Simple Linear Regression: Conditions and Transformations

Sleuth3 Chapter 8

Simple Linear Regression Model and Conditions

- Observations follow a normal distribution with mean that is a linear function of the explanatory variable
- $Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_i, \sigma)$

Conditions: spells "LINE-O"

- **Linear** relationship between explanatory and response variables: $\mu(Y|X) = \beta_0 + \beta_1 X$
    - Read as "The mean of Y for a given value of X"
    - β_0 is intercept: mean response when $X = 0$
    - β_1 is slope: change in mean response when X increases by 1 unit.
    - β_0 and β_1 are **parameters** describing the relationship between X and Y **in the population**
- **Independent** observations (knowing that one observation is above its mean wouldn't give you any information about whether or not another observation is above its mean)
- **Normal** distribution of responses around the line
- **Equal standard deviation** of response for all values of X
    - Denote this standard deviation by σ
- **no Outliers** (not a formal part of the model, but important to check in practice)

Transformations

- Transformations can sometimes help with the following issues:
    - skewed distributions (but skewness is only a problem if it is very serious)
    - lack of equal standard deviation for all values of X
    - outliers (but usually only if this is a side effect of serious skewness)
    - lack of a linear relationship

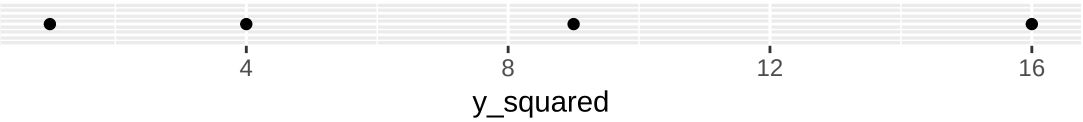
Reminder of the Ladder of Powers

- We start at y and go up or down the ladder.

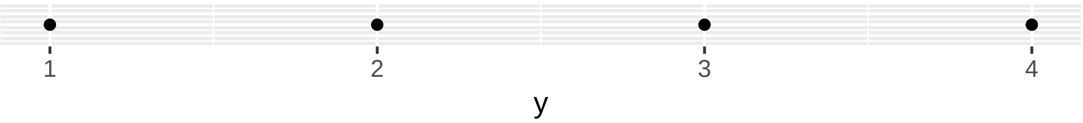
Transformation	R Code	Comments
$\vdots$		
e^y	`exp(y)`	Exactly where on the ladder the exponential transformation belongs depends on the magnitude of the data, but somewhere around here...
y^2	`y^2`	
y		Start here (no transformation)
$\sqrt{y}$	`sqrt(y)`	
$y^{\text{“}0\text{”}}$	`log(y)`	We use $\log(y)$ here
$-1/\sqrt{y}$	`-1/sqrt(y)`	The $-$ keeps the values of y in order
$-1/y$	`-1/y`	
$-1/y^2$	`-1/y^2`	
$\vdots$		

- Which direction?
  - If a variable is skewed right, move it down the ladder (pull down large values)
  - If a variable is skewed left, move it up the ladder (pull up small values)

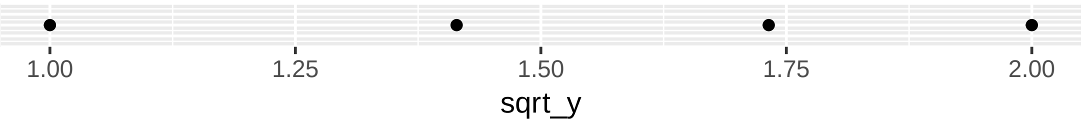
Moved Up 1 Step: spread out points on the right side



