Pol Sci 630: Problem Set 8 Solutions: Dummy Variables and Interactions (Part 2)

Prepared by: Anh Le (anh.le@duke.edu)

Due Date: Friday, Oct 23, 2015, 12 AM (Beginning of Lab)

1 Interaction (8 points)

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! - 8/8 Points” Also briefly point out which, if any, problems were not solved correctly and what the mistake was. See below for more examples.

a)

Download FDI, tariff, and GDP data from WDI for all countries, year 2010 (`indicator = c("BX.KLT.DINV.CD.WD", "TM.TAX.MRCH.SM.AR.ZS", "NY.GDP.MKTP.CD")`). Clean the data as usual. Run the following regression, showing a stargazer result table.

$$\log(FDI) = \beta_0 + \beta_1 tariff + \beta_2 \log(gdp) + \beta_3 tariff \times \log(gdp) \qquad (1)$$

Solution

```
library(WDI)
library(dplyr)

d_wdi_raw <- WDI(indicator = c("BX.KLT.DINV.CD.WD", "TM.TAX.MRCH.SM.AR.ZS",
                               "NY.GDP.MKTP.CD"),
         start = 2010, end = 2010, extra = TRUE)

# Cleaning data with dplyr (rename variables, remove aggregate, log transform)
# Hope its nice syntax motivates you to learn this package
d_wdi <- d_wdi_raw %>%
  rename(fdi = BX.KLT.DINV.CD.WD,
         tariff = TM.TAX.MRCH.SM.AR.ZS,
         gdp = NY.GDP.MKTP.CD) %>%
  filter(region != "Aggregates") %>%
  mutate(loggdp = log(gdp), logfdi = log(fdi))

## Warning in mutate_impl(.data, dots):  NaNs produced

m1 <- lm(logfdi ~ tariff + loggdp + tariff:loggdp, data = d_wdi)
# Altenartively, lm(logfdi ~ tariff * loggdp)
```

```
library(stargazer)

##
## Please cite as:
##
##  Hlavac, Marek (2014).  stargazer:  LaTeX code and ASCII text for
well-formatted regression and summary statistics tables.
##  R package version 5.1.  http://CRAN.R-project.org/package=stargazer

stargazer(m1)
```

Table 1:

	Dependent variable:
	logfdi
tariff	0.141
	(0.313)
loggdp	0.930^{***}
	(0.101)
tariff:loggdp	-0.008
	(0.013)
Constant	-1.537
	(2.556)
Observations	108
R^2	0.761
Adjusted R^2	0.754
Residual Std. Error	1.287 (df = 104)
F Statistic	110.431^{***} (df = 3; 104)
Note:	*p<0.1; **p<0.05; ***p<0.01

b)

Mathematically, what is the marginal effect of tariff on logfdi (i.e. taking partial derivative with regards to tariff)? (Hint: This would be a function of loggdp)

Plugging in the number, what's the marginal effect of tariff on logfdi, holding loggdp at its median value? Note: Use `\Sexpr()` to extract coefficients from the model, do not hand write your calculation.

Solution

$$\frac{\partial}{\partial tariff} logfdi = \beta_1 + \beta_3 loggdp \tag{2}$$

The median value of loggdp is 23.8854648

The marginal effect of tariff, holding loggdp at its median value, is -0.0437074

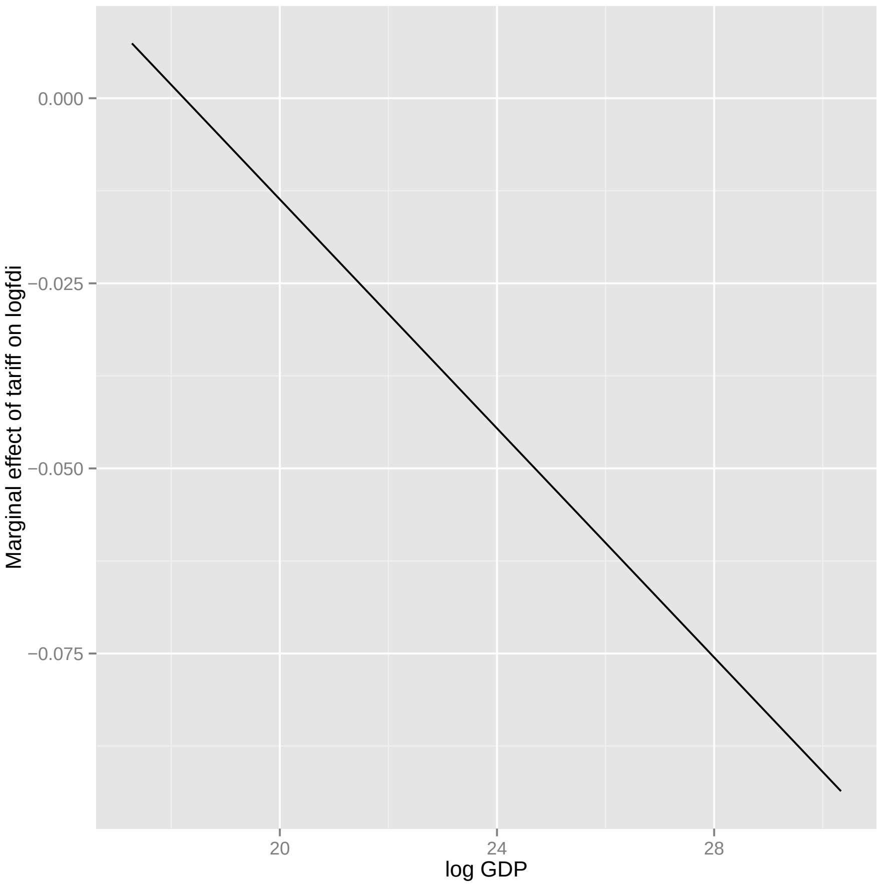
c)

