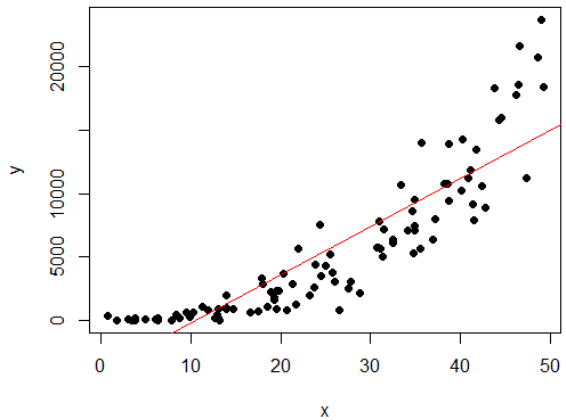
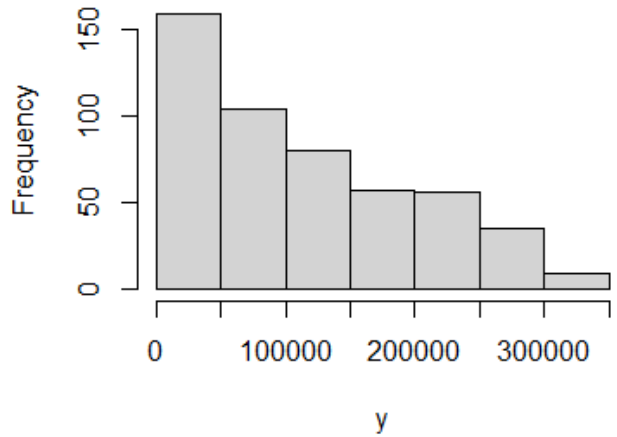
Transformation of variables

Manjari Das

Learning objectives

- Understand the importance of variable transformations in data analysis
  - to meet model assumptions, example linearity
  - to deal with skewed data.





- Interpret residual analysis (post-regression diagnostics) to check model validity.
- Identify associations that are non-linear.
- Get familiar with Box-Cox method to decide on a variable transformation strategy.

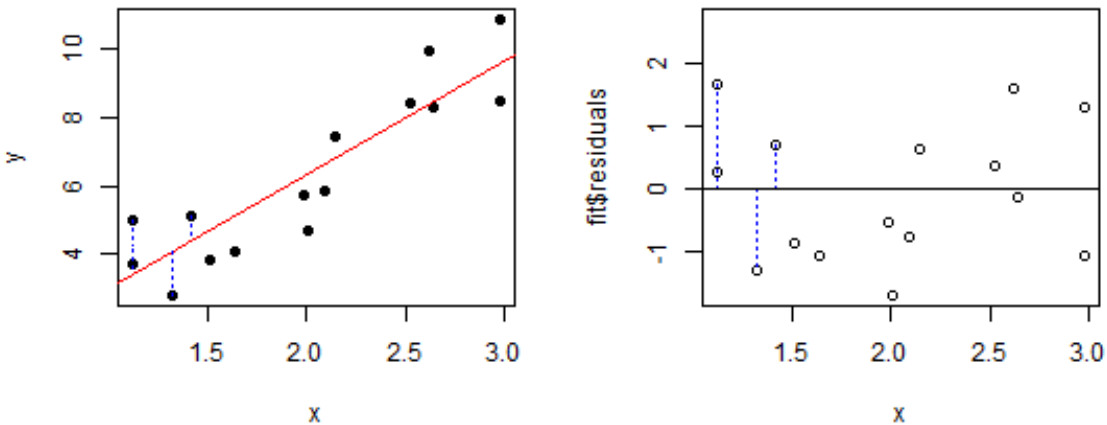
Simple Linear Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

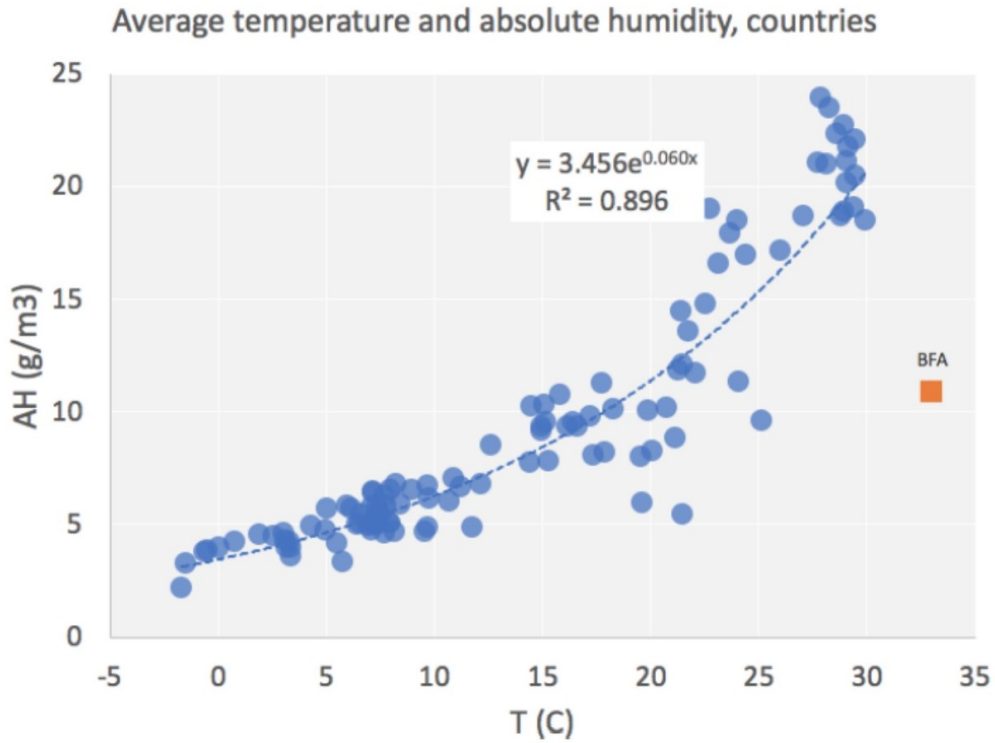
$$yield = \beta_0 + \beta_1 rainfall + \beta_2 temperature + \beta_3 fertilizer + \epsilon$$

The basic rules for linear regression:

1. The dependent variable has a nearly linear relationship with the independent variable(s).
2. The errors are approximately normal $\mathcal{N}(0, \sigma^2)$.
3. The error have no correlation with the covariates.



