First Year Project: Extinction prediciton

2023-02-07

We start by importing the dataset directly from the csv file and saving it to the data variable:

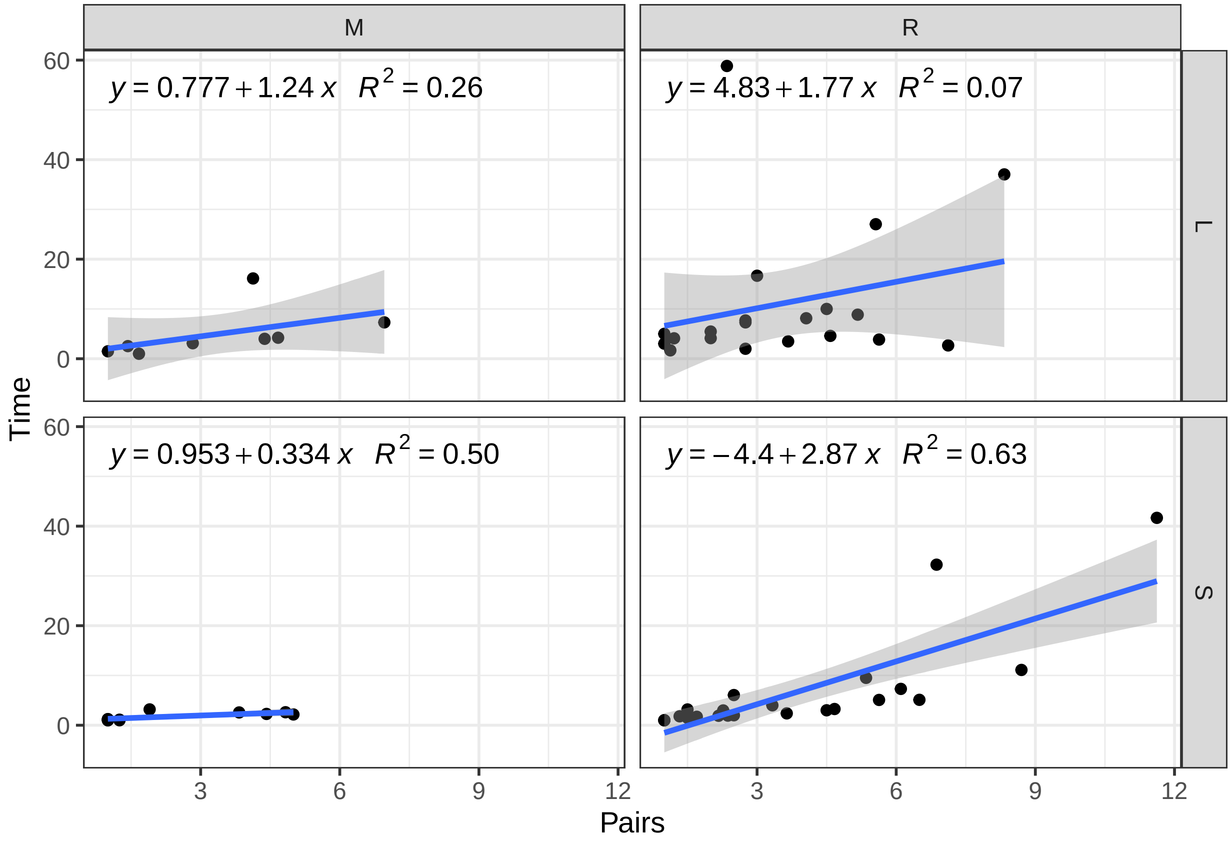
```
head(data <- read.csv('Factors Affecting Extinction.csv', header=T))
```

```
##          Species Time Pairs Size Status
## 1    Sparrowhawk 3.03  1.00    L      R
## 2        Buzzard 5.46  2.00    L      R
## 3        Kestrel 4.10  1.21    L      R
## 4      Peregrine 1.68  1.13    L      R
## 5 Grey_partridge 8.85  5.17    L      R
## 6          Quail 1.49  1.00    L      M
```

