Worksheet 02 - Correllation and Regression Parts 1 and 2

##Shawn Kim

***Directions**: Please upload a PDF to Gradescope that includes both your written responses and corresponding R code inputs/outputs (if requested) for each problem. In code blocks, you can uncomment (delete #) the non-NOTE lines and fill in the blanks (FILL IN).*

Problem 1 Global warming has many indirect effects on climate. For example, summer monsoon winds in the Arabian Sea bring rain to India needed for agriculture. As the climate warms and winter snow cover in Europe and Asia decreases, the land heats up more rapidly in summer which may increase the strength of the monsoon. For this problem we will work with data representing the snow cover (in millions of square kilometers) and windstress (in Newtons per square meter).

```
# Write out data in temporary vectors
snowcover <- c(6.6, 5.9, 6.8, 7.7, 7.9, 7.8, 8.1, 16.6, 18.2, 15.2, 16.2, 17.1, 17.3,
    18.1, 26.6, 27.1, 27.5, 28.4, 28.6, 29.6, 29.4)

windstress <- c(0.125, 0.16, 0.158, 0.155, 0.169, 0.173, 0.196, 0.111, 0.106, 0.143,
    0.153, 0.155, 0.133, 0.13, 0.062, 0.051, 0.068, 0.055, 0.033, 0.029, 0.024)

# Store data in a data frame with column names snow, and wind
climate <- data.frame(snow = snowcover, wind = windstress)

# Remove the temporary variables from the global environment
rm(snowcover, windstress)

# summary(dataframe) gives a summary of each column of the data frame
summary(climate)
```

```
##       snow            wind       
##  Min.   : 5.90   Min.   :0.0240  
##  1st Qu.: 7.90   1st Qu.:0.0620  
##  Median :17.10   Median :0.1300  
##  Mean   :17.46   Mean   :0.1138  
##  3rd Qu.:27.10   3rd Qu.:0.1550  
##  Max.   :29.60   Max.   :0.1960  
```

Problem 1 Part a Based on the description of the problem, Which is the explanatory variable and which is the response variable?

snow cover is the explanatory variable and windstress is the response variable

Problem 1 Part b Give the correlation between snow cover and wind stress.What does the correlation tell us about the association between snow cover and wind stress?

```
# NOTE: Usage for cor is cor(x,y)

# NOTE: To access vectors from your data set, use climate$snow and climate$wind

cor(climate$snow, climate$wind)
```

```
## [1] -0.9179469
```

the correlation value tells us there is a strong inverse relationship between snow cover and windstress, meaning the less snow cover there is, the higher the windstress

Problem 1 Part c Using R, display a scatter plot and determine the equation of the regression line. Plot and label the regression line on the scatter plot. Finally, describe any structure you see within the plot (i.e., what is the form, the direction, and the strength of the scatter plot?).

```
# NOTE: Usage for plot is plot(x, y, xlab = 'x axis label', ylab = 'y axis
# label')

plot(x = climate$snow, y = climate$wind, xlab = "Snow cover", ylab = "Wind stress")

# NOTE: The command lm(formula, data = dataframe) gives a linear model based on
# data. An example formula might be y ~ x. R uses a tilda (~) instead of an
# equals sign for formulas

climate.lm <- lm(wind ~ snow, data = climate)

# NOTE: summary() can also accept a linear model as an input

summary(climate.lm)  # displays intercept and slope of linear model
```

