Lab 2

Y. Samuel Wang

2/4/2022

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
```

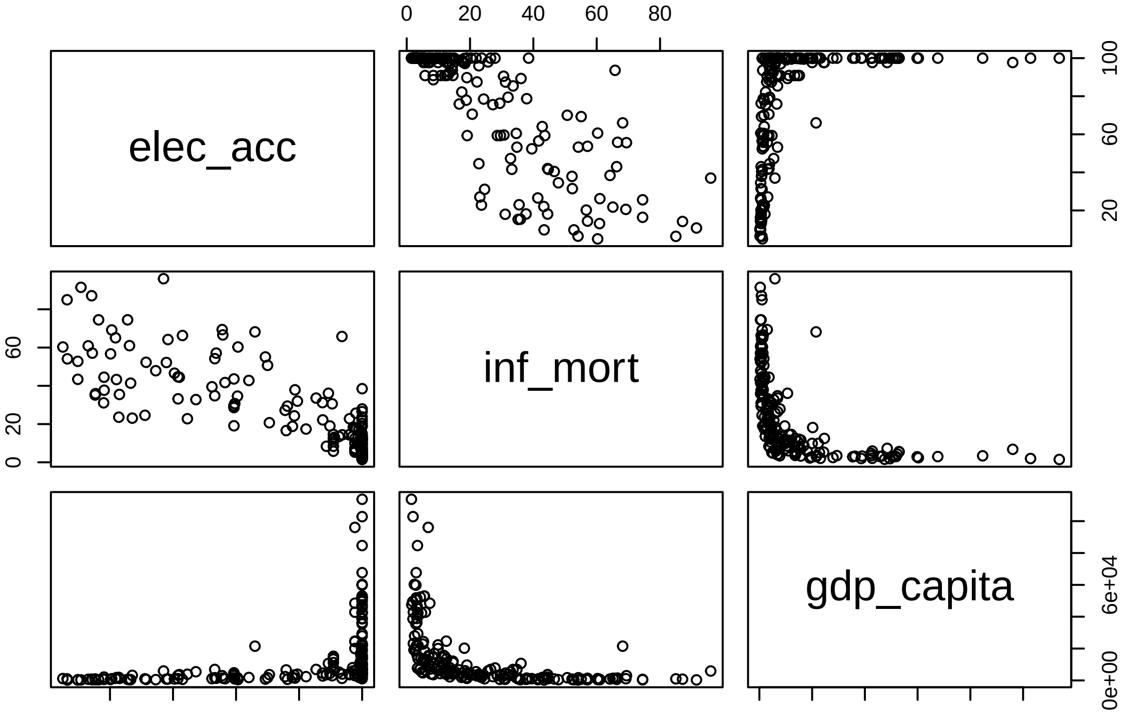
Questions

- 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/



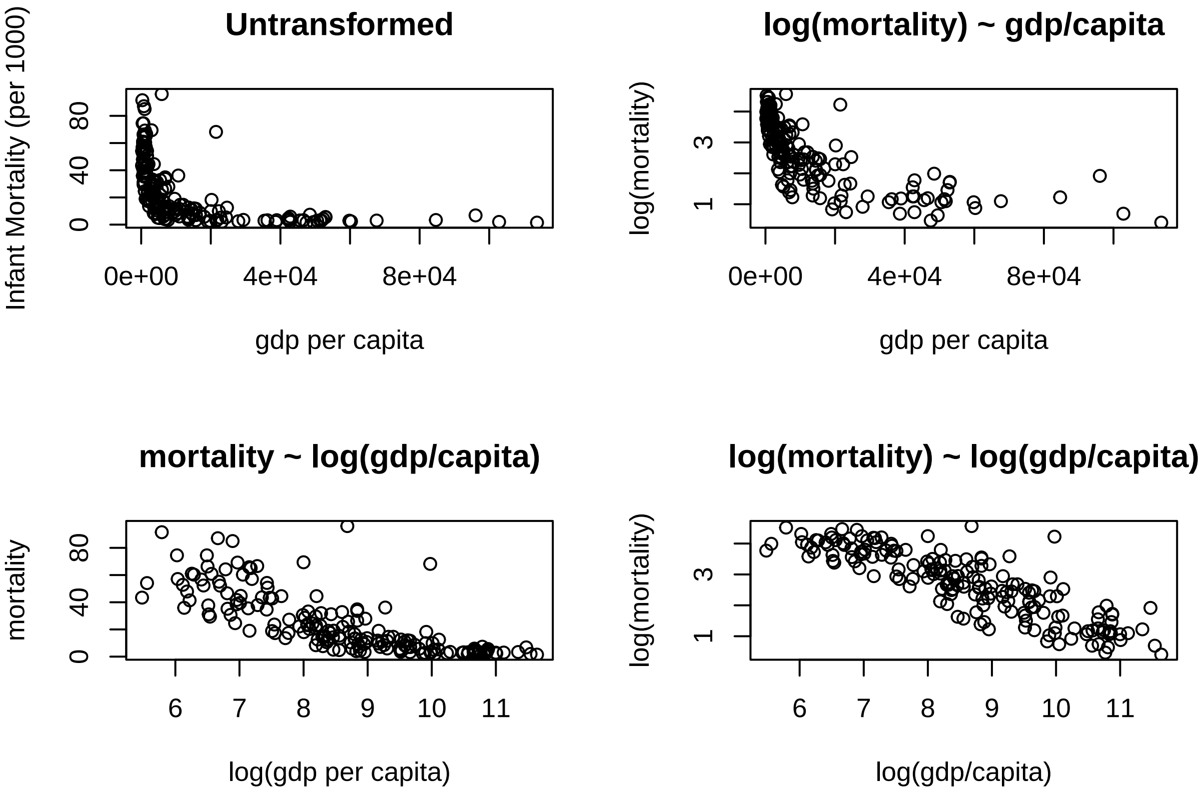
Questions

- 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")
```

```r
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:

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

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

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.

