Lab 2 (cont): Regressing Further

July 1, 2016

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.

```
# To get off website
wb.data <- read.csv("http://www.stat.washington.edu/~ysamwang/notes/world_bank_data.csv")

# If you have downloaded it locally
wb.data <- read.csv("world_bank_data.csv")
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
```

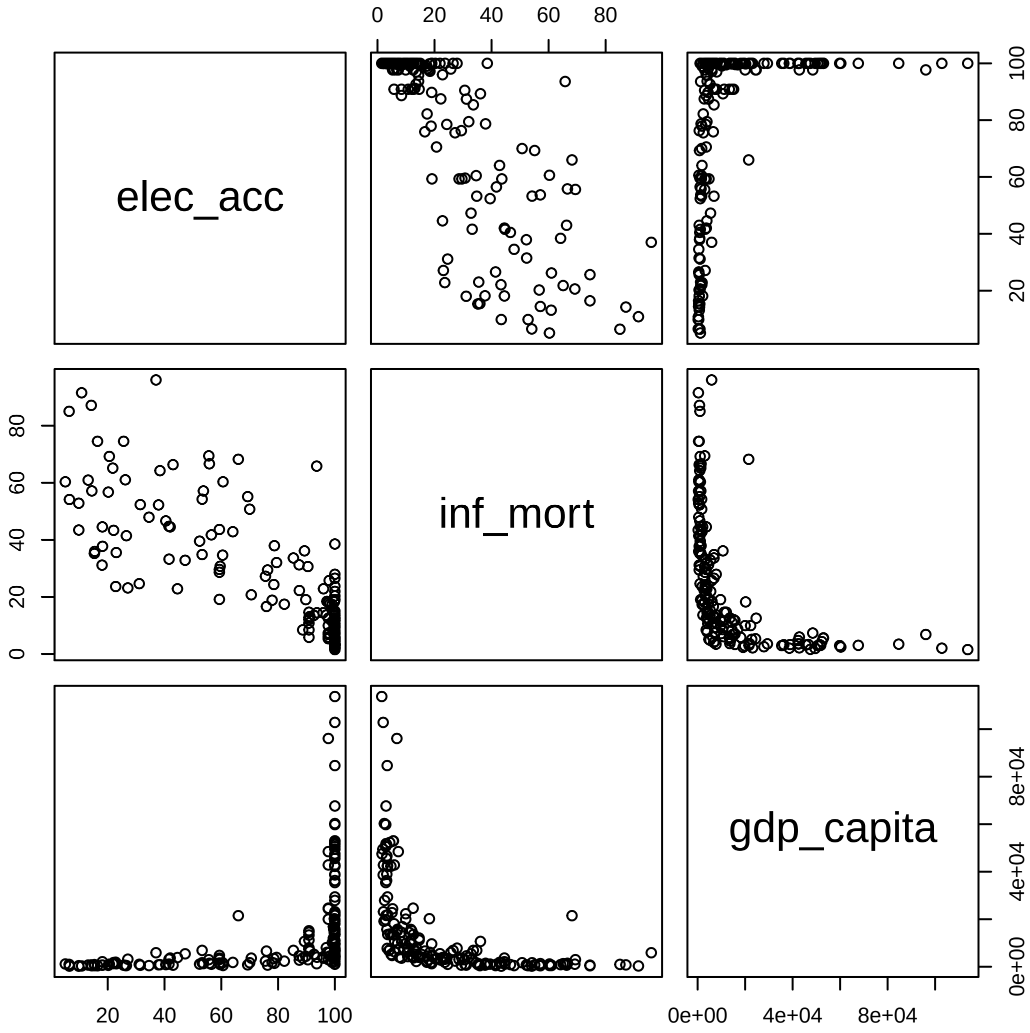
Questions

- 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/



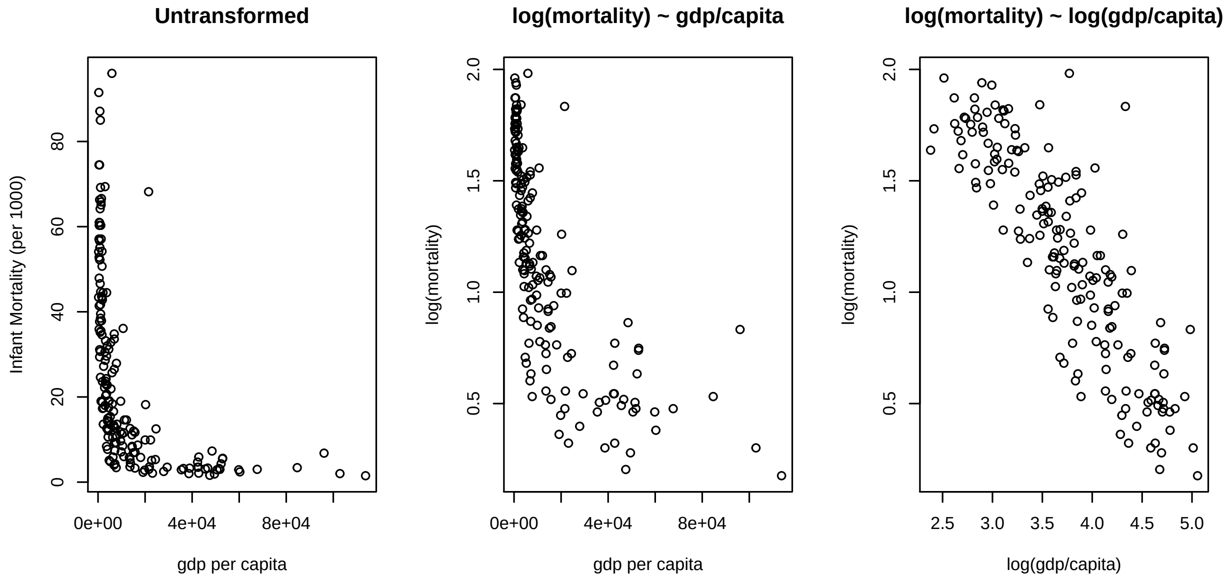
Questions

- 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), but we can specify the base as a second argument. So to take $\log_{10}(200)$ we would use `log(200, 10)`. 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(1,3))

# 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,10),
     main = "log(mortality) ~ gdp/capita",
     xlab = "gdp per capita", ylab = "log(mortality)")

plot(log(wb.data$gdp_capita, 10), log(wb.data$inf_mort, 10),
     main = "log(mortality) ~ log(gdp/capita)",
     xlab = "log(gdp/capita)", ylab = "log(mortality)")
```



Questions

- Which transformation looks most linear?

The transformation that looks most linear requires taking the log of both mortality and gdp per capita. This corresponds to a model of

$$\log(\text{mortality}) = a + b \log(\text{gdp/capita}) + \epsilon \qquad (1)$$

Since we've transformed both variables (not just the y variable), the interpretation changes slightly. We can interpret the slope parameter b as follows- "Every percentage increase in GDP/Capita, is associated a b% increase/decrease in infant mortality." Notice now, that we are talking about percentage changes in the variables of which we have taken the log.

Let's estimate the two models now using the `lm` command.

```
