Pol Sci 630: Problem Set 9 - Data Management and Omitted Variable Bias - Solutions

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

Grading Due Date: Friday, October 30th, 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

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

LDC$regime_class = NA
LDC$regime_class[LDC$polityiv_update2 > 5] = "Democracy"
unique(LDC$regime_class)

## [1] "Democracy" NA

LDC$regime_class[LDC$polityiv_update2 >= -5 & LDC$polityiv_update2 <= 5] = "Anocracy"
unique(LDC$regime_class)

## [1] "Democracy" "Anocracy"  NA

LDC$regime_class[LDC$polityiv_update2 < -5] = "Autocracy"
unique(LDC$regime_class)

## [1] "Democracy" "Anocracy"  "Autocracy" NA

### b

LDC$l1polity = LDC$l1polity + 10
LDC$l1polity_squared = (LDC$l1polity)^2

lm_fdi2 = lm(fdignp ~ l1polity + l1polity_squared + l1signed + l1office + l1gdp_pc +
    l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar + factor(ctylabel) - 1, data = LDC)
summary(lm_fdi2)

##
## Call:
## lm(formula = fdignp ~ l1polity + l1polity_squared + l1signed +
```

```
##     l1office + l1gdp_pc + l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar +
##     factor(ctylabel) - 1, data = LDC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.027  -0.908  -0.124   0.604 152.058
##
## Coefficients:
##                                        Estimate Std. Error t value
## l1polity                             -1.359e-01  1.634e-01  -0.832
## l1polity_squared                      8.910e-03  7.833e-03   1.138
## l1signed                             -2.362e-01  3.101e-01  -0.762
## l1office                             -3.138e-02  2.554e-02  -1.229
## l1gdp_pc                             -1.431e-04  3.028e-04  -0.473
## l1lnpop                               3.489e+00  1.049e+00   3.326
## l1ecris2                              8.556e-02  4.716e-01   0.181
## l1bpc1                               -4.308e-01  3.520e-01  -1.224
## l1avnewtar                           -4.951e-02  1.399e-02  -3.540
## factor(ctylabel)Albania              -4.864e+01  1.538e+01  -3.162
## factor(ctylabel)Algeria              -5.726e+01  1.748e+01  -3.276
## factor(ctylabel)Angola               -1.823e+01  1.734e+01  -1.051
## factor(ctylabel)Argentina            -5.702e+01  1.757e+01  -3.246
## factor(ctylabel)Armenia              -4.769e+01  1.565e+01  -3.047
## factor(ctylabel)Azerbaijan           -3.893e+01  1.642e+01  -2.371
## factor(ctylabel)Bangladesh           -6.255e+01  1.902e+01  -3.289
## factor(ctylabel)Belarus              -5.406e+01  1.669e+01  -3.239
## factor(ctylabel)Benin                -5.103e+01  1.571e+01  -3.248
## factor(ctylabel)Bhutan               -4.434e+01  1.373e+01  -3.230
## factor(ctylabel)Bolivia              -5.113e+01  1.607e+01  -3.182
## factor(ctylabel)Botswana             -4.432e+01  1.414e+01  -3.134
## factor(ctylabel)Brazil               -6.260e+01  1.913e+01  -3.272
## factor(ctylabel)Bulgaria             -5.252e+01  1.635e+01  -3.212
## factor(ctylabel)Burundi              -5.217e+01  1.605e+01  -3.250
## factor(ctylabel)Cambodia             -4.832e+01  1.681e+01  -2.874
```

