Lab 2

Y. Samuel Wang

2/7/2022

Intro

This lab will explore variable transformations and multiple linear regression.

Variable transformations

The World Bank provides valuable data on a number of public health and economic indicators for countries across the globe[1]. Today, we will be looking indicators which might predict infant mortality, which is the number of children (per 1000 births) who die before the age of 1.

Questions

- What factors do you think might affect or correlate with infant mortality?

In particular, we will be looking at 2 specific factors which might correlate well with infant mortality (measured in 2015) - GDP per capita (roughly how much income does the average individual produce) as measured in 2013 and the proportion of the population with access to electricity (as measured in 2012). I have removed countries which were missing data for any of the variables.

```
library("readr")
fileName <- "https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab2/world_bank_data.csv"
wb.data <- read.csv(url(fileName))
head(wb.data)
```

```
##                country  elec_acc inf_mort gdp_capita
## 1              Andorra 100.00000      2.1 42806.5226
## 2          Afghanistan  43.00000     66.3   666.7951
## 3               Angola  37.00000     96.0  5900.5296
## 4              Albania 100.00000     12.5  4411.2582
## 5 United Arab Emirates  97.69783      5.9 42831.0891
## 6            Argentina  99.80000     11.1 14443.0657
```

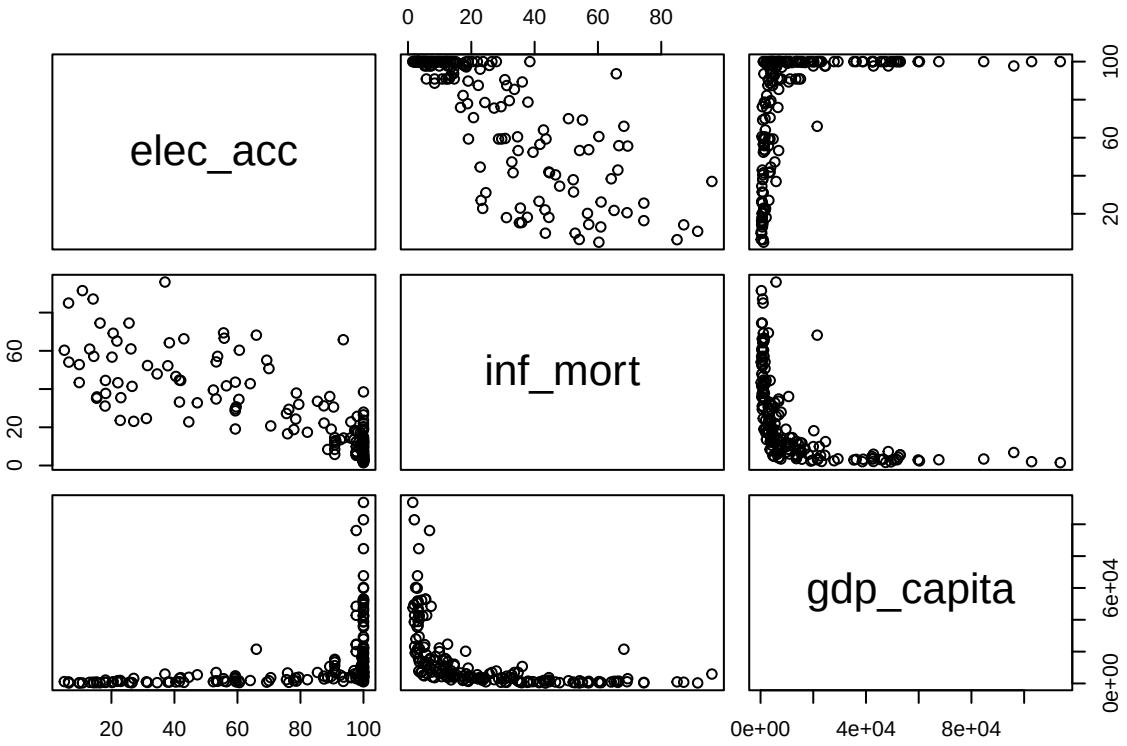
Questions

- What direction do you think the association is between each of these variables?
- What strength do you think the association is between each of these variables?