Deviation from assumptions



[1]

The relation between absolute humidity and absolute temperature is not linear. Its roughly exponential.
If we try the model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

it can never capture the exponential pattern.

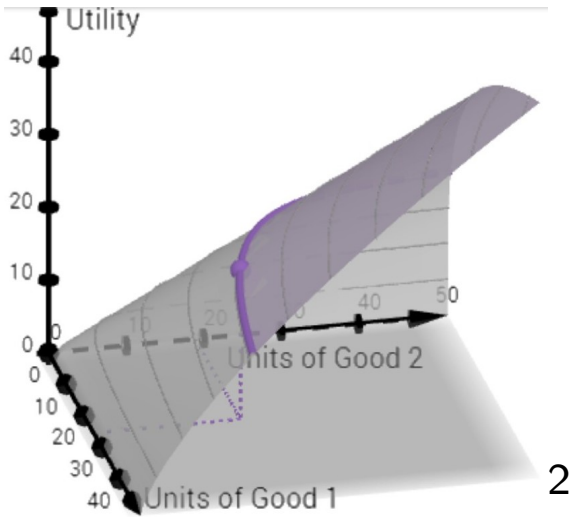
[1] Renato H.L. Pedrosa, 2020, The dynamics of Covid-19

Deviation from assumptions

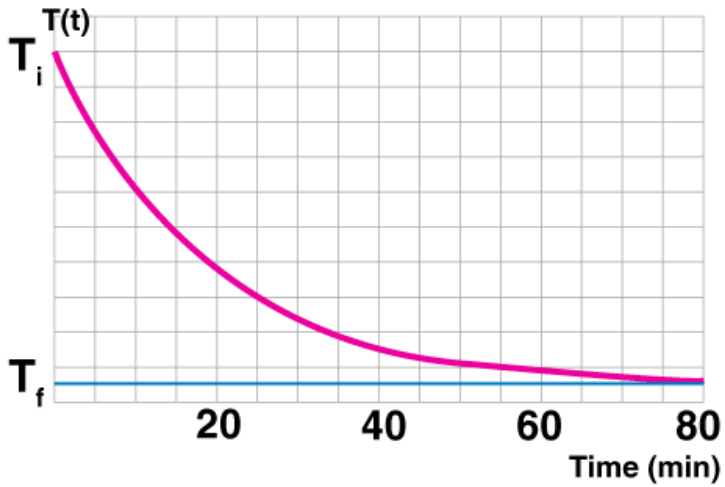
Seldom does linear relationship hold in real data.

Some more examples

1. Economics: Cobb-Douglas utility curve
$$U(x_1, x_2) = x_1^{\alpha} x_2^{\beta}$$



[2]



Newton's Law of Cooling – Temperature vs Time [3]

2. Thermodynamics: Rate of heat loss vs time.
3. Biology: Bacterial growth rate vs time.

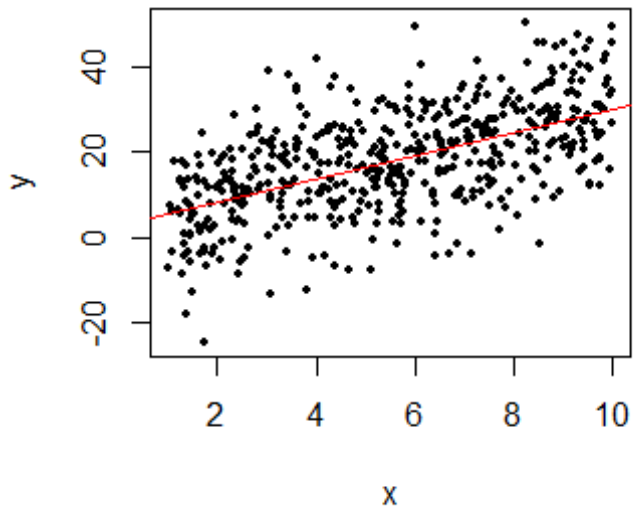
[2] econographs.org

[3] cdn.byjus.com

How to check quality of fit and assumption violation?

Code to fit model

```
> plot(x, y, pch = 19)
> fit <- lm(y~x)
> abline(fit, col = "red")
```



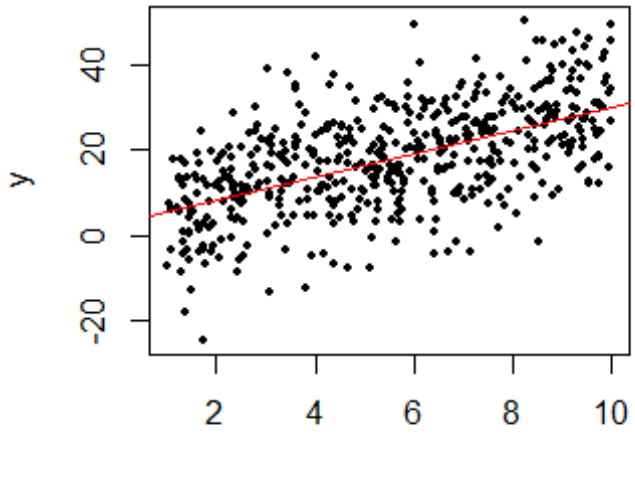
To check model validity we need the following

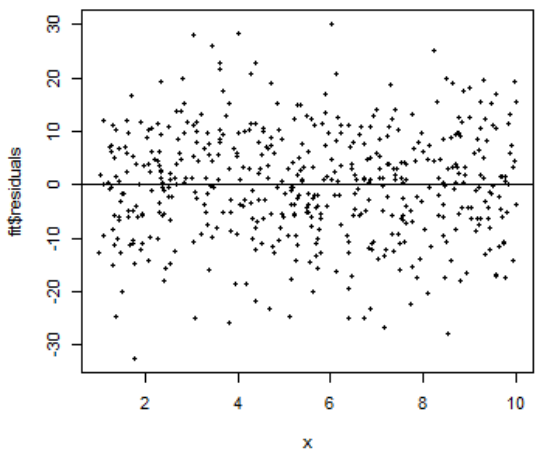
- the residuals are uniform scattered around zero. $\Rightarrow$ zero mean
- the residuals $(Y_i - \widehat{Y}_i)$ are consistently spread across covariates and/or fitted values. $\Rightarrow$ same variance, σ^2
- no visible pattern in the residual vs covariate (or fitted values) plot. $\Rightarrow$ no unmodelled relation
- the residuals are normally distributed. $\Rightarrow N(0, \sigma^2)$