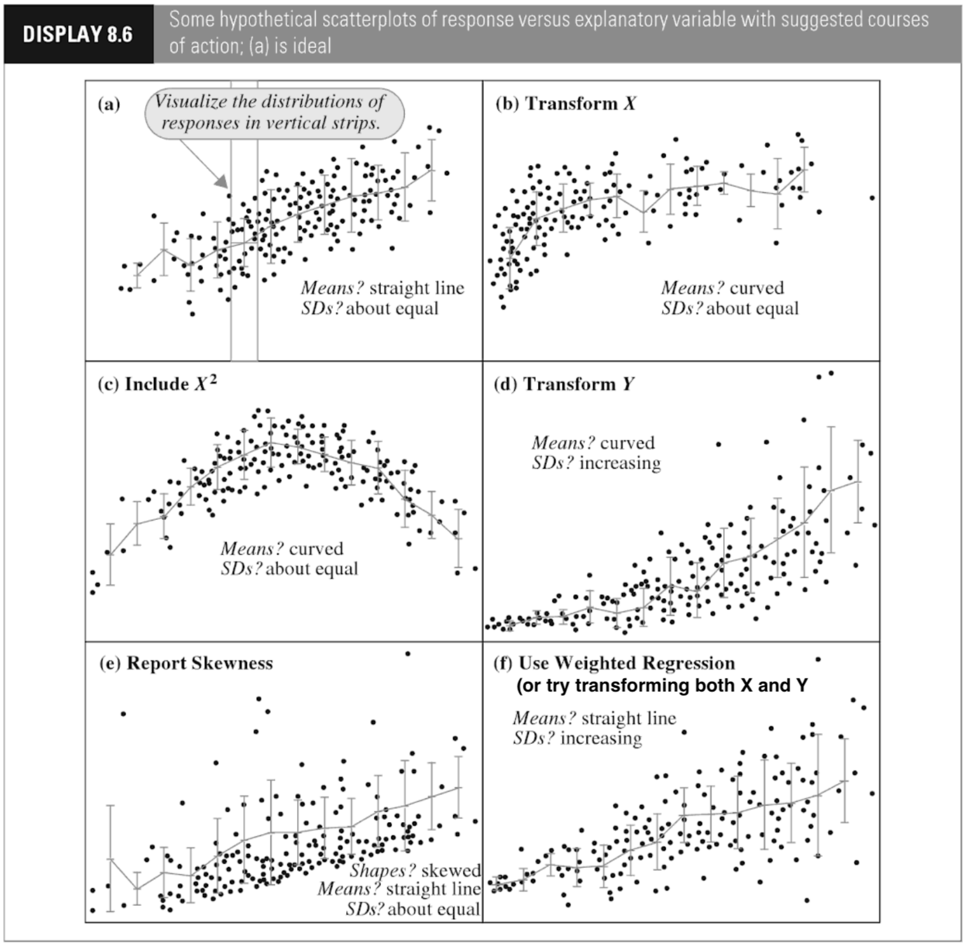
Starting Point: evenly spaced



Moved Down 1 Step: spread out points on the left side



DISPLAY 8.6 Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (a) is ideal



Tukey's Rule of Thumb for Re-Expression

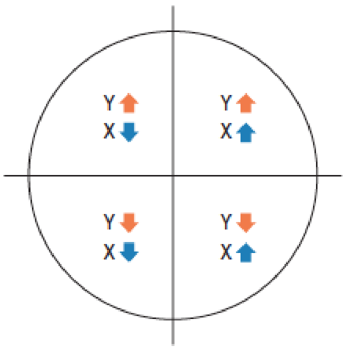


Figure 9.9

Tukey's circle of transformations shows what direction to move on the ladder for each variable depending on what quadrant the scatterplot resembles. For example, if the scatterplot is curved upward like quadrant 3 (bottom left) then either y or *x* should move down the ladder.

Example: Case Study 8.1.1 in Sleuth3

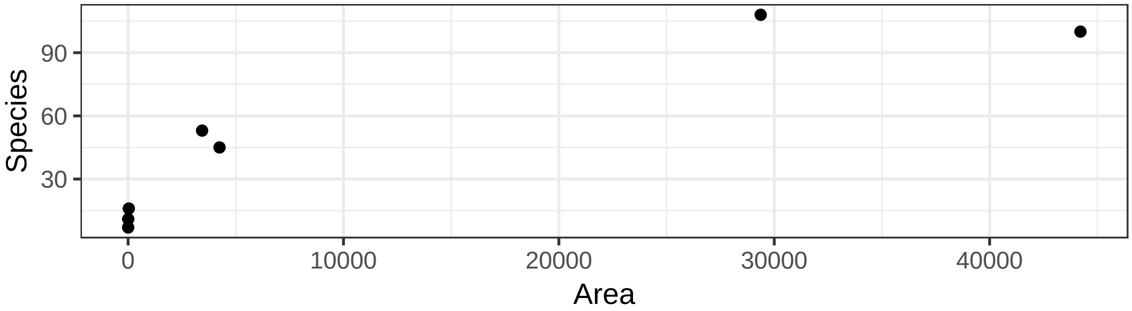
Quote from the book:

> Biologists have noticed a consistent relation between the area of islands and the number of animal and plant species living on them. ... The data [below] are the numbers of reptile and amphibian species and the island areas for seven islands in the west Indies. (Data on species from E. O. Wilson, *The Diversity of Life*, New York: W. W. Norton, 1991; areas from *The 1994 World Almanac, Mahwah, N.J.: Funk & Wagnalls, 1993.).

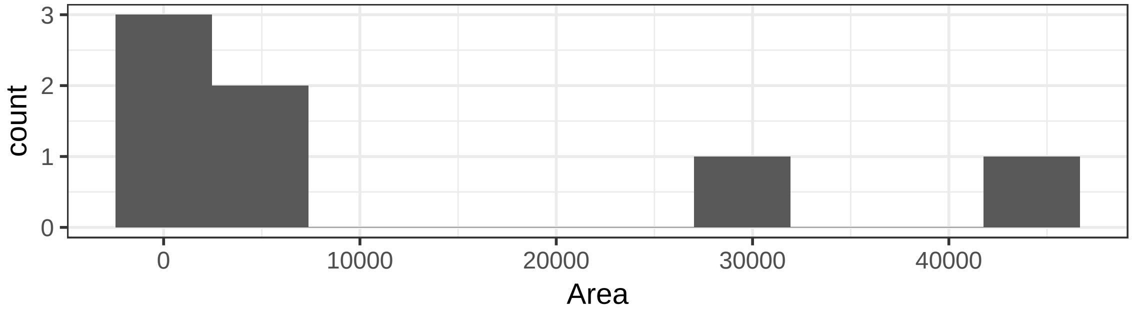
```r
head(island_species, 3)
```

```
## # A tibble: 3 x 2
##    Area Species
##   <dbl>   <dbl>
## 1 44218     100
## 2 29371     108
## 3  4244      45
```