We can use the `pairs` command to plot the many pairs of variables at once. Note that we've excluded the first column here, since that's just the name of countries

```
pairs(wb.data[, -1])
```

[1] You can access the data at http://data.worldbank.org/



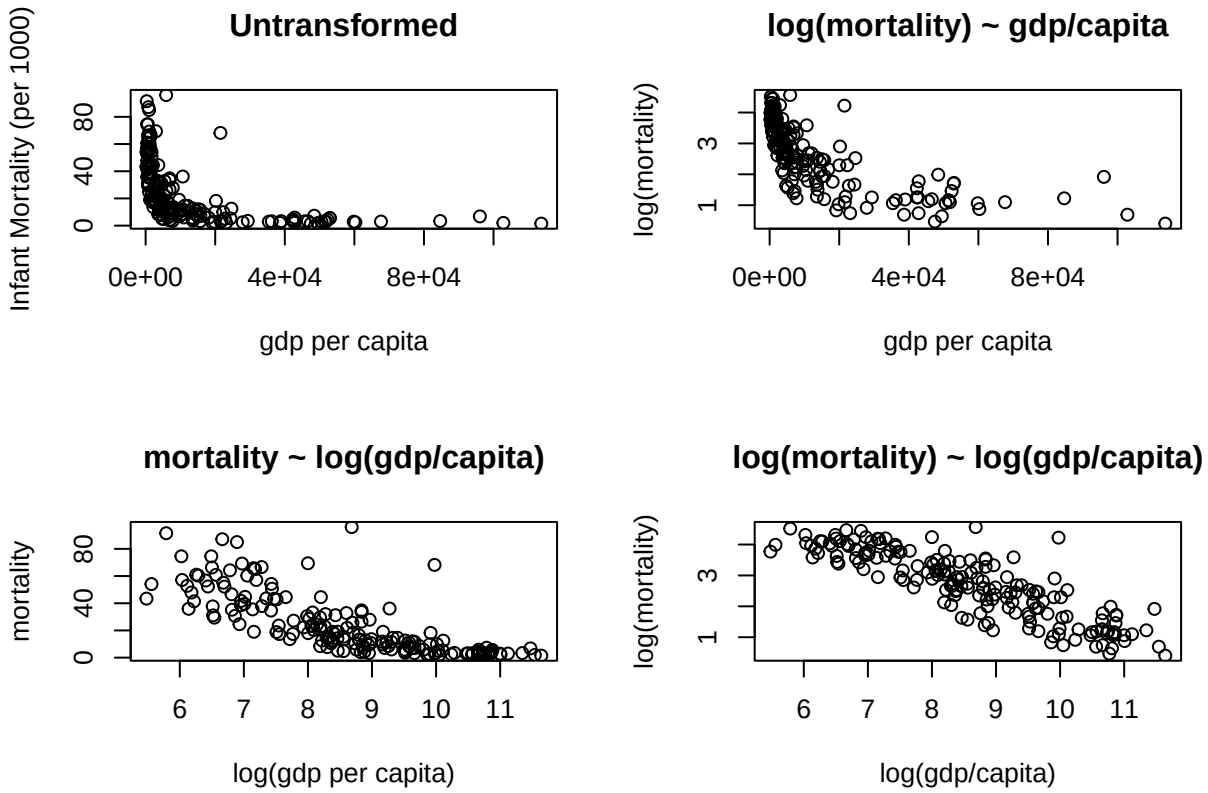
Questions

- Does this look like what you might expect?
- What sticks out?
- Do the relationships look linear?

