

Tutorial 5: Regression Model Interpretation

Jan Vogler (jan.vogler@duke.edu)

September 30, 2016

Today's Agenda

1. Marginal effects and intercepts
2. Hypothesis testing
3. Multiple linear regression
4. Graphical representation
5. Tips for your final paper

1. Marginal effects and intercepts

An essential aspect of all linear models are the marginal effects that predictor variables (independent variables) are estimated to have on the response variable (dependent variable).

Note that the word “effect” may be problematic because it implies causality. However, without any additional assumptions or additional model features, linear models allow us to make statements with respect to correlation only. This means we can't say anything about causality when just having a linear model. So let us be very cautious when we use the word “marginal effect”.

Every linear model has one response variable (dependent variable) and at least one predictor variable (independent variable) plus an intercept.

Let's assume that Y is our response variable and X is our only predictor variable. The model may look like this:

$$Y = 5 + 2X + \text{error}$$

How would we interpret the marginal effect of X?

The interpretation would be: For a 1-point increase in X we expect a 2-point increase in Y.

How would we interpret the intercept?

The intercept is the expected value of Y when X is at a value of 0.

Illustration of the marginal effect interpretation

Let's load another R dataset that can illustrate the interpretation of marginal effects. The “airquality” dataset. According to the documentation, this is “Daily air quality measurements in New York, May to September 1973.”

Here is a short description of each of the variables:

1. Ozone: numeric Ozone (ppb)
2. Solar.R: numeric Solar R (lang)
3. Wind: numeric Wind (mph)
4. Temp: numeric Temperature (degrees F)
5. Month: numeric Month (1-12)
6. Day: numeric Day of month (1-31)

More details can be found here: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/airquality.html>

```
data(airquality)
summary(airquality)
```

```
##      Ozone      Solar.R      Wind      Temp
## Min.   : 1.00   Min.   : 7.0   Min.   : 1.700   Min.   :56.00
## 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
## Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
## Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
## 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
## Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
## NA's   :37      NA's   :7
##      Month      Day
## Min.   :5.000   Min.   : 1.0
## 1st Qu.:6.000   1st Qu.: 8.0
## Median :7.000   Median :16.0
## Mean   :6.993   Mean   :15.8
## 3rd Qu.:8.000   3rd Qu.:23.0
## Max.   :9.000   Max.   :31.0
##
```

Our question is: is there a linear relationship between the Ozone measures and the Solar.R measures?

Let us use linear regression to answer this question:

```
lm1 = lm(Ozone ~ Solar.R, data = airquality)
```

The summary of this linear regression will return a t-value and a p-value for the intercept and all coefficients.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Ozone ~ Solar.R, data = airquality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -48.292 -21.361  -8.864  16.373 119.136
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.59873    6.74790   2.756 0.006856 **
## Solar.R       0.12717    0.03278   3.880 0.000179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.33 on 109 degrees of freedom
## (42 observations deleted due to missingness)
## Multiple R-squared:  0.1213, Adjusted R-squared:  0.1133
## F-statistic: 15.05 on 1 and 109 DF, p-value: 0.0001793
```

How would we interpret the finding with respect to the linear relationship between the two variables? The interpretation would look like this:

There is a positive linear relationship between Ozone and Solar.R. For a 1-point increase in Solar.R, we would expect a 0.13 increase in Ozone (in a multivariate model we would have to add: “holding all other variables constant”).

Furthermore (already going into the next topic): The associated t-value is 3.880. This t-value implies a p-value of 0.0002. This $p < 0.001$ corresponds to a type-1 error rate of $\alpha < 0.001$, meaning that the relationship is significant at all common levels of statistical significance.

How do we interpret the R-squared statistic? Our model explains a proportion of the total variation in the dependent variable. The R-squared statistic returns this proportion. How well does our model do?

2. Hypothesis testing

Let us use another dataset to conduct some hypothesis tests.

We will look at data from an article that was published in the journal “International Organization”, the leading journal in the field of international relations. The article was written by Helen Milner and Keiko Kubota.

The article deals with the effect that democratization has on trade barriers. The authors believe that democratization has a negative effect on trade barriers in developing countries (that are scarce in capital). Their theory is based on the Stolper Samuelson theorem and the selectorate model by Bueno de Mesquita et al.

Let us try to emulate their test. In order to load their dataset you need to use the following command: `install.packages(“foreign”)`

Note that the working directory you set depends on where you have the file on your computer.

