

# Statistical Modeling

CH.5 - Transformations

SS 2021 | | Prof. Dr. Buchwitz

Wir geben Impulse

1 Organizational Information

2 Transformations

# Course Contents

Session	Topic
1	Simple Linear Regression
2	Multiple Linear Regression
3	Regression Diagnostics
4	Qualitative Variables as Predictors
5	Transformation of Variables
6	Weighted Least Squares
7	Correlated Errors
8	Analysis of Collinear Data
9	Working with Collinear Data
10	Variable Selection Procedures
11	Logistic Regression
12	Further Topics

1 Organizational Information

2 Transformations

- Data do not always come in suitable form so that they can be analysed right away and often need to be transformed before carrying out an analysis.
- Transformations are necessary because the original variables of the model using these variables, violates one or more of the standard regression assumptions.
- Transformations are usually applied to accomplish objectives such as to **ensure linearity**, to **achieve normality** or to **stabilize the variance**.
- It is common practice to fit a linear regression model to the transformed rather than the original variables.

# Linearity and Non-Linearity

- As mentioned before we consider a model to be linear when the parameters in the model enter in a linear fashion, even if the predictors occur nonlinearly. **All following models are linear.**

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon$$

$$Y = \beta_0 + \beta_1 \sqrt{X} + \epsilon$$

- The following model is **non-linear** as the regression parameter  $\beta_1$  does not enter linearly.

$$Y = \beta_1 + e^{\beta_1 X} + \epsilon$$

## Transformations may be necessary for a variety of reasons:

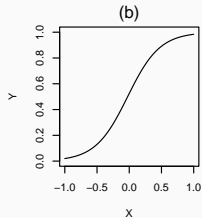
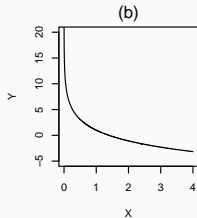
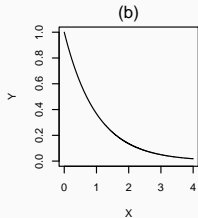
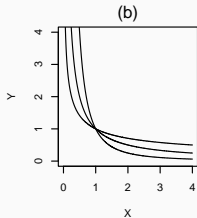
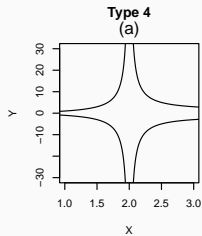
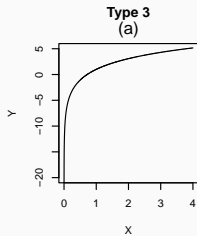
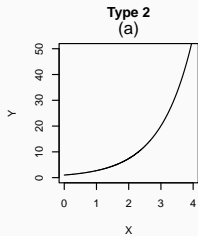
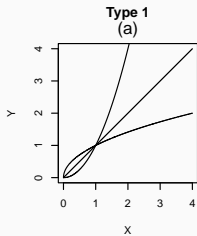
- 1 Theoretical considerations may specify that the relationship between two variables is nonlinear.
- 2 The response variable  $Y$  may have a probability distribution whose variance is related to the mean. When relating  $Y$  and  $X$ , then the variance of  $Y$  will change with  $X$ . The distribution of  $Y$  is often non-normal. This invalidates the standard tests of significance. The unequal variance also leads to inefficient (not smallest variance) estimates of the error term. Transformations that stabilize variances are coincidentally also good normalizing transforms.
- 3 When there is no reason to suspect that a transformation is required, the evidence to apply a transformation comes from inspecting the residuals from a fit with the original variables.

## Transformations to achieve Linearity

- One of the standard assumptions in regression analysis is the linearity of the formed model.
- When analyzing the scatter plot of  $Y$  against  $X_j$  data may appear to be nonlinear.
- The following transformations can be chosen based on the pattern of the Y-X-Scatterplot to linearize the relationship so that linear regression can be applied.