The relationship between electricity and infant mortality looks roughly linear, but the relationship between GDP per capita and infant mortality does not. Let's see how we might transform the data. The `log` function by default returns the natural log (base e). Let's plot a few transformations and see what makes the relationship linear.

```
# using the par(mfrow = c(r, c)) puts multiple
# plots together. The plots are arranged so
# that there are r rows and c columns

par(mfrow = c(2,2))

# first argument is the X variable, second argument is the Y variable
# main specifies the title, xlab specifies the x axis label
# and ylab specifies the y axis label
plot(wb.data$gdp_capita, wb.data$inf_mort, main = "Untransformed",
     xlab = "gdp per capita", ylab = "Infant Mortality (per 1000)")

plot(wb.data$gdp_capita, log(wb.data$inf_mort),
     main = "log(mortality) ~ gdp/capita",
     xlab = "gdp per capita", ylab = "log(mortality)")

plot(log(wb.data$gdp_capita), wb.data$inf_mort,
     main = "mortality ~ log(gdp/capita)",
     xlab = "log(gdp per capita)", ylab = "mortality")
```

```
plot(log(wb.data$gdp_capita), log(wb.data$inf_mort),
     main = "log(mortality) ~ log(gdp/capita)",
     xlab = "log(gdp/capita)", ylab = "log(mortality)")
```



The plots correspond to the models:

$$E(\text{mortality} \mid \text{gdp/capita}) = b_0 + b_1 \text{gdp/capita}$$

$$E(\log(\text{mortality}) \mid \text{gdp/capita}) = b_0 + b_1 \text{gdp/capita}$$

$$E(\text{mortality} \mid \text{gdp/capita}) = b_0 + b_1 \log(\text{gdp/capita})$$

$$E(\log(\text{mortality}) \mid \text{gdp/capita}) = b_0 + b_1 \log(\text{gdp/capita})$$

Questions

- Which transformation looks most linear?
- How do we interpret the b_1 parameter in each model?

The transformation that looks most linear requires taking the log of both mortality and gdp per capita. We can estimate the transformed and untransformed models now using the `lm` command.

```
# Untransformed data
untransformed.reg <- lm(inf_mort ~ gdp_capita, data = wb.data)

summary(untransformed.reg)
```

```
##
## Call:
## lm(formula = inf_mort ~ gdp_capita, data = wb.data)
```

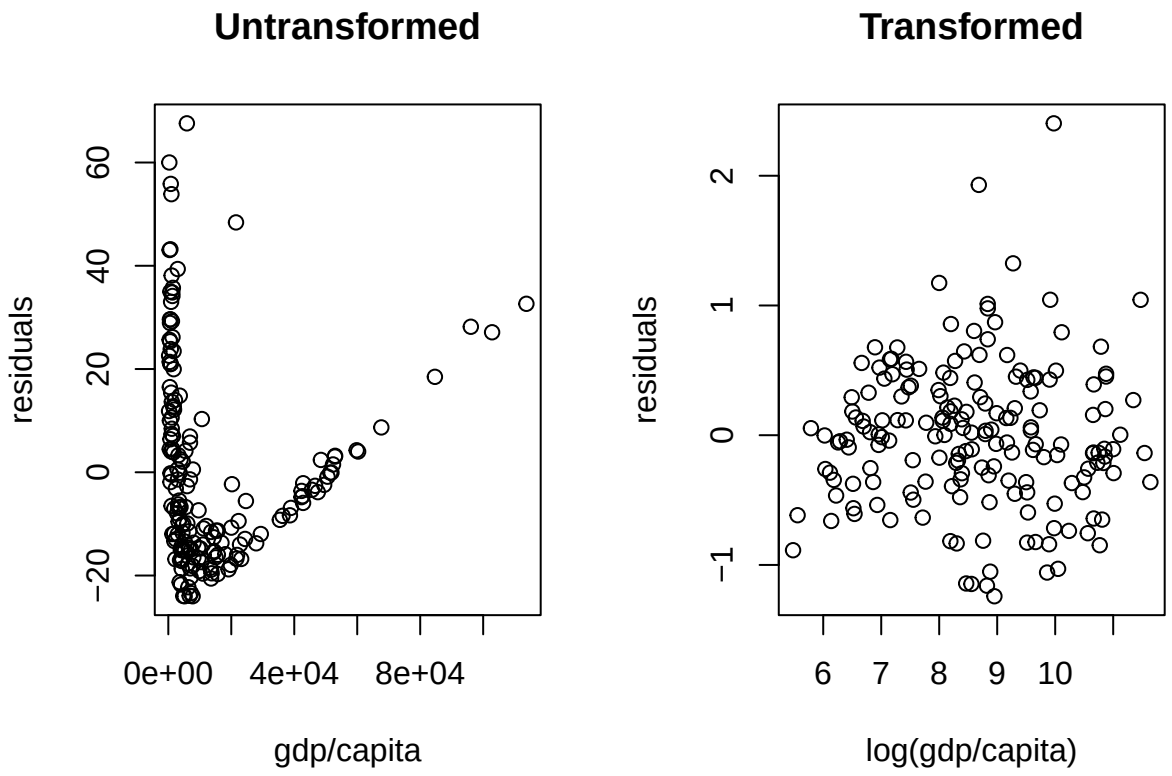
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.011 -14.633  -5.749   8.625  67.583
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.168e+01  1.743e+00  18.171  < 2e-16 ***
## gdp_capita  -5.523e-04  7.093e-05  -7.787 5.68e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.07 on 176 degrees of freedom
## Multiple R-squared:  0.2562, Adjusted R-squared:  0.252
## F-statistic: 60.63 on 1 and 176 DF,  p-value: 5.678e-13
```

