Lab 14 - Bivariate Regression & Interpretation

Your name here

November 28, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 14 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select the main focal relationship you're interested in exploring for your poster project.

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

The response variable is the RGDP for 4 country, and the explanatory variable is TOTALCAS, which is the sum of nkill & nwound. I beleive that their exists a positive relationship between TOTALCAS and RGDP.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
library(ggplot2)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(readxl)


gtd_filter <- c("country_txt", "nkill", "iyear", "nwound")
gtd_full <- read_csv("GTD FULL DB.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   eventid = col_double(),
##   iyear = col_integer(),
##   imonth = col_integer(),
##   iday = col_integer(),
##   extended = col_integer(),
##   country = col_integer(),
##   region = col_integer(),
##   latitude = col_double(),
##   longitude = col_double(),
##   specificity = col_integer(),
##   vicinity = col_integer(),
##   crit1 = col_integer(),
##   crit2 = col_integer(),
```

```
##   crit3 = col_integer(),
##   doubtterr = col_integer(),
##   alternative = col_integer(),
##   multiple = col_integer(),
##   success = col_integer(),
##   suicide = col_integer(),
##   attacktype1 = col_integer()
##   # ... with 44 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 1)
```

```
## Warning: 246 parsing failures.
## row # A tibble: 5 x 5 col     row       col               expected actual              file expected
## ... ................. ... ..................................................................... ........
## See problems(...) for more details.
```

```
IMF_Real_GDP_Growth <- read_excel("IMF Real GDP Growth.xls")


#Set the ranges for the data using filter
plotTerror <- filter(gtd_full, country_txt == c("Nigeria", "Niger", "Chad", "Cameroon"))
```

```
## Warning in country_txt == c("Nigeria", "Niger", "Chad", "Cameroon"): longer
## object length is not a multiple of shorter object length
```

```
plotTerror <- select(plotTerror,gtd_filter)

#Fix the GDP growth data
temp <- gather(IMF_Real_GDP_Growth, key = Year, value = RGDP, ... = -X__1)

#Rename columns to setup keys for join
names(temp)[names(temp) == 'X__1'] <- 'Countries'
names(plotTerror)[names(plotTerror) == 'country_txt'] <- 'Countries'
names(plotTerror)[names(plotTerror) == 'iyear'] <- "Year"

#Join the two datasets

joinT <- merge(plotTerror, temp, by = c("Countries", "Year"))

#Clean joinT data
#Get rid of NA's
#joinT <- joinT[complete.cases(joinT),]
#Create new catagory - totalcas
joinT[is.na(joinT)] <- 0
joinT <- mutate(joinT, totalcas = (nkill + nwound))

test <- joinT %>%
  select( -nwound) %>%
  group_by(Countries, Year) %>%
  mutate(TOTALCAS = sum(totalcas)) %>%
  select(-nkill, -totalcas) %>%
  distinct() %>%
  ungroup() %>%
```

```
  mutate(Year = as.character(Year)) %>%
  right_join(temp, by = c("Countries", "Year"))

test[is.na(test)] <- 0
test <- select(test, -RGDP.x)
test <- arrange(test, Countries, Year)
test$Year <- as.numeric(as.character(test$Year))

testLM <- lm(RGDP.y ~ TOTALCAS, data = test)
summary(testLM)
```

```
##
## Call:
## lm(formula = RGDP.y ~ TOTALCAS, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8793  -2.3793  -0.1793   1.6684  29.3207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.279e+00  4.280e-01   9.998   <2e-16 ***
## TOTALCAS    4.744e-05  1.206e-03   0.039    0.969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.737 on 126 degrees of freedom
## Multiple R-squared:  1.227e-05,  Adjusted R-squared:  -0.007924
## F-statistic: 0.001546 on 1 and 126 DF,  p-value: 0.9687
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

The direction is "positive", the magnitude is miniscule, and the p value is .96.

d. What is the meaning of the model intercept?

In this case, it doesn't really mean much.

e. How well does the bivariate model fit the data? How is this information calculated?

Not well at all, R^2 was -.007. (R^2 is how we check for fittament)

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

No, there is not. RGDP is simply based on entirely too many factors. For 3 of the countries the model is negative, and one is positive. Nigeria is a country of 200 million people and who's economy is based almost entirely on oil production.

2. Select a different focal relationship related to your project. This could be:

- **A different response and a different explanatory variable**
- **A different response and the same explanatory variable**
- **The same response and a different explanatory variable**

```
bymonthtest <- filter(gtd_full, imonth > 0)
bymonth <- lm(nkill ~ imonth, data = bymonthtest)
```

```r
summary(bymonth)
```

```
##
## Call:
## lm(formula = nkill ~ imonth, data = bymonthtest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##   -2.47   -2.39   -2.31   -0.42 1497.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.295587   0.060994  37.636   <2e-16 ***
## imonth      0.014150   0.008334   1.698   0.0895 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.33 on 160657 degrees of freedom
##   (9671 observations deleted due to missingness)
## Multiple R-squared:  1.794e-05,  Adjusted R-squared:  1.172e-05
## F-statistic: 2.883 on 1 and 160657 DF,  p-value: 0.08953
```

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

The response variable is the number of people killed in attack and the explanatory variable is the month.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

There seems to be a positive, very small, statistically significant association.

d. What is the meaning of the model intercept?

It has no meaning in this datset.

e. How well does the bivariate model fit the data? How is this information calculated?

It doesn't fit very well at all.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

I suppose I think that there will be more deaths in the summer than in winter, but a regression like this doesn't really show us much about that.