```
## factor(ctylabel)Cameroon                -5.341e+01  1.671e+01  -3.196
## factor(ctylabel)CentralAfricanRepublic -4.939e+01  1.526e+01  -3.235
## factor(ctylabel)Chad                    -5.145e+01  1.611e+01  -3.193
## factor(ctylabel)Chile                   -5.280e+01  1.669e+01  -3.164
## factor(ctylabel)China                   -6.837e+01  2.161e+01  -3.164
## factor(ctylabel)Colombia                -5.796e+01  1.765e+01  -3.283
## factor(ctylabel)Comoros                 -4.258e+01  1.323e+01  -3.217
## factor(ctylabel)Congo                   -4.767e+01  1.495e+01  -3.189
## factor(ctylabel)CostaRica               -4.861e+01  1.513e+01  -3.213
## factor(ctylabel)Coted'Ivoire            -5.290e+01  1.673e+01  -3.163
## factor(ctylabel)Croatia                 -4.894e+01  1.581e+01  -3.096
## factor(ctylabel)DominicanRepublic       -5.137e+01  1.602e+01  -3.207
## factor(ctylabel)Ecuador                 -5.356e+01  1.642e+01  -3.263
## factor(ctylabel)Egypt                   -5.739e+01  1.821e+01  -3.151
## factor(ctylabel)ElSalvador              -5.175e+01  1.575e+01  -3.285
## factor(ctylabel)EquatorialGuinea        -1.025e+01  1.325e+01  -0.773
## factor(ctylabel)Estonia                 -4.181e+01  1.451e+01  -2.882
## factor(ctylabel)Ethiopia                -5.923e+01  1.830e+01  -3.237
## factor(ctylabel)Fiji                    -4.236e+01  1.366e+01  -3.102
## factor(ctylabel)Gabon                   -4.290e+01  1.402e+01  -3.060
## factor(ctylabel)Gambia                  -4.400e+01  1.391e+01  -3.164
## factor(ctylabel)Georgia                 -4.530e+01  1.660e+01  -2.729
## factor(ctylabel)Ghana                   -5.470e+01  1.689e+01  -3.239
## factor(ctylabel)Guatemala               -5.256e+01  1.627e+01  -3.230
## factor(ctylabel)Guinea                  -5.208e+01  1.619e+01  -3.218
## factor(ctylabel)GuineaBissau            -4.540e+01  1.424e+01  -3.188
## factor(ctylabel)Guyana                  -3.964e+01  1.390e+01  -2.851
## factor(ctylabel)Haiti                   -5.228e+01  1.618e+01  -3.231
## factor(ctylabel)Honduras                -5.059e+01  1.564e+01  -3.235
## factor(ctylabel)Hungary                 -5.118e+01  1.648e+01  -3.106
## factor(ctylabel)India                   -7.061e+01  2.107e+01  -3.351
## factor(ctylabel)Indonesia               -6.279e+01  1.962e+01  -3.201
## factor(ctylabel)Iran                    -5.920e+01  1.822e+01  -3.249
## factor(ctylabel)Jamaica                 -4.856e+01  1.500e+01  -3.236
```

```
## factor(ctylabel)Jordan                -4.818e+01  1.530e+01  -3.148
## factor(ctylabel)Kazakhstan            -5.137e+01  1.710e+01  -3.004
## factor(ctylabel)Kenya                 -5.609e+01  1.740e+01  -3.223
## factor(ctylabel)Korea                 -5.814e+01  1.790e+01  -3.249
## factor(ctylabel)KyrgyzRepublic        -4.708e+01  1.587e+01  -2.967
## factor(ctylabel)Laos                  -4.803e+01  1.588e+01  -3.024
## factor(ctylabel)Latvia                -4.446e+01  1.513e+01  -2.938
## factor(ctylabel)Lesotho               -4.348e+01  1.475e+01  -2.947
## factor(ctylabel)Lithuania             -4.874e+01  1.560e+01  -3.124
## factor(ctylabel)Madagascar            -5.428e+01  1.667e+01  -3.257
## factor(ctylabel)Malawi                -5.266e+01  1.644e+01  -3.203
## factor(ctylabel)Malaysia              -5.172e+01  1.686e+01  -3.068
## factor(ctylabel)Mali                  -5.285e+01  1.636e+01  -3.230
## factor(ctylabel)Mauritania            -4.705e+01  1.491e+01  -3.156
## factor(ctylabel)Mauritius             -4.666e+01  1.410e+01  -3.310
## factor(ctylabel)Mexico                -5.938e+01  1.858e+01  -3.196
## factor(ctylabel)Moldova               -4.931e+01  1.582e+01  -3.117
## factor(ctylabel)Mongolia              -4.886e+01  1.506e+01  -3.245
## factor(ctylabel)Morocco               -5.593e+01  1.749e+01  -3.198
## factor(ctylabel)Mozambique            -5.425e+01  1.704e+01  -3.184
## factor(ctylabel)Nepal                 -5.665e+01  1.712e+01  -3.308
## factor(ctylabel)Nicaragua             -4.865e+01  1.553e+01  -3.133
## factor(ctylabel)Niger                 -5.271e+01  1.625e+01  -3.244
## factor(ctylabel)Nigeria               -5.918e+01  1.890e+01  -3.131
## factor(ctylabel)Oman                  -4.650e+01  1.497e+01  -3.106
## factor(ctylabel)Pakistan              -6.220e+01  1.898e+01  -3.277
## factor(ctylabel)Panama                -4.640e+01  1.498e+01  -3.098
## factor(ctylabel)PapuaNewGuinea        -4.864e+01  1.548e+01  -3.143
## factor(ctylabel)Paraguay              -4.981e+01  1.551e+01  -3.213
## factor(ctylabel)Peru                  -5.567e+01  1.715e+01  -3.245
## factor(ctylabel)Philippines           -5.965e+01  1.840e+01  -3.242
## factor(ctylabel)Poland                -5.726e+01  1.787e+01  -3.204
## factor(ctylabel)Romania               -5.611e+01  1.738e+01  -3.229
## factor(ctylabel)Russia                -6.290e+01  1.931e+01  -3.257
```