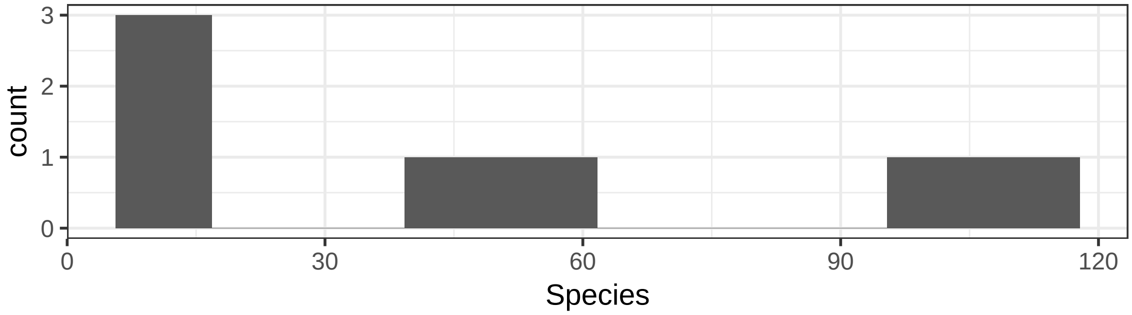
```r
ggplot(data = island_species, mapping = aes(x = Area, y = Species)) +
  geom_point() +
  theme_bw()
```



```r
ggplot(data = island_species, mapping = aes(x = Area)) +
  geom_histogram(bins = 10) +
  theme_bw()
```



```r
ggplot(data = island_species, mapping = aes(x = Species)) +
  geom_histogram(bins = 10) +
  theme_bw()
```

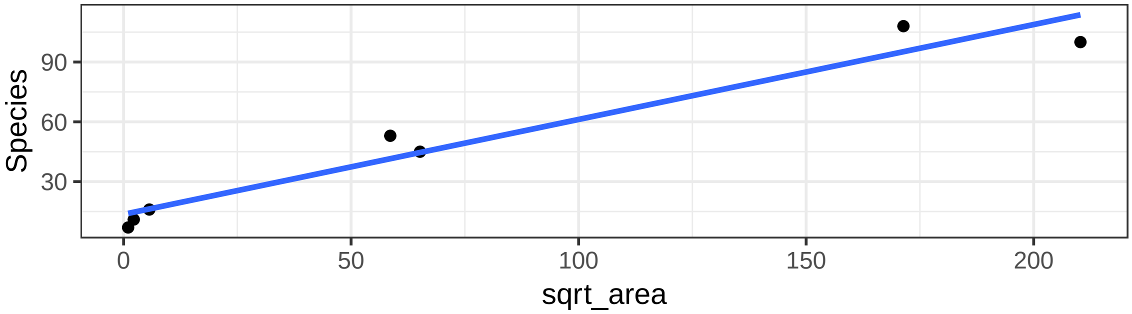


The relationship is non-linear. Let's try a transformation. Pulling in the extreme values for `Area` could help with the non-linearity. Let's move down the ladder for that variable.

Square root of area vs Species

```
island_species <- island_species %>%
  mutate(
    sqrt_area = sqrt(Area)
  )
```

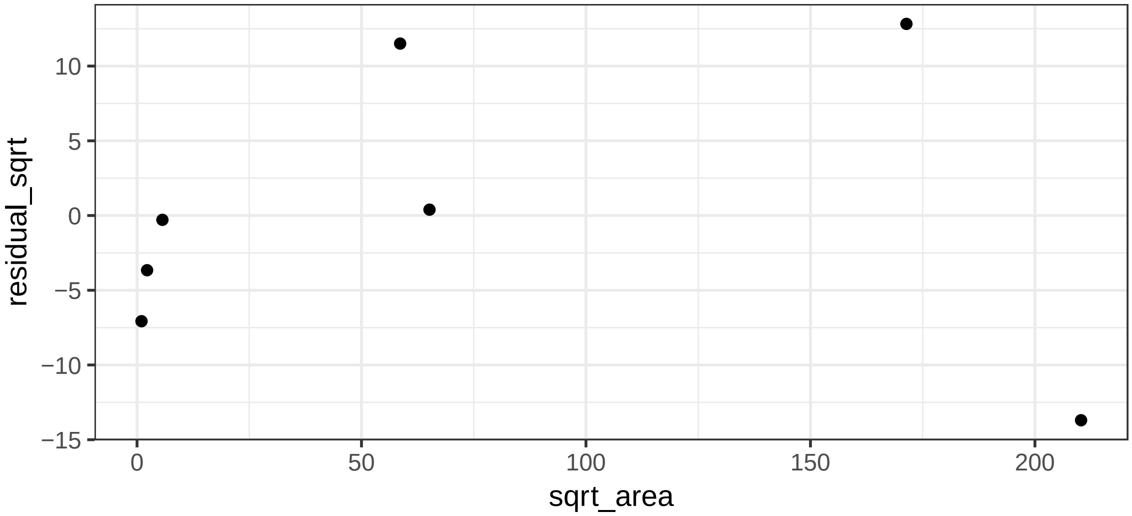
```
ggplot(data = island_species, mapping = aes(x = sqrt_area, y = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



Looks better… let's check residuals plots.