# regression with electricity as X variable
# note that since we specify the data frame, we can directly use
# the variable names in the formulat
# alternatively, we could have used
# lm(wb.data£inf_mort ~ wb.data£elec_acc)
elec.reg <- lm(inf_mort ~ elec_acc, data = wb.data)

summary(elec.reg)

##
## Call:
## lm(formula = inf_mort ~ elec_acc, data = wb.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.174  -7.488  -2.608   5.109  51.260
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.05763    2.61897   26.37   <2e-16 ***
## elec_acc    -0.58245    0.03141  -18.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.87 on 176 degrees of freedom
## Multiple R-squared:  0.6614,Adjusted R-squared:  0.6595
## F-statistic: 343.8 on 1 and 176 DF,  p-value: < 2.2e-16

# regression with log(gdp_capita) as X variable
gdp.reg <- lm(log(inf_mort) ~ log(gdp_capita), data = wb.data)

summary(gdp.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 electricity accesss. Using these quantities, we can calculate the r^2 value.

```
ss.total <- 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
a.hat <- elec.reg$coeff[1]
b.hat <- elec.reg$coeff[2]
inf_mort.hat <- a.hat + b.hat * wb.data$elec_acc

# Calculate SS_regression
ss.regression <- sum((inf_mort.hat - mean(wb.data$inf_mort))^2)

# Calculate SS_error
ss.error <- sum((inf_mort.hat - wb.data$inf_mort)^2)

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

## [1] 86092.8

ss.total

## [1] 86092.8

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

## [1] 0.6614182

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

## [1] 0.6614182
```

If you look back to the summary of the regression above, you'll notice that this is the value reported as multiple R squared. Note, that we can also get the predicted values and residuals directly from the regression object

```
# residuals
resid <- elec.reg$residuals

# Predicted values (also called fitted values)
predicted <- elec.reg$fitted.values
```