```
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/")
# Sets the working directory
library(foreign)
# Allows you to read more data formats
LDC = read.dta("LDC_IO_replication.dta")
summary(LDC)
```

```
##      country      ctylabel      date      gatt_wto_new
## Min.      :186.0 Length:5370 Min.      :1970 Min.      :0.0000
## 1st Qu.:423.0 Class :character 1st Qu.:1977 1st Qu.:0.0000
## Median :628.0 Mode  :character Median :1984 Median :0.0000
## Mean   :605.9      Mean   :1984 Mean   :0.4747
## 3rd Qu.:816.0      3rd Qu.:1992 3rd Qu.:1.0000
## Max.    :968.0      Max.    :1999 Max.    :1.0000
##                                     NA's    :698
##      aclpn      bpc1      dopen_wacz2      ecris2
## Min.      :0.0000 Min.      :0.000 Min.      :0.0000 Min.      :0.0000
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000
## Median :0.0000 Median :1.000 Median :0.0000 Median :0.0000
## Mean   :0.3002 Mean   :0.591 Mean   :0.3097 Mean   :0.0641
## 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000
## Max.    :1.0000 Max.    :1.000 Max.    :1.0000 Max.    :1.0000
## NA's    :1183  NA's    :2734 NA's    :2580  NA's    :1967
```

```

##      fdignp      gdp_pc_95d      l1aclpn      l1bpc1
## Min.   :-27.2356   Min.    :    0.0   Min.    :0.0000   Min.    :0.0000
## 1st Qu.:  0.0361   1st Qu.:  442.9   1st Qu.:0.0000   1st Qu.:0.0000
## Median :  0.6644   Median : 1266.5   Median :0.0000   Median :1.0000
## Mean    :  1.8962   Mean    : 2885.5   Mean    :0.2924   Mean    :0.5909
## 3rd Qu.:  2.0829   3rd Qu.: 3002.4   3rd Qu.:1.0000   3rd Qu.:1.0000
## Max.    :184.5647   Max.    :44164.5   Max.    :1.0000   Max.    :1.0000
## NA's    :2294      NA's    :1679      NA's    :1341      NA's    :2735
##      l1ecris2      newtar      polityiv_update2      signed
## Min.    :0.0000   Min.    :  0.00   Min.    : -10.000   Min.    :0.0000
## 1st Qu.:0.0000   1st Qu.: 10.95   1st Qu.: -7.000   1st Qu.:0.0000
## Median :0.0000   Median : 17.00   Median : -6.000   Median :0.0000
## Mean    :0.0641   Mean    : 20.54   Mean    : -2.074   Mean    :0.1465
## 3rd Qu.:0.0000   3rd Qu.: 27.00   3rd Qu.:  6.000   3rd Qu.:0.0000
## Max.    :1.0000   Max.    :102.20   Max.    : 10.000   Max.    :1.0000
## NA's    :1967      NA's    :4463      NA's    :2003      NA's    :1362
##      yrsoffic      usheg      l1usheg      l1fiveop
## Min.    : 0.000   Min.    :0.2434   Min.    :0.2434   Min.    :10.20
## 1st Qu.: 2.000   1st Qu.:0.2574   1st Qu.:0.2574   1st Qu.:10.90
## Median : 5.000   Median :0.2663   Median :0.2655   Median :12.35
## Mean    : 8.431   Mean    :0.2696   Mean    :0.2683   Mean    :12.03
## 3rd Qu.:12.000   3rd Qu.:0.2785   3rd Qu.:0.2784   3rd Qu.:12.72
## Max.    :44.000   Max.    :0.3083   Max.    :0.2988   Max.    :13.20
## NA's    :2361      NA's    :179       NA's    :358
##      l1gdp_pc      avsw      avnewtar      l1avsw
## Min.    :    0   Min.    :0.1398   Min.    : 0.00   Min.    :0.1398
## 1st Qu.:  442   1st Qu.:0.1505   1st Qu.: 0.00   1st Qu.:0.1505
## Median : 1266   Median :0.1720   Median :17.43   Median :0.1613
## Mean    : 2888   Mean    :0.3097   Mean    :14.91   Mean    :0.2974
## 3rd Qu.: 2999   3rd Qu.:0.5269   3rd Qu.:24.37   3rd Qu.:0.5054
## Max.    :44165   Max.    :0.6667   Max.    :30.52   Max.    :0.6559
## NA's    :1823      NA's    :179
##      l1avnewtar      lnpop      l1lnpop      l1office
## Min.    : 0.00   Min.    :10.57   Min.    :10.62   Min.    : 0.000
## 1st Qu.: 0.00   1st Qu.:13.86   1st Qu.:13.86   1st Qu.: 2.000
## Median :18.73   Median :15.32   Median :15.31   Median : 5.000
## Mean    :15.01   Mean    :15.11   Mean    :15.10   Mean    : 8.431
## 3rd Qu.:24.37   3rd Qu.:16.40   3rd Qu.:16.39   3rd Qu.:12.000
## Max.    :30.52   Max.    :20.95   Max.    :20.94   Max.    :44.000
## NA's    :179     NA's    :490     NA's    :661     NA's    :2361
##      l1partyage2000      l1fdi      l1polity      l2polity
## Min.    : 0.00   Min.    : -27.2356   Min.    : -10.000   Min.    : -10.00
## 1st Qu.: 10.00   1st Qu.:  0.0269   1st Qu.: -7.000   1st Qu.: -7.00
## Median : 19.50   Median :  0.6382   Median : -6.000   Median : -7.00
## Mean    : 24.18   Mean    :  1.7931   Mean    : -2.215   Mean    : -2.36
## 3rd Qu.: 32.00   3rd Qu.:  1.9904   3rd Qu.:  6.000   3rd Qu.:  5.00
## Max.    :183.00   Max.    :184.5647   Max.    : 10.000   Max.    : 10.00
## NA's    :3284     NA's    :2423     NA's    :2124     NA's    :2246
##      l3polity      l1signed      milit2      sp2
## Min.    : -10.000   Min.    :0.0000   Min.    :0.0000   Min.    :0.0000
## 1st Qu.: -7.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
## Median : -7.000   Median :0.0000   Median :0.0000   Median :0.0000
## Mean    : -2.512   Mean    :0.1511   Mean    :0.1119   Mean    :0.1959
## 3rd Qu.:  5.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000

```

