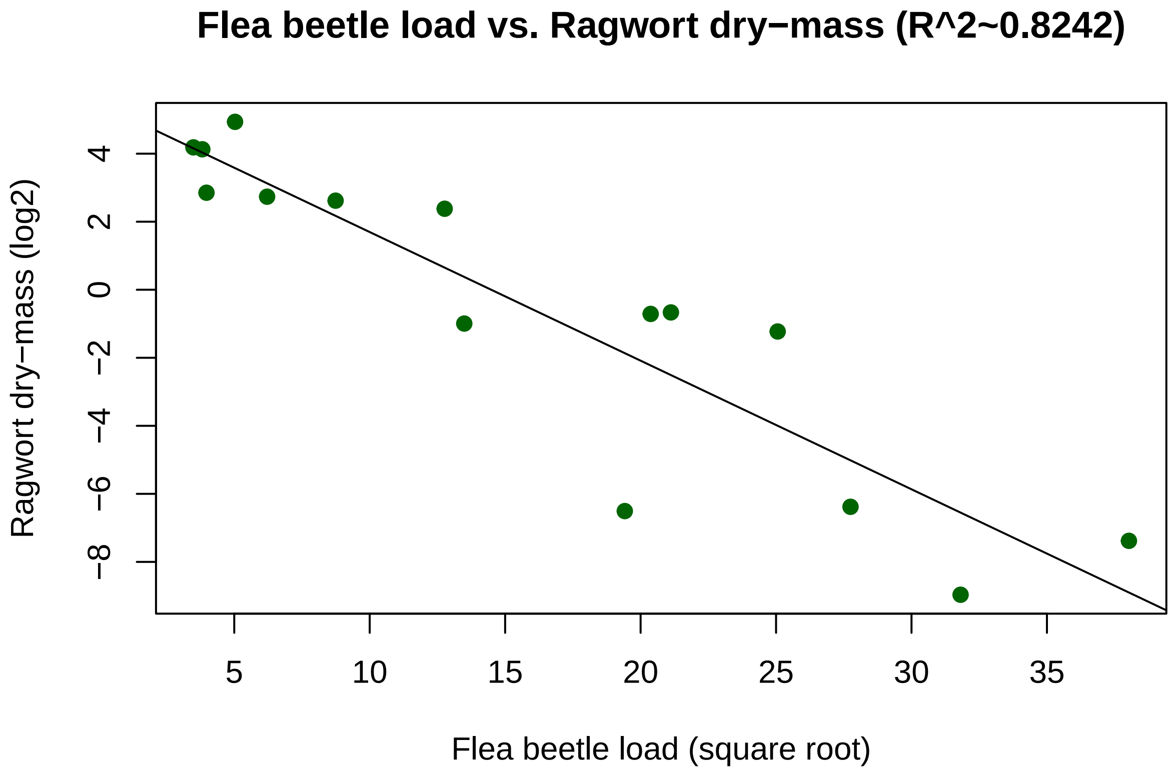
FIYEP Week 03 Exercises

Part one

1. Biological Pest Control. In a study of the effectiveness of biological control of the exotic weed tangsy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragwort. Harvesting the plots the next season, they measured the average dry mass of ragwort remaining (grams/ plant) and the flea beetle load (beetles/ gram of ragwort dry mass) to see if the ragwort plants in plots with high flea beetle loads were smaller as a result of herbivore by the beetles (data from P. McEvoy and C. Cox, ”Successful Biological Control of Ragwort, Senecio jacobaea, by Internatioal insects in Orego”, Ecological Applications 1(4) (1991): 430-42). Data set can be found in this weeks folder Data Sets.