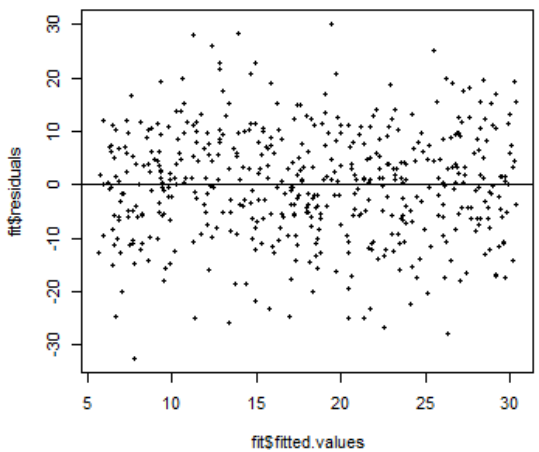
How to check quality of fit and assumption violation?

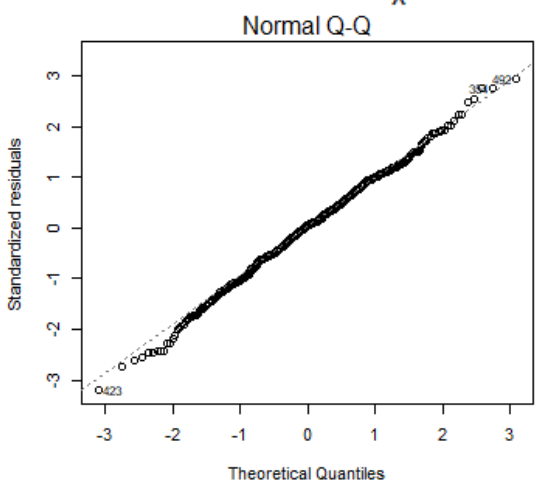
Code for residual analysis

```
 #residual vs covariate
> plot(x, fit$residuals)
 #residual vs fitted and Q-Q plot
> plot(fit, which = c(1,2))
```









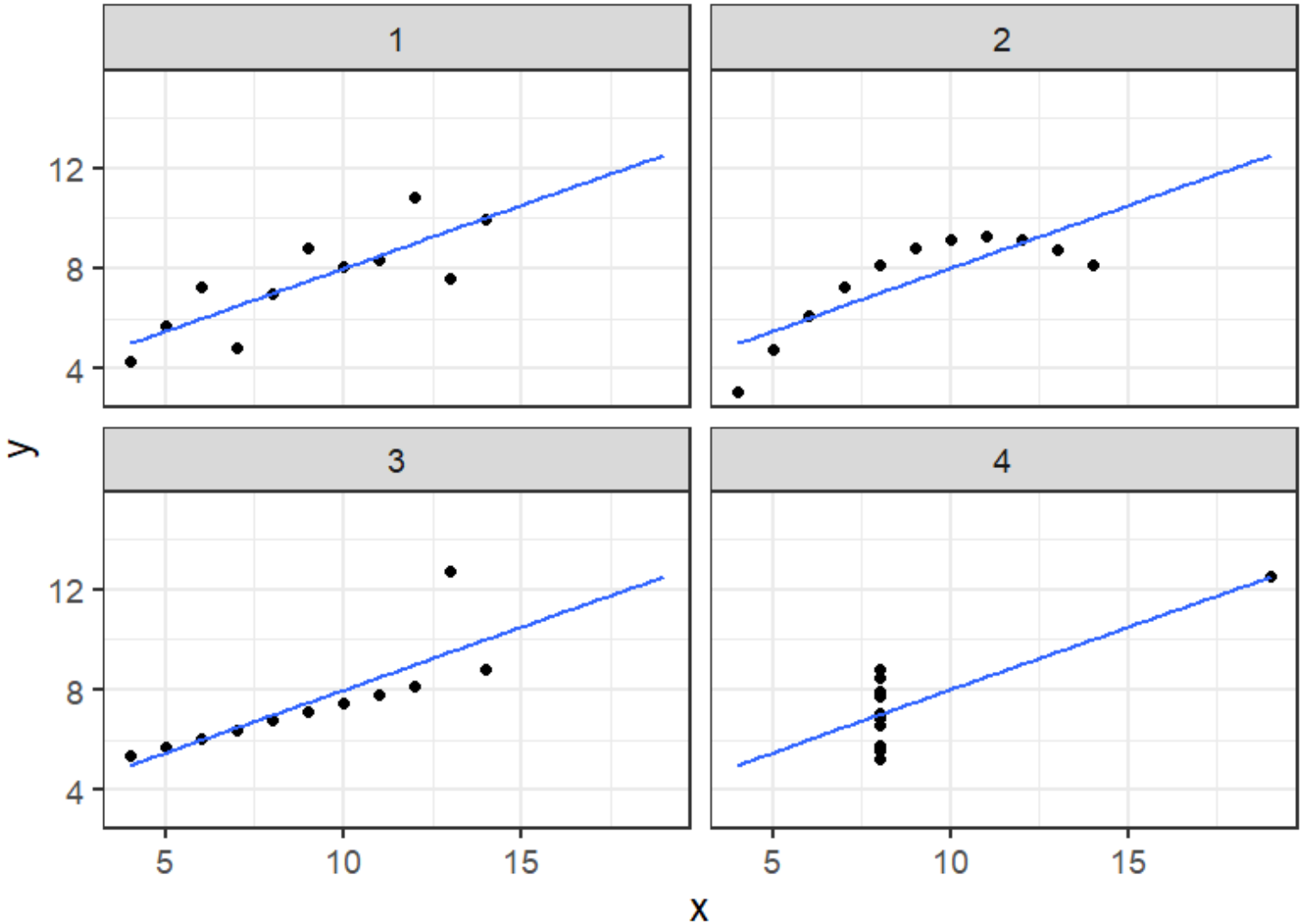
Check residuals' pattern

Expect: points spread symmetrically around $y = 0$

Check normality

Expect: points align with $y = x$

Linear model applied to non-linear data (Anscombe's)