```
## Max. : 10.000 Max. :1.0000 Max. :1.0000 Max. :1.0000
## NA's :2371 NA's :1517
## pers2 l1milit2 l1sp2 dictator1
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :1.000
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:2.000
## Median :0.0000 Median :0.0000 Median :0.0000 Median :5.000
## Mean :0.1665 Mean :0.1135 Mean :0.1986 Mean :4.737
## 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:8.000
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :8.000
## NA's :179 NA's :179 NA's :1157
## l1dictator1 yr70 yr80 l1ssch
## Min. :1.000 Min. :0.0000 Min. :0.0000 Min. :0.0140
## 1st Qu.:2.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.4562
## Median :5.000 Median :0.0000 Median :0.0000 Median :0.8519
## Mean :4.708 Mean :0.3333 Mean :0.3333 Mean :1.0411
## 3rd Qu.:8.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.4652
## Max. :8.000 Max. :1.0000 Max. :1.0000 Max. :4.4422
## NA's :1315 NA's :3140
## closedyr _spline1 _spline2 _spline3
## Min. : 0.000 Min. : -24389 Min. : -7854.0 Min. : -9030.0
## 1st Qu.: 0.000 1st Qu.: -3375 1st Qu.: -2048.3 1st Qu.: -1629.3
## Median : 7.000 Median : -343 Median : -260.2 Median : -165.6
## Mean : 8.691 Mean : -3075 Mean : -1388.8 Mean : -1340.9
## 3rd Qu.:15.000 3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.: 0.0
## Max. :29.000 Max. : 0 Max. : 0.0 Max. : 0.0
## NA's :2580 NA's :2580 NA's :2580 NA's :2580
## l1gatt_wto_new
## Min. :0.000
## 1st Qu.:0.000
## Median :0.000
## Mean :0.468
## 3rd Qu.:1.000
## Max. :1.000
## NA's :868
```

For information on the meaning of the variables see “LDCcodebook.pdf”.

Let’s have a look at our data.

We need a package for generating a scatterplot matrix that allows us to see relationships in our matrix.

```
install.packages(“car”)
```

```
library(car)

LDC2=as.data.frame(LDC[,c("l1polity", "l1signed", "l1office", "l1gdp_pc", "l1lnpop",
                          "l1ecris2", "l1bpc1", "l1avnewtar")])

cor(LDC2)
```

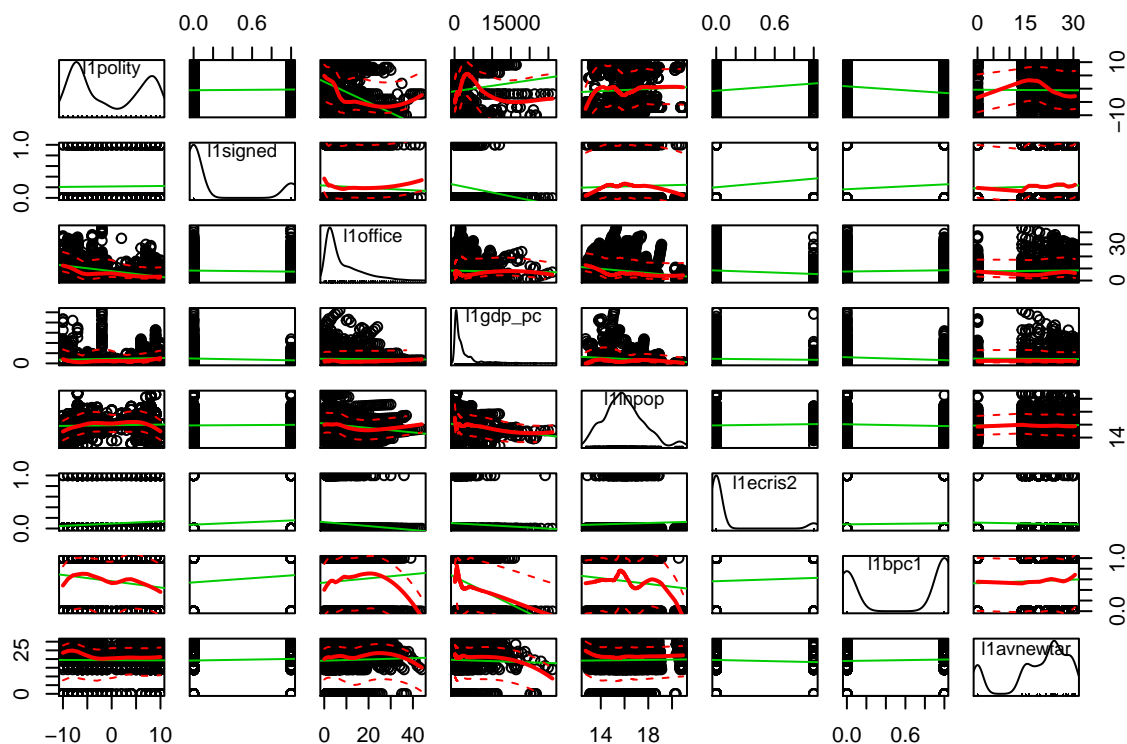
```
##          l1polity l1signed l1office l1gdp_pc l1lnpop l1ecris2 l1bpc1
## l1polity          1      NA      NA      NA      NA      NA      NA
## l1signed          NA          1      NA      NA      NA      NA      NA
## l1office          NA          NA          1      NA      NA      NA      NA
## l1gdp_pc          NA          NA          NA          1      NA      NA      NA
## l1lnpop           NA          NA          NA          NA          1      NA      NA
```

