC$l1bpc1, na.rm = T),
    from = -10, to = 10, ylim = c(-5, 45), xlab = "Polity IV Score (t-1)", ylab = "Tarif
    main = "Tariff Level as Function of the Polity IV Score (t-1)", lwd = 2)
for (i in 1:1000) {
    curve(coef(model.sims)[i, 1] + coef(model.sims)[i, 2] * x + coef(model.sims)[i,
        3] * mean(LDC$l1signed, na.rm = T) + coef(model.sims)[i, 4] * mean(LDC$l1office,
        na.rm = T) + coef(model.sims)[i, 5] * mean(LDC$l1gdp_pc, na.rm = T) +
        coef(model.sims)[i, 6] * mean(LDC$l1lnpop, na.rm = T) + coef(model.sims)[i,
        7] * mean(LDC$l1ecris2, na.rm = T) + coef(model.sims)[i, 8] * mean(LDC$l1bpc1,
        na.rm = T), add = TRUE, col = "gray80")
}
curve(coef(lm4)[1] + coef(lm4)[2] * x + coef(lm4)[3] * mean(LDC$l1signed, na.rm = T) +
    coef(lm4)[4] * mean(LDC$l1office, na.rm = T) + coef(lm4)[5] * mean(LDC$l1gdp_pc,
    na.rm = T) + coef(lm4)[6] * mean(LDC$l1lnpop, na.rm = T) + coef(lm4)[7] *
    mean(LDC$l1ecris2, na.rm = T) + coef(lm4)[8] * mean(LDC$l1bpc1, na.rm = T),
    col = "black", lwd = 2, add = TRUE)
```



Problem 4

```
quantile(LDC$l1polity, probs = c(0.25, 0.75), na.rm = T)
```

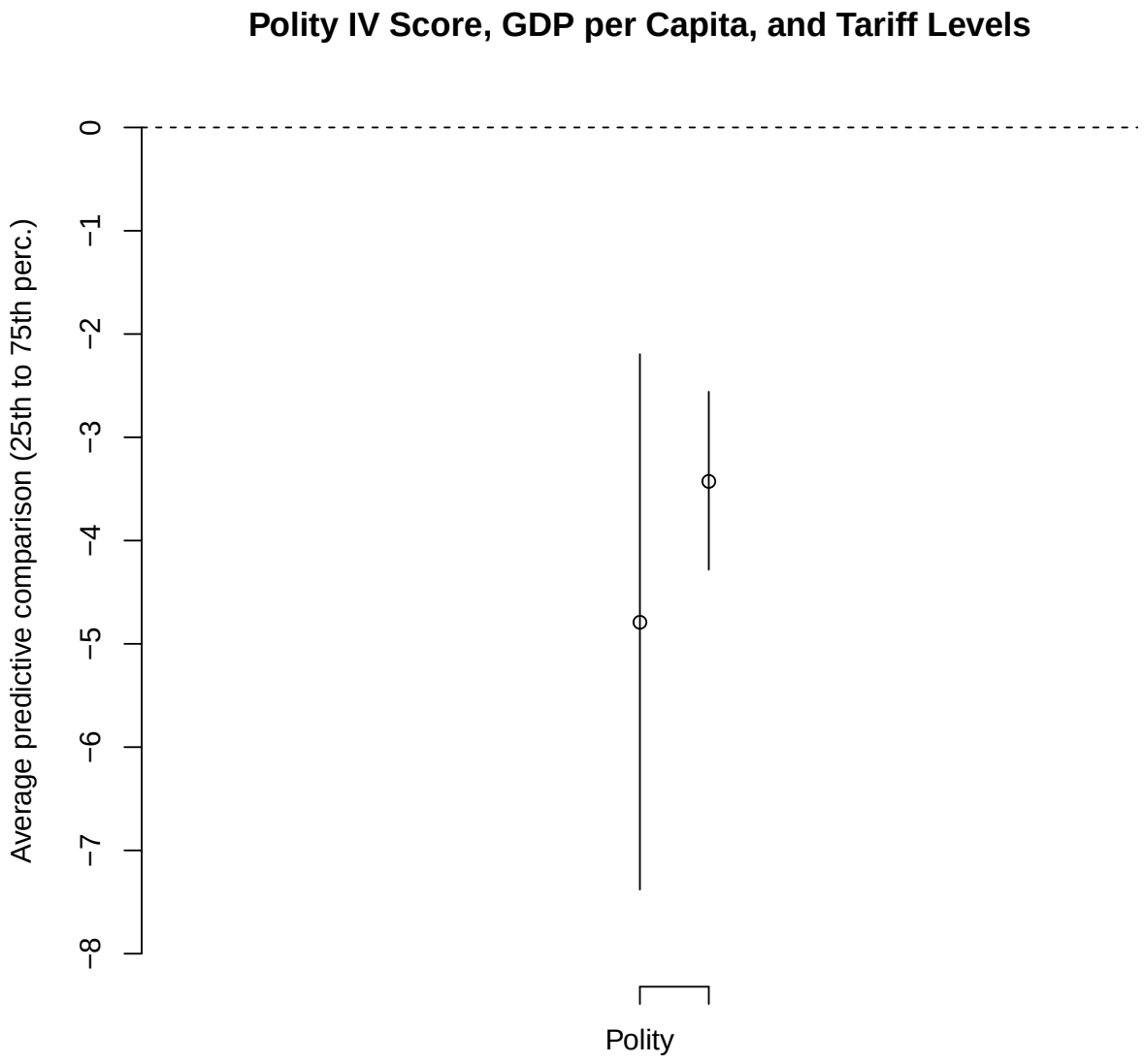
```
## 25% 75%
##  -7   8
```