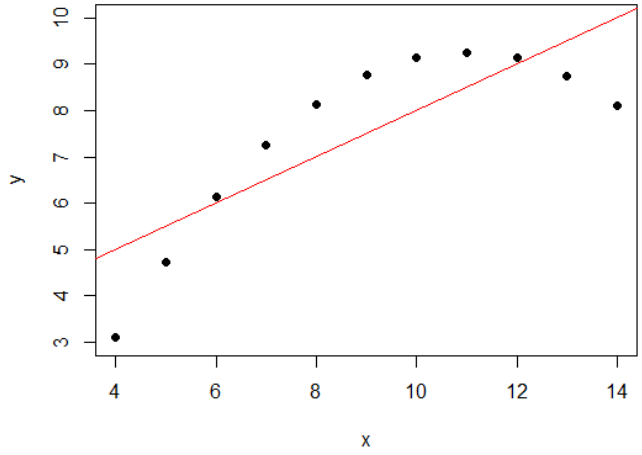
All four examples have the same $R^2 = 0.67$ and correlation $= 0.816$.

Moral of the story: Always plot the data to check fit.

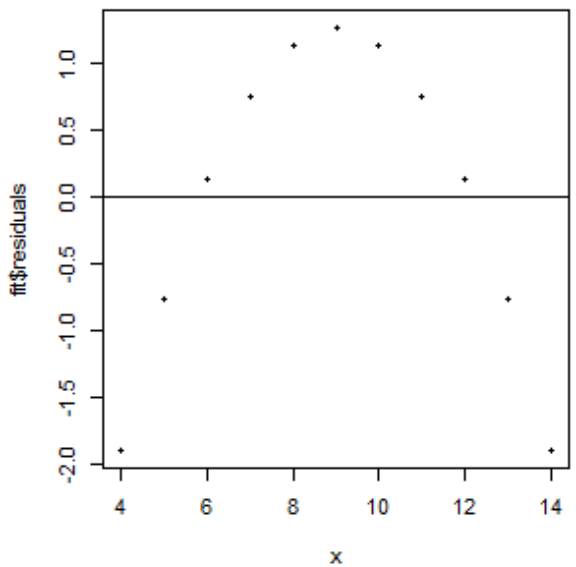
Diagnostic plots for Anscombe's second dataset

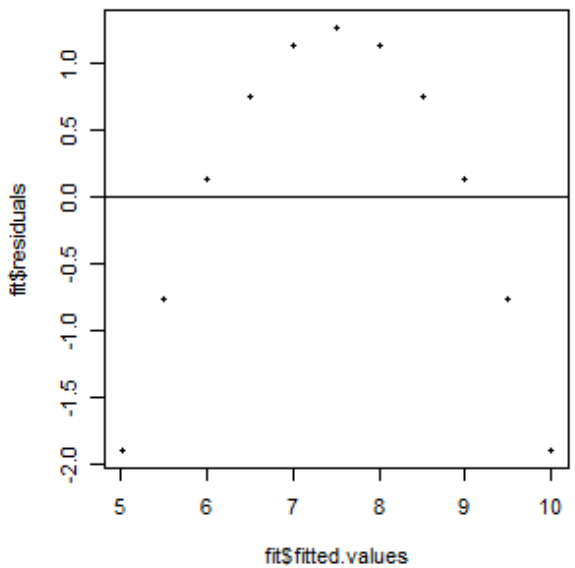
```
> plot(x, y, pch = 19)
> fit <- lm(y~x)
> abline(fit, col = "red")

> plot(x, fit$residuals)
> plot(fit, which = c(1,2))
```

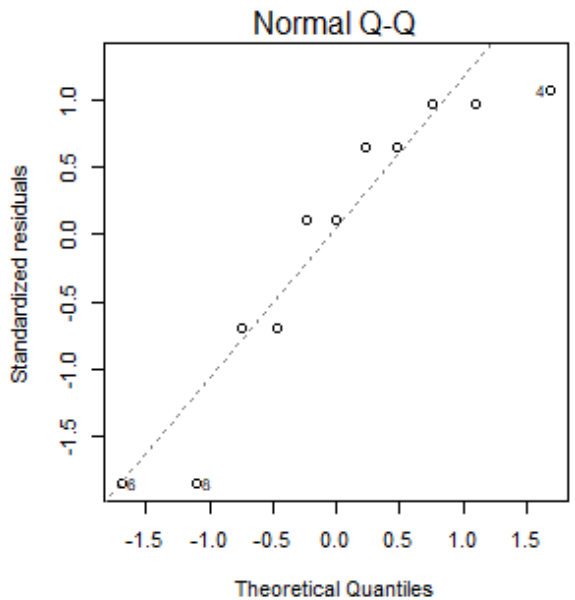


Check residuals' pattern





Check normality

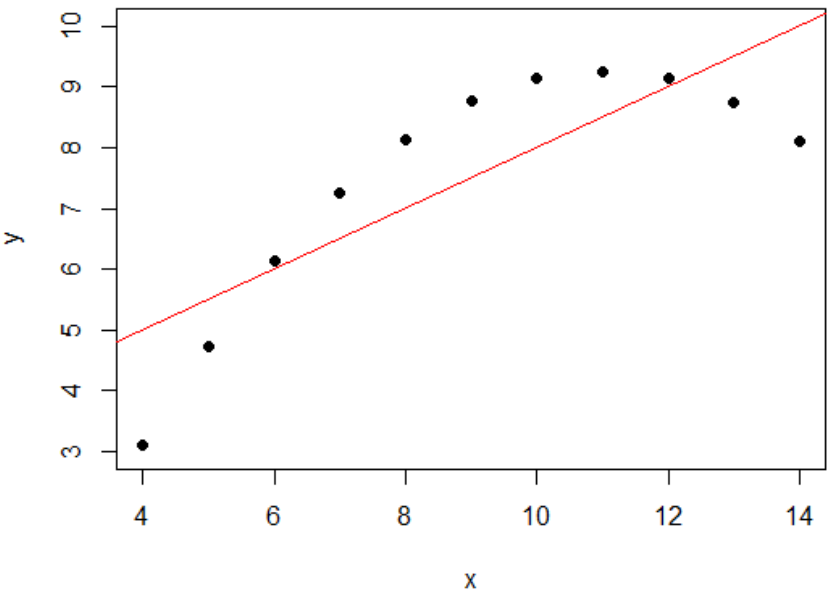


Dangers of fitting linear model to non-linear data

$\widehat{y} = 0.3001 + 0.5x$

Fit says y increases with x.