Using `ggplot2`, plot the marginal effect of tariff on logfdi (y-axis) against different values of loggdp (x-axis). (Hint: Create a data frame, in which one variable is the values of loggdp, the other variable is the corresponding marginal effect given that value of loggdp. This data frame is the data that makes up your plot. The plot is just a line.)

Solution

```r
library(ggplot2)
# pd is my shorthand for plot data
loggdp <- sort(d_wdi$loggdp)
marginaleffect <- m1$coefficients['tariff'] + loggdp * m1$coefficients['tariff:loggdp']
pd <- data.frame(loggdp = loggdp,
                 marginaleffect = marginaleffect)

ggplot(data = pd) +
  geom_line(aes(x = loggdp, y = marginaleffect)) +
  labs(y = "Marginal effect of tariff on logfdi", x = "log GDP")
```



d)

With log fdi on the y-axis, tariff on the x-axis, plot the effect of tariff on log fdi when log gdp is at the 25%, 50%, and 75% percentile. Brownie point if you do this in ggplot2.

(Hint: The plot should have 3 lines, each according to a value of log gdp. This is the plot you saw in last lab, with confidence interval included)

Solution

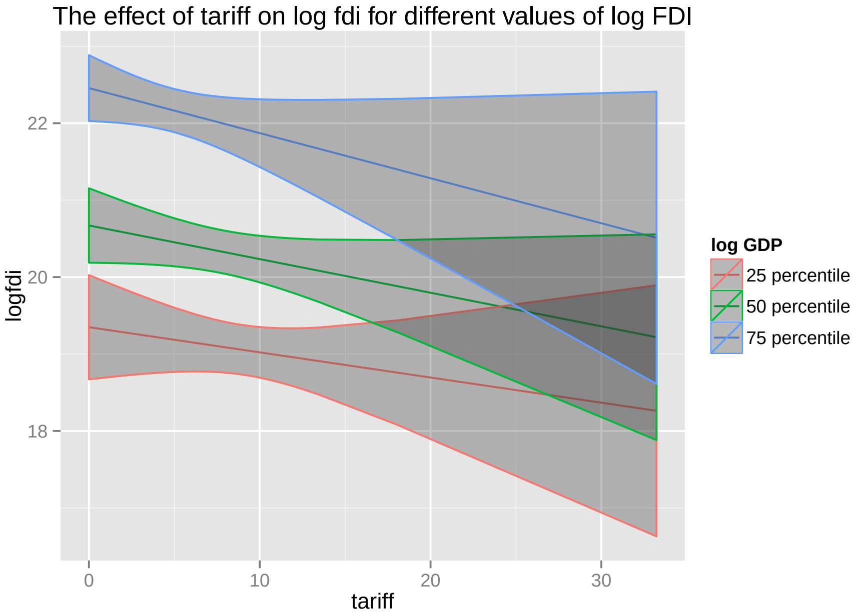
```
library(dplyr)
library(ggplot2)

loggdp_quantiles <- quantile(d_wdi$loggdp, probs = c(0.25, 0.5, 0.75), na.rm=TRUE)


newdata <- data.frame(tariff = rep(d_wdi$tariff, times = 3),
                  loggdp = rep(loggdp_quantiles, each = nrow(d_wdi)))
pred <- predict(m1, newdata = newdata, se.fit = TRUE)

pd <- newdata %>%
  mutate(logfdi = pred$fit,
         ymin = logfdi - 1.96 * pred$se.fit,
         ymax = logfdi + 1.96 * pred$se.fit)
ggplot(data = pd) +
  geom_line(aes(x = tariff, y = logfdi, color = factor(loggdp))) +
  geom_ribbon(aes(x = tariff, ymin = ymin, ymax = ymax,
                  color = factor(loggdp)), alpha = 0.3) +
  labs(title = "The effect of tariff on log fdi for different values of log FDI") +
  scale_color_discrete(name = 'log GDP',
                       labels = c('25 percentile', '50 percentile', '75 percentile'))

## Warning:  Removed 282 rows containing missing values (geom_path).
```



e) Interpretation

Interpret the result (i.e. statistical significance and effect size) using information from both the table and the two plots.