(1) Initial Plotting

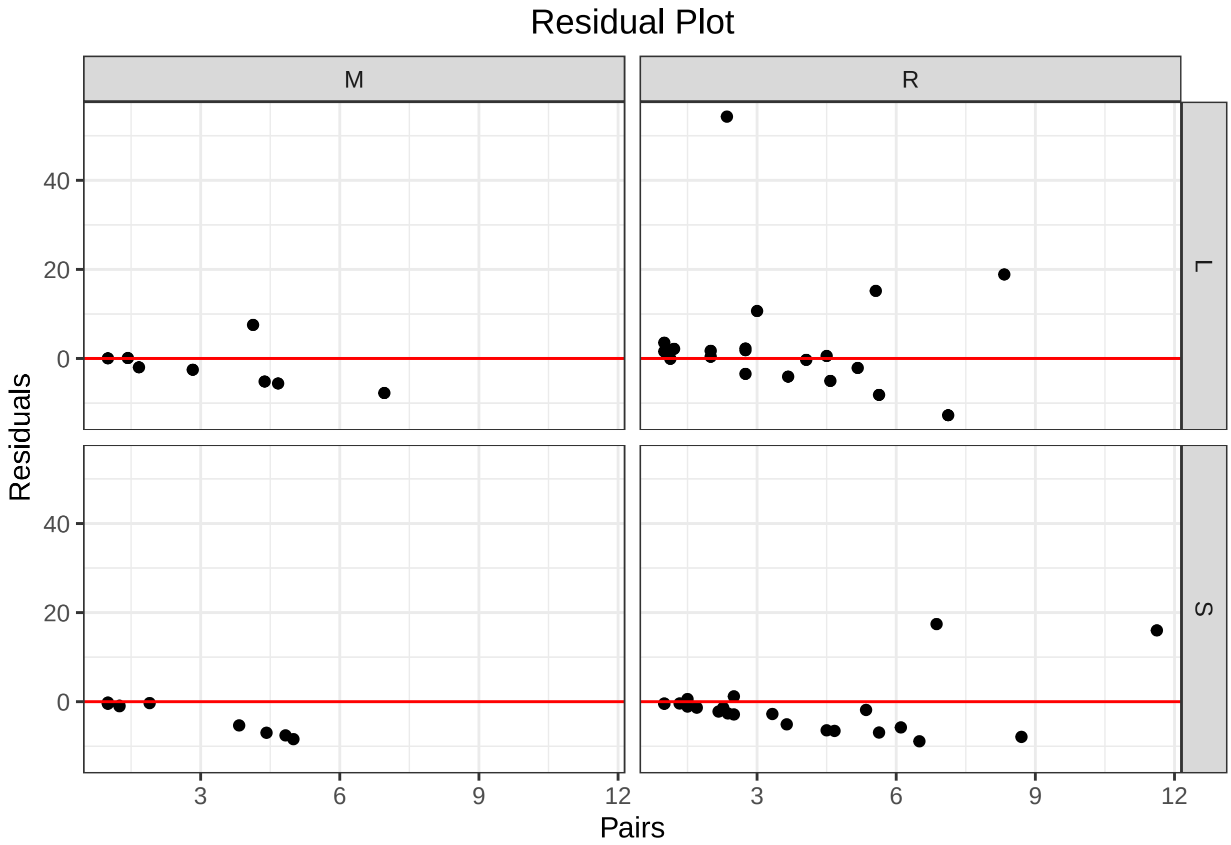
There are four different combinations of Size and Status, LR, LM, SR and SM. If we want to find the correlation between *extinction time* as a function of *pairs*, we can make a regression line with *pairs* as the predictor value and *extinction time* as the predicted value.

```
ggplot(data, aes(x = Pairs, y = Time)) +
  geom_point() +
  facet_grid(Size ~ Status) +
  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) +

  theme_bw()
```