```
## factor(ctylabel)Rwanda                -5.203e+01  1.613e+01  -3.226
## factor(ctylabel)Senegal               -5.173e+01  1.608e+01  -3.216
## factor(ctylabel)SierraLeone           -5.104e+01  1.566e+01  -3.259
## factor(ctylabel)SouthAfrica           -5.868e+01  1.803e+01  -3.255
## factor(ctylabel)SriLanka              -5.559e+01  1.698e+01  -3.275
## factor(ctylabel)Swaziland             -3.991e+01  1.400e+01  -2.850
## factor(ctylabel)Syria                 -5.380e+01  1.672e+01  -3.217
## factor(ctylabel)Tanzania              -5.669e+01  1.771e+01  -3.201
## factor(ctylabel)Thailand              -5.903e+01  1.815e+01  -3.252
## factor(ctylabel)Togo                  -4.827e+01  1.541e+01  -3.133
## factor(ctylabel)Trinidad&Tobago       -4.267e+01  1.415e+01  -3.017
## factor(ctylabel)Tunisia               -5.075e+01  1.630e+01  -3.112
## factor(ctylabel)Turkey                -6.036e+01  1.814e+01  -3.327
## factor(ctylabel)Uganda                -5.492e+01  1.712e+01  -3.208
## factor(ctylabel)Ukraine               -5.954e+01  1.833e+01  -3.249
## factor(ctylabel)Uruguay               -4.993e+01  1.519e+01  -3.288
## factor(ctylabel)Venezuela             -5.623e+01  1.697e+01  -3.314
## factor(ctylabel)Zambia                -5.020e+01  1.631e+01  -3.078
## factor(ctylabel)Zimbabwe              -5.316e+01  1.644e+01  -3.234
##                                       Pr(>|t|)
## l1polity                              0.405792
## l1polity_squared                      0.255484
## l1signed                              0.446403
## l1office                              0.219355
## l1gdp_pc                              0.636568
## l1lnpop                               0.000901 ***
## l1ecris2                              0.856053
## l1bpc1                                0.221207
## l1avnewtar                            0.000412 ***
## factor(ctylabel)Albania               0.001595 **
## factor(ctylabel)Algeria               0.001078 **
## factor(ctylabel)Angola                0.293357
## factor(ctylabel)Argentina             0.001196 **
## factor(ctylabel)Armenia               0.002351 **
```

```
## factor(ctylabel)Azerbaijan             0.017864 *
## factor(ctylabel)Bangladesh             0.001028 **
## factor(ctylabel)Belarus                0.001226 **
## factor(ctylabel)Benin                  0.001187 **
## factor(ctylabel)Bhutan                 0.001266 **
## factor(ctylabel)Bolivia                0.001491 **
## factor(ctylabel)Botswana               0.001759 **
## factor(ctylabel)Brazil                 0.001093 **
## factor(ctylabel)Bulgaria               0.001346 **
## factor(ctylabel)Burundi                0.001180 **
## factor(ctylabel)Cambodia               0.004105 **
## factor(ctylabel)Cameroon               0.001421 **
## factor(ctylabel)CentralAfricanRepublic 0.001240 **
## factor(ctylabel)Chad                   0.001437 **
## factor(ctylabel)Chile                  0.001587 **
## factor(ctylabel)China                  0.001585 **
## factor(ctylabel)Colombia               0.001049 **
## factor(ctylabel)Comoros                0.001321 **
## factor(ctylabel)Congo                  0.001455 **
## factor(ctylabel)CostaRica              0.001340 **
## factor(ctylabel)Coted'Ivoire           0.001594 **
## factor(ctylabel)Croatia                0.001997 **
## factor(ctylabel)DominicanRepublic      0.001367 **
## factor(ctylabel)Ecuador                0.001128 **
## factor(ctylabel)Egypt                  0.001656 **
## factor(ctylabel)ElSalvador             0.001042 **
## factor(ctylabel)EquatorialGuinea       0.439418
## factor(ctylabel)Estonia                0.004010 **
## factor(ctylabel)Ethiopia               0.001233 **
## factor(ctylabel)Fiji                   0.001958 **
## factor(ctylabel)Gabon                  0.002250 **
## factor(ctylabel)Gambia                 0.001586 **
## factor(ctylabel)Georgia                0.006429 **
## factor(ctylabel)Ghana                  0.001224 **
```