Solution

There does not appear to be any statistically significant interactive effect, as shown both in the table and the fact that the three lines in the second plot are almost parallel.

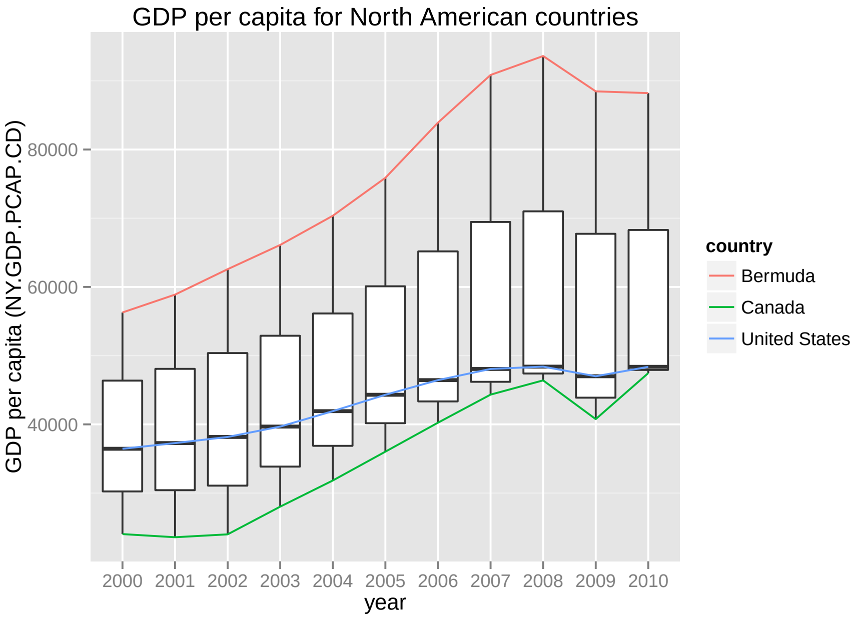
Looking at the table, loggdp appears to have a statistically significant positive effect on logfdi when tariffs are small.

Looking at the first plot, the marginal effect of tariff on logfdi is statistically insignificant when loggdp=0, and it appears that it remains modest across changes in loggdp.

2 ggplot2 (4 points)

Plot this. Note: DO NOT look at the .Rnw code that generates the plot.

If you did try but couldn't figure out, you can look at the code for hints, but then you have to add comments to explain what the code does.



3 ANOVA (4 points)

a)

Load the diamond dataset in R (`data(diamonds)). With \verb`price' as the dependent variable, run 1) one-way ANOVA on `cut`; 2) two-way ANOVA on `cut` and `clarity` and their interaction.

Interpret the table (i.e. which factor is important in determining the diamond's price?)

Solution

```
data("diamonds")
summary(aov(price ~ cut, data = diamonds))

##                Df    Sum Sq   Mean Sq F value Pr(>F)
## cut             4 1.104e+10 2.760e+09   175.7 <2e-16 ***
## Residuals   53935 8.474e+11 1.571e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

`Cut` is a significant factor, judging from the large F-stat and small p-value

```
summary(aov(price ~ cut + clarity + cut:clarity, data = diamonds))

##                Df    Sum Sq   Mean Sq F value Pr(>F)
## cut             4 1.104e+10 2.760e+09 180.157 <2e-16 ***
## clarity         7 1.891e+10 2.701e+09 176.305 <2e-16 ***
## cut:clarity    28 2.647e+09 9.452e+07   6.169 <2e-16 ***
## Residuals   53900 8.259e+11 1.532e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cut, clarity, and their interaction are all statistically significant factors

b)

What is the expected price for a diamond that has:

- Ideal cut, VS1 clarity
- Very Good cut, VVS1 clarity

Note: You'll have to do dummy regression like last homework, not ANOVA. Also you need to remove the order from the factors `cut` and `clarity`. See this SO answer. Remember to use `\Sexpr{}`, don't write down answers by hand (see how I use `\Sexpr{}` from last solution if confused.)

Solution

```
# Change variables from ordered factor to unordered factor
diamonds <- diamonds %>%
  mutate(cut = factor(cut, ordered = FALSE),
         clarity = factor(clarity, ordered = FALSE))

# Store the model
m3 <- lm(price ~ cut + clarity + cut:clarity, data = diamonds)

# Store the model coefficients
coef <- m3$coefficients
```

Expected price for a diamond that has:

- Ideal cut, VS1 clarity: 3489.7444971
- Very Good cut, VVS1 clarity: 2459.4410646