# Transformations to achieve Linearity



## Transformations to achieve Linearity

Function	Transformation	Linear Form	Type
$Y = \alpha X^\beta$	$Y' = \log(Y), X' = \log(X)$	$Y' = \log(\alpha) + \beta X'$	Type 1
$y = \alpha e^{\beta X}$	$Y' = \ln(Y)$	$Y' = \ln(\alpha) + \beta X$	Type 2
$Y = \alpha + \beta \log(X)$	$X' = \log(X)$	$Y = \alpha + \beta X'$	Type 3
$Y = \frac{X}{\alpha X - \beta}$	$Y' = \frac{1}{Y}, X' = \frac{1}{X}$	$Y' = \alpha - \beta X'$	Type 4 a
$Y = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$	$Y' = \ln\left(\frac{Y}{1-Y}\right)$	$Y' = \alpha + \beta X$	Type 4 b

Not every curvature is linearizable! Depending on the observed patterns it may be necessary to choose a different estimation method, which we do not discuss here.

## Example: Bacteria Data

P168

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

### Your turn

$N_t$  Number of surviving bacteria after X-ray exposure of time  $t$ .

$t$  Exposure time to X-rays in minutes.

## Example: Bacteria Data

- The bacteria data was collected to test the “Single-Hit” Hypothesis. The underlying theory (not discussed) states that there is a single vital center in each bacteria that needs to be hit by a X-Ray to inactivate the organism.
- If the theory is applicable the number of surviving bacteria  $\eta_t$  should relate to the exposure time to X-ray  $t$  by

$$\eta_t = \eta_0 e^{\beta_1 \cdot t}$$

- The parameters are  $\eta_0$  and  $\beta_1$  relate to physical quantities.  $\eta_0$  is the number of bacteria at the start of the experiment and  $\beta_1$  is the destruction (decay) rate.

## Example: Bacteria Data

- The relation between  $\eta_t$  and  $t$  cannot be estimated using OLS directly. Therefore we need to apply a transformation by taking logarithms of both sides

$$\ln(\eta_t) = \ln(\eta_0 e^{\beta_1 \cdot t}) = \ln(\eta_0) + \beta_1 t = \beta_0 + \beta_1 t$$

- The presented equation is deterministic as it contains no error. Introducing the error in the linearized equation in an *additive* way, the (transformed) error must occur in multiplicative form in the original equation ( $\epsilon_t = \ln(\epsilon'_t)$ ).

$$\ln(\eta_t) = \beta_0 + \beta_1 t + \epsilon_t \quad \rightarrow \quad \eta_t = \eta_0 e^{\beta_1 \cdot t} \epsilon'_t$$

## Example: Bacteria Data

```
mod1 <- lm(N_t ~ 1 + t, data = P168)      # Inadequate Model  
mod2 <- lm(log(N_t) ~ 1 + t, data = P168) # Adequate Model
```

```
texreg::texreg(list(mod1,mod2), custom.model.names = c("Nt", "log(Nt)"))
```

	Nt	log(Nt)
(Intercept)	259.58*** (22.73)	5.97*** (0.06)
t	-19.46*** (2.50)	-0.22*** (0.01)
R <sup>2</sup>	0.82	0.99
Adj. R <sup>2</sup>	0.81	0.99
Num. obs.	15	15

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$

**Table 3:** Statistical models

## Example: Bacteria Data

- The estimate of the intercept in the equation is the best linear unbiased estimate of  $\ln(\eta_0)$ . Given we are interested in  $\hat{\beta}_0$ , the backtransformation  $e^{\hat{\beta}_0}$  **is not an unbiased estimate of  $\eta_0$ !**

```
exp(coef(mod2)[1]) # Not an unbiased estimate!
```

```
## (Intercept)
##          392.7
```

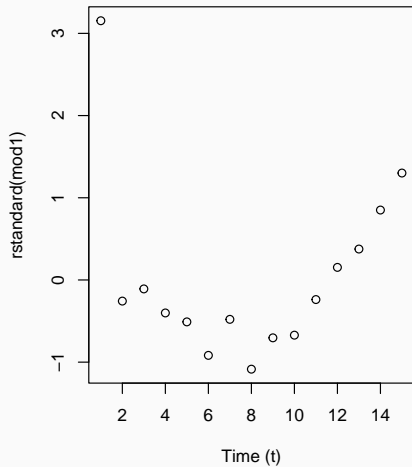
- To obtain a (nearly) unbiased estimate the correction  $\hat{\eta}_0 = \exp(\hat{\beta}_0 - \frac{1}{2}\text{Var}(\hat{\beta}_0))$  can be applied.

```
exp(coef(mod2)[1] - 0.5 * coef(summary(mod2))[, "Std. Error"][1]^2)
```