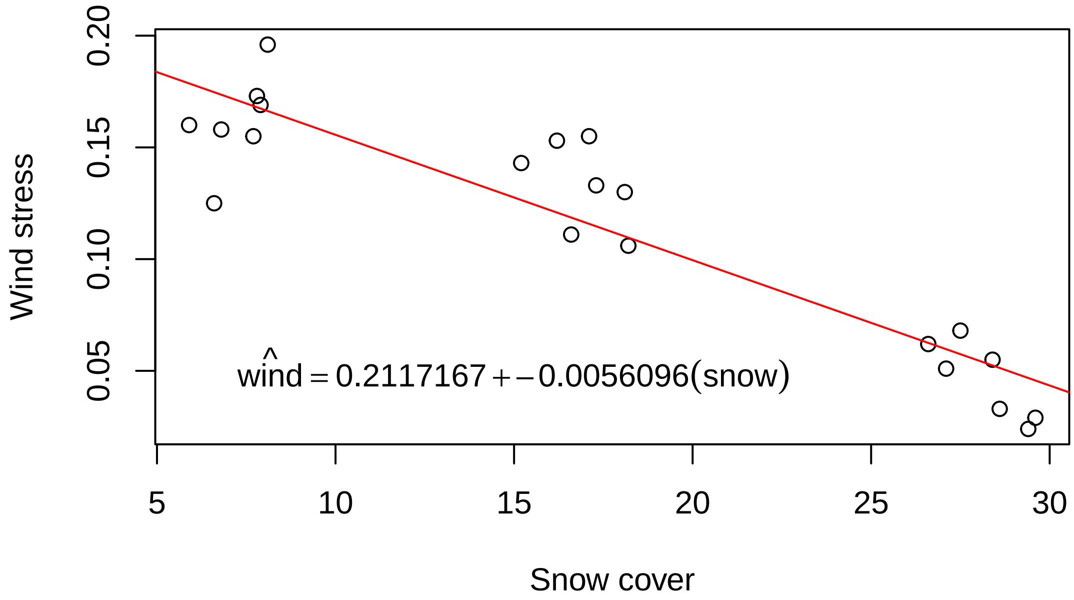
```
##
## Call:
## lm(formula = wind ~ snow, data = climate)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.049693 -0.015571 -0.000501  0.016550  0.039208
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.2117167  0.0108251   19.56 4.78e-14 ***
## snow        -0.0056096  0.0005562  -10.09 4.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02191 on 19 degrees of freedom
## Multiple R-squared:  0.8426, Adjusted R-squared:  0.8343
## F-statistic: 101.7 on 1 and 19 DF,  p-value: 4.582e-09
```

```
# NOTE: abline(a, b) draws a line of the form y = a + bx on your current graph
# NOTE: abline() can also accept a linear model
```

```r
abline(climate.lm, col = "red")  # plot regression line on scatter plot

# NOTE: Next we label the regression line on the scatter plot. Replace FILL IN
# ALPHA and FILL IN BETA with the numerical values from the linear model

par(new = TRUE)
text(15, 0.05, expression(hat(wind) == 0.2117167 + -0.0056096(snow)))
```



the scatter plot shows the data grouped in 3 places, namely at the start, middle, and end of the range for snow cover, the direction of the plot is a negative slope which is consistent with the negative correlation, and the strength of scatter plot is strong since we can clearly see the inverse relationship between snow cover and windstress

Problem 1 Part d What is the predicted wind stress when snow cover is 12 million square kilometers? Complete this calculation in R and by hand.

NOTE: The R code below will help to do the calculation in R.
How to show by-hand calculations? You may 1) add in R code that shows the algebra clearly, 2) type it out as text outside/below of the R chunk, or 3) by inserting an image [See the bottom of this RMD file].

```r
# NOTE: The command predict(linear model, data frame with newdata) takes in two
# arguments. newdata is entered as a data frame since the linear model is
# stored as a data frame

predict(climate.lm, newdata = data.frame(snow = 12))
```

```
##         1
## 0.1444012
```

```r
# manual calculation in R:

# wind = 0.2117167 - 0.0056096(snow)

# wind_12 = 0.2117167 - 0.0056096(12)
0.2117167 - 0.0056096 * (12)
```

```
## [1] 0.1444015
```

Problem 1 Part e The R code below shows how to calculate the residual for each data point. Notice that the first observation has the largest negative residual. Confirm R's calculation of this residual by hand. In addition, interpret this residual value in the context of the problem by explaining what the value represents.