(2) Residual plot of raw (not transformed) data

```
ggplot(data,
  mapping = aes(x = Pairs,
                y = resid(lm(Time ~ Pairs, data = data)))) +
  geom_point() +
  facet_grid(Size ~ Status) +
  geom_hline(yintercept = 0, color = "red") +
  xlab("Pairs") +
  ylab("Residuals") +
  labs(title = "Residual Plot") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "bottom")
```

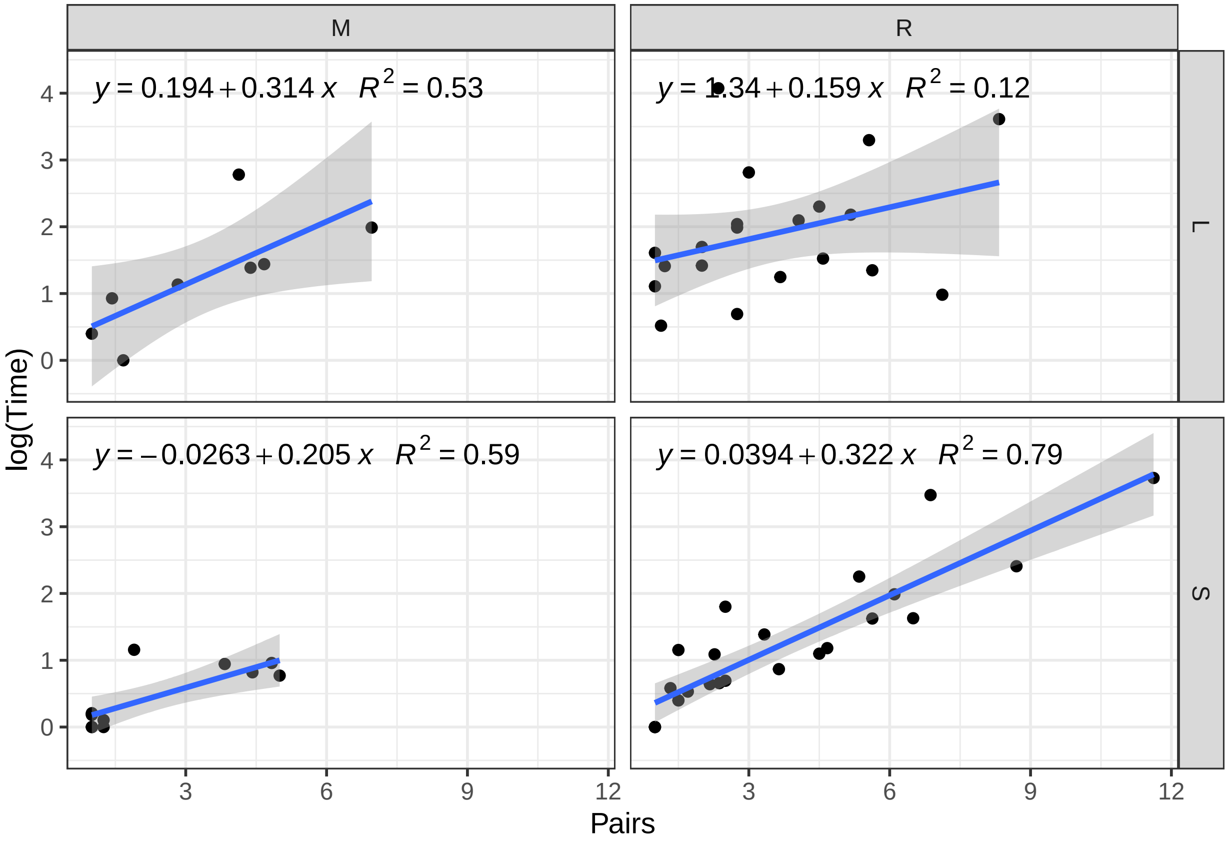


The data points on all the residual plots have a downward trend, not perfectly following the regression line.