```
# regression with transformed data
transformed.reg <- lm(log(inf_mort) ~ log(gdp_capita), data = wb.data)

summary(transformed.reg)
```

```
##
## Call:
## lm(formula = log(inf_mort) ~ log(gdp_capita), data = wb.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24132 -0.34865 -0.00525  0.34525  2.40377
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.11682    0.24882   32.62   <2e-16 ***
## log(gdp_capita) -0.63135    0.02848  -22.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5554 on 176 degrees of freedom
## Multiple R-squared:  0.7363, Adjusted R-squared:  0.7348
## F-statistic: 491.3 on 1 and 176 DF,  p-value: < 2.2e-16
```

We can also look at the residuals plotted against GDP/Capita for both models. What does this suggest about how each model fits our data?

```
par(mfrow = c(1,2))
plot(wb.data$gdp_capita, untransformed.reg$residuals, main = "Untransformed",
     xlab = "gdp/capita", ylab = "residuals")
plot(log(wb.data$gdp_capita), transformed.reg$residuals, main = "Transformed",
     xlab = "log(gdp/capita)", ylab = "residuals")
```



Questions

- Compare the r^2 from both regressions. What does this suggest about which explanatory variable is a better predictor of infant mortality?
- Why do you think this is true?
- Note that we aren't exactly comparing apples to apples here because one regression has log(mortality) as the response while the other uses mortality untransformed. Is there a way you could make the comparison more fair?
- Which model would you use if you are trying to predict infant mortality for a country not in the data set? Which model would you use if you are trying to explain to a collaborator? Which model would you use if you are trying to test if infant mortality is associated with gdp/capita?

Multiple Linear Regression

The rest of today's lab will have less instruction, so it is on you, as a budding statistician to provide a bit of creativity and apply what we have learned so far. In addition, we will use this data set for the module 2 assessment.

We will be looking at recent data from the UK Brexit vote. If you, like many Britons[2], aren't familiar with what the European Union is, you can read more about the whole story here http://www.vox.com/2016/6/17/11963668/brexit-uk-eu-explained.

In particular, the response variable we will be using is the percentage of individuals who voted to remain in the European Union in each local authority. We will be looking at several explanatory variables including

- Percentage of individuals born in the UK
- Percentage of individuals with no formal education beyond compulsory education
- Percentage of individuals working in manufacturing
- Percentage of individuals working in finance
- Percentage of individuals over the age of 60
- Percentage of individuals between the ages of 20 and 35

Each row in the data represents a local authority/distict in either England or Wales. The Brexit vote took place in 2016, and the explanatory variables were collected in the 2011 census. Local Authorities with missing data have been removed.

