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. See below for more examples.

Use the following scheme to assign points:
Nothing correct: 0 points
25 % or more correct: 1 point
50 % or more correct: 2 points
75 % or more correct: 3 points
100 % or more correct: 4 points
Correct bonus problems can add points, but not beyond the maximum of 4 points

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 levels of significance can be found in the tutorial notes.

Problem 3

0.1 a)

0.2 b)

What can we say about causality? Nothing really. There are two primary reasons for this:

First and foremost, linear regression does not per se tell us anything about causality - it primarily measures correlation between variables.

Second, we do not have any theory regarding the relationship of Education on the other covariates and so we cannot make any causal claims that are grounded in theory. In particular, there might be a mutual influence between Education and the other variables that we regress it on. This phenomenon is called "endogeneity" and there are various ways to deal with it that you will learn about in the class.

In short, we can't say anything about causality here.

Probability Theory: Linear Model Interpretation

Problem 4

a)

b)