Reality: y starts decreasing if $x > 11$.

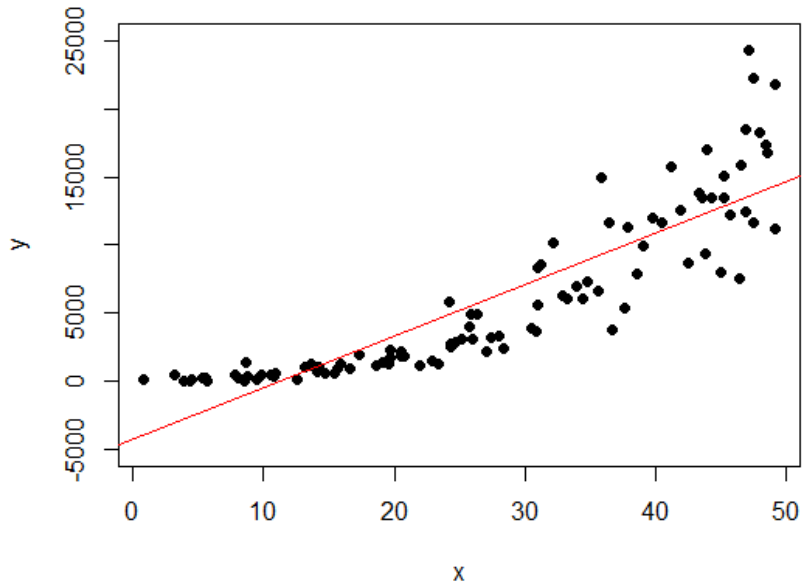


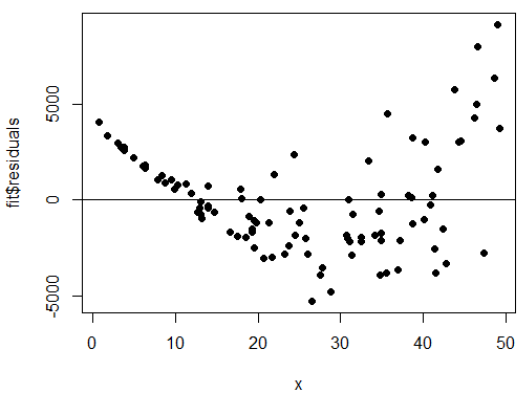
1. Wrong inference.
2. Harmful in policy making.
3. Loss of resources.

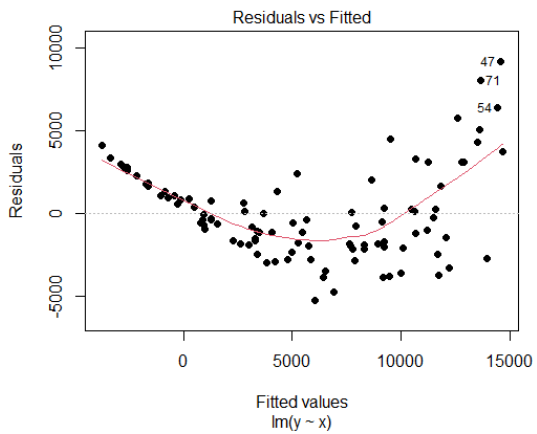
Another non-linear data example

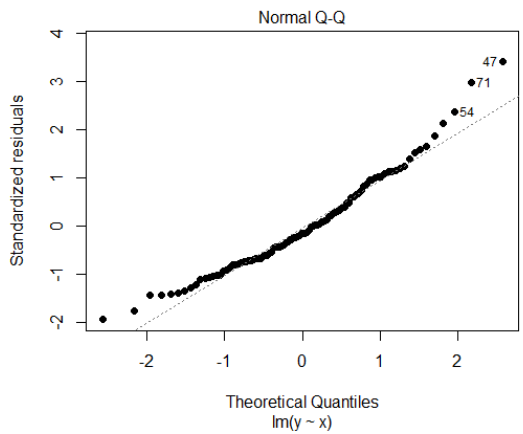
```
> plot(x, y, pch = 19)
> fit <- lm(y~x)
> abline(fit, col = "red")

> plot(x, fit$residuals)
> plot(fit, which = c(1,2))
```





