

Pol Sci 630: Problem Set 5 - Regression Model

Interpretation - Solutions

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

Grading Due Date: Friday, September 25th, 12.15 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 correctly solved problems, assign the point value that is stated in the problem (see original problem set for exact point values). 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

data(swiss)
summary(swiss)

##      Fertility      Agriculture      Examination      Education
##  Min.       :35.00   Min.        : 1.20   Min.        : 3.00   Min.        : 1.00
## 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
## Median :70.40   Median :54.10   Median :16.00   Median : 8.00
## Mean  :70.14   Mean  :50.66   Mean  :16.49   Mean  :10.98
## 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
## Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##      Catholic      Infant.Mortality
##  Min.       : 2.150   Min.       :10.80
## 1st Qu.: 5.195   1st Qu.:18.15
## Median :15.140   Median :20.00
## Mean  :41.144   Mean  :19.94
## 3rd Qu.:93.125   3rd Qu.:21.70
## Max.   :100.000   Max.       :26.60

#### b

lm1 = lm(Education ~ Fertility + Agriculture + Examination + Catholic + Infant.Mortality
         data = swiss)

summary(lm1)

##
## Call:
## lm(formula = Education ~ Fertility + Agriculture + Examination +
##      Catholic + Infant.Mortality, data = swiss)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3949  -2.3716  -0.2856   2.8108  11.2985
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    32.74414     8.87888   3.688 0.000657 ***
## Fertility      -0.40851     0.08585  -4.758 2.43e-05 ***
## Agriculture    -0.16242     0.04488  -3.619 0.000804 ***
## Examination     0.41980     0.16339   2.569 0.013922 *
## Catholic        0.10023     0.02150   4.663 3.29e-05 ***
## Infant.Mortality 0.20408     0.28390   0.719 0.476305
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.907 on 41 degrees of freedom
## Multiple R-squared:  0.7678, Adjusted R-squared:  0.7395
## F-statistic: 27.12 on 5 and 41 DF,  p-value: 5.223e-12
```

c) In order to get full points on this problem, you need an interpretation for each of the 5 variables.

The interpretation would look like this for *Fertility*:

There is a negative linear relationship between *Fertility* and *Education*. For a 1-point increase in Fertility, we expect a 0.41-point decrease in Education, holding all other variables constant. The t-value is -4.758. This t-value implies a p-value of 2.43×10^{-5} . This $p < 0.001$ corresponds to a type-1 error rate of $\alpha < 0.001$, meaning that the statistical relationship is significant at all common levels of statistical significance.

The other variables are interpreted accordingly. *Agriculture* and *Catholic* are significant at all common levels of statistical significance as well. Please note that *Examination* is significant at a level of $p < 0.05$, $\alpha < 0.05$, and *Infant.Mortality* is not significant at common levels of statistical significance.

The R-squared statistic shows us that our model explains 76.78 percent (multiple R-squared) or 73.95 percent (adjusted R-squared) of the variation in the dependent variable

— depending on whether one is penalized for introducing further variables into the model.

Problem 3

a) In this task you have to formulate a hypothesis regarding the relationship of several political and economic factors and the level of FDI inflows. For example, you could claim that economic crises generally lead to a lower inflow of foreign investment because countries that experience crises are less attractive to investors. In a well-known paper in International Organization, Nathan Jensen made the claim that democratic institutions can make more credible commitments to upholding property rights, which means that foreign investors trust democratic governments more than authoritarian governments. Regardless of which variable you choose, your hypothesis should look similar to this one:

Hypothesis: If a country experiences an economic crisis, it will experience less foreign direct investment (as percentage of GDP) than a country that does not experience an economic crisis.

Or, alternatively:

Hypothesis: Countries with higher levels of democracy experience more foreign direct investment (as percentage of GDP) than countries with lower levels of democracy.

```
### b
setwd("C:/Users/Jan/OneDrive/Documents/GitHub/ps630_lab/W5")
library(foreign)
LDC = read.dta("LDC_IO_replication.dta")

lm_fdi = lm(fdignp ~ l1polity + l1signed + l1office + l1gdp_pc + l1lnpop + l1ecris2 +
            l1bpc1 + l1avnewtar, data = LDC)
summary(lm_fdi)