(3) Transformations

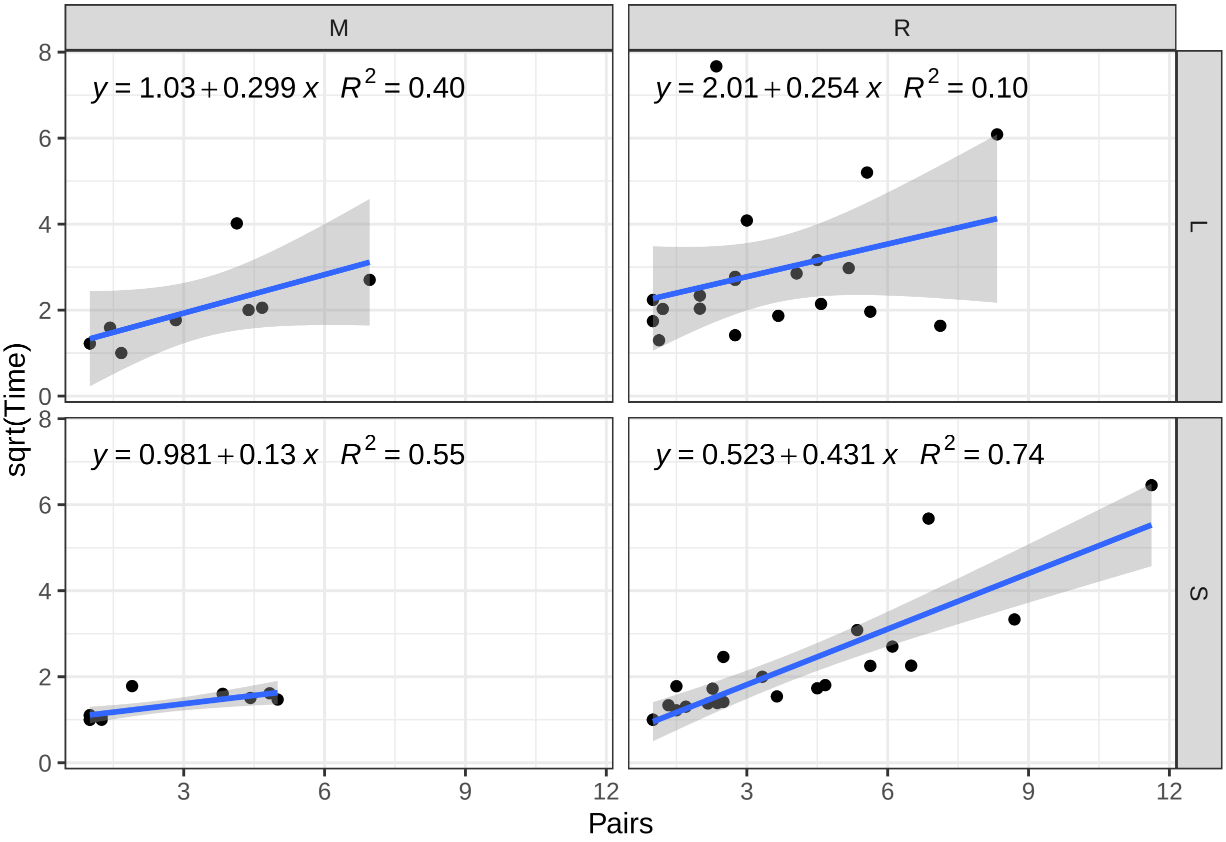
We now try different transformations on the data to see if we can get a better fitted regression line. We try 3 different transformation: log(“time”), sqrt(“time”) and 1/(“time”).

log("time")