```
lm_fit <- lm(Species ~ sqrt_area, data = island_species)
island_species <- island_species %>%
  mutate(
    residual_sqrt = residuals(lm_fit)
  )

ggplot(data = island_species, mapping = aes(x = sqrt_area, y = residual_sqrt)) +
  geom_point() +
  theme_bw()
```

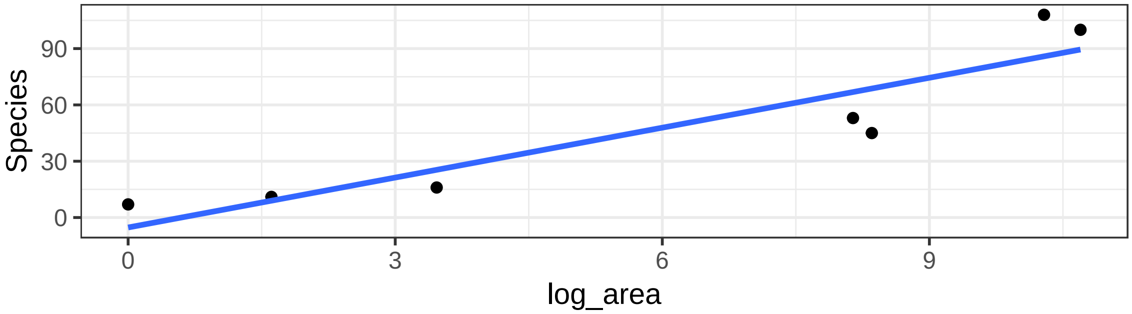


What do we see?

Log of area vs. Species

```
island_species <- island_species %>%
  mutate(
    log_area = log(Area)
  )
```

```
ggplot(data = island_species, mapping = aes(x = log_area, y = Species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



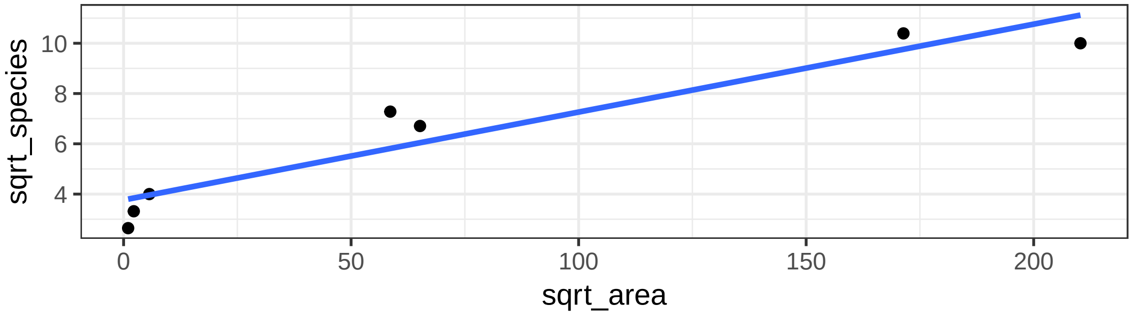
Too far?

Let's go back to square root of `Area`, and try taking a log of `Species` to address the non-constant variance.

Square root of Area vs. square root of Species

```
island_species <- island_species %>%
  mutate(
    sqrt_species = sqrt(Species)
  )
```

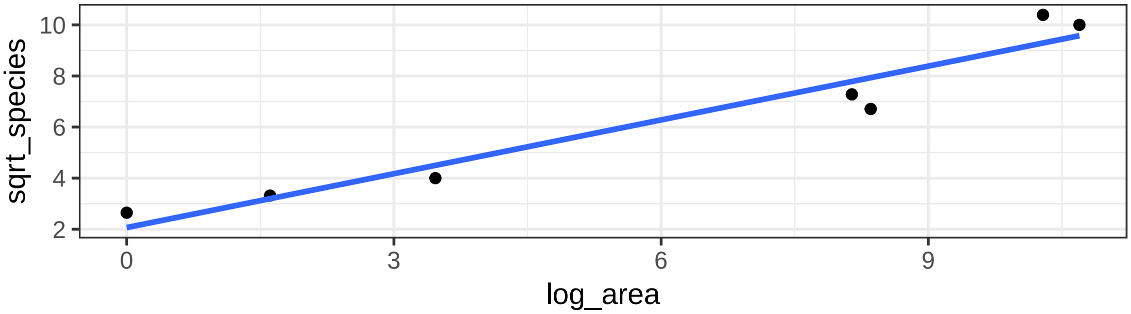
```
ggplot(data = island_species, mapping = aes(x = sqrt_area, y = sqrt_species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



Variability around the trend looks better now, but the curvature is worse. Should we try adjusting area down again, now that we're looking at the square root of species as our response?

Log of area vs. square root of species

```
ggplot(data = island_species, mapping = aes(x = log_area, y = sqrt_species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

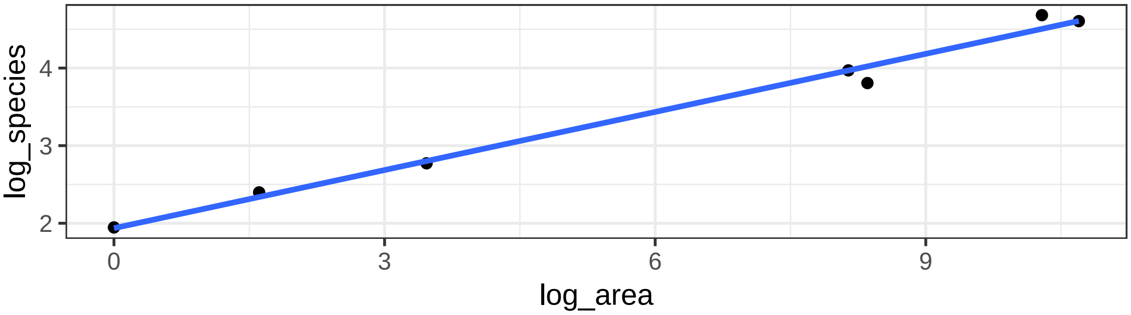


This does seem like it's gone too far in terms of transforming area. On the other hand, moving a step down on the ladder for species might help with this non-linearity.