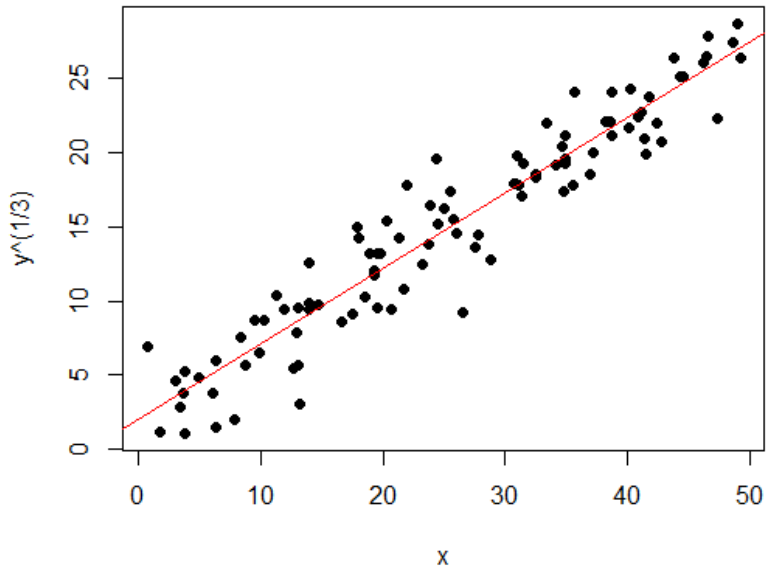


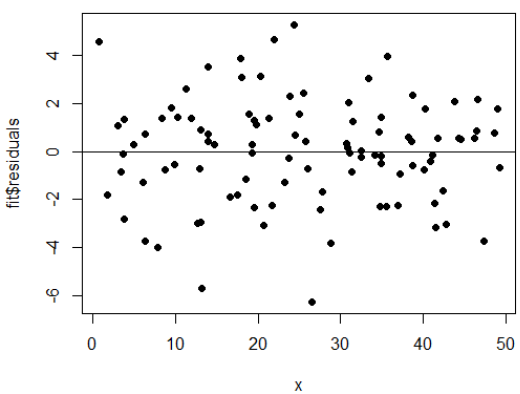


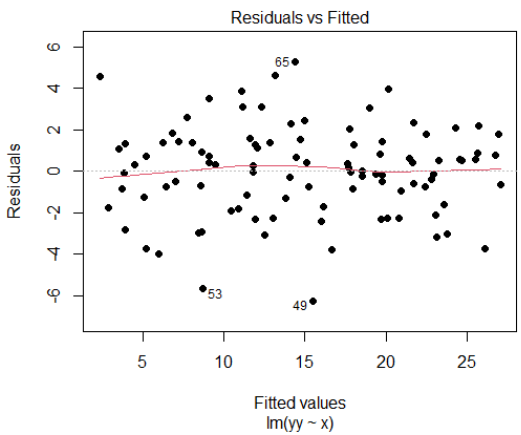
After transforming $f(y) = y^{\frac{1}{3}}$ on y: $y \longrightarrow y^{1/3}$

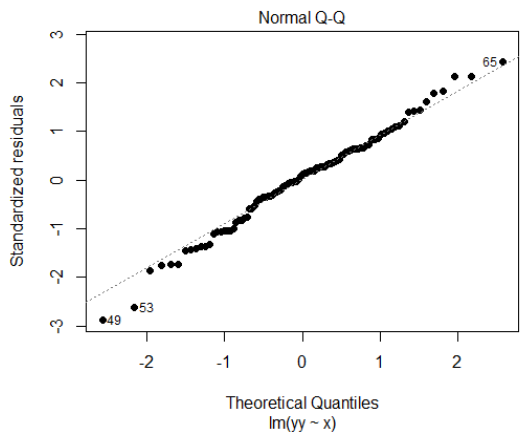
```
> yt <- y^(1/3)
> plot(x, y^(1/3), pch = 19)
> fit <- lm(yt~x)
> abline(fit, col = "red")

> plot(x, fit$residuals)
> plot(fit, which = c(1,2))
```









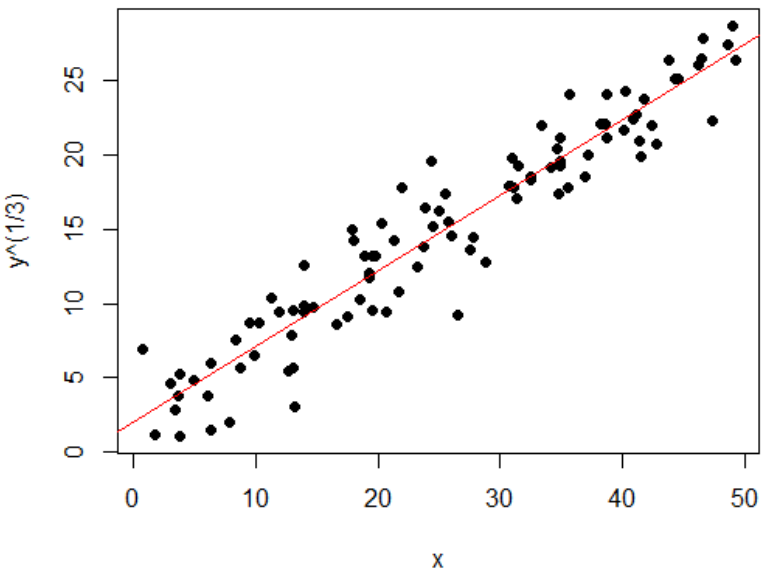
After transforming $f(y) = y^{\frac{1}{3}}$ on y: $y \longrightarrow y^{1/3}$

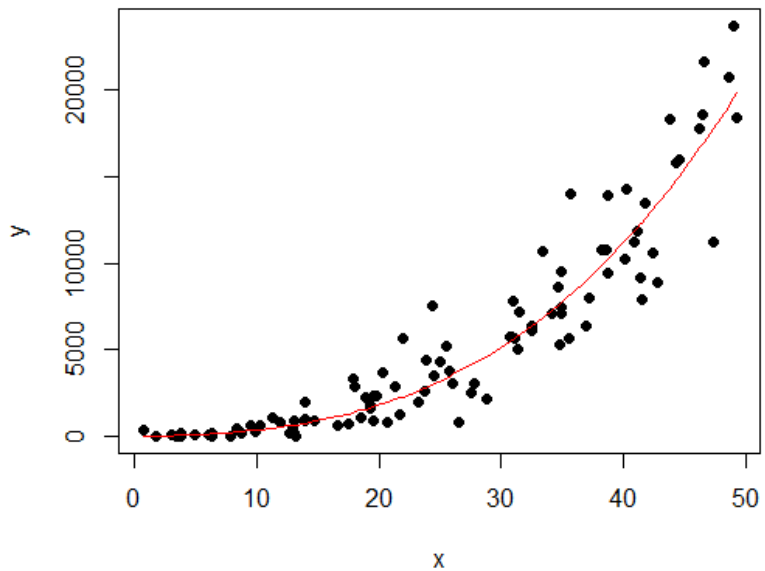
To get back fitted value for original data, apply the reverse transformation $f^{-1}(y) = y^3$ on the fitted values.

```
> y.fit <- (fit$fitted.values)^3
```

The following code plots the fitted line for the original data after applying f^{-1}.