```
## l1ecris2      NA      NA      NA      NA      NA      1      NA
## l1bpc1        NA      NA      NA      NA      NA      NA      1
## l1avnewtar    NA      NA      NA      NA      NA      NA      NA
##              l1avnewtar
## l1polity      NA
## l1signed      NA
## l1office      NA
## l1gdp_pc      NA
## l1lnpop       NA
## l1ecris2      NA
## l1bpc1        NA
## l1avnewtar    1
```

```
LDC3=na.omit(LDC2)
cor(LDC3)
```

```
##              l1polity    l1signed    l1office    l1gdp_pc    l1lnpop
## l1polity      1.000000000  0.01499208 -0.42901753  0.09129002  0.04404210
## l1signed      0.014992081  1.000000000 -0.04303356 -0.11240587  0.02321607
## l1office      -0.429017535 -0.04303356  1.000000000 -0.01936245 -0.17457461
## l1gdp_pc      0.091290020 -0.11240587 -0.01936245  1.000000000 -0.14082411
## l1lnpop       0.044042096  0.02321607 -0.17457461 -0.14082411  1.000000000
## l1ecris2      0.108433533  0.11269712 -0.10331828 -0.04446845  0.03580805
## l1bpc1        -0.171762855  0.10895350  0.06396247 -0.22176992 -0.08570695
## l1avnewtar    -0.008790383  0.03967319  0.03040646 -0.02357807  0.01262803
##              l1ecris2    l1bpc1    l1avnewtar
## l1polity      0.10843353 -0.17176286 -0.008790383
## l1signed      0.11269712  0.10895350  0.039673186
## l1office      -0.10331828  0.06396247  0.030406457
## l1gdp_pc      -0.04446845 -0.22176992 -0.023578065
## l1lnpop       0.03580805 -0.08570695  0.012628033
## l1ecris2      1.000000000  0.03517381 -0.036064786
## l1bpc1        0.03517381  1.000000000  0.042851289
## l1avnewtar    -0.03606479  0.04285129  1.000000000
```

```
scatterplotMatrix(~ l1polity + l1signed + l1office + l1gdp_pc + l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar)
```



The results above indicate that there generally is a low level of multicollinearity among our variables.

Let us start with a simple model that is easy to interpret:

```
simple = lm(newtar ~ l1polity, data = LDC)
summary(simple)
```

```
##
## Call:
## lm(formula = newtar ~ l1polity, data = LDC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.425  -9.200  -3.425   5.275  80.475
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.92495    0.52865  41.474 < 2e-16 ***
## l1polity     -0.30001    0.07293  -4.113 4.3e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.9 on 804 degrees of freedom
## (4564 observations deleted due to missingness)
## Multiple R-squared:  0.02061,    Adjusted R-squared:  0.01939
## F-statistic: 16.92 on 1 and 804 DF, p-value: 4.298e-05
```

What can we conclude from these statistics? What can we say about the hypothesis that there is a linear relationship between “l1polity” and “newtar”? What is the total variation that is explained by our model?

If there’s too much information in this type of summary, try another one. We need another package: `install.packages(“arm”)`

```
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
```

```
## arm (Version 1.9-1, built: 2016-8-21)
```

```
## Working directory is C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/ps630_f16/W5
```

```
##
```

```
## Attaching package: 'arm'
```

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

```
##
```

```
##      logit
```

```
display(simple)
```

```
## lm(formula = newtar ~ l1polity, data = LDC)
```

```
##           coef.est coef.se
```

```
## (Intercept) 21.92      0.53
```

```
## l1polity    -0.30      0.07
```

```
## ---
```

```
## n = 806, k = 2
```

```
## residual sd = 14.90, R-Squared = 0.02
```

As you can see, this is narrowed down to just a few pieces of information. Sometimes reducing the amount of information that is displayed can be very useful.

3. Multiple linear regression

In the vast majority of cases there are good reasons to include multiple predictor variables.

The most important reasons to do so are:

1. Potential omitted variable bias
2. Theoretical reasons
3. Reviewers that demand you to include them