```
fileName <- "https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lab2/uk_data.csv"
brexit.data <- read.csv(url(fileName))
head(brexit.data)
```

```
##              geography   uk_born    no_edu       mfct    finance   over_60
## 1           Darlington 0.9475295 0.2481226 0.09997144 0.03835639 0.2263366
## 2        County Durham 0.9675338 0.2750048 0.13156555 0.02221647 0.2360407
## 3           Hartlepool 0.9721932 0.3065959 0.11676861 0.02089125 0.2211066
## 4        Middlesbrough 0.9178539 0.2989068 0.08121437 0.02489596 0.1934081
## 5       Northumberland 0.9717588 0.2387226 0.09236833 0.02368942 0.2635178
## 6 Redcar and Cleveland 0.9776663 0.2842061 0.10318700 0.01957270 0.2520029
##   over_20_less_than35 pct_remain
## 1           0.1926604     0.4382
## 2           0.1937371     0.4245
## 3           0.1911049     0.3043
## 4           0.2263821     0.3452
## 5           0.1636817     0.4589
## 6           0.1780406     0.3381
```

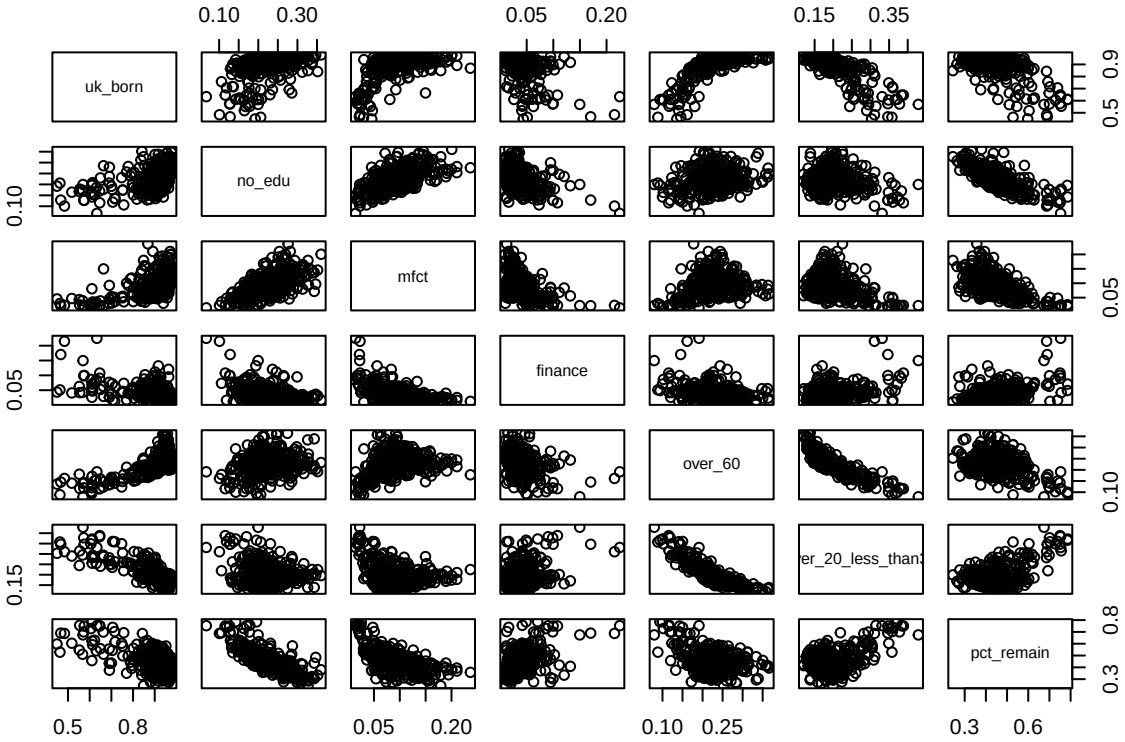
Questions

- What direction do you think the association is between each of these variables?
- What strength do you think the association is between each of these variables?

Again, we'll use the `pairs` command to plot the many pairs of variables at once. Note that we've excluded the first column here, since that's just the name of local authority