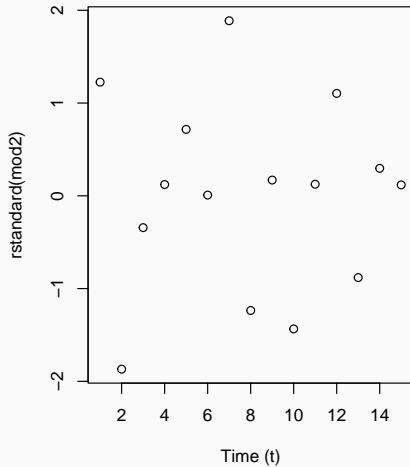
```
## (Intercept)
##          392
```

## Example: Bacteria Data

**Residuals without Transformation**



**Residuals after Transformation**



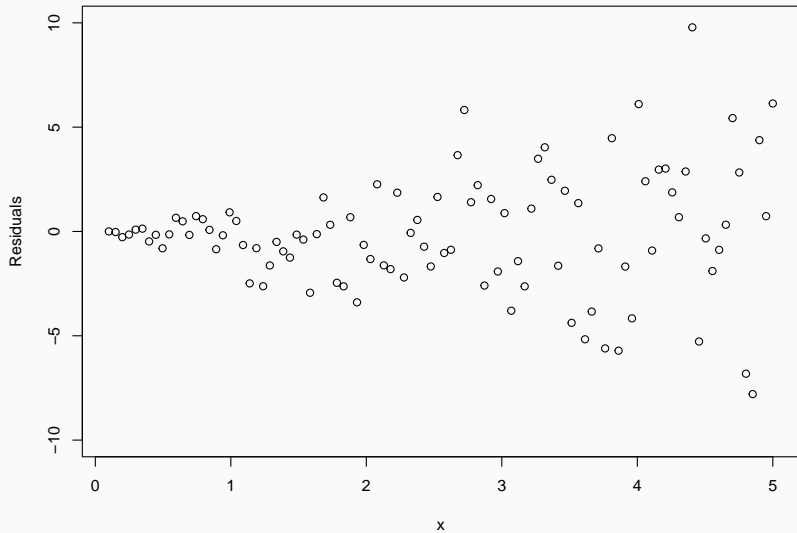


# Heteroscedasticity

# Heteroscedasticity

Constancy of error variance is one of the assumptions of least squares theory. If the error variance is not constant the error is said to be **heteroscedastic**. It is detected by graphs of the residuals against **all** predictors, which usually show a funnel (increase or decrease with  $X$ ).

# Heteroscedasticity



- Heteroscedasticity causes parameter estimates which lack precision in a theoretical sense. The estimated standard errors of the coefficients are often understated, giving a false sense of accuracy.
- The assumed normal distribution has the property that its mean and variance independent in the sense that one is not a function of the other. This is not the case for e.g. the Binomial or Poisson distributions.
- Heteroscedasticity can easily be removed by means of suitable transformations, given that the probability distribution of the response is known.
- The discussed transformations **stabilize the variance** and make the distribution of the transformed variable **closer to the normal distribution**.

# Example: Detection of heteroscedastic Errors

P176

##	X	Y
## 1	294	30
## 2	247	32
## 3	267	37
## 4	358	44
## 5	423	47
## 6	311	49
## 7	450	56
## 8	534	62
## 9	438	68
## 10	697	78
## 11	688	80
## 12	630	84
## 13	709	88
## 14	627	97
## 15	615	100
## 16	999	109
## 17	1022	114
## 18	1015	117
## 19	700	106
## 20	850	128
## 21	980	130
## 22	1025	160
## 23	1021	97
## 24	1200	180
## 25	1250	112
## 26	1500	210
## 27	1650	135

## Your turn

X Number of supervised workers.

Y Number of Supervisors.