```
main = lm(newtar ~ l1polity + l1signed + l1office + l1gdp_pc + l1lnpop + l1ecris2 +
  l1bpc1 + l1avnewtar, data = LDC)
summary(main)
```

```
##
## Call:
## lm(formula = newtar ~ l1polity + l1signed + l1office + l1gdp_pc +
##     l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar, data = LDC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.286  -7.694  -2.175   4.490  65.008
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.901e+01  5.912e+00  -8.289 6.03e-16 ***
## l1polity     -2.053e-01  8.347e-02  -2.460 0.014151 *
## l1signed      4.758e-01  1.099e+00   0.433 0.665332
## l1office     -1.759e-01  6.989e-02  -2.516 0.012083 *
## l1gdp_pc     -1.281e-03  1.495e-04  -8.564 < 2e-16 ***
## l1lnpop      3.693e+00  3.217e-01  11.478 < 2e-16 ***
## l1ecris2     -5.736e+00  1.517e+00  -3.780 0.000171 ***
## l1bpc1       4.564e-01  9.681e-01   0.471 0.637462
## l1avnewtar   7.103e-01  8.413e-02   8.442 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.16 on 685 degrees of freedom
## (4676 observations deleted due to missingness)
## Multiple R-squared:  0.3781, Adjusted R-squared:  0.3708
## F-statistic: 52.05 on 8 and 685 DF,  p-value: < 2.2e-16
```

In the multiple linear regression, how would our expectation for the average tariff level change if our Polity Score increased from -10 to 10 and we had an economic crisis?

How well does our model do compared to the simple linear regression? Do we observe an improvement in the total variation that is explained by our model?

Again, it would be possible to reduce the amount of information with another command:

```
display(main)
```

```
## lm(formula = newtar ~ l1polity + l1signed + l1office + l1gdp_pc +
##     l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar, data = LDC)
##              coef.est coef.se
## (Intercept)  -49.01      5.91
## l1polity      -0.21      0.08
## l1signed       0.48      1.10
## l1office      -0.18      0.07
## l1gdp_pc       0.00      0.00
## l1lnpop       3.69      0.32
## l1ecris2      -5.74      1.52
## l1bpc1        0.46      0.97
```

```
## llavnewtar    0.71    0.08
## ---
## n = 694, k = 9
## residual sd = 12.16, R-Squared = 0.38
```

We can access different elements of our model. Let's have a look at what those are:

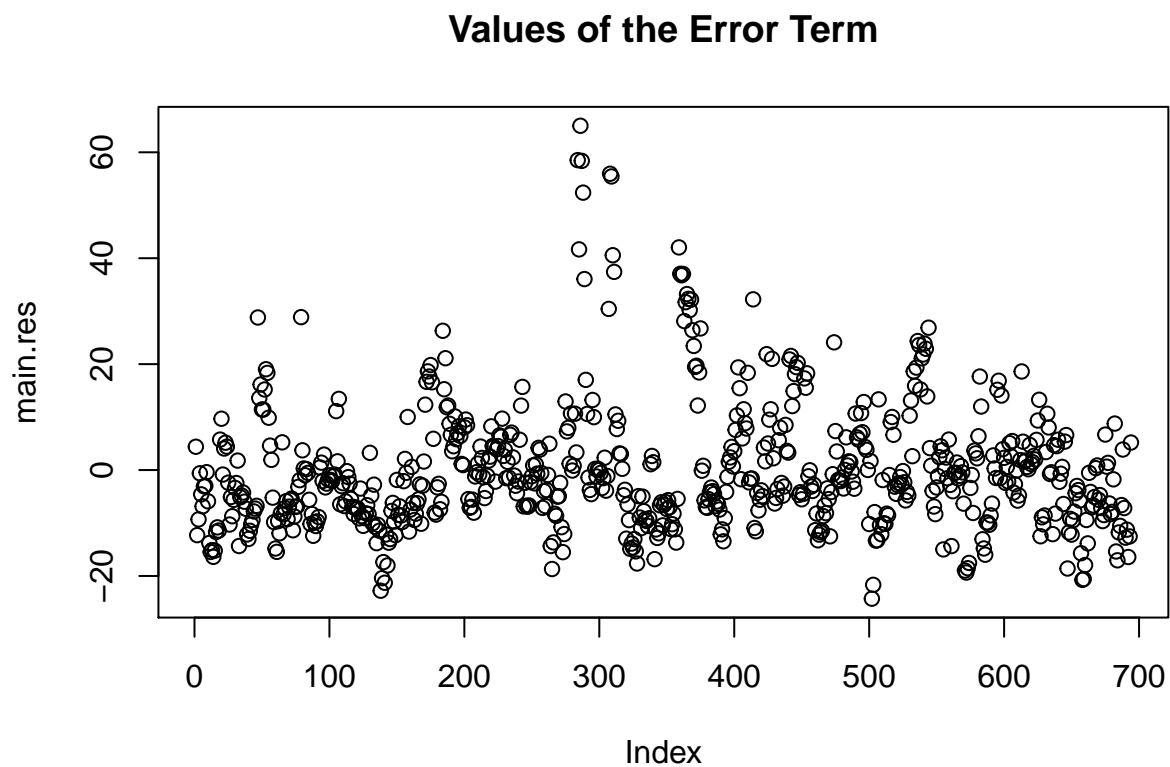
```
names(main)
```

