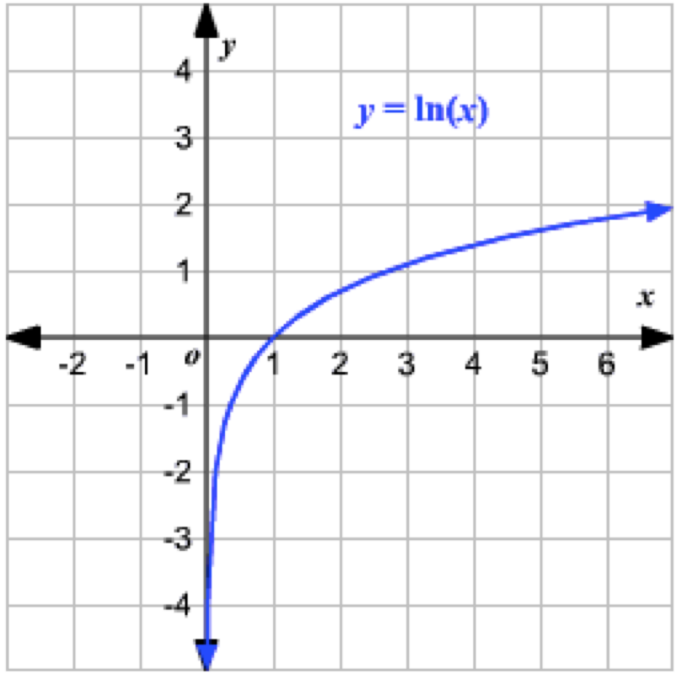
Linear Regression

log-trasformation

Logarithm

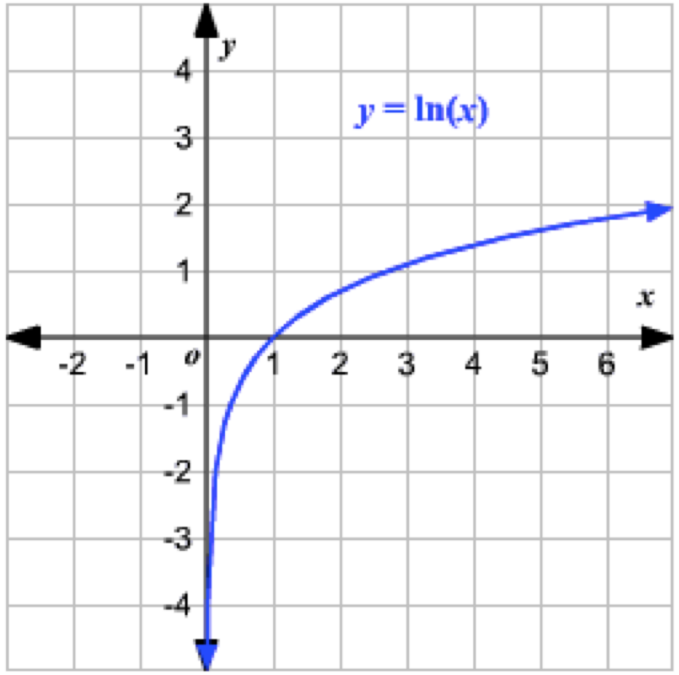
- Have you ever read articles where instead of using a variable of interest they used the logarithm of that variable?



Logarithm

- **Have you ever read articles where instead of using a variable of interest they used the logarithm of that variable?**

