Transformations

Matt Ryan

School of Mathematical Sciences, University of Adelaide

Semester 2 2021

Transformations

Curvilinear relationships of all sorts can be found in every field. Many of these non-linear models can still be fitted using the linear regression approach, provided the data can be initially "linearised" by a suitable transformation.

> *A regression function is linearisable if we can transform it into a function linear in the (unknown) parameters via transformations of the predictor variables and/or the original parameters and a monotone transformation of the response.*

Transformation can be applied to both the predictors and the response. Commonly used transformations

exponential e^Y square root $\sqrt{Y}$

logarthmic $\log(Y)$ reciprocal $\frac{1}{Y}$

Exponential regression

Many populations of plant or animals tend to grow at exponential rates. If Y denotes the size of a population at time x, we may use the model

$$Y = \alpha e^{\beta x}.$$

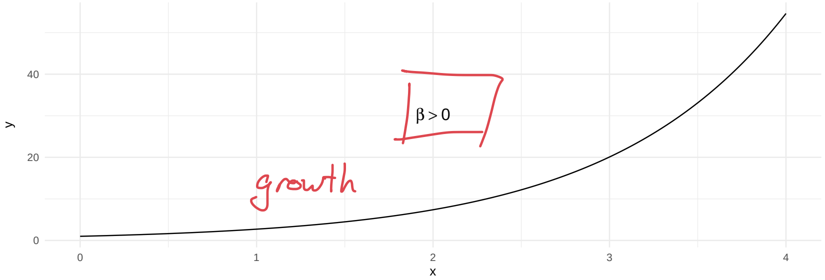
easy to fit our model now $\hat{\beta}_0, \hat{\beta}_1$ using least squares

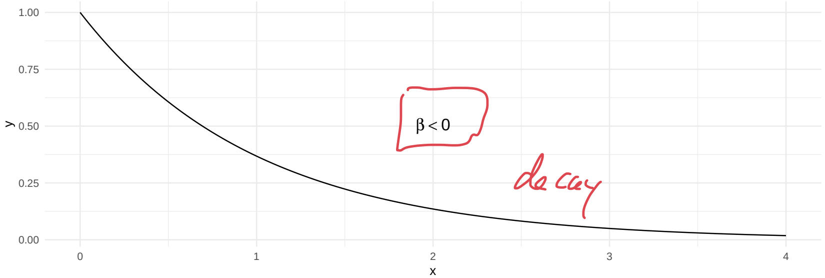
$$\underbrace{\log Y}_{Y^*} = \underbrace{\log \alpha}_{\beta_0} + \underbrace{\beta}_{\beta_1} x$$

$$\left\{ Y^* = \beta_0 + \beta_1 x + \varepsilon \right\} \Rightarrow \hat{\alpha} = \exp(\hat{\beta}_0), \quad \hat{\beta} = \hat{\beta}_1$$

Note: fitting like this is different to fitting just the exponential model

Exponential regression





Example 4.7

The data below represents the number of surviving bacteria (in hundreds) in an experiment with marine bacterium following exposure to X-rays. The response (y) is the bacteria count and the predictor (x) is time intervals

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	355	211	197	166	142	106	104	60	56	38	36	32	21	19	15

* a) Fit a linear regression to the data, plot the residuals.

* b) Fit an exponential regression to the data, plot the residuals.

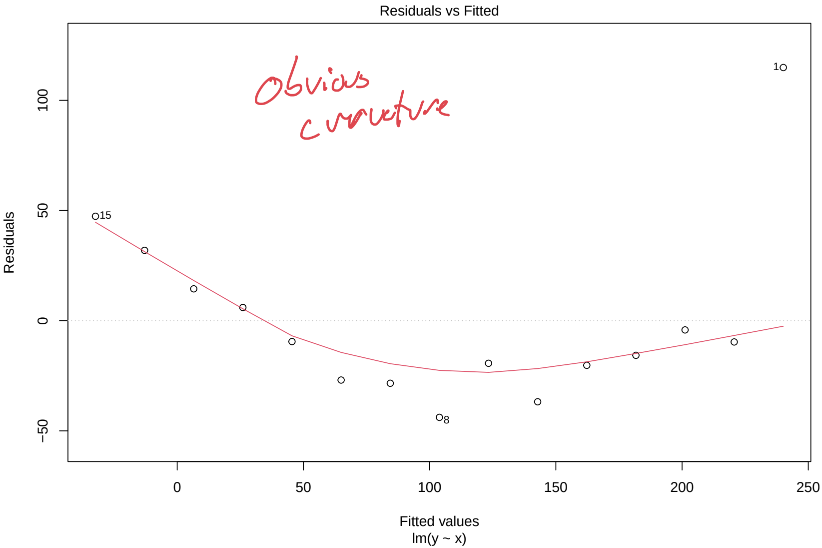
Example 4.7 Solution

```
df <- tibble(x = 1:15,
             y = c(355, 211, 197, 166, 142, 106, 104,
                   60, 56, 38, 36, 32, 21, 19, 15))
linear <- lm(y ~ x, data = df)
summary(linear)
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.867 -23.599  -9.652  10.223 114.883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   259.58      22.73  11.420 3.78e-08 ***
## x             -19.46       2.50  -7.786 3.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.83 on 13 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8098
## F-statistic: 60.62 on 1 and 13 DF,  p-value: 3.006e-06
```