```
## [1] "coefficients" "residuals"    "effects"      "rank"
## [5] "fitted.values" "assign"       "qr"          "df.residual"
## [9] "na.action"    "xlevels"     "call"        "terms"
## [13] "model"
```

4. Graphical representation

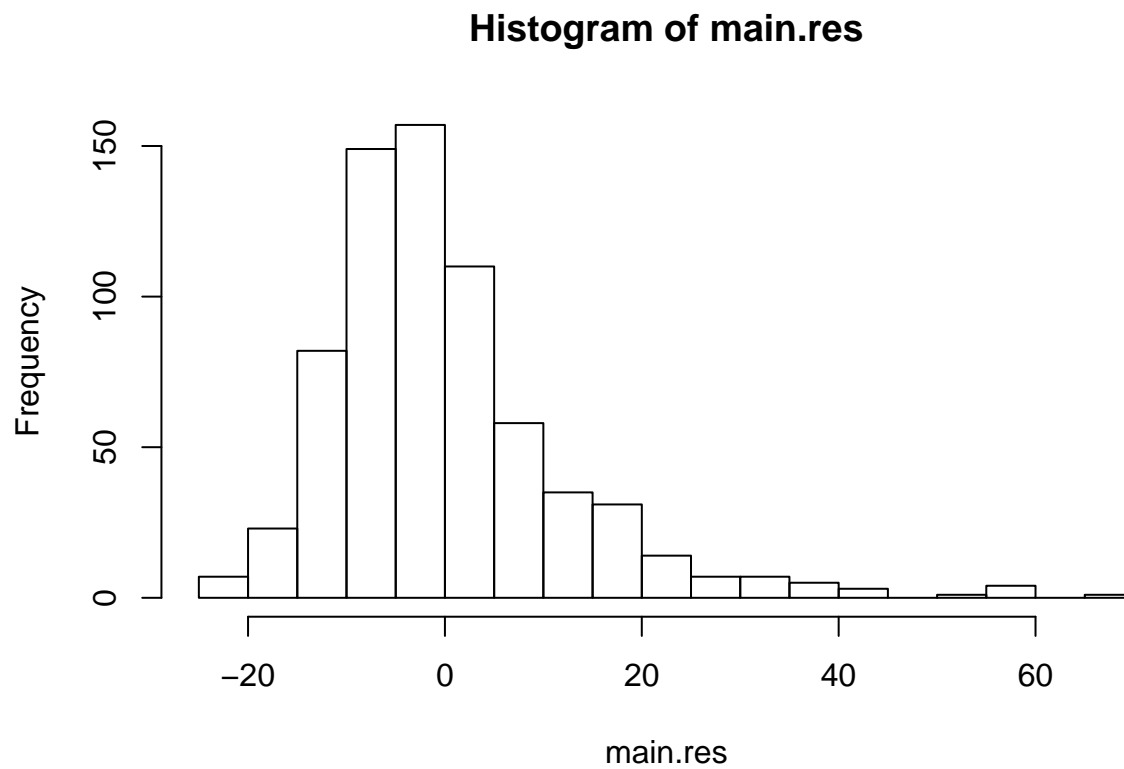
Let us first have a look at the distribution of errors in our model.