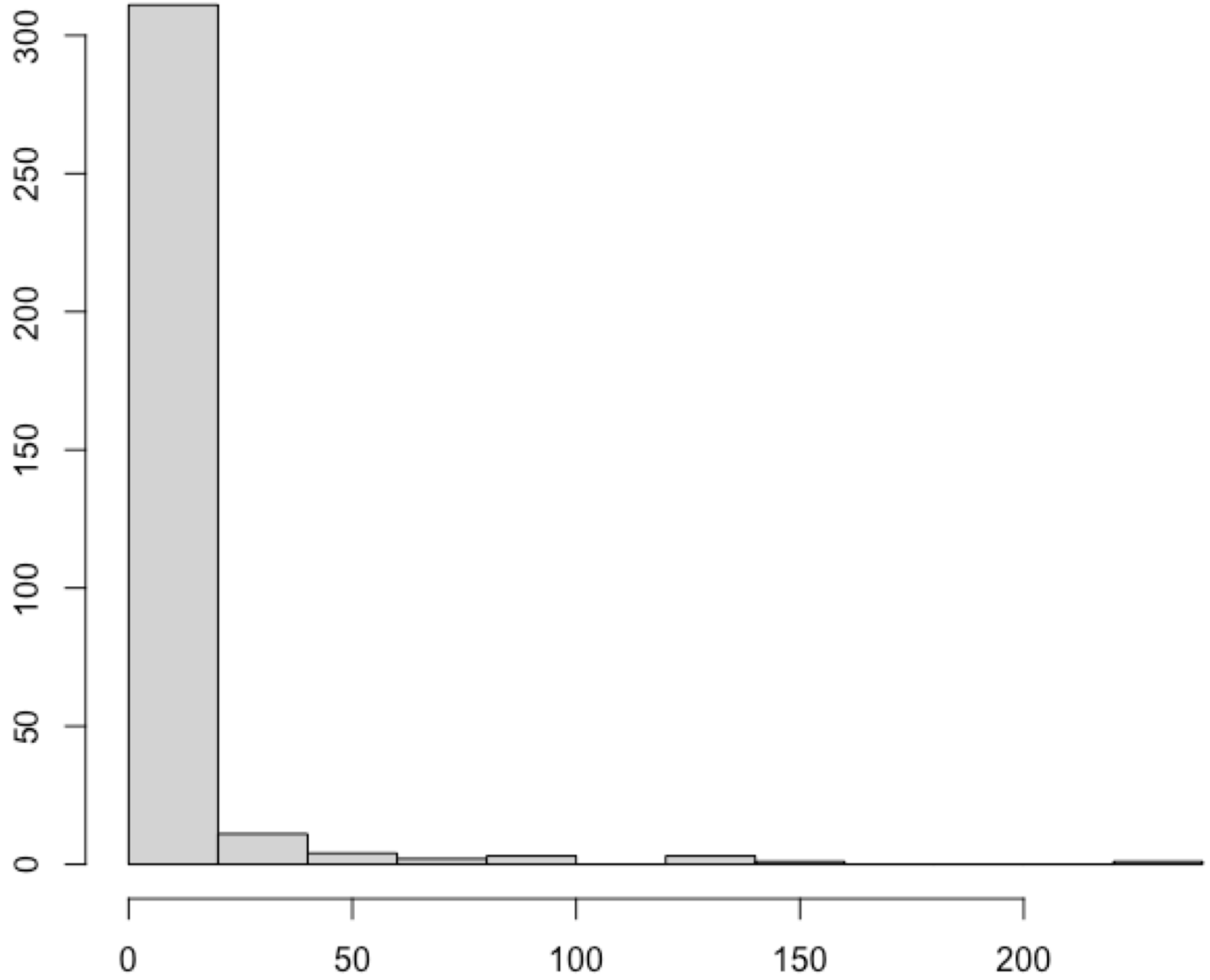
Logarithmic transformation is a convenient means of transforming a highly skewed variable into a more normalized dataset.



Example

- Length of hospital stay is an important indicator used to assess care of health.
- A database (uciPL) contains the record of 336 pediatrics admissions in two (0=Porto, 1=Lisboa) intensive care units.
- We want to compare the time of hospitalization (tempo_int) of Porto and Lisboa units (P1).
- The variable PRISM is an indicator of the patient's severity at the time of admission.