$\hat{\beta}_0 = 259.58, \quad \hat{\beta}_1 = -19.46$

Example 4.7 Solutions



Example 4.7 Solutions

```r
exponential <- lm(log(y) ~ x, data = df)
summary(exponential)
```

```
##
## Call:
## lm(formula = log(y) ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18445 -0.06189  0.01253  0.05201  0.20021
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.973160   0.059778   99.92  < 2e-16 ***
## x           -0.218425   0.006575  -33.22 5.86e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.11 on 13 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.9875
## F-statistic:  1104 on 1 and 13 DF,  p-value: 5.86e-14
```

$\log \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\Rightarrow \hat{y} = \hat{\alpha} e^{\hat{\beta} x}$

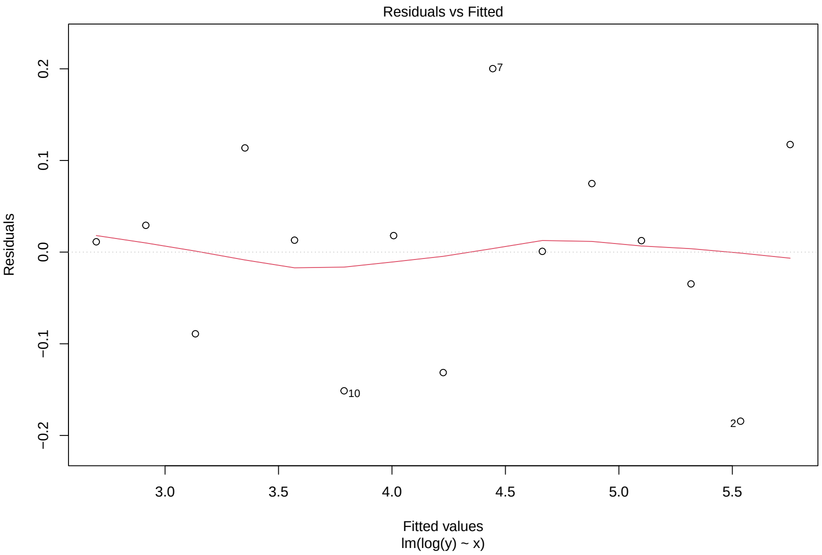
where $\hat{\alpha} = \exp(5.973)$

$\hat{\beta} = -0.2184$

$\hat{\beta}_0 = 5.973$

$\hat{\beta}_1 = -0.2184$

Example 4.7 Solutions



Power regression

In biological sciences it is sometimes possible to relate the weight (or volume) of an organism to some linear measurement such as length (or weight). If Y denotes the weight and x denotes the length, then the model

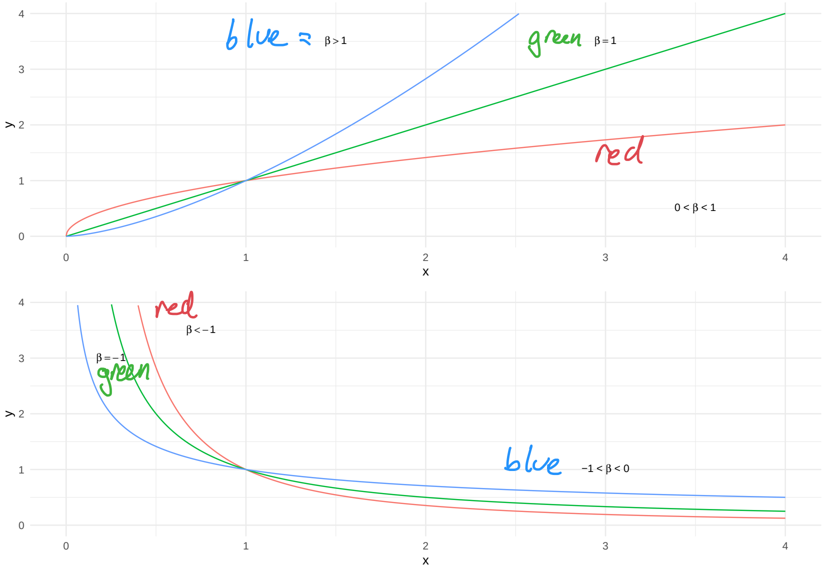
$$Y = \alpha x^{\beta}$$

is often applicable. This model is also known as an *allomertic equation* .