## Example: Detection of heteroscedastic Errors

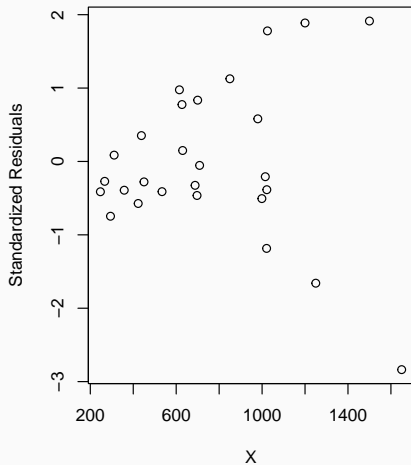
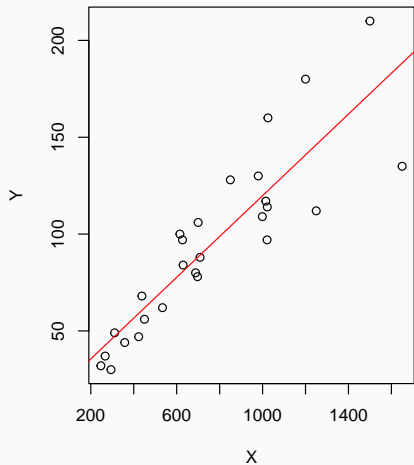
```
mod <- lm(Y ~ 1 + X, data=P176)
texreg::texreg(mod)
```

	Model 1
(Intercept)	14.45 (9.56)
X	0.11*** (0.01)
R <sup>2</sup>	0.78
Adj. R <sup>2</sup>	0.77
Num. obs.	27

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$

**Table 4:** Statistical models

## Example: Detection of heteroscedastic Errors



- In many applications unequal error variance is observed in a for where the variance increases when the predictor variable increases.
- Based on this empirical observation, we can hypothesize that the standard deviation of the residuals is **proportional** to  $X$ .

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\text{with } \text{Var}(\epsilon_i) = k^2 x_i^2 \quad \text{and} \quad k > 0$$



## Removal of heteroscedastic Errors

- Given a proportional relationship between the standard deviation and the predictor indicates that it is beneficial to divide both sides of the regression equation by  $x_i$ :

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\epsilon_i}{x_i}$$

- Defining a new set of variables and coefficients

$$Y' = \frac{Y}{X}, \quad X' = \frac{1}{X}, \quad \beta'_0 = \beta_1, \quad \epsilon' = \frac{\epsilon}{X}$$

yields the new following form:

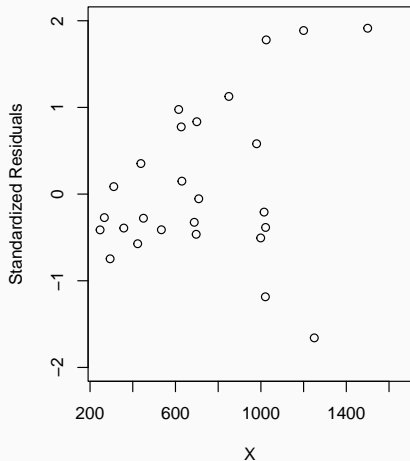
$$y'_i = \beta'_0 + \beta'_1 x'_i + \epsilon'_i$$

- For the transformed model the  $\text{Var}(\epsilon'_i) = k^2$ . If our assumption about the error term fits the model properly, we must work with the transformed variables  $Y/X$  (response) and  $1/X$  (predictor).

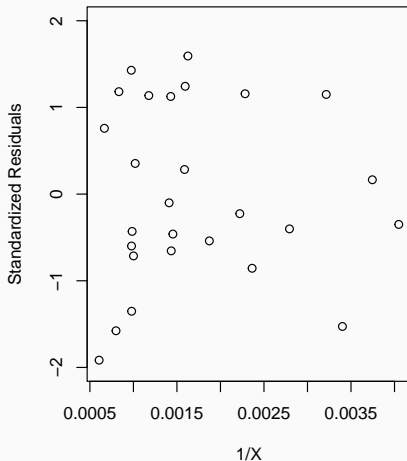
$$\text{Transformed: } \frac{\hat{Y}}{X} = \hat{\beta}'_0 + \frac{\hat{\beta}'_1}{X} \quad \text{Original: } \hat{Y} = \hat{\beta}'_1 + \hat{\beta}'_0 X$$

# Removal of heteroscedastic Errors

**Before Transformation**



**After Transformation**



# Removal of heteroscedastic Errors

```
mod2 <- lm(I(Y/X) ~ 1 + I(1/X), data=P176)
summary(mod2)
```

```
##
## Call:
## lm(formula = I(Y/X) ~ 1 + I(1/X), data = P176)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0415 -0.0138 -0.0050  0.0247  0.0354
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.121      0.009   13.45  6e-13 ***
## I(1/X)         3.803      4.570    0.83   0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0227 on 25 degrees of freedom
## Multiple R-squared:  0.027, Adjusted R-squared: -0.012
## F-statistic: 0.693 on 1 and 25 DF, p-value: 0.413
```

## Note

The results here are expressed in terms of the transformed variables ( $Y/X$  and  $1/X$ ) so measures except the coefficient estimates (their SD, t-values, ...) like  $R^2$  cannot simply be interpreted.