```
## factor(ctylabel)Guatemala              0.001265 **
## factor(ctylabel)Guinea                 0.001319 **
## factor(ctylabel)GuineaBissau           0.001460 **
## factor(ctylabel)Guyana                 0.004418 **
## factor(ctylabel)Haiti                  0.001259 **
## factor(ctylabel)Honduras               0.001242 **
## factor(ctylabel)Hungary                0.001932 **
## factor(ctylabel)India                  0.000823 ***
## factor(ctylabel)Indonesia              0.001400 **
## factor(ctylabel)Iran                   0.001182 **
## factor(ctylabel)Jamaica                0.001236 **
## factor(ctylabel)Jordan                 0.001675 **
## factor(ctylabel)Kazakhstan             0.002704 **
## factor(ctylabel)Kenya                  0.001294 **
## factor(ctylabel)Korea                  0.001182 **
## factor(ctylabel)KyrgyzRepublic         0.003056 **
## factor(ctylabel)Laos                   0.002534 **
## factor(ctylabel)Latvia                 0.003349 **
## factor(ctylabel)Lesotho                0.003256 **
## factor(ctylabel)Lithuania              0.001818 **
## factor(ctylabel)Madagascar             0.001152 **
## factor(ctylabel)Malawi                 0.001390 **
## factor(ctylabel)Malaysia               0.002189 **
## factor(ctylabel)Mali                   0.001262 **
## factor(ctylabel)Mauritania             0.001630 **
## factor(ctylabel)Mauritius              0.000953 ***
## factor(ctylabel)Mexico                 0.001422 **
## factor(ctylabel)Moldova                0.001858 **
## factor(ctylabel)Mongolia               0.001199 **
## factor(ctylabel)Morocco                0.001411 **
## factor(ctylabel)Mozambique             0.001484 **
## factor(ctylabel)Nepal                  0.000961 ***
## factor(ctylabel)Nicaragua              0.001763 **
## factor(ctylabel)Niger                  0.001206 **
```