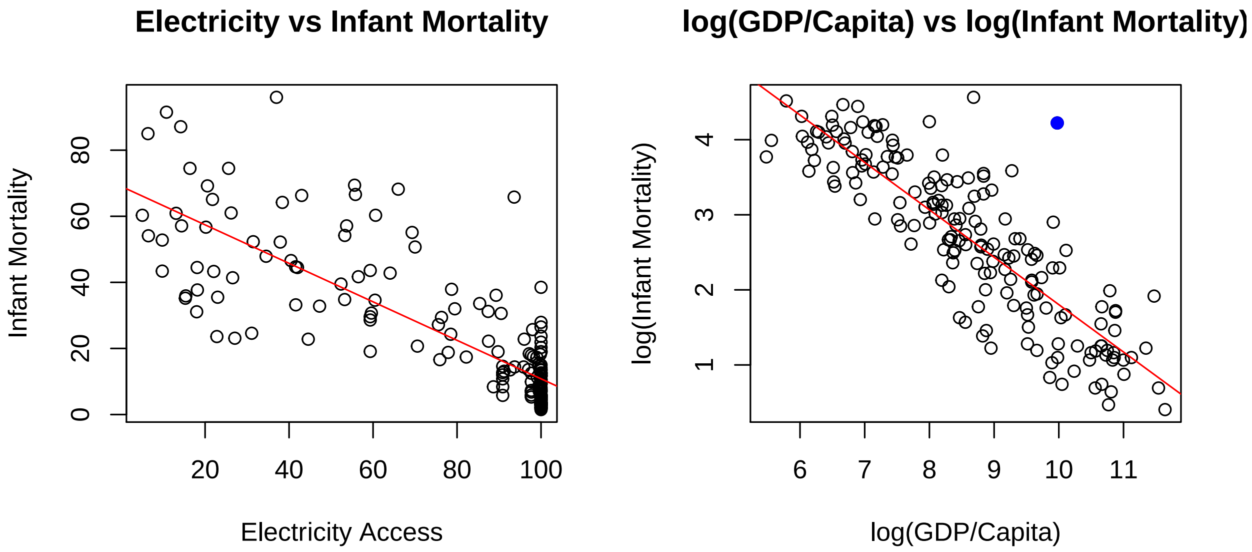
Questions

- Compare the r^2 from the gdp regression to the r^2 of the electricity regression. 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.

Let's plot the data and the best fit line for each explanatory variable. For the gdp plot, we have highlighted Equatorial Guinea in blue which has a much higher infant mortality rate than we would expect based on their GDP per capita.

```
par(mfrow = c(1,2))
plot(wb.data$elec_acc, wb.data$inf_mort,
     main = "Electricity vs Infant Mortality",
     xlab = "Electricity Access",
     ylab = "Infant Mortality")
```

```r
abline(a = elec.reg$coeff[1], b = elec.reg$coeff[2], col = "red")

plot(log(wb.data$gdp_capita), log(wb.data$inf_mort),
     main = "log(GDP/Capita) vs log(Infant Mortality)",
     xlab = "log(GDP/Capita)",
     ylab = "log(Infant Mortality)")
abline(a = gdp.reg$coeff[1], b = gdp.reg$coeff[2], col = "red")

# the points command can plot on existing frames
# col specifies the color, and pch specifies the shape of the mark
points(log(wb.data$gdp_capita)[65], log(wb.data$inf_mort)[65], col = "blue", pch = 19)
```



Questions

- According to Wikipedia, Equatorial Guinea has the highest GDP per capita in Sub-Saharan Africa due to significant oil production. However, that wealth is concentrated in a few elites, and not distributed evenly. Instead of GDP per capita (which is a mean), what might be an even better way to predict infant mortality?

Lab Assignment

Go to `http://guessthecorrelation.com/` and guess at least 20 times (If you loose all your hearts, you may have to play multiple games). Once you have at least 20 guesses, go back to the main menu, click on settings, and then click on download your game data.

From the catalyst site, download the lab_2_assignment.R file and make sure the data from guessthecorrelation.com and the assignment file are saved in the same folder. Then, set your working directory (where R looks for all your files) to the source directory by going to top menu in Rstudio and selecting Session > Set Working Directory > To Source File Location.

Then fill out the lab_2_assignment.R form and turn it into the dropbox on the catalyst site. To save a plotted image for question 2, click on the Export button above your plot, and select Save as Image.