(a) Use scatterplots of the raw data, along with trial and error, to determine transformations of $Y = ragwortdrymass$ and of $X = Fleabeetleload$ that will produce an approximate linear relationship (Search the net of how to make transformations of a variable).
Solution:

```
beetle_data <- read.csv('data/Biological Pest Control.csv')
beetle_trans <- data.frame(
  Load = sqrt(beetle_data$Load),
  Mass = log2(beetle_data$Mass)
)

beetle_relation <- lm(Mass ~ Load, data=beetle_trans)
R_squared <- summary(beetle_relation)$r.squared
# R-squared tirals
# | X | Y | R^2 |
# |---|---|-----|
# | X | Y | 0.3298 |
# | X^2 | Y | 0.1562 |
# | 1/X | Y | 0.552 |
# | log2(X) | Y | 0.6133652 |
# | X^X | Y | NAN |
# | sqrt(X) | Y | 0.4863435 |
# | X | Y^2 | 0.1620256 |
# | X^2 | Y^2 | 0.07228518 |
# | 1/X | Y^2 | 0.2888418 |
# | log2(X) | Y^2 | 0.329719 |
# | X^X | Y^2 | NAN |
# | sqrt(X) | Y^2 | 0.2503664 |
# | X | 1/Y | 0.4457879 |
# | X^2 | 1/Y | 0.3894913 |
# | 1/X | 1/Y | 0.1031819 |
# | log2(X) | 1/Y | 0.2617375 |
# | X^X | 1/Y | NAN |
# | sqrt(X) | 1/Y | 0.3846922 |
# | X | log2(Y) | 0.7568126 |
# | X^2 | log2(Y) | 0.5357336 |
# | 1/X | log2(Y) | 0.4654737 |
# | log2(X) | log2(Y) | 0.7581041 |
# | X^X | log2(Y) | NAN |
# | sqrt(X) | log2(Y) | 0.8242301 |
# | X | Y^Y | 0.05078722 |
# | X^2 | Y^Y | 0.0228336 |
# | 1/X | Y^Y | 0.03246237 |
# | log2(X) | Y^Y | 0.08502914 |
# | X^X | Y^Y | NAN |
# | sqrt(X) | Y^Y | 0.07477563 |
# | X | sqrt(Y) | 0.5312318 |
# | X^2 | sqrt(Y) | 0.2854013 |
# | 1/X | sqrt(Y) | 0.6607908 |
# | log2(X) | sqrt(Y) | 0.8166384 |
# | X^X | sqrt(Y) | NAN |
# | sqrt(X) | sqrt(Y) | 0.7118863 |

plot(beetle_trans$Load, beetle_trans$Mass,
     main=sprintf("Flea beetle load vs. Ragwort dry-mass (R^2~%s)", round(R_squared, digits=4)),
     xlab="Flea beetle load (square root)", ylab="Ragwort dry-mass (log2)",
     col="darkgreen", pch=19
     )
abline(beetle_relation)
```