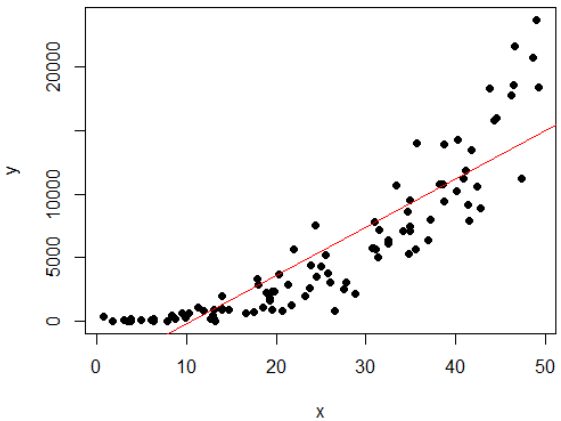
```
> plot(x, y, pch = 19)
> lines(x, fit$fitted.values^3, col = "red")
```



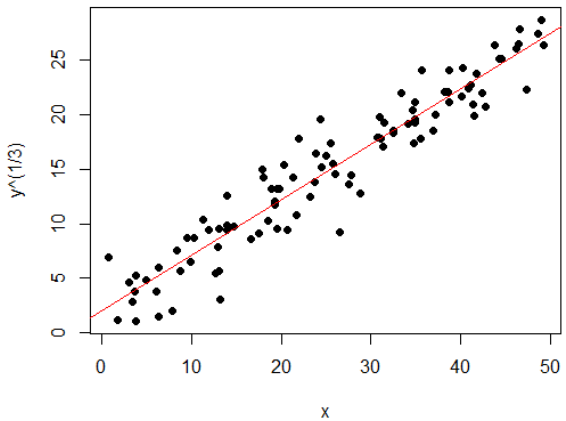


Reasons to apply transformation

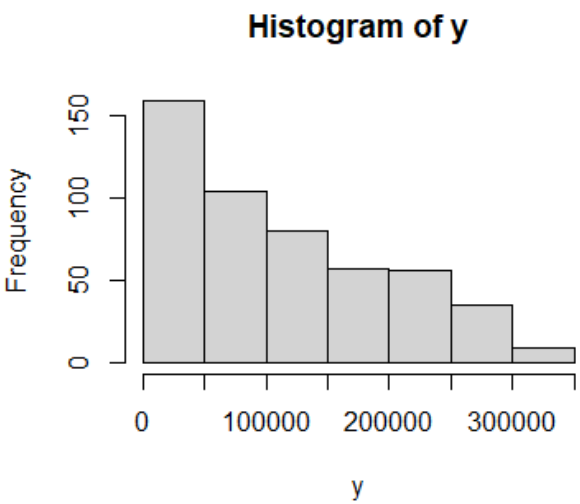
In model fitting: for linearity



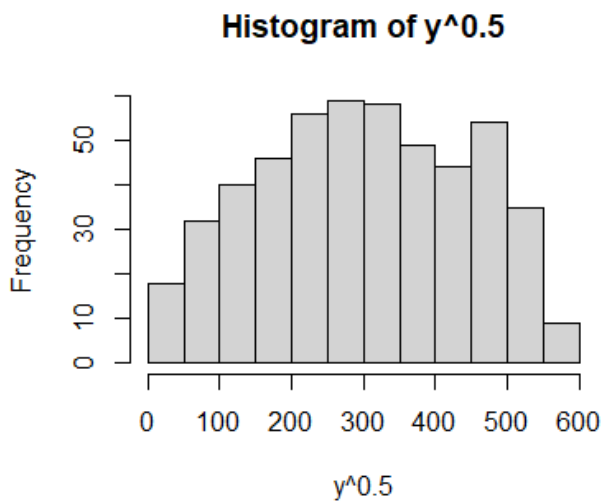
Apply
$yt = y^{1/3}$



On skewed data: for symmetric spread or confidence interval



Apply
$yt = y^{0.5}$

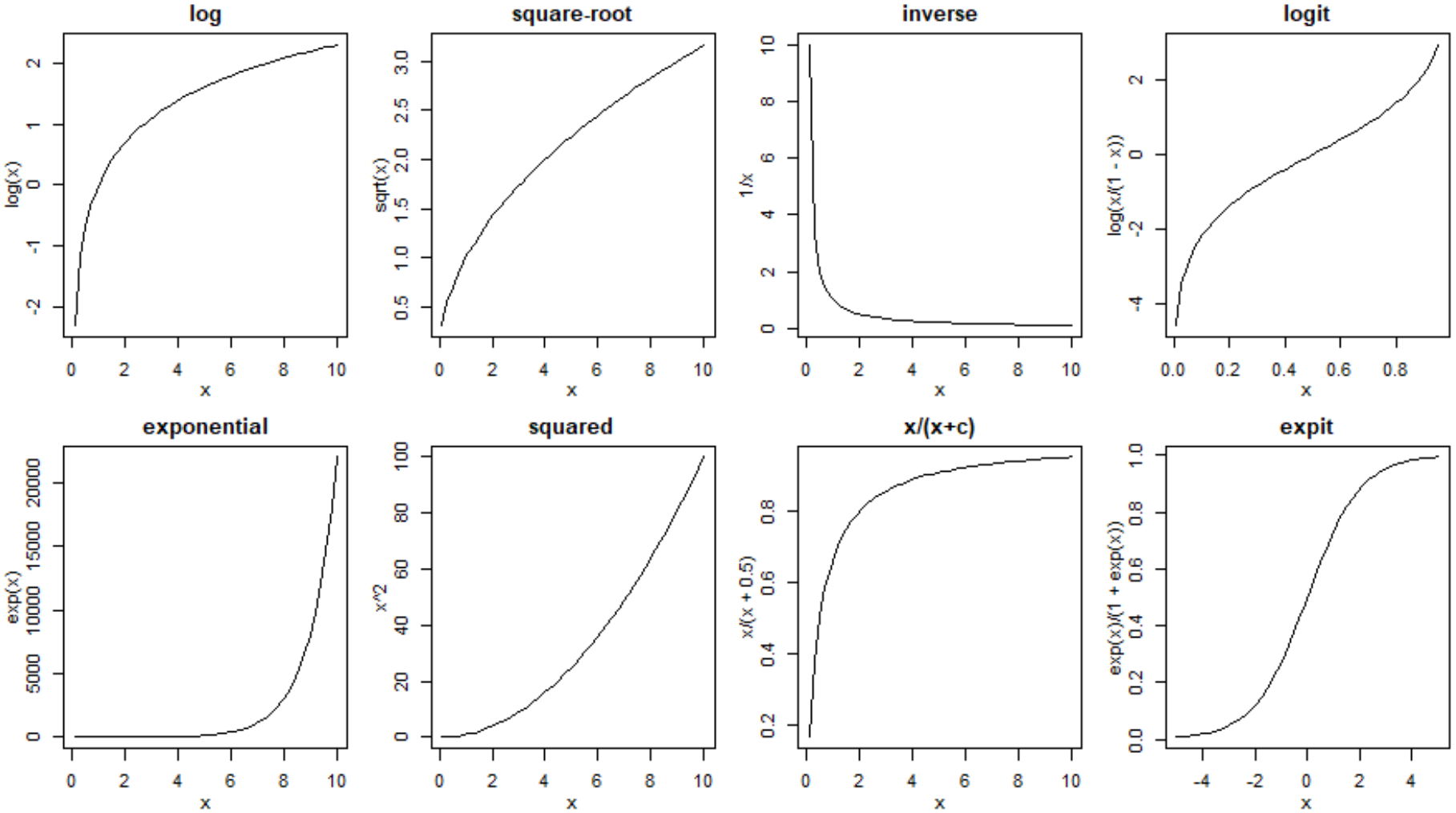


Reason for variable transformation

- Make non-linear to linear for *regression*.
  - Transform both x's and Y.
  - Transform only x.
  - Transform only Y.
- Make skewed data unskewed or bell-shaped for or better spread and avoid outliers.

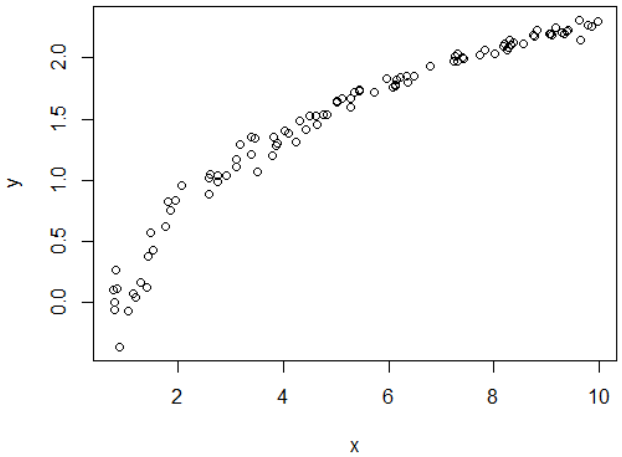
Common transformations (example)

$log(x)$, $exp(x)$, $\sqrt{x}$, x^2, $1/x$, $x/(x+c)$, *logit* and *expit*.



Play activity

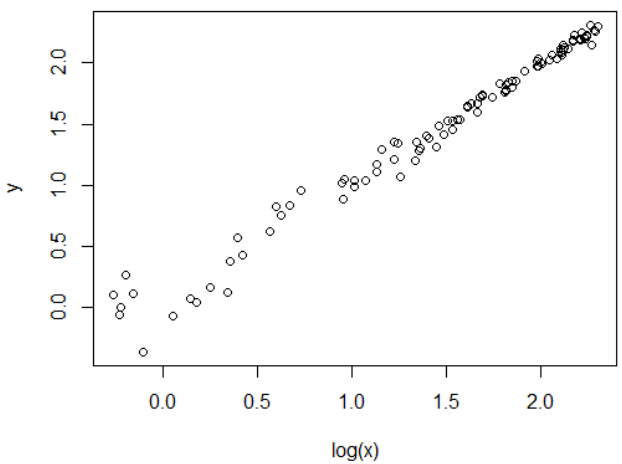
An example: the relationship is roughly like *log*

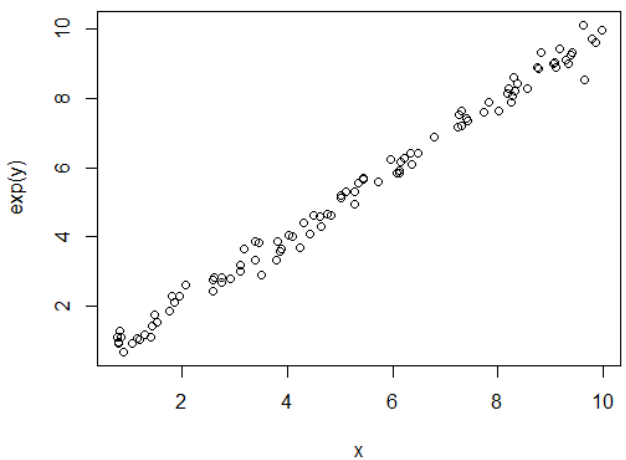


$y \propto log(x)$.

Two options to transform:

1. $x \longrightarrow log(x)$
2. $y \longrightarrow exp(y)$





Open Rstudio

Run `shiny::runGitHub("shiny_apps", "mqnjqrid")`

Cobb-Douglas example

Cobb-Douglas utility function:

$$Original: U(x_1, x_2) = x_1^{\alpha} x_2^{\beta}$$

Hence, the new model is