Log of area vs. log of species

```
island_species <- island_species %>%
  mutate(
    log_species = log(Species)
  )
```

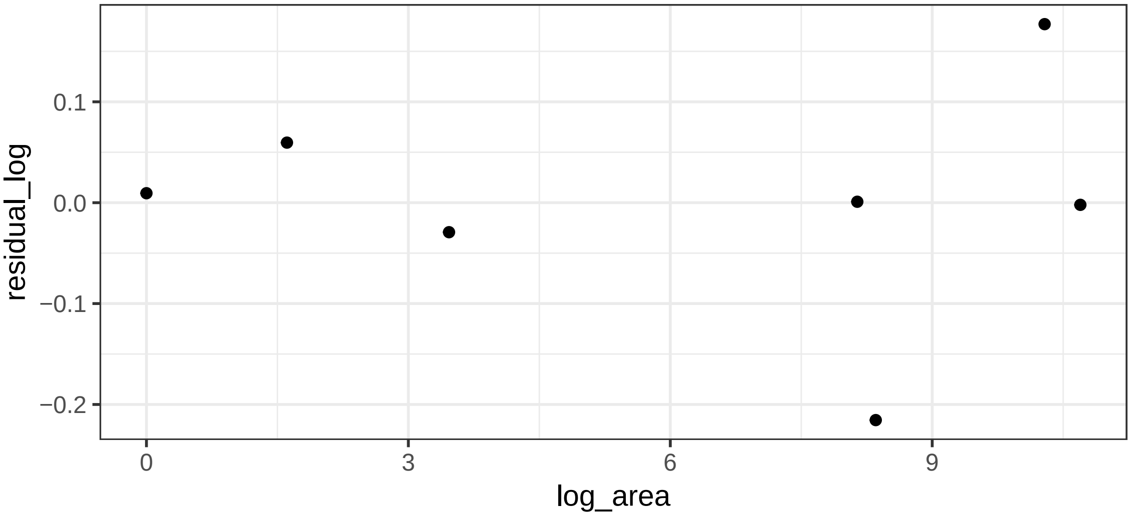
```
ggplot(data = island_species, mapping = aes(x = log_area, y = log_species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



Looks pretty good, but let's check a residuals plot again.

```
lm_fit <- lm(log_species ~ log_area, data = island_species)
island_species <- island_species %>%
  mutate(
    residual_log = residuals(lm_fit)
  )

ggplot(data = island_species, mapping = aes(x = log_area, y = residual_log)) +
  geom_point() +
  theme_bw()
```



Our book says this is ok, and it might be, but I have doubts…

```
island_species <- island_species %>%
  mutate(
    area_grouped = ifelse(log_area < 6, "Small Area", "Large Area")
  )

island_species %>% select(Area, log_area, area_grouped)
```

```
## # A tibble: 7 x 3
##    Area log_area area_grouped
##   <dbl>    <dbl> <chr>
## 1 44218    10.7  Large Area
## 2 29371    10.3  Large Area
## 3  4244     8.35 Large Area
## 4  3435     8.14 Large Area
## 5    32     3.47 Small Area
## 6     5     1.61 Small Area
## 7     1     0    Small Area
```

```
island_species %>%
  group_by(area_grouped) %>%
  summarize(sd_residual_log = sd(residual_log))
```

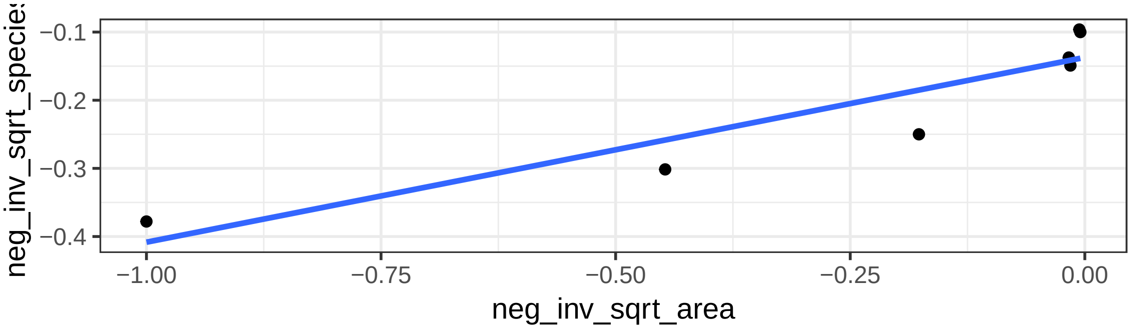
```
## # A tibble: 2 x 2
##   area_grouped sd_residual_log
##   <chr>                  <dbl>
## 1 Large Area            0.161
## 2 Small Area            0.0445
```

What if we go down another step for each?

-1/Square root transformation for both