$$\Rightarrow \underbrace{\log y}_{Y^*} = \underbrace{\log \alpha}_{\beta_0} + \underbrace{\beta}_{\beta_1} \underbrace{\log x}_{x^*}$$

$$Y^* \simeq \beta_0 + \beta_1 x^*$$

Power regression



Example 4.8

Alligator	$x = \ln(l)$	$y = \ln(W)$
1	3.87	4.87
2	3.61	3.93
3	4.33	6.46
4	3.43	3.33
5	3.81	4.38
6	3.83	4.70
7	3.46	3.50
8	3.76	4.50
9	3.50	3.58
10	3.58	3.64
11	4.19	5.90
12	3.78	4.43
13	3.71	4.38
14	3.73	4.42
15	3.78	4.25

power regression

$W = \alpha \, l^{\beta}$

$\Rightarrow$

$\log W = \log \alpha + \beta \log(l)$

$\Rightarrow \; y = \beta_0 + \beta_1 x$

Example 4.8

We want to:

a) Fit a power regression model to the data.
b) Find a 90% prediction interval for W if $\log(\ell) = 4$.

$\shortparallel$

$\exp(y)$

Example 4.8 - Solution

a) log d
log W

```
df <- tibble(
  x = c(3.87, 3.61, 4.33, 3.43, 3.81, 3.83, 3.46, 3.76,
        3.50, 3.58, 4.19, 3.78, 3.71, 3.73, 3.78),
  y = c(4.87, 3.93, 6.46, 3.33, 4.38, 4.70, 3.50, 4.50,
        3.58, 3.64, 5.90, 4.43, 4.38, 4.42, 4.25)
)
power_regression <- lm(y ~ x, data = df)
summary(power_regression)
```

$\hat{y} = -8.4761 + 3.4311x$

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24348 -0.03186  0.03740  0.07727  0.12669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.4761     0.5007  -16.93 3.08e-10 ***
## x             3.4311     0.1330   25.80 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1229 on 13 degrees of freedom
## Multiple R-squared:  0.9808, Adjusted R-squared:  0.9794
## F-statistic: 665.8 on 1 and 13 DF,  p-value: 1.495e-12
```

Example 4.8 - Solution

newdata

model

newdata

```
x0 <- tibble(x = 4)
(PI.y <- predict(power_regression, newdata = x0,
                interval = "prediction", level = 0.9))
```

```
##        fit      lwr      upr
## 1 5.248326 5.016355 5.480297
```

90% PI for Y.