```
pairs(brexit.data[, -1])
```

[2]Google searches for "What is the EU" spiked in the UK. Unfortunately, the spike occured after the vote had occured. {http://www.npr.org/sections/alltechconsidered/2016/06/24/480949383/britains-google-searches-for-what-is-the-eu-spike-after-brexit-vote



Questions

- Does this look like what you might expect?
- What sticks out?
- Do the relationships look roughly linear?

Multivariate Regression

When there are multiple variables, we still use the regular `lm` command, but we need to specify more variables in our formula. Notice now on the right hand side of the $\sim$, we have multiple variables which are separated by the + sign. We can add additional variables simply by using the + sign.

```
output <- lm(pct_remain ~ uk_born + no_edu, data = brexit.data)
summary(output)
```

```
##
## Call:
## lm(formula = pct_remain ~ uk_born + no_edu, data = brexit.data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.132640 -0.035044 -0.005769  0.030399  0.206090
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.01641    0.02606   39.00   <2e-16 ***
## uk_born     -0.32934    0.03220  -10.23   <2e-16 ***
## no_edu      -1.19710    0.06604  -18.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.05556 on 341 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.6799
## F-statistic: 365.3 on 2 and 341 DF,  p-value: < 2.2e-16
```

We can see from the summary of our model that the estimated model is

$$\hat{y}_i = \hat{b}_0 + \hat{b}_{\text{uk born}} x_{i,\text{uk born}} + \hat{b}_{\text{no edu}} x_{i,\text{no edu}}$$

where $b_{uk\_born} = -.33$ and $b_{no\_edu} = -1.20$.

We can get the residuals and fitted values from the `lm` objects, and we can look at the values for specific geographic areas. For instance, "Eden" is the 23 row in the list. We can see that by using the `which` function. The function returns the index for which the statement evalues to "TRUE." This means the 23rd element of geography vector is equal to "Eden." In the residual and fitted values vector, the 23rd element corresponds to the values for "Eden"

```
which(brexit.data$geography == "Eden")
```

```
## [1] 23
```

```
output$residuals[23]
```

```
##          23
## 0.03981415
```

```
output$fitted.values[23]
```

```
##        23
## 0.4269859
```

Questions

- How would you interpret each of the estimated coefficients above?
- Does the magnitude (size) of the coefficients agree with what you would've guessed?

Now is your chance to explore the data yourself. Using the form above, fit a regression and include variables which you think might be associated with the percentage of people voting to remain in the EU. As you fit your models, check to make sure that the associations are roughly linear, and take a log transformation if necessary.

Try fitting multiple models (at least 3 or 4) and think about what makes sense to investigate and what variables might need transformations.

Questions

- Look at the r^2 value for each model. As you include more variables, what happens to the r^2 value? Does this always happen?
- When you include more variables, how do the regression coefficients change for the existing variables?

After you are done, discuss your findings with your neighbor and pat yourself on the back. Congratulations, you're on your way to being a statistician!

Questions

Questions to discuss with your neighbor.

- How did you decide which variables to include and which variables not to include?
- What is the proper interpretation of your regression coefficients?
- What are the signs of each of the coefficients?
- What are the relative sizes of the coefficients?

- Does this make sense with what we know about the world?
- What would we need to be careful about in interpreting these models?
- What other variables (that weren't available) would also be good to include?