Time of hospitalization



Linear model

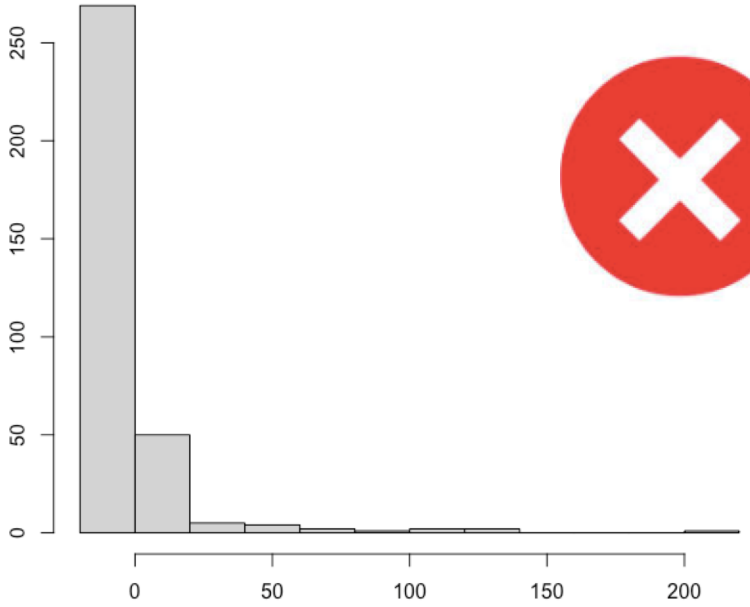
- Length of hospital stay = 6.9358 + 3.1025 x UCI unit + 0.1801 x PRISM

```
lm(formula = uciPL$tempo_int ~ uciPL$P1 + uciPL$PRISM)

Residuals:
    Min      1Q  Median      3Q     Max
-14.941  -6.939  -4.928  -1.158 210.881

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9358     2.4648   2.814  0.00519 **
uciPL$P1      3.1025     2.4075   1.289  0.19839
uciPL$PRISM   0.1801     0.1385   1.300  0.19451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear model

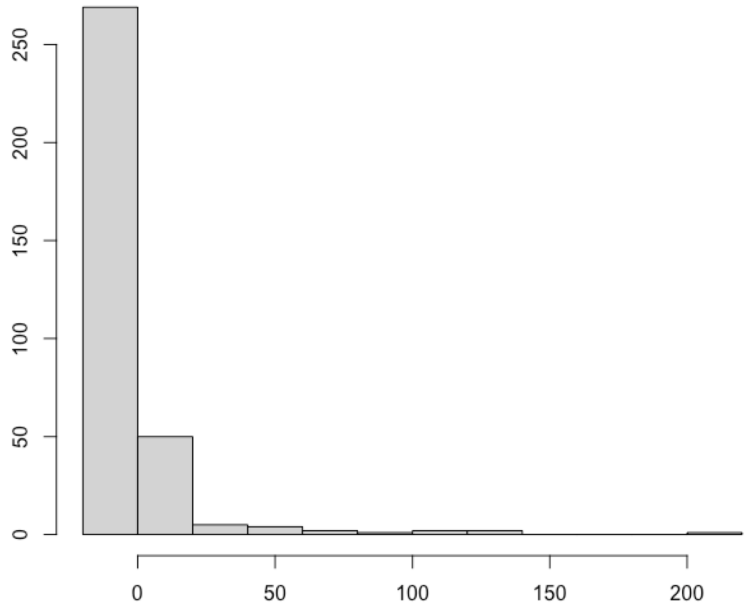