```
## factor(ctylabel)Nigeria                0.001775 **
## factor(ctylabel)Oman                   0.001934 **
## factor(ctylabel)Pakistan               0.001073 **
## factor(ctylabel)Panama                 0.001983 **
## factor(ctylabel)PapuaNewGuinea         0.001703 **
## factor(ctylabel)Paraguay               0.001343 **
## factor(ctylabel)Peru                   0.001198 **
## factor(ctylabel)Philippines            0.001210 **
## factor(ctylabel)Poland                 0.001383 **
## factor(ctylabel)Romania                0.001268 **
## factor(ctylabel)Russia                 0.001151 **
## factor(ctylabel)Rwanda                 0.001280 **
## factor(ctylabel)Senegal                0.001325 **
## factor(ctylabel)SierraLeone            0.001141 **
## factor(ctylabel)SouthAfrica            0.001160 **
## factor(ctylabel)SriLanka               0.001082 **
## factor(ctylabel)Swaziland              0.004424 **
## factor(ctylabel)Syria                  0.001322 **
## factor(ctylabel)Tanzania               0.001399 **
## factor(ctylabel)Thailand               0.001171 **
## factor(ctylabel)Togo                   0.001765 **
## factor(ctylabel)Trinidad&Tobago        0.002596 **
## factor(ctylabel)Tunisia                0.001889 **
## factor(ctylabel)Turkey                 0.000897 ***
## factor(ctylabel)Uganda                 0.001363 **
## factor(ctylabel)Ukraine                0.001183 **
## factor(ctylabel)Uruguay                0.001033 **
## factor(ctylabel)Venezuela              0.000941 ***
## factor(ctylabel)Zambia                 0.002124 **
## factor(ctylabel)Zimbabwe               0.001248 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.963 on 1546 degrees of freedom
```

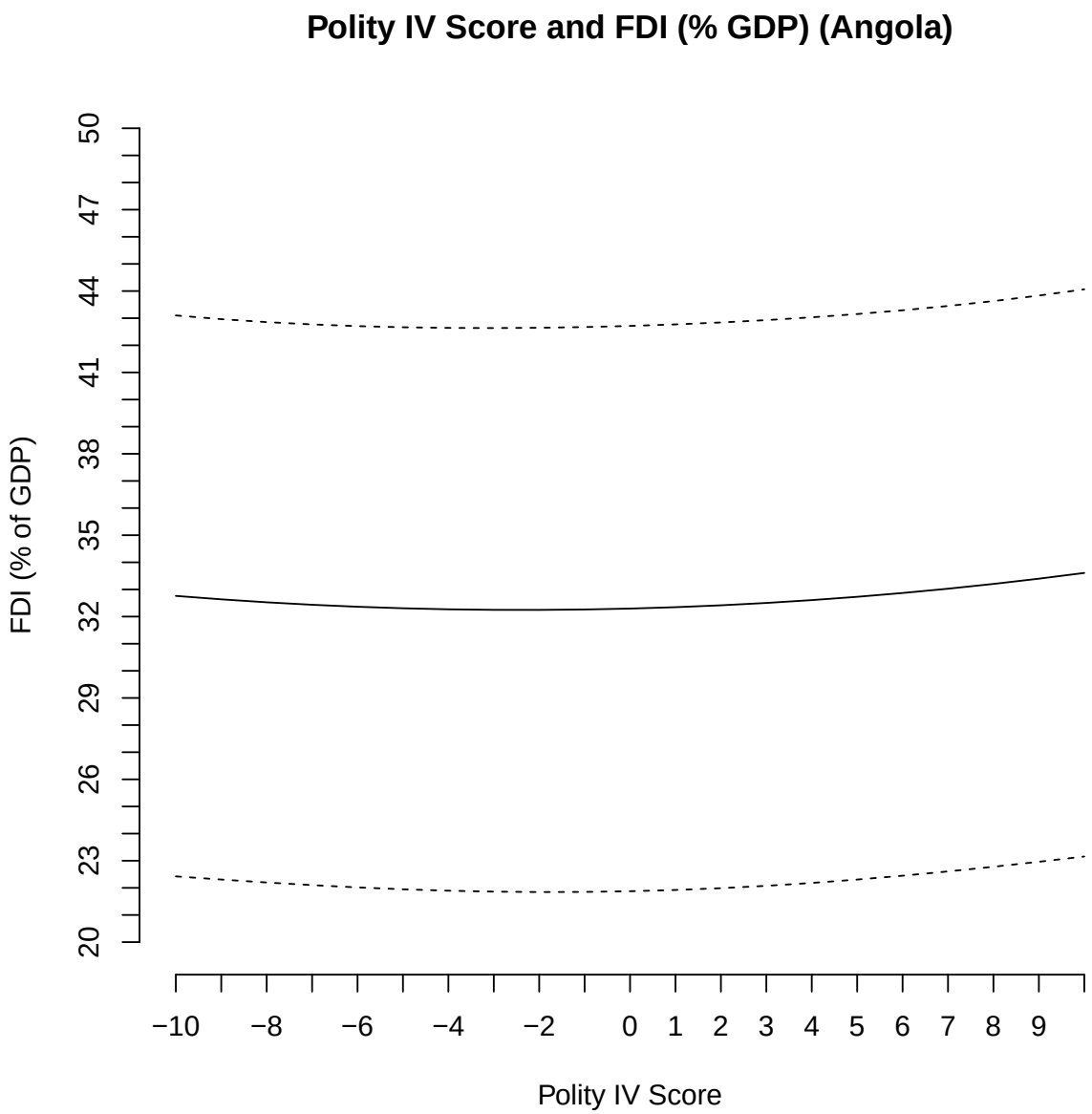
```
##   (3712 observations deleted due to missingness)
## Multiple R-squared:  0.3616,Adjusted R-squared:  0.3153
## F-statistic: 7.817 on 112 and 1546 DF,  p-value: < 2.2e-16

nd <- data.frame(l1polity = seq(0, 20, by = 1), l1polity_squared = seq(0, 20,
    by = 1)^2, l1signed = rep(0.1511, 21), l1office = rep(8.431, 21), l1gdp_pc = rep(288
    21), l1lnpop = rep(15.1, 21), l1ecris2 = rep(0.0641, 21), l1bpc1 = rep(0.5909,
    21), l1avnewtar = rep(14.91, 21), ctylabel = rep("Angola", 21))

pred.p1 <- predict(lm_fdi2, type = "response", se.fit = TRUE, newdata = nd)
pred.table <- cbind(pred.p1$fit, pred.p1$se.fit)