```
main.res = resid(main)
plot(main.res, main = "Values of the Error Term")
```

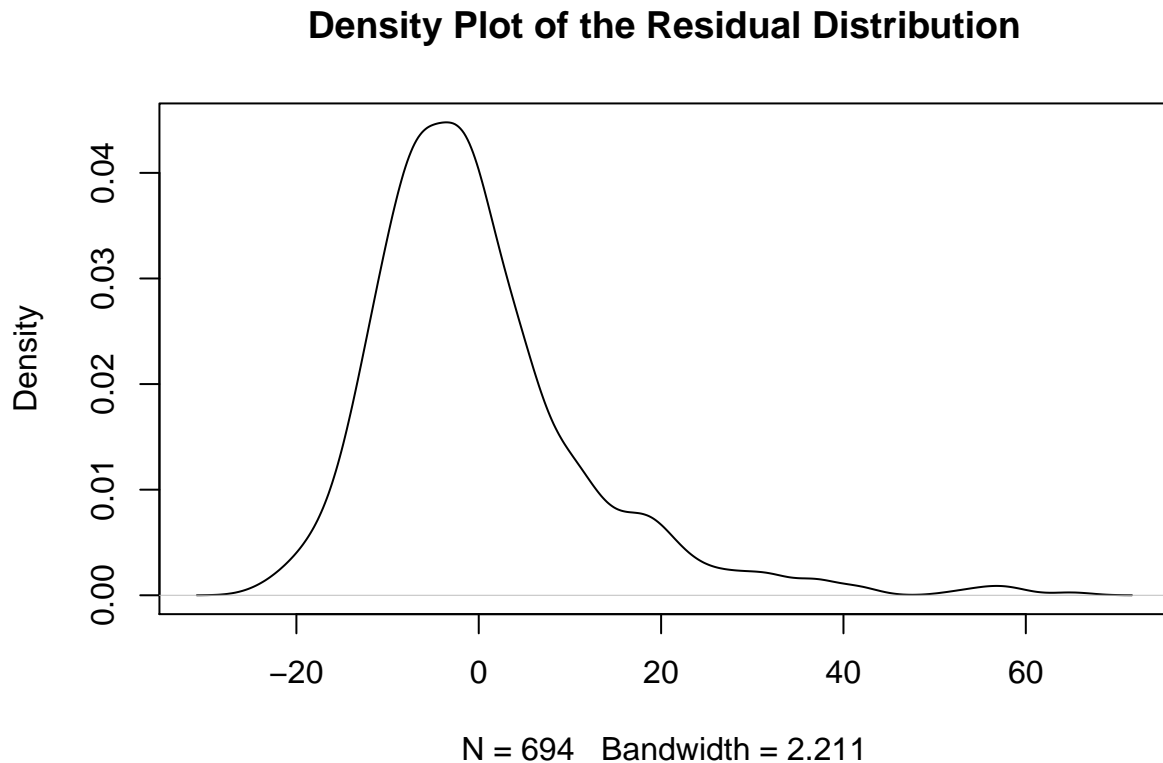


Let us look at the distribution of the error term:

```
hist(main.res, breaks = 20)
```



```
res.density = density(main.res)  
plot(res.density, main = "Density Plot of the Residual Distribution")
```



The distribution of the errors is approximately normal. If this condition is met, then more precise statements about the distribution of the coefficients can be made (they're also normal). Also, under these conditions, OLS is equivalent to a maximum likelihood approach.

Plotting predicted values

Let us plot some predicted values with confidence intervals for our multiple regression.

In order to do that we first create a dataframe that contains different values for our main predictor variable and the average values for all variables.

```
nd = data.frame(l1polity = seq(-10, 10, by = 1), l1signed = rep(0.1511, 21),
  l1office = rep(8.431, 21), l1gdp_pc = rep(2888, 21), l1lnpop = rep(15.1,
    21), l1ecris2 = rep(0.0641, 21), l1bpc1 = rep(0.5909, 21), l1avnewtar = rep(14.91,
    21))
```

Note: Alternatively, we could also use the following code:

```
nd = data.frame(l1polity = seq(-10, 10, by = 1), l1signed = rep(mean(LDC$l1signed,
  na.rm = T), 21), l1office = rep(mean(LDC$l1office, na.rm = T), 21), l1gdp_pc = rep(mean(LDC$l1gdp_p
  na.rm = T), 21), l1lnpop = rep(mean(LDC$l1lnpop, na.rm = T), 21), l1ecris2 = rep(mean(LDC$l1ecris2,
  na.rm = T), 21), l1bpc1 = rep(mean(LDC$l1bpc1, na.rm = T), 21), l1avnewtar = rep(mean(LDC$l1avnewta
  na.rm = T), 21))
nd
```

```
##      l1polity  l1signed l1office l1gdp_pc  l1lnpop  l1ecris2  l1bpc1
## 1      -10 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
```

```

## 2      -9 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 3      -8 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 4      -7 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 5      -6 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 6      -5 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 7      -4 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 8      -3 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 9      -2 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 10     -1 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 11      0 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 12      1 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 13      2 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 14      3 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 15      4 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 16      5 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 17      6 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 18      7 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 19      8 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 20      9 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
## 21     10 0.1510511 8.431373 2887.762 15.10395 0.06406112 0.5908918
##      l1avnewtar
## 1      15.01104
## 2      15.01104
## 3      15.01104
## 4      15.01104
## 5      15.01104
## 6      15.01104
## 7      15.01104
## 8      15.01104
## 9      15.01104
## 10     15.01104
## 11     15.01104
## 12     15.01104
## 13     15.01104
## 14     15.01104
## 15     15.01104
## 16     15.01104
## 17     15.01104
## 18     15.01104
## 19     15.01104
## 20     15.01104
## 21     15.01104

```

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