```r
island_species <- island_species %>%
  mutate(
    neg_inv_sqrt_area = -1/sqrt(Area),
    neg_inv_sqrt_species = -1/sqrt(Species)
  )
```

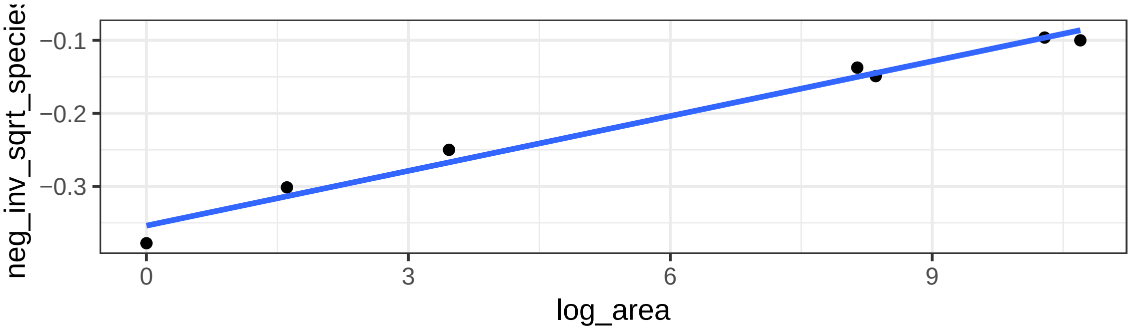
```r
ggplot(data = island_species, mapping = aes(x = neg_inv_sqrt_area, y = neg_inv_sqrt_species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



Definitely worse. Area is skewed left now, and we've put ourselves in the lower right corner of Tukey's circle. Back up a step for area?

log area vs. -1/Square root of species

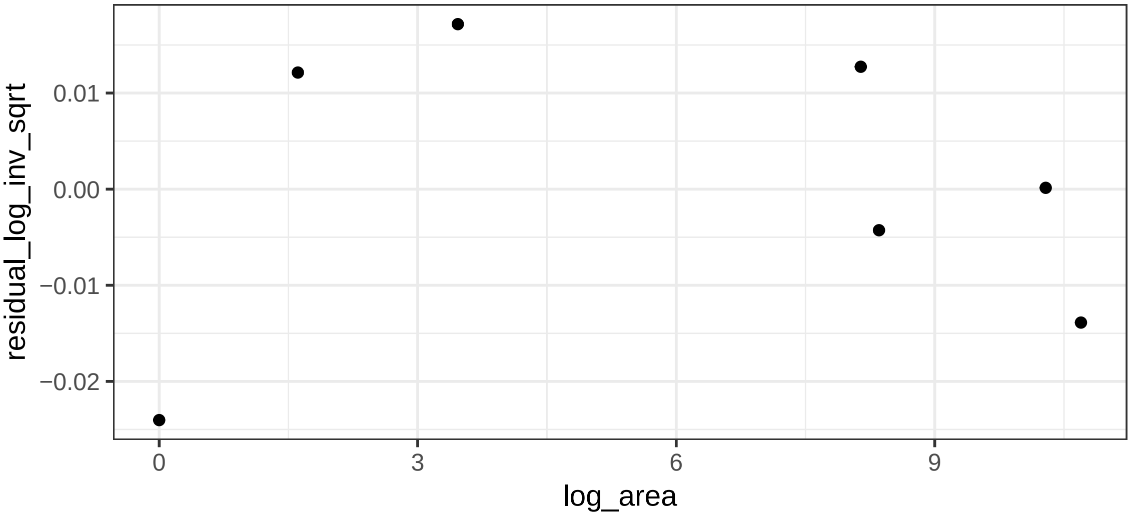
```r
ggplot(data = island_species, mapping = aes(x = log_area, y = neg_inv_sqrt_species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



OK? Let's check a residuals plot.

```r
lm_fit <- lm(neg_inv_sqrt_species ~ log_area, data = island_species)
island_species <- island_species %>%
  mutate(
    residual_log_inv_sqrt = residuals(lm_fit)
  )

ggplot(data = island_species, mapping = aes(x = log_area, y = residual_log_inv_sqrt)) +
  geom_point() +
  theme_bw()
```



Variances look about equal! Perhaps slight indications of non-linearity, but it's pretty good!

Double check standard deviations:

```
island_species %>%
  group_by(area_grouped) %>%
  summarize(residual_log_inv_sqrt = sd(residual_log_inv_sqrt))
```

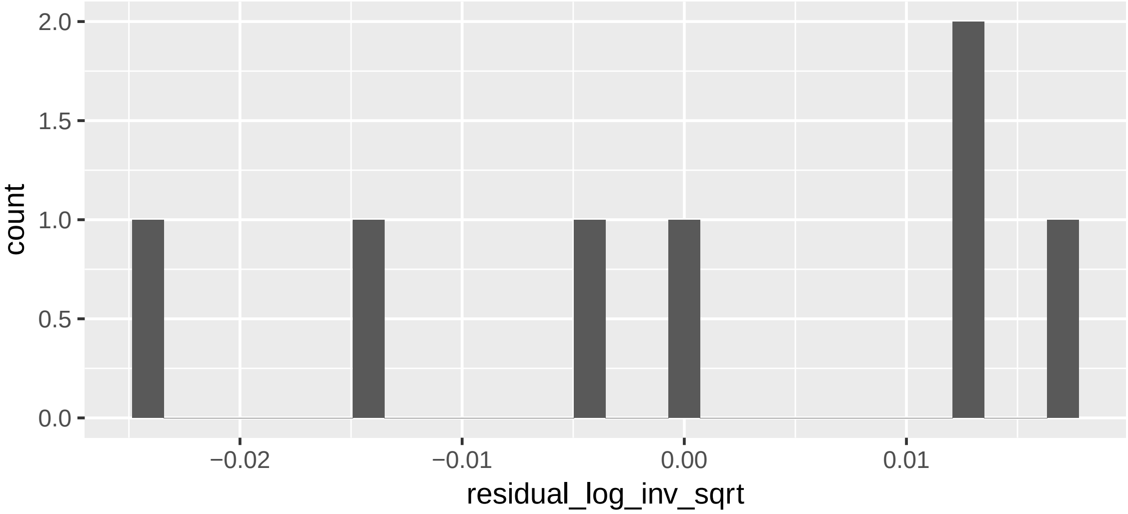
```
## # A tibble: 2 x 2
##   area_grouped residual_log_inv_sqrt
##   <chr>                        <dbl>
## 1 Large Area                  0.0110
## 2 Small Area                  0.0225
```

OK, now the small area group has a standard deviation that's twice the large area group. But that's better than 4, and our sample sizes are small (so small differences in standard deviations could be due to random noise). I might be willing to go with this.

Residuals approximately normal?