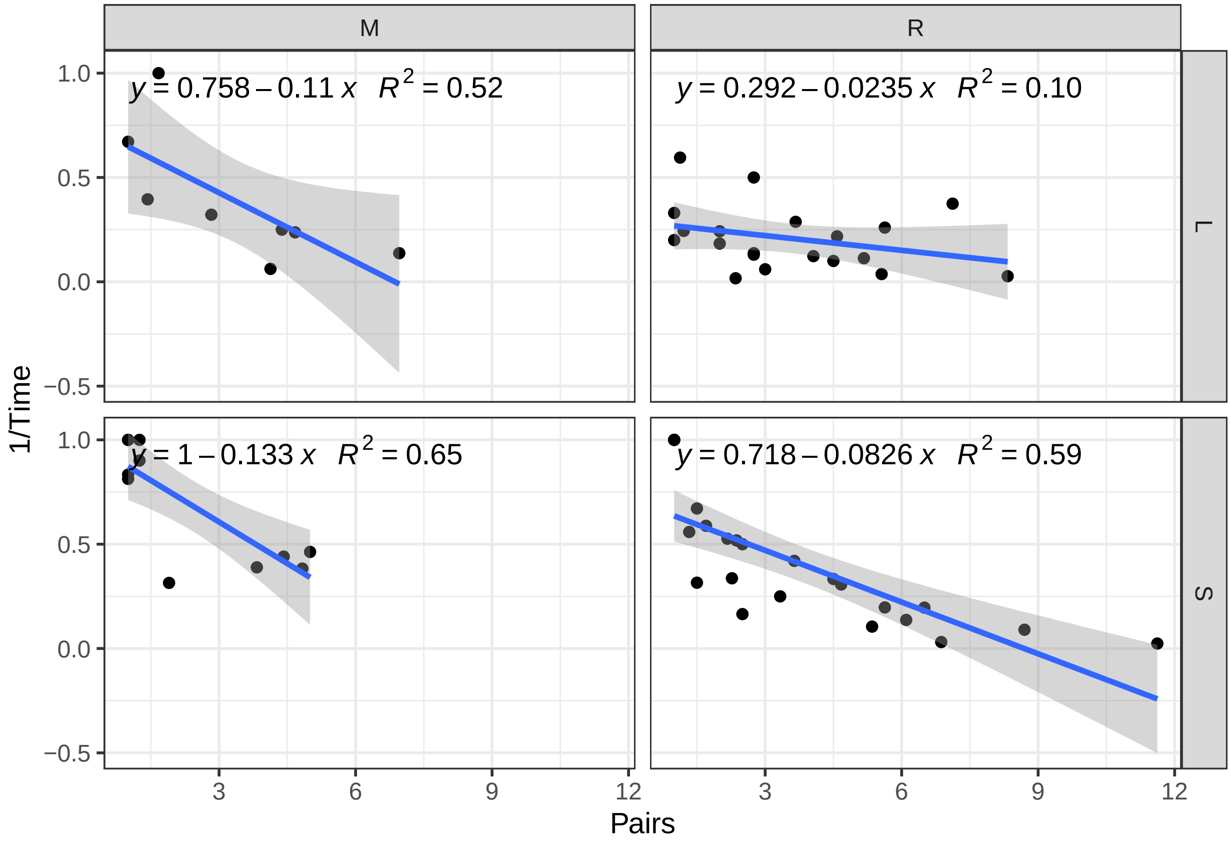


M
R
L
S
$y = 0.194 + 0.314\,x \quad R^2 = 0.53$
$y = 1.34 + 0.159\,x \quad R^2 = 0.12$
$y = -0.0263 + 0.205\,x \quad R^2 = 0.59$
$y = 0.0394 + 0.322\,x \quad R^2 = 0.79$
log(Time)
0
1
2
3
4
Pairs
3
6
9
12

sqrt("time")