```r
# 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 calculate the SS_{total} $\text{SS}_{regression}$ and SS_{errors} for the regression with untransformed data Using these quantities, we can calculate the R^2 value.

```
ss.total.untransformed <- sum((wb.data$inf_mort - mean(wb.data$inf_mort))^2)

# Get the estimated coefficients from the regression
# elec.reg$coeff gets a vector the regression coefficients
# The first element is the y intercept, and the second element is the slope

# Calculate SS_regression
ss.regression.untransformed <- sum((untransformed.reg$fitted.values - mean(wb.data$inf_mort))^2)

# Calculate SS_error
ss.error.untransformed <- sum((untransformed.reg$fitted.values - wb.data$inf_mort)^2)

# Check that ss.regression + ss.error = ss.total
```

```r
ss.regression.untransformed + ss.error.untransformed
```

```
## [1] 86092.8
```

```r
ss.total.untransformed
```

```
## [1] 86092.8
```

```r
# r^2 the long way
ss.regression.untransformed / ss.total.untransformed
```

```
## [1] 0.2562274
```

```r
# r^2 the short way
cor(wb.data$inf_mort, wb.data$gdp_capita)^2
```

```
## [1] 0.2562274
```

We can also calculate the same quantities for the transformed model

```r
ss.total.transformed <- sum((log(wb.data$inf_mort) - mean(log(wb.data$inf_mort)))^2)

# Get the estimated coefficients from the regression
# elec.reg$coeff gets a vector the regression coefficients
# The first element is the y intercept, and the second element is the slope

# Calculate SS_regression
ss.regression.transformed <- sum((transformed.reg$fitted.values - mean(log(wb.data$inf_mort)))^2)

# Calculate SS_error
ss.error.transformed <- sum((transformed.reg$fitted.values - log(wb.data$inf_mort))^2)

# Check that ss.regression + ss.error = ss.total
ss.regression.transformed + ss.error.transformed
```

```
## [1] 205.8664
```

```r
ss.total.transformed
```

```
## [1] 205.8664
```

```r
# r^2 the long way
ss.regression.transformed / ss.total.transformed
```

```
## [1] 0.7362674
```

```r
# r^2 the short way
cor(log(wb.data$inf_mort), log(wb.data$gdp_capita))^2
```

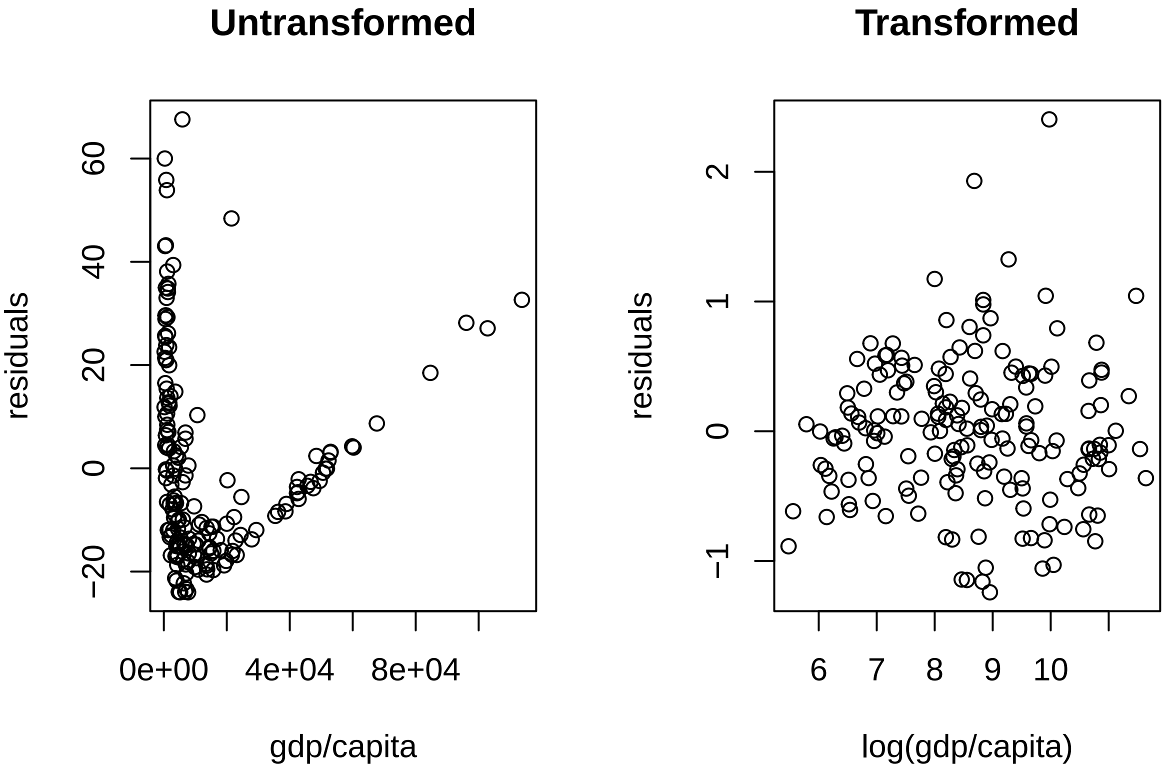
```
## [1] 0.7362674
```

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.

We can also look at the residuals plotted against GDP/Capita for both models

```
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")
```


Untransformed
residuals
0e+00
4e+04
8e+04
gdp/capita
Transformed
residuals
log(gdp/capita)