```
ggplot(data = island_species, mapping = aes(x = residual_log_inv_sqrt)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

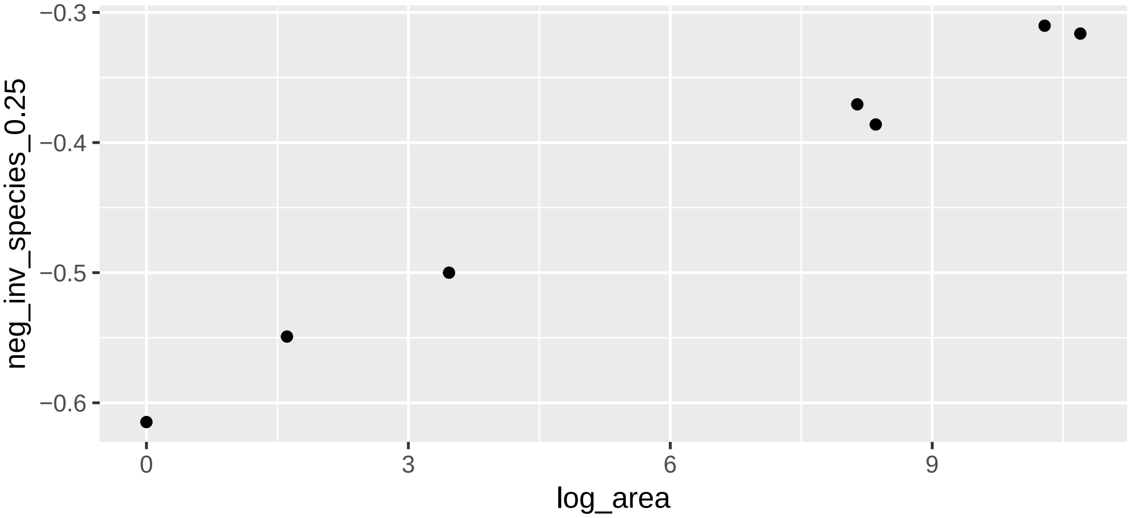


Not terrible, certainly not so skewed that I'm worried.

One last try... Log area vs. -1/(species^0.25)

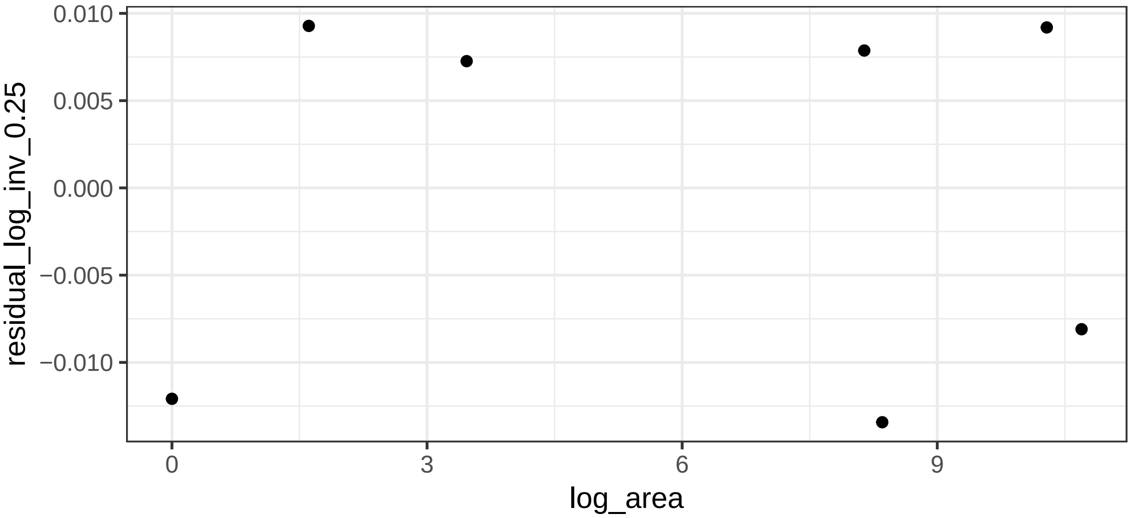
```
island_species <- island_species %>%
  mutate(
    neg_inv_species_0.25 = -1/(Species^0.25)
  )
```

```
ggplot(data = island_species, mapping = aes(x = log_area, y = neg_inv_species_0.25)) +
  geom_point()
```



```
lm_fit <- lm(neg_inv_species_0.25 ~ log_area, data = island_species)
island_species <- island_species %>%
  mutate(
    residual_log_inv_0.25 = residuals(lm_fit)
  )

ggplot(data = island_species, mapping = aes(x = log_area, y = residual_log_inv_0.25)) +
  geom_point() +
  theme_bw()
```



Variances look about equal! No indications of non-linearity!