fit <- pred.p1$fit
low <- pred.p1$fit - 2 * pred.p1$se.fit
high <- pred.p1$fit + 2 * pred.p1$se.fit
cis <- cbind(fit, low, high)

plot(pred.p1$fit, type = "l", ylim = c(20, 50), main = "Polity IV Score and FDI (% GDP)
    xlab = "Polity IV Score", ylab = "FDI (% of GDP)", axes = FALSE)
axis(1, at = seq(1, 21), labels = seq(-10, 10, 1))
axis(2, at = seq(20, 50), labels = seq(20, 50))
matlines(cis[, c(2, 3)], lty = 2, col = "black")
```



Note: It is possible to take the mean values from a subset of the dataset that only contains complete cases (with values of all independent variables available). However, the effect of the Polity IV Score will not be affected by different values of the control variables because we keep those values constant in either case. Therefore, it is fine but not necessary to take the mean values from a subset with complete cases.

Problem 2

```
### a
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/")
library(readstata13)

## Warning:  package 'readstata13' was built under R version 3.2.2

na_data = read.dta13("na_data.dta")
summary(na_data)

##  countrycode             year           v_c
##  Length:10624       Min.   :1950   Min.   :0.000e+00
##  Class :character   1st Qu.:1973   1st Qu.:7.600e+02
##  Mode  :character   Median :1986   Median :1.296e+04
##                     Mean   :1985   Mean   :8.369e+06
##                     3rd Qu.:1999   3rd Qu.:2.230e+05
##                     Max.   :2011   Max.   :4.053e+09
##                                    NA's   :561
##       v_i                 v_g                v_x
##  Min.   :     -7427   Min.   :        0   Min.   :0.000e+00
##  1st Qu.:       251   1st Qu.:      171   1st Qu.:4.260e+02
##  Median :      4319   Median :     3197   Median :5.520e+03
##  Mean   :   4418919   Mean   :  1658085   Mean   :4.787e+06
##  3rd Qu.:     70963   3rd Qu.:    49433   3rd Qu.:8.595e+04
##  Max.   :2433863510   Max.   :667440135   Max.   :1.955e+09
##  NA's   :561          NA's   :561         NA's   :519
##       v_m                v_gdp                q_c
##  Min.   :0.000e+00   Min.   :0.000e+00   Min.   :2.000e+00
##  1st Qu.:5.320e+02   1st Qu.:1.188e+03   1st Qu.:7.238e+03
##  Median :6.670e+03   Median :2.077e+04   Median :1.263e+05
##  Mean   :4.538e+06   Mean   :1.473e+07   Mean   :1.202e+07
##  3rd Qu.:1.094e+05   3rd Qu.:3.398e+05   3rd Qu.:9.104e+05
##  Max.   :2.194e+09   Max.   :7.427e+09   Max.   :2.343e+09
##  NA's   :519         NA's   :519         NA's   :561
```

```
##       q_i                  q_g                 q_x
##  Min.   :    -39042   Min.   :        6   Min.   :0.000e+00
##  1st Qu.:      2543   1st Qu.:     1865   1st Qu.:3.730e+03
##  Median :     39263   Median :    27296   Median :4.116e+04
##  Mean   :   6616470   Mean   :  2764949   Mean   :7.191e+06
##  3rd Qu.:    279383   3rd Qu.:   224138   3rd Qu.:2.672e+05
##  Max.   :1098261440   Max.   :372916568   Max.   :1.453e+09
##  NA's   :561          NA's   :561         NA's   :519
##       q_m                q_gdp                pop
##  Min.   :7.000e+00   Min.   :9.000e+00   Min.   :4.608e+03
##  1st Qu.:5.002e+03   1st Qu.:1.252e+04   1st Qu.:7.673e+05
##  Median :4.708e+04   Median :2.065e+05   Median :4.951e+06
##  Mean   :6.188e+06   Mean   :2.188e+07   Mean   :3.177e+07
##  3rd Qu.:3.579e+05   3rd Qu.:1.424e+06   3rd Qu.:1.614e+07
##  Max.   :1.223e+09   Max.   :3.903e+09   Max.   :1.324e+09
##  NA's   :519         NA's   :519         NA's   :459
##        xr                xr2               v_gfcf
##  Min.   :    0.00   Min.   :    0.00   Min.   :0.000e+00
##  1st Qu.:    0.90   1st Qu.:    0.91   1st Qu.:3.900e+02
##  Median :    2.57   Median :    2.64   Median :7.822e+03
##  Mean   :  220.40   Mean   :  221.12   Mean   :4.854e+06
##  3rd Qu.:   28.58   3rd Qu.:   31.64   3rd Qu.:1.101e+05
##  Max.   :31900.00   Max.   :31900.00   Max.   :2.378e+09
##  NA's   :459        NA's   :459        NA's   :2370
##      q_gfcf
##  Min.   :3.000e+00
##  1st Qu.:1.826e+03
##  Median :4.209e+04
##  Mean   :6.118e+06
##  3rd Qu.:3.044e+05
##  Max.   :1.004e+09
##  NA's   :2390
```