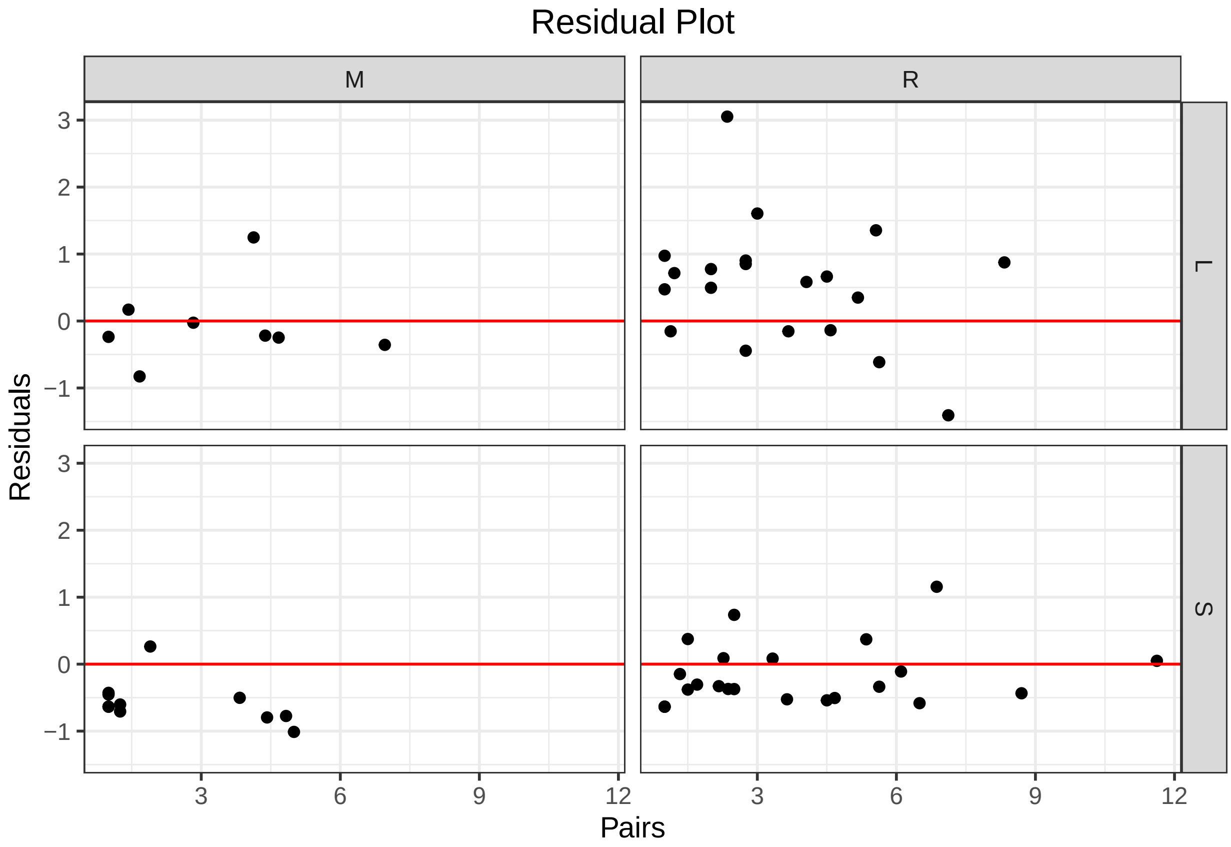

M
R
L
S
$y = 1.03 + 0.299\,x \quad R^2 = 0.40$
$y = 2.01 + 0.254\,x \quad R^2 = 0.10$
$y = 0.981 + 0.13\,x \quad R^2 = 0.55$
$y = 0.523 + 0.431\,x \quad R^2 = 0.74$
sqrt(Time)
0
2
4
6
8
3
6
9
12
Pairs

1/("time")



We can see that doing a natural logarithm transformation on time provides the best fitting line of any of the transformation, and the fit is even better than the base case with no transformation.
The risidual plot for the log("time") transformation is shown below.

```
ggplot(data,
  mapping = aes(x = Pairs,
                y = resid(lm(log(Time) ~ Pairs, data = data)))) +
  geom_point() +
  facet_grid(Size ~ Status) +
  geom_hline(yintercept = 0, color = "red") +
  xlab("Pairs") +
  ylab("Residuals") +
  labs(title = "Residual Plot") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "bottom")
```



As we can see, the downward trend of the points is no longer apparent.