- Linear regression models with heteroscedastic errors can also be fitted by a method called the *weighted least squares* (WLS), where parameter estimates are obtained by minimizing **weighted sum of squares** of residuals.
- The weights in that case are chosen to be inversely proportional to the variance of the errors. In the discussed example, this means

$$\text{WLS: } \sum \frac{1}{x_i^2} (y_i - \beta_0 - \beta_1 x_i)^2 \quad \text{OLS: } \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

# Weighted Least Squares

```
mod.wls <- lm(Y ~ 1 + X, weights = 1/X^2, data=P176)
summary(mod.wls)
```

```
##
## Call:
## lm(formula = Y ~ 1 + X, data = P176, weights = 1/X^2)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0415 -0.0138 -0.0050  0.0247  0.0354
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.803      4.570     0.83   0.41
## X              0.121      0.009    13.45 6e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0227 on 25 degrees of freedom
## Multiple R-squared:  0.879, Adjusted R-squared:  0.874
## F-statistic: 181 on 1 and 25 DF, p-value: 6.04e-13
```

## Note

Performing OLS on the transformed variables  $Y/X$  and  $1/X$  is equivalent to the shown WLS Model.

- The most widely used transformation is the logarithmic transformation, where  $\ln(Y)$  is used as response instead of  $Y$ .

$$\ln(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

- This transformation is particularly useful for variables, where the standard deviation is large compared to the mean.
- Working on a **log scale** has the effect of dampening variability and reducing asymmetry and also reduces heteroskedasticity.
- Results obtained on a log scale are sometimes **harder to interpret** than on the original scale and original variables.

# Logarithmic Transformation

```
mod1 <- lm(log(Y) ~ 1 + X, data=P176)
texreg::texreg(mod1, digits = 8,
               custom.model.names = "log(Y)")
```

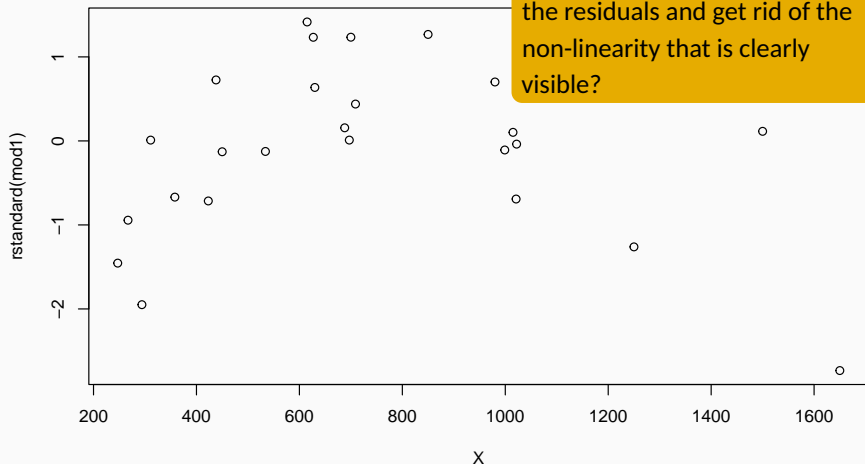
	log(Y)
(Intercept)	3.51502316*** (0.11106702)
X	0.00120408*** (0.00013155)
R <sup>2</sup>	0.77016652
Adj. R <sup>2</sup>	0.76097318
Num. obs.	27
*** $p < 0.001$ ; ** $p < 0.01$ ; * $p < 0.05$	

**Table 5:** Statistical models



# Logarithmic Transformation

```
plot(P176$X, rstandard(mod1), xlab="X")
```



## Your turn

What could you do to improve the residuals and get rid of the non-linearity that is clearly visible?