```
### b
na_data$gdpgrowth = NA
```

```r
for (i in 2:length(na_data$q_gdp)) {
    if (na_data$countrycode[i] == na_data$countrycode[i - 1]) {
        na_data$gdpgrowth[i] = (na_data$q_gdp[i]/na_data$q_gdp[i - 1] - 1) *
            100
    }
}

summary(na_data$gdpgrowth)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -66.120   1.372   4.038   3.997   6.776 205.000     728

### c
na_data$date = na_data$year

LDC$countrycode = NA
LDC$countrycode[LDC$ctylabel == "Turkey"] = "TUR"
LDC$countrycode[LDC$ctylabel == "SouthAfrica"] = "ZAF"
LDC$countrycode[LDC$ctylabel == "Mexico"] = "MEX"

merged_data = merge(LDC, na_data, by = c("countrycode", "date"))

newmodel = lm(newtar ~ l1polity + gdpgrowth + factor(countrycode) - 1, data = merged_dat
summary(newmodel)

##
## Call:
## lm(formula = newtar ~ l1polity + gdpgrowth + factor(countrycode) -
##     1, data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4339 -3.7056 -0.2797  4.1416 11.7623
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## l1polity                  -1.4270     0.2978  -4.792 3.40e-05 ***
## gdpgrowth                  0.4709     0.2907   1.620    0.115
## factor(countrycode)MEX    29.4030     3.2636   9.009 2.06e-10 ***
## factor(countrycode)TUR    43.4783     5.0282   8.647 5.41e-10 ***
## factor(countrycode)ZAF    34.8150     5.1067   6.818 8.89e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.694 on 33 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.9176,Adjusted R-squared:  0.9051
## F-statistic: 73.52 on 5 and 33 DF,  p-value: < 2.2e-16
```

Problem 3

a) As grader, please make sure that the person who has submitted the homework has answered all questions of this problem. This includes a brief explanation (2-3 sentences) of the student's theory and a reference to both the dependent and key independent variable.

Make sure that the students has done the following things:

1. The student has explained which datasets contain the variables and provided an overview of how the variables are coded there.

2. If and only if there were variables that were nominal or ordinal scale or coded as characters, the student has recognized that these variables have to be transformed to be used for a linear regression. Nominal variables have to be introduced as factors (dummies). Ordinal variables have to be either used as factors (dummies) or have to be assigned numerical values. Similarly, variables that are coded as characters have to be either introduced as factors (dummies) or recoded as numerical variables.

3. The students has briefly discussed the numbers of units and the time periods covered by the datasets. Note that the discussion does not have to be extensive. (See problem set for an example)

b) As grader, please make sure that the person who has submitted the homework has answered all questions of this problem. This includes a brief theoretical explanation for the importance of at least two control variables that the student suggests to use for the final paper. For all control variables there should be a brief reference to either literature that has explained the theoretical impact of the variable, the concept of omitted variable bias, or both.

Important: when a reference to the concept of omitted variable bias is made (as the justification for the inclusion of a control variable), it is most important that the student has recognized that the **variable in question must have an influence on both the dependent and the independent variable**. Otherwise we cannot speak of the phenomenon of OVB.

Make sure that the students has done the following things:

1. The student has explained which datasets contain the variables and provided an overview of how the variables are coded there.

2. If and only if there were variables that were nominal or ordinal scale or coded as characters, the student has recognized that these variables have to be transformed to be used for a linear regression. Nominal variables have to be introduced as factors (dummies). Ordinal variables have to be either used as factors (dummies) or have to be assigned numerical values. Similarly, variables that are coded as characters have to be either introduced as factors (dummies) or recoded as numerical variables.

3. The students has addressed potential differences (if there are any) in the time periods and units covered. For example, data for the control variables may be available only for OECD countries while the data for the dependent variable may only be available for developing countries. Another example would be that data for the control variables may be available on a quarterly basis while data for the dependent variable may be available on an annual basis.