NOTE: Again, include details of your calculation by either by including the algebra inside the chunk as R code or outside/below the chunk as text or by inserting an image.

```
# NOTE: The command resid(linear model) computes the residuals based on the
# linear model.

climate_residual = resid(climate.lm)
climate_residual
```

```
##             1             2             3             4             5
## -0.0496931450 -0.0186198843 -0.0155712195 -0.0135225547  0.0015993708
##             6             7             8             9            10
##  0.0050384080  0.0297212963 -0.0075968695 -0.0036214654  0.0165496520
##            11            12            13            14            15
##  0.0321592795  0.0392079443  0.0183298698  0.0198175719 -0.0005005939
##            16            17            18            19            20
## -0.0086957801  0.0105480709  0.0025967357 -0.0182813388 -0.0166717112
##            21
## -0.0227936367
```

```
# manual calcuation in R

# residual = windstress_actual - windstress_expected

# windstress_expected = (0.2117167 - 0.0056096*snow)

# residual_firstdatapoint = 0.125 - (0.2117167 - 0.0056096 * 6.6)

climate$wind[1] - (0.2117167 - 0.0056096 * climate$snow[1])
```

```
## [1] -0.04969334
```

the fact that the first data point has the largest negative residual means that this data point has the lowest actual value of windstress compared to predicted value of windstress given that snowcover is 6.6 million square kilometers

Probem 2 Consider the data set set **mammals** downloaded from the **"MASS"** library. The data set shows the **body** (kg) and **brain** (g) mass of 62 species of mammals.

```
# load a built in library in R
library("MASS")

# NOTE: The command head() shows the first few rows of a data frame NOTE:
# Similarly, the command tail() shows the last few rows of a data frame
head(mammals)
```

```
##                    body brain
## Arctic fox        3.385  44.5
## Owl monkey        0.480  15.5
## Mountain beaver   1.350   8.1
## Cow             465.000 423.0
## Grey wolf        36.330 119.5
## Goat             27.660 115.0
```

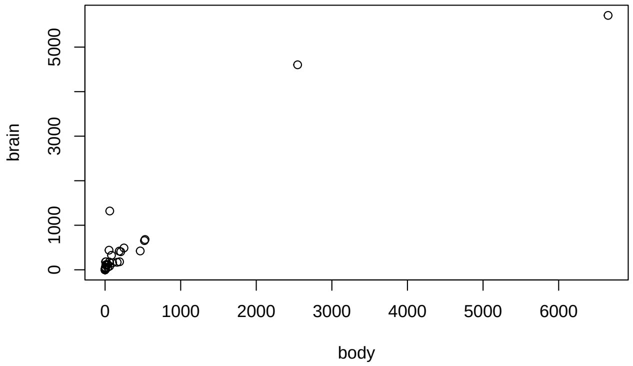
```
# NOTE: You can also access rows by their name
mammals["Human", ]
```

```
##       body brain
## Human   62  1320
```

Problem 2 Part a) The R code below produces the scatter plot of the **body** and **brain** mass of 62 species of mammals. Describe the scatter plot by commenting on its form, direction, and strength. Would it be appropriate to use a least squares regression line as a fit for this data?

NOTE: Enter your response outside/below the chunk as text.

```
plot(mammals)
```



The scatter plot is mostly concentrated at the lower range of body weight with a few outliers on the extremely heavy end, it's direction looks to be a positive slope, a least squares regression line would not be appropriate because the scatter plot is very concentrated towards the lower end of the body weight spectrum with a few outliers, suggesting that a logarithmic scale would work better here for the regression line

Problem 2 Part b) In the R chunk below are scatter plots showing different transforms for the data of the **body** and **brain**. Which one is most appropriate transformation for linear regression? Explain why.