```
quantile(LDC$l1gdp_pc, probs = c(0.25, 0.75), na.rm = T)
```

```
##       25%       75%
##  458.8198 3225.3915

quantile(LDC$l1fdi, probs = c(0.25, 0.75), na.rm = T)

##       25%       75%
## 0.2185015 1.8839107

# Polity IV Score

d.l1polity <- array(NA, c(1000, length(LDC$newtar)))
m.l1polity <- array(NA, 1000)

for (i in 1:1000) {
    d.l1polity[i, ] <- (coef(model.sims)[i, 1] + coef(model.sims)[i, 2] * 8 +
        coef(model.sims)[i, 3] * LDC$l1signed + coef(model.sims)[i, 4] * LDC$l1office +
        coef(model.sims)[i, 5] * LDC$l1gdp_pc + coef(model.sims)[i, 6] * LDC$l1lnpop +
        coef(model.sims)[i, 7] * LDC$l1ecris2 + coef(model.sims)[i, 8] * LDC$l1bpc1) -
        (coef(model.sims)[i, 1] + coef(model.sims)[i, 2] * -7 + coef(model.sims)[i,
            3] * LDC$l1signed + coef(model.sims)[i, 4] * LDC$l1office + coef(model.sims)
            5] * LDC$l1gdp_pc + coef(model.sims)[i, 6] * LDC$l1lnpop + coef(model.sims)[
            7] * LDC$l1ecris2 + coef(model.sims)[i, 8] * LDC$l1bpc1)
    m.l1polity[i] <- mean(d.l1polity[i, ])
}

mean(m.l1polity)

## [1] -4.791132

sd(m.l1polity)

## [1] 1.337032

quantile(m.l1polity, probs = c(0.025, 0.16, 0.84, 0.975))

##      2.5%       16%       84%     97.5%
## -7.380059 -6.074565 -3.481395 -2.195121
```

```r
# GDP per Capita

d.l1gdp_pc <- array(NA, c(1000, length(LDC$newtar)))
m.l1gdp_pc <- array(NA, 1000)

for (i in 1:1000) {
    d.l1gdp_pc[i, ] <- (coef(model.sims)[i, 1] + coef(model.sims)[i, 2] * LDC$l1polity +
        coef(model.sims)[i, 3] * LDC$l1signed + coef(model.sims)[i, 4] * LDC$l1office +
        coef(model.sims)[i, 5] * 3225.3915 + coef(model.sims)[i, 6] * LDC$l1lnpop +
        coef(model.sims)[i, 7] * LDC$l1ecris2 + coef(model.sims)[i, 8] * LDC$l1bpc1) -
        (coef(model.sims)[i, 1] + coef(model.sims)[i, 2] * LDC$l1polity + coef(model.sim
            3] * LDC$l1signed + coef(model.sims)[i, 4] * LDC$l1office + coef(model.sims)
            5] * 458.8198 + coef(model.sims)[i, 6] * LDC$l1lnpop + coef(model.sims)[i,
            7] * LDC$l1ecris2 + coef(model.sims)[i, 8] * LDC$l1bpc1)
    m.l1gdp_pc[i] <- mean(d.l1gdp_pc[i, ])
}

mean(m.l1gdp_pc)

## [1] -3.427087

sd(m.l1gdp_pc)

## [1] 0.4298

quantile(m.l1gdp_pc, probs = c(0.025, 0.16, 0.84, 0.975))

##      2.5%       16%       84%     97.5%
## -4.282437 -3.828239 -3.015397 -2.558968

# Plot

plot(1:2, c(mean(m.l1polity), mean(m.l1gdp_pc)), type = "p", ylim = c(-8, 0),
    xlim = c(0, 2), xlab = "", main = "Polity IV Score, GDP per Capita, and Tariff Level
    ylab = "Average predictive comparison (25th to 75th perc.)", asp = 1.5,
    axes = FALSE)
```

```
axis(1, at = c(1, 2), labels = c("Polity", "GDP"))
axis(2, at = seq(-8, 0, by = 1))
abline(h = 0, lty = 2)
segments(1, quantile(m.l1polity, probs = c(0.025)), 1, quantile(m.l1polity,
    probs = c(0.975)))
segments(2, quantile(m.l1gdp_pc, probs = c(0.025)), 2, quantile(m.l1gdp_pc,
    probs = c(0.975)))
```



Interpretation: although the Polity IV Score appears to have the larger average substantive effect (when looking at changes from the 25th to the 75th percentile value), there is also greater uncertainty about its impact. The mean effect of GDP per capita changes (for the same percentile values) is smaller but uncertainty about the effect is smaller as well. Considering both the means and the confidence intervals, we cannot make any definitive statements about which variable has the greater substantive impact, but the Polity-IV score is likely to have a larger one considering these results.

Note: It is fine to say that there is no significant visible difference if you are grading someone's homework and the person has arrived at this conclusion.