```
(PI.w <- exp(PI.y))
```

```
##        fit      lwr     upr
## 1 190.2475 150.8603 239.918
```

since W = exp(Y)

90% for W.

A warning

Back-transforming a prediction interval makes good sense, ***but back-transforming a confidence interval does not!***

See Workshop 9 (Week 10) for a discussion of this.

End Video 1

Begin Video 2

Examples of transformations

1. logarithmic model: $Y = \alpha + \beta \overbrace{\boxed{\log(x)}}^{x^*}$
2. logistic model: $Y = \dfrac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$
3. $Y = \dfrac{\cancel{Y}\,X}{\alpha X - \beta}$

$$\log\left(\frac{Y}{1-Y}\right) = \alpha + \beta x$$

$$\underbrace{\frac{1}{Y}}_{Y^*} = \frac{\alpha X - \beta}{X} = \alpha - \beta \underbrace{\frac{1}{X}}_{X^*}$$

Further examples of linerisable functions

$$Y = \alpha\beta^{x} \qquad Y = \alpha e^{\frac{\beta}{x}}$$

$$Y = \alpha + \frac{\beta}{x} \qquad Y = \frac{\alpha}{\beta + x}$$

$$Y = \alpha + \beta x^{n} \qquad Y = \frac{1}{\alpha + \beta e^{-x}}$$

$$Y = e^{-\alpha x_1 e^{-\frac{\beta}{x_2}}}$$

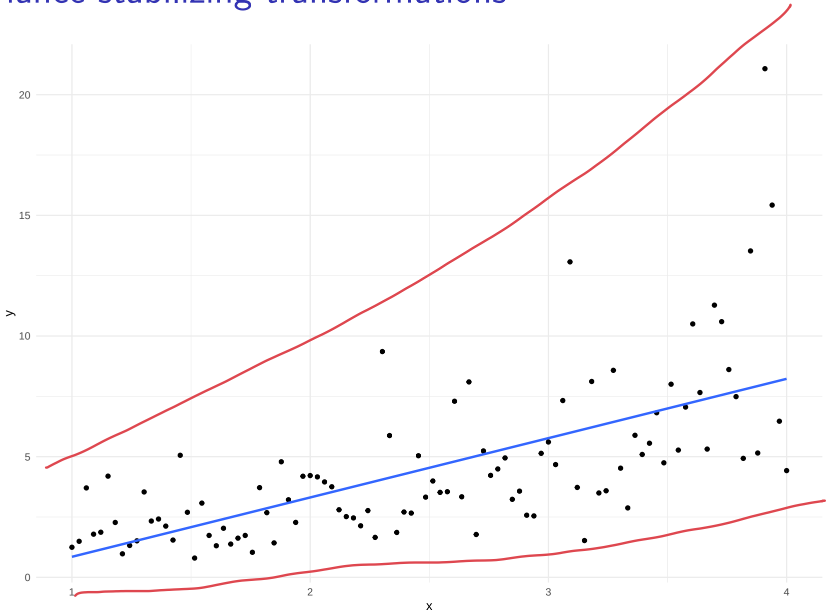
Tute 4 (week 10)

Why transform?

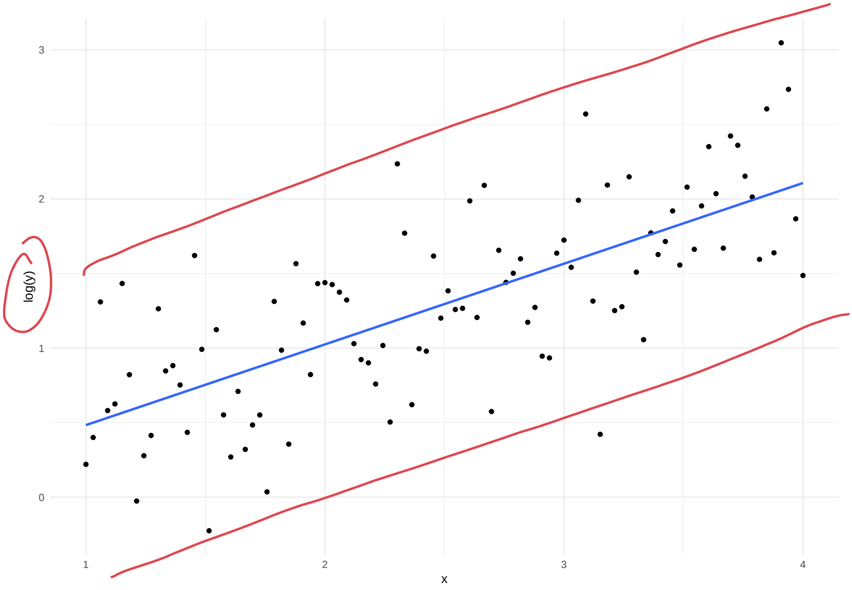
Transformations can be used when the model in terms of the original variables violates one or more of the standard regression assumptions.

- **Linearity**: Theory, scatter plots of the data, or residuals may suggest a non-linear relationship.
- **Non-constant variance**: The response variable Y may have a distribution whose variance is related to the mean. If the mean is related to the predictors, then the variance of Y will change with X. loose accuracy on our estimates.

Variance-stabilizing transformations



Variance-stabilizing transformations



Box-Cox transformations

What happens if normality isn't valid? One method of dealing with this is through *Box-Cox transformations*.

Transform Y into $Y^{(\lambda)}$, where

$$\boxed{Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0\,, \\ \log(Y) & \text{if } \lambda = 0\,. \end{cases}}$$

$*$

SSE does not depend on λ

Fix our X

our relationship $y^{(\lambda)} = \beta_0 + \beta_1 x + \varepsilon$

Which λ?

might do now

1. We can try many different λ values, and choose the value that minimises the SSE.
2. We can use a special thing called *log-likelihood*, which you will see later in the course.
3. This method will be covered in depth in SM III.

a little bit later