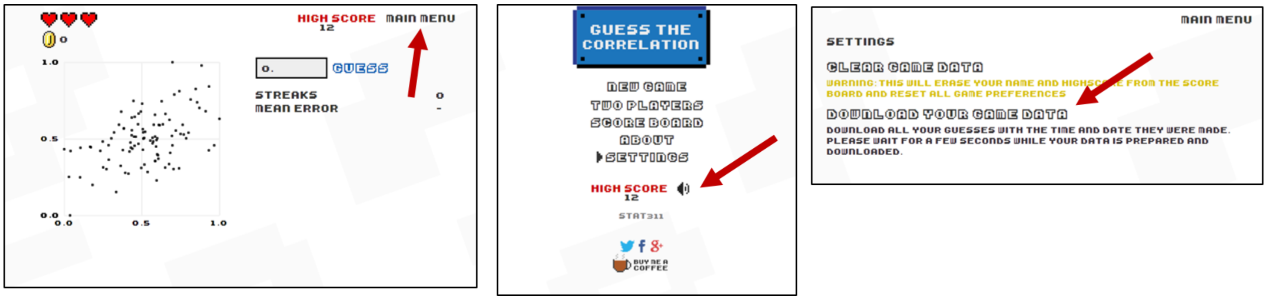
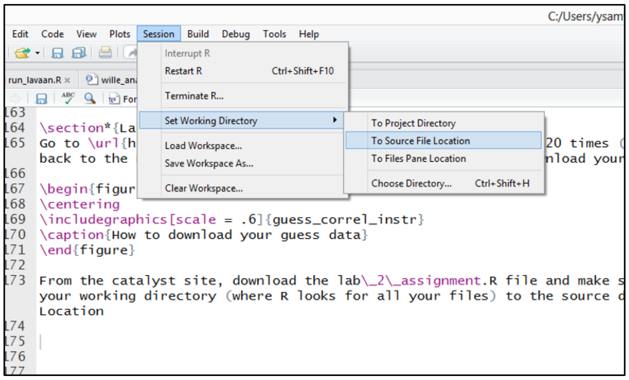
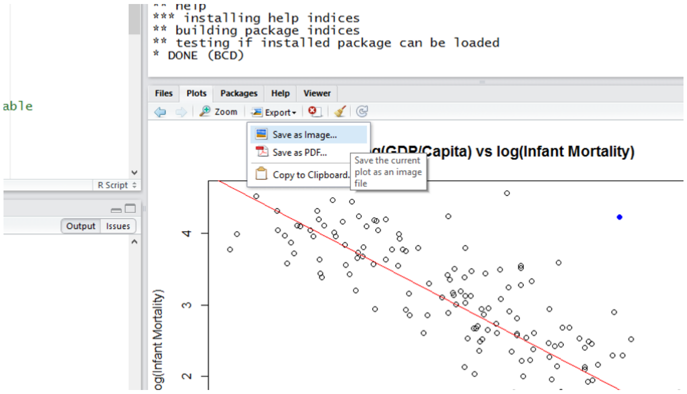

HIGH SCORE 12
MAIN MENU
0
1.0
0.5
0.0
0.0
0.5
1.0
0.
GUESS
STREAKS
MEAN ERROR
0
-
GUESS THE CORRELATION
NEW GAME
TWO PLAYERS
SCORE BOARD
ABOUT
SETTINGS
HIGH SCORE 12
STAT311
BUY ME A COFFEE
MAIN MENU
SETTINGS
CLEAR GAME DATA
WARNING: THIS WILL ERASE YOUR NAME AND HIGHSCORE FROM THE SCORE BOARD AND RESET ALL GAME PREFERENCES
DOWNLOAD YOUR GAME DATA
DOWNLOAD ALL YOUR GUESSES WITH THE TIME AND DATE THEY WERE MADE. PLEASE WAIT FOR A FEW SECONDS WHILE YOUR DATA IS PREPARED AND DOWNLOADED.

Figure 1: How to download your guess data



```

\section*{La
Go to \url{h                                                 20 times (
back to the                                                  nload your

\begin{figur
\centering
\includegraphics[scale = .6]{guess_correl_instr}
\caption{How to download your guess data}
\end{figure}

From the catalyst site, download the lab\_2\_assignment.R file and make s
your working directory (where R looks for all your files) to the source d
Location

```

Figure 2: How to set your working directory



```
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (BCD)
```

Figure 3: How to save your plot