##
## Call:
## lm(formula = fdignp ~ l1polity + l1signed + l1office + l1gdp_pc +
##      l1lnpop + l1ecris2 + l1bpc1 + l1avnewtar, data = LDC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -17.943  -1.537  -0.724   0.358 181.394
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.051e+01  1.542e+00   6.814 1.33e-11 ***
## l1polity      3.976e-02  2.316e-02   1.717  0.0862 .
## l1signed     -4.802e-01  3.347e-01  -1.435  0.1516
## l1office     -1.222e-02  1.948e-02  -0.628  0.5304
## l1gdp_pc     -2.231e-05  9.383e-05  -0.238  0.8121
## l1lnpop      -4.945e-01  9.033e-02  -5.474 5.07e-08 ***
## l1ecris2      8.480e-01  4.806e-01   1.765  0.0778 .
## l1bpc1       -2.411e-02  2.956e-01  -0.082  0.9350
## l1avnewtar   -3.311e-02  1.482e-02  -2.234  0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.665 on 1649 degrees of freedom
## (3712 observations deleted due to missingness)
## Multiple R-squared:  0.0272, Adjusted R-squared:  0.02248
## F-statistic: 5.764 on 8 and 1649 DF,  p-value: 2.883e-07
```

Let us interpret our findings for the two hypotheses above:

Regarding *l1polity*: For a 1-point increase in the Polity IV Score, we would expect a 0.0397 (0.04) increase in the level of foreign direct investment as percentage of GDP, holding all other variable constant. The associated p-value of 0.0862 means that this relationship is statistically significant at $p < 0.1$ ($\alpha < 0.1$) but not at $p < 0.05$ ($\alpha < 0.05$). This means that there is some support for the hypothesis that democracy leads to higher levels of foreign investment, although the evidence is not as strong as we might have expected.

Regarding *l1ecris2*: If a country experiences an economic crisis, we would expect a 0.848 (0.85) increase in the level of foreign direct investment as percentage of GDP, holding all other variable constant. The p-value of 0.0778 means that this relationship is statistically significant at $p < 0.1$ ($\alpha < 0.1$) but not at $p < 0.05$ ($\alpha < 0.05$). This means that the empirical evidence speaks against the hypothesis that there is a negative effect of economic crises on foreign direct investment. This result may appear counter intuitive and could be the result of omitted variable bias or endogeneity.

The R-squared statistic shows us that our model explains 2.72 percent (multiple R-squared) or 2.25 percent (adjusted R-squared) of the variation in the dependent variable — depending on whether one is penalized for introducing further variables into the model.

Finally, we show the effect of the Polity IV Score on FDI graphically:

```
### We create a new dataframe with the average values for every variable and
### vary Polity IV
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(14.9,
    21))

pred.p1 <- predict(lm_fdi, 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)