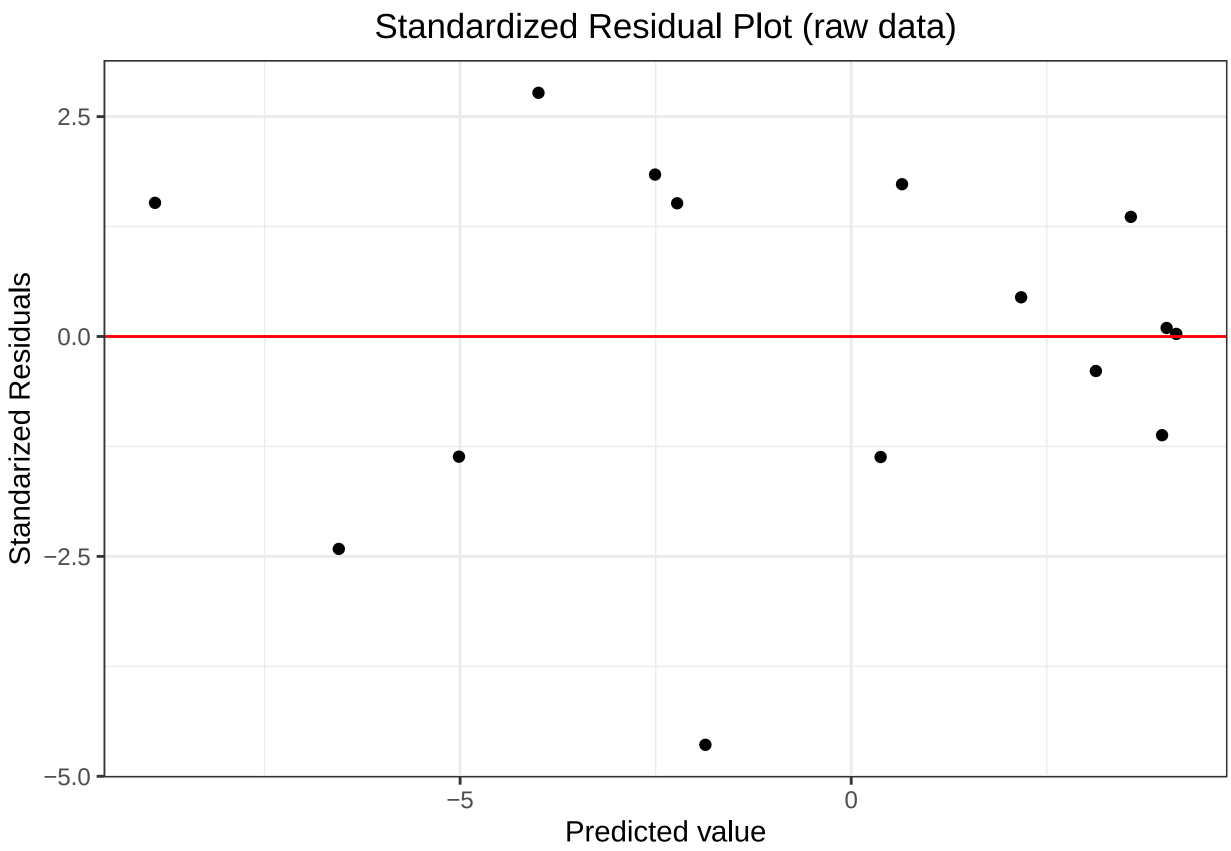


After trying every single combination of the functions $f(x) = x, g(x) = x^2, h(x) = 1/x, i(x) = log2(x), j(x) = x^x$ and $k(x) = \sqrt(x)$, the single best R-value was $\sqrt(x), \log_2(y)$ with and R^2 value of ~ 0.824.

(b) For a linear regression model on the transformed scale; calculate residuals and fitted values.

(c) Look at the residual plot. Do you want to try other transformations? What do you suggest? *Solution*:

```
model <- lm(Mass ~ Load, data = beetle_trans)
y_hat <- predict(model, newdata = beetle_trans)
ggplot(beetle_trans,
  mapping = aes(x = y_hat,
                y = resid(lm(Mass ~ Load, data = beetle_trans)))) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  xlab("Predicted value") +
  ylab("Standarized Residuals") +
  labs(title = "Standardized Residual Plot (raw data)") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "bottom")
```



There is no reason to try any other transformations, as this is the single best fitted line for all common transformation.

2. Ecosystem Decay. In the following we will consider a data set from a study about the effect of Amazon forest clearing. The publication is from 1984 where there is a requirement in Brazil that at least 50 % of the land in any development project remain in forest and tree cover. As a consequence of this requirement, "islands" of forest of various sizes remain in otherwise cleared areas. In the data set Ecosystem Decay you will find a table with the number of butterfly species in such islands. Analyze the role of area in the distribution of number of butterfly species. Where should such an analysis begin, what should be in such an analysis, what should the order of such statistical methods be? (Take some notes, you will use this on Thursday).
   Solution:
   First we'll look at the data.

```
fly_data <- read.csv('data/Ecosystem Decay.csv')
unique(fly_data$Area)
```