- Length of hospital stay = 6.9358 + 3.1025 x UCI unit + 0.1801 x PRISM

- And are the assumptions met?
  - hist(mymodel$residuals)

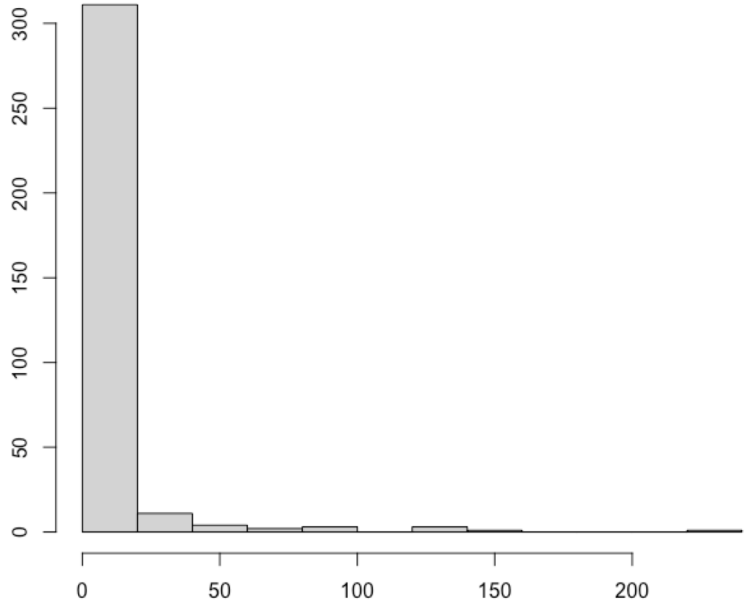


Linear model

- Residuals histogram

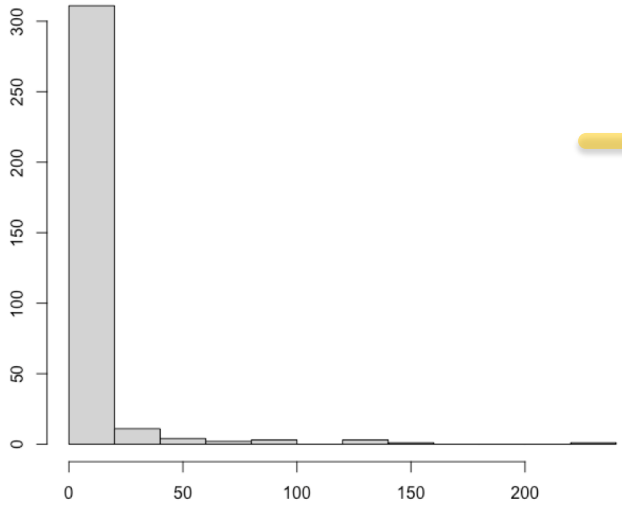


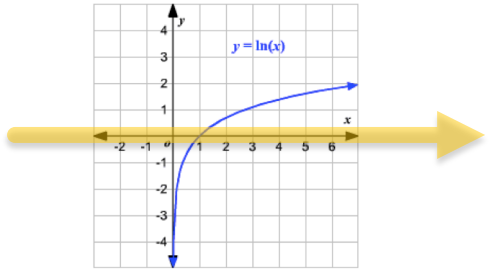
- Dependent variable histogram



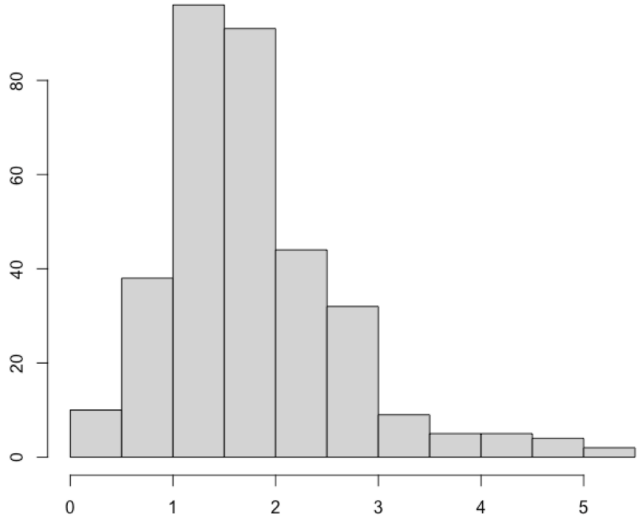
Right-Skewed Histogram and log transformation

- Time of hospitalization





- LN(Time of hospitalization)



Log Linear model

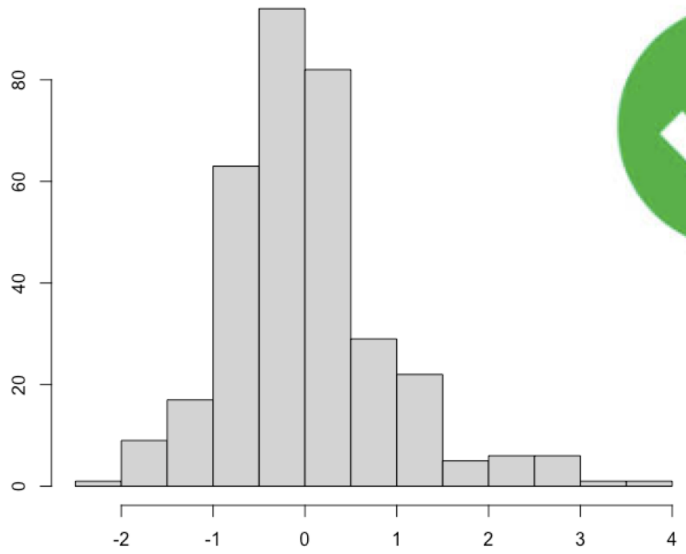
- LN(Length of hospital stay) = 1.515619 + 0.180659 x UCI unit + 0,012892 x PRISM