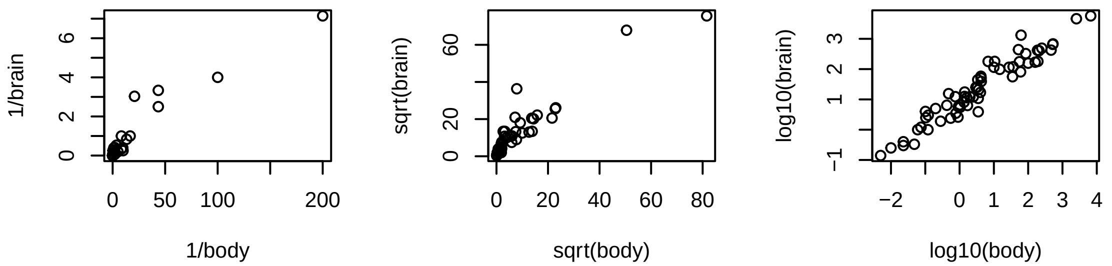
NOTE: Enter your response outside/below the chunk as text.

```r
# NOTE: par() creates one figure with three subplots across one row and three
# columns
par(mfrow = c(1, 3))

# NOTE: Reciprocal transform: 1/body vs 1/brain
plot(1/mammals$body, 1/mammals$brain, xlab = "1/body", ylab = "1/brain")

# NOTE: Square roots: sqrt(body) vs sqrt(brain)
plot(sqrt(mammals$body), sqrt(mammals$brain), xlab = "sqrt(body)", ylab = "sqrt(brain)")

# NOTE: Logrithmic: log10(body) vs log10(brain)
plot(log10(mammals$body), log10(mammals$brain), xlab = "log10(body)", ylab = "log10(brain)")
```



the third graph (log scale) would be the most appropriate transformation for linear regression because we can clearly see the positive correlation between the explanatory and response variables from the even spread of data throughout the scatter plot

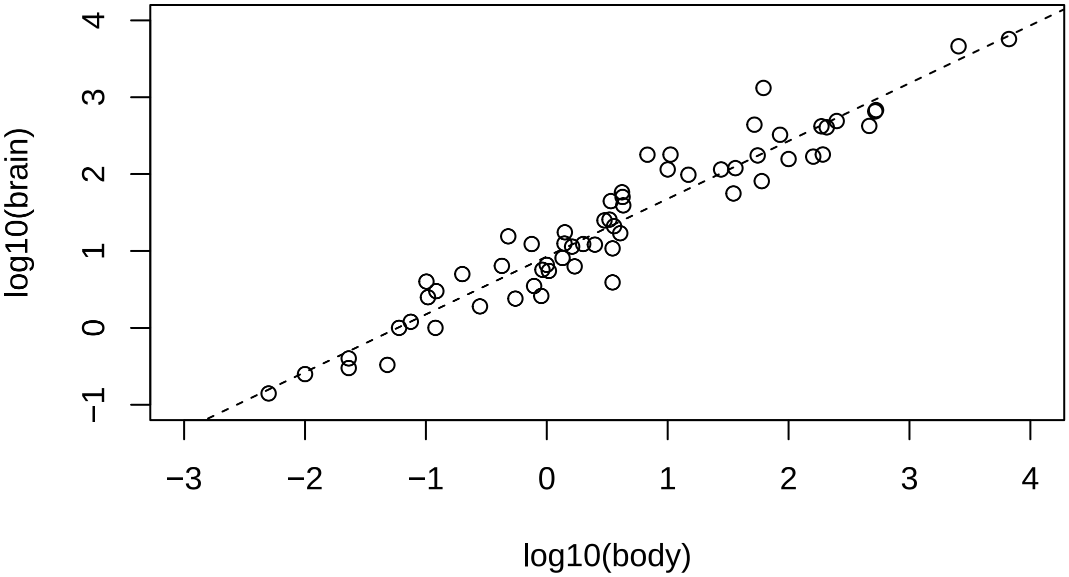
Problem 2 Part c) Using R, determine the regression line of the transformed data and state the equation of the regression line. Produce a fresh scatter plot using your chosen transformation and add the regression line to that plot.

```r
# NOTE: The line below binds the two new columns of the transformed data to the
# mammals data
mammals2 = cbind(mammals, tbody = log10(mammals$body), tbrain = log10(mammals$brain))

# NOTE: Then we use R to calculate the linear model (remember the 1st entry for
# lm() is 'y ~ x')
mammals.lm = lm(tbrain ~ tbody, data = mammals2)
summary(mammals.lm)
```

```
##
## Call:
## lm(formula = tbrain ~ tbody, data = mammals2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74503 -0.21380 -0.02676  0.18934  0.84613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.92713    0.04171   22.23   <2e-16 ***
## tbody        0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3015 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
# NOTE: Now for the plot with the linear model
plot(x = mammals2$tbody, y = mammals2$tbrain, xlab = "log10(body)", ylab = "log10(brain)",
    xlim = c(-3, 4), ylim = c(-1, 4))
par(new = TRUE)
abline(mammals.lm, lty = 2)
```