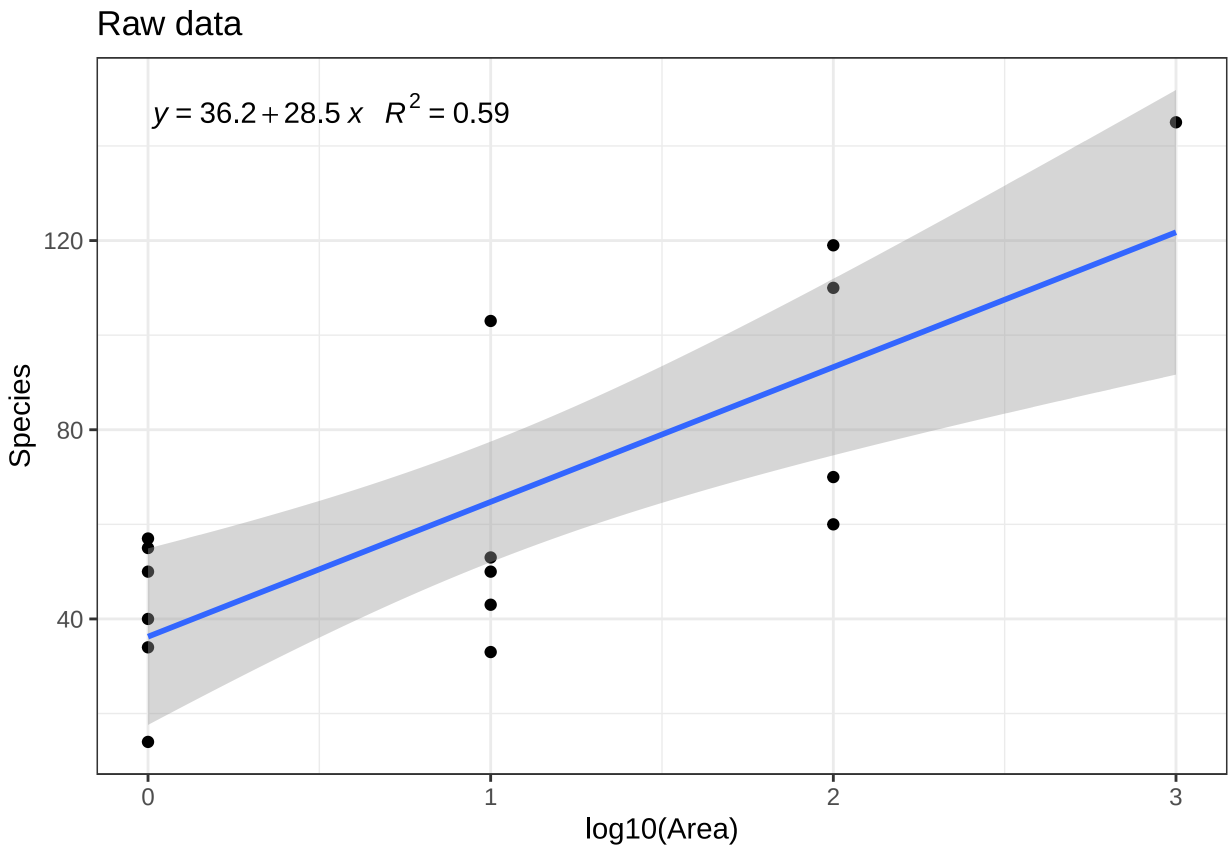
```
## [1]    1   10  100 1000
```

```
sort(unique(fly_data$Species))
```

```
##  [1]  14  33  34  40  43  50  53  55  57  60  70 103 110 119 145
```