Double check standard deviations:

```
island_species %>%
  group_by(area_grouped) %>%
  summarize(residual_log_inv_0.25 = sd(residual_log_inv_0.25))
```

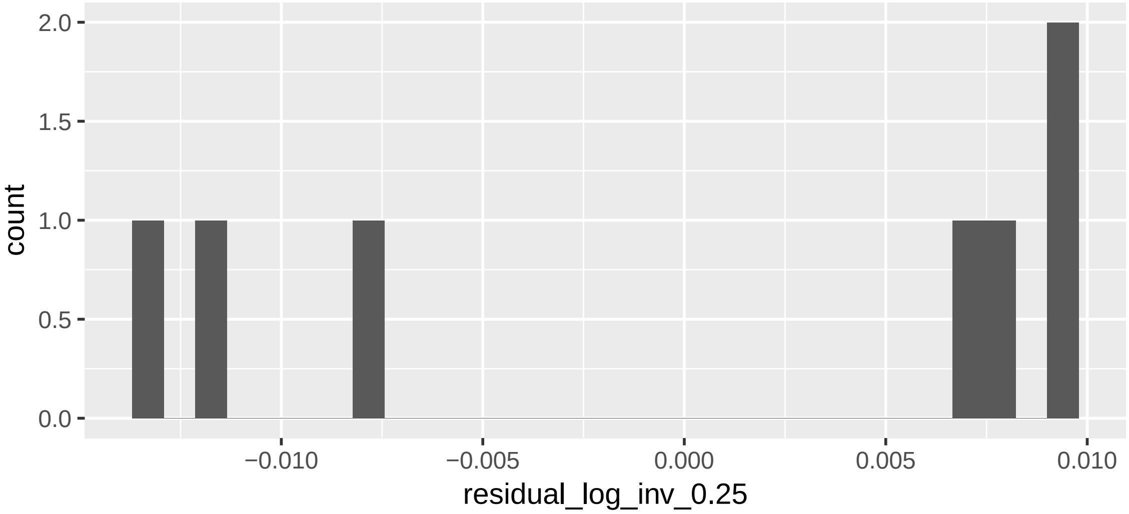
```
## # A tibble: 2 x 2
##   area_grouped residual_log_inv_0.25
##   <chr>                        <dbl>
## 1 Large Area                  0.0114
## 2 Small Area                  0.0118
```

Residual standard deviations look good now.

Residuals approximately normal?

```
ggplot(data = island_species, mapping = aes(x = residual_log_inv_0.25)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



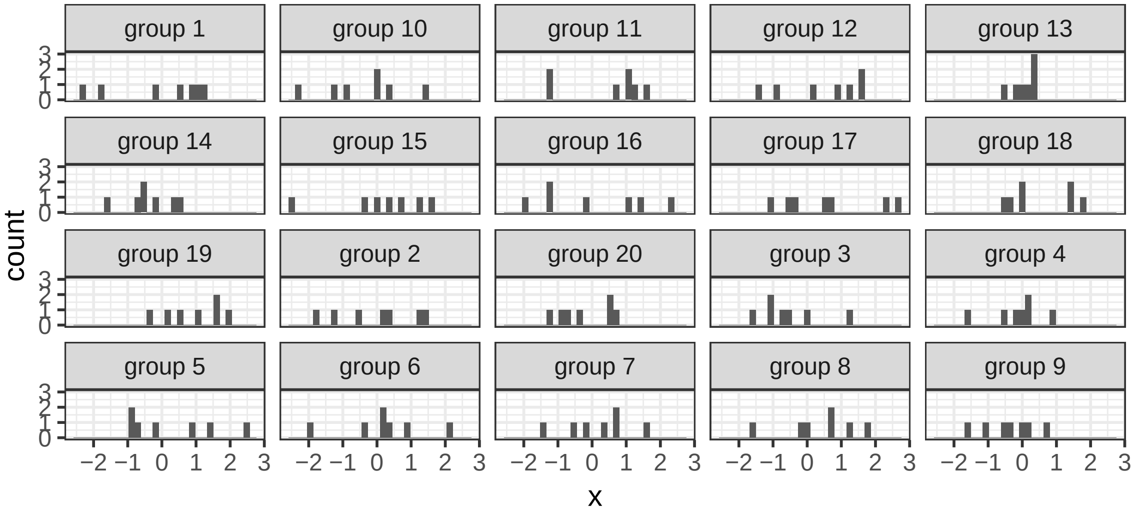
Bimodal? Skewed Left?

What might 7 observations from a normal distribution look like?

```
fake_data <- data.frame(
  x = rnorm(7 * 20),
  group = rep(paste0("group ", 1:20), each = 7)
)

ggplot(data = fake_data, mapping = aes(x = x)) +
  geom_histogram() +
  facet_wrap( ~ group) +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


group 1
group 10
group 11
group 12
group 13
group 14
group 15
group 16
group 17
group 18
group 19
group 2
group 20
group 3
group 4
group 5
group 6
group 7
group 8
group 9
count
x