NOTE: Remember to state the regression line: *Replace the words INTERCEPT and SLOPE with their appropriate numerical values from the regression line of the transformed variables.*

The regression line is

$$\widehat{\log 10(\text{brain})} = 0.92713 + 0.75169 \cdot \log 10(\text{body})$$

Problem 2 Part d) For an animal with a **body** that weighs 1.0 kg, what does the linear model predict for the mass of that animal (i.e., the **brain**)?

NOTE: To complete this task, you can use R or show your calculations by hand. If using R, enter your code in the R chunk below by filling in the blank denoted with `FILL IN` and uncommenting all non-NOTE lines. Remember, the linear model will predict the value of log10(**brain**) and we need the value of **brain**, so you will need to do an additional calculation after `predict()`.

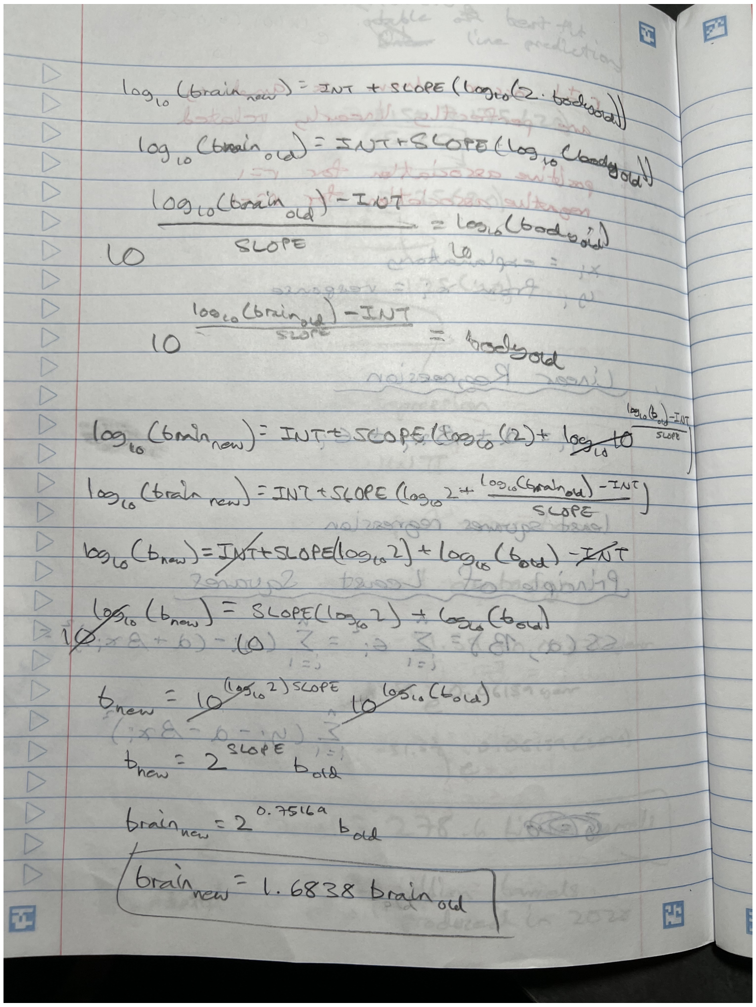
```
# NOTE: Use predict(linear model, newdata = data.frame(tbody = 'transformed
# mass'))
tbrain_pred = predict(mammals.lm, newdata = data.frame(tbody = log(1)))
(predbrain <- 10^tbrain_pred)
```

```
##       1
## 8.45526
```

Problem 2 Part e) On average, how does brain size change with a doubling of body size? Show your full calculations by hand.