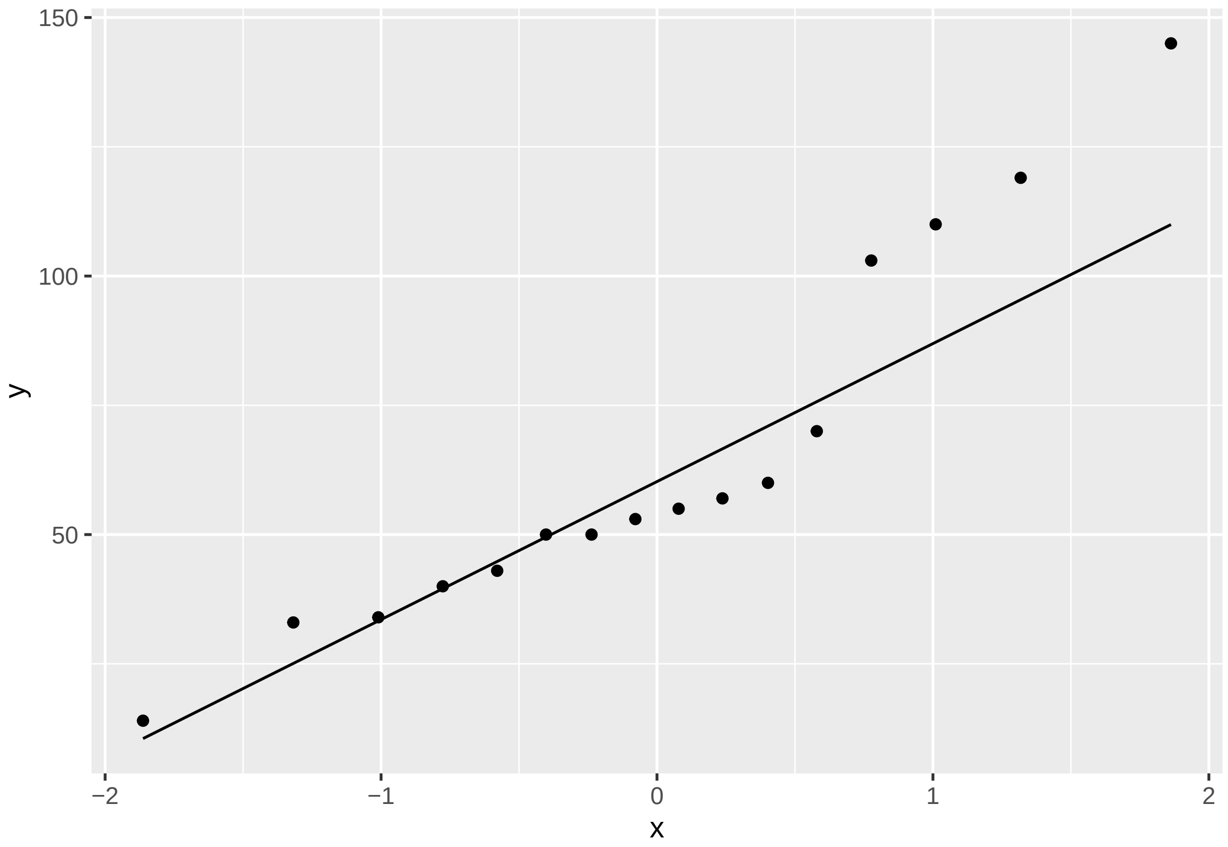
It seems as though the area values follow a $\log_1 0$ scale, and as such, plots should follow this. Let's plot the data with `log10(x)`.

```r
ggplot(fly_data, aes(x = log10(Area), y = Species)) +
  geom_point() +
  theme(legend.position = "top") +
  geom_smooth(method = "lm", formula = y ~ x) +

  stat_poly_eq(formula = y ~ x,
  aes(label = paste(after_stat(eq.label), after_stat(rr.label), sep = "~~~")),
  parse = TRUE) +

  labs(title = "Raw data") +
  theme_bw()
```



As we can see, there is a very clear correlation between area and number of butterfly species. We can go further and analyse the qq-plot:

```r
ggplot(fly_data, aes(sample = Species)) +
  stat_qq() + stat_qq_line()
```



It seems qq-plot is not perfectly linear, which means we might be missing something in the data. What that is, I am not sure :D.