# Logarithmic Transformation

```
mod1 <- lm(log(Y) ~ 1 + X, data=P176)
mod2 <- lm(log(Y) ~ 1 + X + I(X^2), data=P176)
texreg::texreg(list(mod1,mod2), digits = 8,
                  custom.model.names = c("log(Y)", "log(Y)"))
```

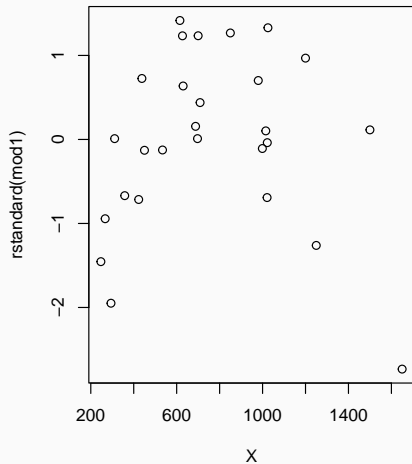
	log(Y)	log(Y)
(Intercept)	3.51502316*** (0.11106702)	2.85160036*** (0.15664013)
X	0.00120408*** (0.00013155)	0.00311267*** (0.00039893)
X <sup>2</sup>		-0.00000110*** (0.00000022)
R <sup>2</sup>	0.77016652	0.88569267
Adj. R <sup>2</sup>	0.76097318	0.87616706
Num. obs.	27	27

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$

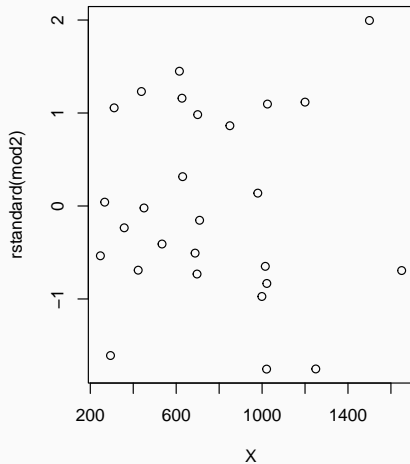
**Table 6:** Statistical models

# Logarithmic Transformation

Without quadratic Term



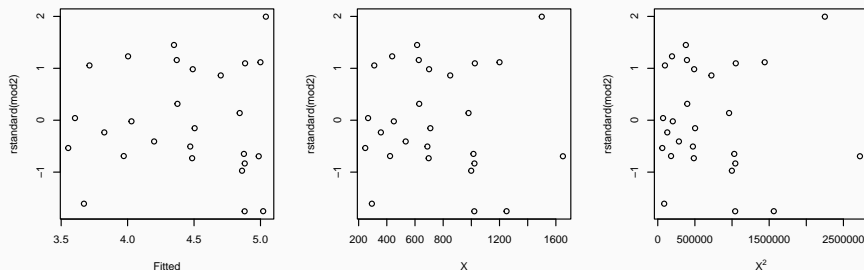
With quadratic Term



# Logarithmic Transformation

- Residuals for the model  $\ln(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$  appear satisfactory. There is no appearance of heteroscedasticity or non-linearity in the residuals.

Diagnostic Plots for Model with quadratic Term



Applying different transformation may yield multiple acceptable candidates, which all may be used as final models.

- The common transformations  $\ln(Y)$ ,  $1/Y$  and  $\sqrt{Y}$  can be seen as special cases of the so called **power transformation**.

$$Y^\lambda$$

- It is also common to use the Box-Cox-Transformation  $(Y^\lambda - 1)/\lambda$  which approaches  $\log(Y)$  as  $\lambda$  approaches 0.
- Reciprocal ( $\lambda = -1$ ), square root ( $\lambda = 0.5$ ) and logarithmic transformation ( $\lambda = 0$ ) can all be modeled within this framework.
- Choosing transformations based on empirical evidence to achieve normality and/or to stabilize the error variance may require **experimentation** with different power transforms.
- Typical values for  $\lambda$  are between -2 and 2 and should be sufficient for most practical use cases.