NOTE: To complete this task, you will need to solve $\log_{10}(\textbf{brain}_{new}) = INTERCEPT + SLOPE \cdot \log_{10}(2 \cdot \textbf{body}_{old})$ for $\textbf{brain}_{new}$, where the values of INTERCEPT and SLOPE are taken from part (c). Upload an image of your work by replacing "upload_image.jpg" with your appropriately titled .jpg file in the R chunk below.

Hint: *You will need to substitute* $\log_{10}(\textbf{brain}_{old}) = INTERCEPT + SLOPE \cdot \log_{10}(\textbf{body}_{old})$ *at some point in your work.*



$$\log_{10}(\text{brain}_{new}) = \text{INT} + \text{SLOPE}(\log_{10}(2 \cdot \text{body}_{old}))$$

$$\log_{10}(\text{brain}_{old}) = \text{INT} + \text{SLOPE}(\log_{10}(\text{body}_{old}))$$

$$10^{\frac{\log_{10}(\text{brain}_{old}) - \text{INT}}{\text{SLOPE}}} = 10^{\log_{10}(\text{body}_{old})}$$

$$10^{\frac{\log_{10}(\text{brain}_{old}) - \text{INT}}{\text{SLOPE}}} = \text{body}_{old}$$

$$\log_{10}(\text{brain}_{new}) = \text{INT} + \text{SLOPE}\left(\log_{10}(2) + \cancel{\log_{10}10}^{\frac{\log_{10}(b_{old}) - \text{INT}}{\text{SLOPE}}}\right)$$

$$\log_{10}(\text{brain}_{new}) = \text{INT} + \text{SLOPE}\left(\log_{10}2 + \frac{\log_{10}(\text{brain}_{old}) - \text{INT}}{\text{SLOPE}}\right)$$

$$\log_{10}(b_{new}) = \cancel{\text{INT}} + \text{SLOPE}(\log_{10}2) + \log_{10}(b_{old}) - \cancel{\text{INT}}$$

$$\cancel{10}^{\cancel{\log_{10}}(b_{new})} = \text{SLOPE}(\log_{10}2) + \log_{10}(b_{old})$$

$$b_{new} = \cancel{10}^{(\cancel{\log_{10}}2)\text{SLOPE}} \; \cancel{10}^{\cancel{\log_{10}}(b_{old})}$$

$$b_{new} = 2^{\text{SLOPE}} \; b_{old}$$

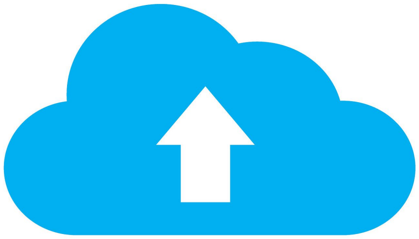
$$\text{brain}_{new} = 2^{0.7516a} \; b_{old}$$

$$\boxed{\text{brain}_{new} = 1.6838 \text{ brain}_{old}}$$

To complete some tasks, you will need to include details of your calculation. Upload an image of your work by replacing "upload_image.jpg" with your appropriately titled .jpg or .pdf file in the R chunk below.

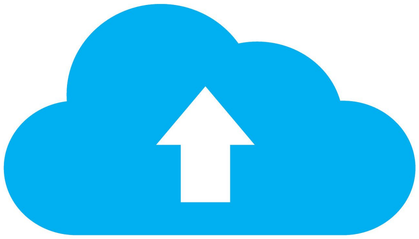
Option 1

Make sure that the image file (see D2L for the placeholder image seen in the PDF files) is in the same folder as the RMD file. Copy and paste the next two lines where you want to include the graphic. Then uncomment them.



Option 2

Make sure that the image file is in the same folder as the RMD file. Copy and paste the following line into the place you want it.



Final note: In the RMD file, pressing enter once does not guarantee a new line.
In order to ensure text begins on a new line you can press enter twice
or use the newline command seen in the RMD file throughout this note.
This type of effort in formatting is appreciated!