4. The students has addressed differences in the coding of time periods and units. For example, the names of countries may be coded as full names in one dataset while another dataset uses 3-letter isocodes to refer to countries. Another example would be that time in one dataset could be coded in the format YYYY-MM (Y = year, M = month) while it could be coded in the format YY-MM in another dataset.

Statistical Theory: Omitted Variable Bias

Problem 4

a) In this section, each student is expected to write down the mathematical formula of the linear regression that he or she intends to estimate in their final project. For example, the student could be interested in how the regime type of a country influences its military expenditures.

In this case the regression could look like this:

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$

Where the variables represent the following concepts:

Y	Military Expenditures (Percent of GDP)
X_1	Regime Type (Polity IV)
X_2	External Military Threat
X_3	Militaristic Ideology
X_4	Size of the Arms Industry
X_5	No. of Armed Conflicts in the Last Decade

Note that you have to include the ϵ for the error term if you use "Y" in your notation. Alternatively, you can also use $\hat{Y}$ (Y-hat) in your notation. In the latter version, however, you are not allowed to include the error term because the *predicted values of Y* (denoted by $\hat{Y}$) solely depend on the linear combination of the independent variables, not the error term. Recall that the error term is in fact the difference between Y and $\hat{Y}$. So the alternative regression would look like this:

$\hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

Note that it is also perfectly fine if you write out the variable names instead of using placeholder variables such as Y, X_1, or X_2. If you use placeholder names, it would be necessary to add some explanation like above to make it clear to the reader which real-world measurements/concepts you are referring to.

b) The concept of omitted variable bias requires that the omitted variable is correlated with both the dependent variable and the independent variable of interest. In the case above, for example, one might argue that militaristic ideology has an impact on both military expenditures and regime type. The argument here could be that a high level of militaristic

ideology supports authoritarian regimes and thus has a negative effect on regime type measured through the Polity IV Score. Additionally, we would expect that military ideology has a positive effect on military expenditures as percentage of GDP. Mathematically these statements would mean:

$Cov(X_1, X_3) < 0$ and $Cov(Y, X_3) > 0$

What would happen if we omit the variable X_3 from the regression? We begin with two regressions:

1. $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$
2. $Y = \alpha + \lambda_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_2$

Note that the second regression has X_3 omitted and that we therefore expect to get a different coefficient for X_1, namely λ_1 instead of β_1.

$$\lambda_1 = \frac{Cov(X_1, Y)}{Var(X_1)}.$$

Assuming that X_3 has some impact on Y, we know that Y can be rewritten as a linear function of it (and the other variables that we have in the model). So:

$$\lambda_1 = \frac{Cov(X_1, \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon)}{Var(X_1)}$$

We can rewrite this as:

$$\lambda_1 = \beta_1 * \frac{Var(X_1)}{Var(X_1)} + \beta_2 * \frac{Cov(X_1, X_2)}{Var(X_1)} + \beta_3 * \frac{Cov(X_1, X_3)}{Var(X_1)} + \beta_4 * \frac{Cov(X_1, X_4)}{Var(X_1)} + \beta_5 * \frac{Cov(X_1, X_5)}{Var(X_1)} + \frac{Cov(X_1, \epsilon)}{Var(X_1)}$$

Recall that the task asks you to assume that there is omitted variable bias for **one of the control variables only**. In this case, there would not be any correlation between X_1 and the other variables, implying that their covariances would be theoretically zero. Note that the covariance between X_1 and the error term is also theoretically zero if the condition holds that there is omitted variable bias for only one variable. It then follows that:

$$\lambda_1 = \beta_1 + \beta_3 * \frac{Cov(X_1, X_3)}{Var(X_1)}$$

Notice that the $\beta_3 > 0$ because $Cov(Y, X_3) > 0$. However, $Cov(X_1, X_3) < 0$, meaning that:

$$\lambda_1 = \beta_1 + \text{Positive Term} * \frac{NegativeTerm}{Var(X_1)}$$

Because the variance of any variable is positive as long as there is more than one value, meaning that $Var(X_1) > 0$, the coefficient of X_1 would be biased downwards.

If you are grading someone else's problem set, please note that if we have the following variables:

Y	Dependent Variable
X_1	Key Independent Variable
X_2	Potentially Omitted Variable

The following happens if you leave X_2 out of the linear regression:

	$Cov(X1, X2) > 0$	$Cov(X1, X2) < 0$
$Cov(Y, X2) > 0$	upward bias of X_1 coefficient	downward bias of X_1 coefficient
$Cov(Y, X2) < 0$	downward bias of X_1 coefficient	upward bias of X_1 coefficient