# Example: Brain Data

P184

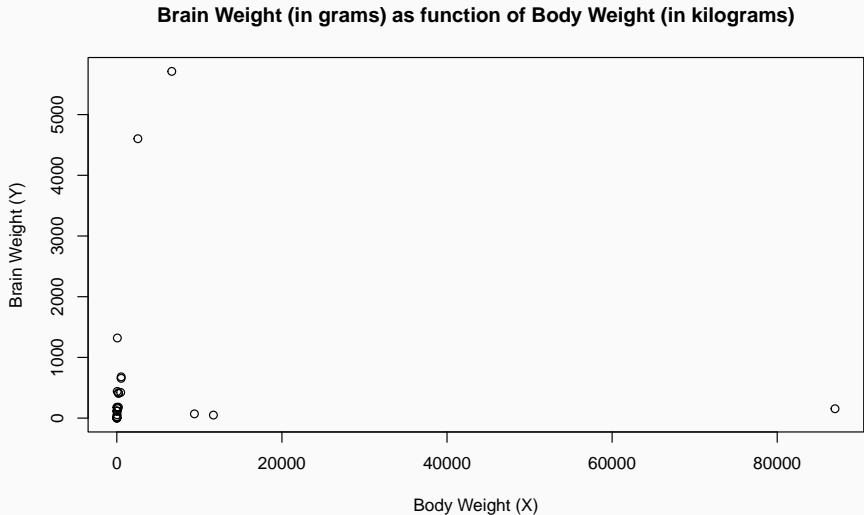
##	BrainWeight	BodyWeight
## Mountain beaver	8.1	1.350
## Cow	423.0	465.000
## Graywolf	119.5	36.330
## Goat	115.0	27.660
## Guineapig	5.5	1.040
## Diplodocus	50.0	11700.000
## Asian elephant	4603.0	2547.000
## Donkey	419.0	187.100
## Horse	655.0	521.000
## Potar monkey	115.0	10.000
## Cat	25.6	3.300
## Giraffe	680.0	529.000
## Gorilla	406.0	207.000
## Human	1320.0	62.000
## African elephant	5712.0	6654.000
## Triceratops	70.0	9400.000
## Rhesus monkey	179.0	6.800
## Kangaroo	56.0	35.000
## Hamster	1.0	0.120
## Mouse	0.4	0.023
## Rabbit	12.1	2.500
## Sheep	175.0	55.500
## Jaguar	157.0	100.000
## Chimpanzee	440.0	52.160
## Brachiosaurus	154.5	87000.000
## Rat	1.9	0.280
## Mole	3.0	0.122
## Pig	180.0	192.000

## Your turn

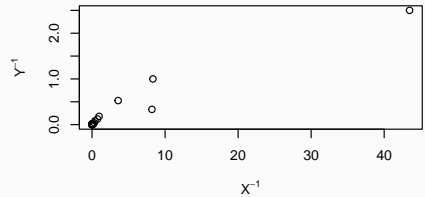
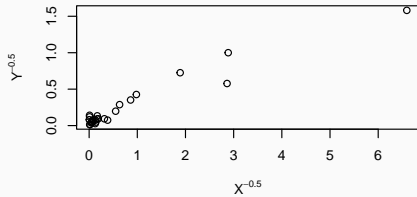
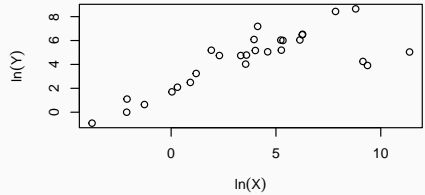
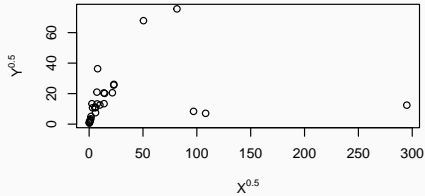
BrainWeight Brain Weight of the animal in grams.

BodyWeight Body Weight of the respective animal in kilograms.

## Example: Brain Data



# Power Transformation





- The logarithmic transformation ( $\lambda = 0$ ) is the most appropriate one for the data.
- In the case of  $\lambda = 0$  the relationship looks linear, but three data points (dinosaurs) deviate from the other observations.
- In the example the power transformation has been applied to  $X$  and  $Y$  simultaneously and with the same value of  $\lambda$ . In practice it may be more appropriate to raise each value to a different power, choose the values **independently** or transform only a single variable.
- Heteroscedasticity and non-linearity can be diagnosed by checking the residuals of the model. The final model (with applied transformations) should not show evidence of heteroscedasticity or deterministic patterns.