```
lm(formula = uciPL$lntempo ~ uciPL$P1 + uciPL$PRISM)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1602 -0.5588 -0.1448  0.4220  3.6290

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.515619   0.105344  14.387   <2e-16 ***
uciPL$P1    0.180659   0.102892   1.756   0.0800 .
uciPL$PRISM 0.012892   0.005921   2.177   0.0302 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Linear model

- LN(Length of hospital stay) = 1.515619 + 0.180659 x UCI unit + 0,012892 x PRISM

- And are the assumptions met now?
  - hist(mymodel2$residuals)



Log Linear model

- LN(Length of hospital stay) = 1.515619 + 0.180659 x UCI unit + 0,012892 x PRISM

however the interpretation of the coefficients is not very intuitive:

for Children from the same UCI unit (fixing the unit) the log of lengh of stay increases on average 0,012892 for na increase of 1 unit in PRISM (risk of mortality at the intensive care admission)

Log Linear model

- For PORTO:

LN(Length of hospital stay) = 1.515619 + 0,012892 x PRISM

however the interpretation of the coefficients is not very intuitive:

for Children from the Porto's UCI unit the log of lengh of stay increases on average 0,012892 for na increase of 1 unit in PRISM (risk of mortality at the intensive care admission)

Geometric mean ratio

$\log(\theta x) = \beta 0 + \beta 1\, x$
from our standard interpretation of regression slope parameters, we know that every 1 unit difference in X is associated with a β1 unit difference in log(θx):

$\log(\theta a+1) - \log(\theta a) = (\beta 0 + \beta 1 \times (a + 1)) - (\beta 0 + \beta 1 \times a) = \beta 1.$

We do not find it very convenient to talk about log(θ), however.
Hence we back transform to obtain statements about the ratio of θ across groups.

$e^{(\log(\theta a+1) - \log(\theta a))} = e^{\log(\frac{\theta a+1}{\theta a})} = \frac{\theta a+1}{\theta a} = e^{\beta 1}$ **geometric mean ratio**

The interpretation is similar to odds ratio.

So we find that every 1 unit difference in X is associated with a $e^{\beta 1}$ fold change in θ

Geometric mean ratio

- For PORTO:

LN(Length of hospital stay) = 1.515619 + 0.012892 x PRISM

for Children from the Porto's UCI unit every 1 unit difference in PRISM is associated with a $e^{0.012892}$ = 1,013 fold change in length of stay