```

pred.p1 = predict(main, type = "response", se.fit = TRUE, newdata = nd)

pred.table = cbind(pred.p1$fit, pred.p1$se.fit)
pred.table

```

```

##      [,1]      [,2]
## 1  14.27863 1.3372462
## 2  14.07333 1.2741682

```

```
## 3 13.86803 1.2135547
## 4 13.66272 1.1557934
## 5 13.45742 1.1013331
## 6 13.25212 1.0506874
## 7 13.04682 1.0044334
## 8 12.84152 0.9632040
## 9 12.63622 0.9276694
## 10 12.43092 0.8985055
## 11 12.22562 0.8763485
## 12 12.02032 0.8617391
## 13 11.81502 0.8550644
## 14 11.60972 0.8565097
## 15 11.40442 0.8660344
## 16 11.19912 0.8833773
## 17 10.99382 0.9080905
## 18 10.78852 0.9395926
## 19 10.58322 0.9772273
## 20 10.37792 1.0203162
## 21 10.17262 1.0681995
```

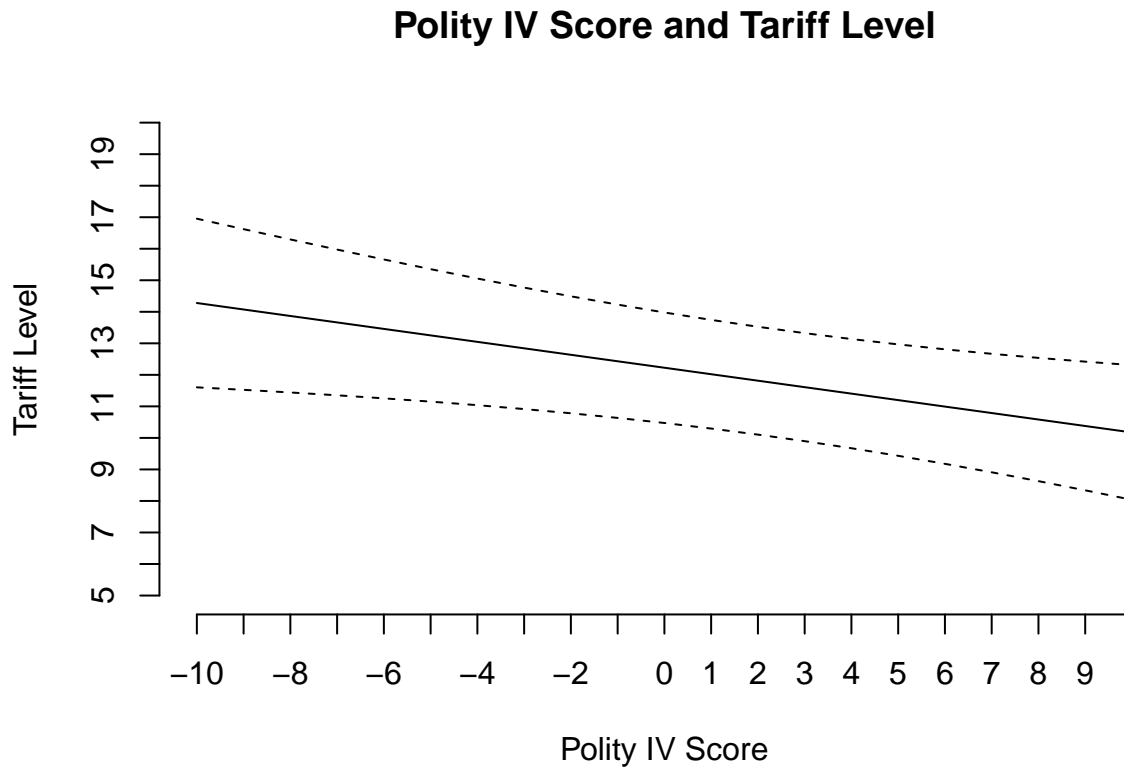
Finally, we create the plot:

```
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)

cis ### To extract the values
```

```
##      fit      low      high
## 1 14.27863 11.604133 16.95312
## 2 14.07333 11.524989 16.62166
## 3 13.86803 11.440916 16.29513
## 4 13.66272 11.351138 15.97431
## 5 13.45742 11.254758 15.66009
## 6 13.25212 11.150749 15.35350
## 7 13.04682 11.037957 15.05569
## 8 12.84152 10.915115 14.76793
## 9 12.63622 10.780884 14.49156
## 10 12.43092 10.633912 14.22793
## 11 12.22562 10.472925 13.97832
## 12 12.02032 10.296844 13.74380
## 13 11.81502 10.104893 13.52515
## 14 11.60972  9.896702 13.32274
## 15 11.40442  9.672352 13.13649
## 16 11.19912  9.432366 12.96588
## 17 10.99382  9.177639 12.81000
## 18 10.78852  8.909335 12.66770
## 19 10.58322  8.628765 12.53767
## 20 10.37792  8.337287 12.41855
## 21 10.17262  8.036220 12.30902
```

```
plot(pred.pi$fit, type = "l", ylim = c(5, 20), main = "Polity IV Score and Tariff Level",
     xlab = "Polity IV Score", ylab = "Tariff Level", axes = FALSE)
axis(1, at = seq(1, 21), labels = seq(-10, 10, 1))
axis(2, at = seq(5, 20), labels = seq(5, 20))
matlines(cis[, c(2, 3)], lty = 2, col = "black")
```



5. Tips for your final paper

1. Start working on it early.
2. Consult with your professors and TAs.
3. Try to find a comprehensive dataset in your area of interest.
4. Work on it throughout the semester and try to include new things that you've learned.
5. Make sure that you use all the tools you've learned: interpret your findings carefully and visualize them.
6. Annotate your code extensively and explain what you did.