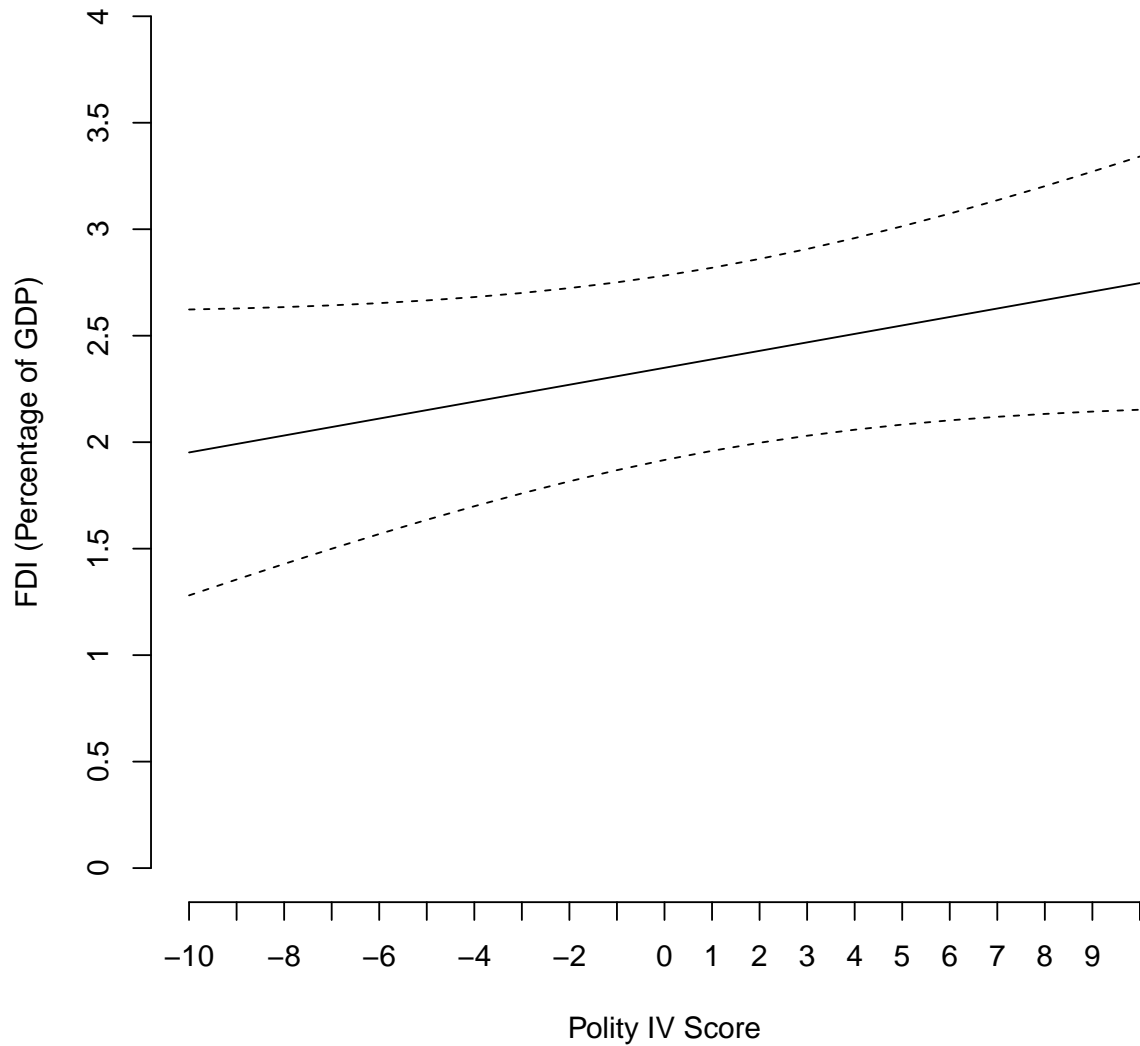
cis ### To extract the values

##      fit      low      high
## 1  1.951736 1.280473 2.622999
## 2  1.991497 1.355123 2.627871
## 3  2.031258 1.428231 2.634285
## 4  2.071019 1.499528 2.642509
## 5  2.110780 1.568698 2.652861
## 6  2.150540 1.635376 2.665705
## 7  2.190301 1.699152 2.681450
## 8  2.230062 1.759582 2.700542
## 9  2.269823 1.816208 2.723438
## 10 2.309584 1.868594 2.750574
## 11 2.349345 1.916368 2.782322
```

```
## 12 2.389106 1.959272 2.818939
## 13 2.428867 1.997201 2.860532
## 14 2.468627 2.030216 2.907039
## 15 2.508388 2.058538 2.958239
## 16 2.548149 2.082513 3.013786
## 17 2.587910 2.102565 3.073255
## 18 2.627671 2.119150 3.136192
## 19 2.667432 2.132719 3.202144
## 20 2.707193 2.143692 3.270693
## 21 2.746954 2.152446 3.341461

plot(pred.p1$fit, type = "l", ylim = c(0, 4), main = "Polity IV Score and FDI",
      xlab = "Polity IV Score", ylab = "FDI (Percentage of GDP)", axes = FALSE)
axis(1, at = seq(1, 21), labels = seq(-10, 10, 1))
axis(2, at = seq(0, 4, by = 0.5), labels = seq(0, 4, by = 0.5))
matlines(cis[, c(2, 3)], lty = 2, col = "black")
```

Polity IV Score and FDI



Linear Regression Models Interpretation Questions

Problem 3

a) If we do not include polynomials of higher order, OLS regression can adequately model linear relationships between one dependent variable (response variable) and multiple independent variables (predictor variables). The reason for this is that our model assumes that

for every independent variable there is only one slope coefficient that is constant for all values of that independent variable.

b) A probabilistic linear relationship means that two variables vary with each other in a way that knowing one of them (let's say X) allows us to make relatively precise predictions about the average of the other variable (let's say Y) — but not about its exact values. A probabilistic relationship therefore means that there are other factors that cause some variation in Y and that cannot be explained through X , even though X is systematically related to Y .

A deterministic linear relationship means that two variables vary with each other in a way that knowing one of them (X) allows us to make precise predictions about the exact value of the other variable (Y). A deterministic relationship therefore means that X is the sole factor that influences the value of Y .

OLS can model both relationships. The difference is that in probabilistic relationships the error term will correct for the difference in prediction and true value. Although most relationships modeled by OLS are probabilistic, OLS can also display deterministic relationships. If a relationship is truly deterministic, then the error term would be zero but the linear model would still be a meaningful representation of the underlying relationship.

c) OLS regression does not per se tell us anything about causality. OLS regression primarily measures linear relationships between two variables and can give us an answer to the question how two variables are correlated with each other. However, without a strong theory, OLS does not allow us to make statements regarding causality. There could be reverse causality, meaning that the response variable in our model has a causal effect on the predictor variable. There could be endogeneity, meaning that there is mutual causal influence of response and predictor variables. Finally, there could be omitted variable bias, meaning that a third variable influences both the predictor and the response variable.

Statistical Theory: Linear Regression Models

Problem 4

a)

b)