$$Y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \epsilon.$$

```
fit <- lm(y ~ x1tilde + x2tilde)
fit <- lm(y ~ 0 + x1tilde + x2tilde)  #no intercept model
```

Estimated utility

$$\widehat{U(x_1, x_2)} = exp(\widehat{Y}) = exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 log(x_1) + \widehat{\beta}_2 log(x_2)\right\}.$$

Cobb-Douglas example

Cobb-Douglas utility function:

$$Original: U(x_1, x_2) = x_1^{\alpha} x_2^{\beta}$$

Hence, the new model is

$$Y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \epsilon.$$

```
fit <- lm(y ~ x1tilde + x2tilde)
fit <- lm(y ~ 0 + x1tilde + x2tilde)  #no intercept model
```

Estimated utility

$$\widehat{U(x_1, x_2)} = exp(\widehat{Y}) = exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 log(x_1) + \widehat{\beta}_2 log(x_2)\right\}.$$

Cobb-Douglas example

Cobb-Douglas utility function:

$$\begin{aligned} \textit{Original} &: U(x_1, x_2) = x_1^{\alpha} x_2^{\beta} \\ \textit{Transformed} &: log\{U(x_1, x_2)\} = \alpha log(x_1) + \beta log(x_2). \end{aligned}$$

Hence, the new model is

$$Y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \epsilon.$$

```
fit <- lm(y ~ x1tilde + x2tilde)
fit <- lm(y ~ 0 + x1tilde + x2tilde)  #no intercept model
```

Estimated utility

$$\widehat{U(x_1, x_2)} = exp(\widehat{Y}) = exp\left\{\widehat{\beta}_0 + \widehat{\beta}_1 log(x_1) + \widehat{\beta}_2 log(x_2)\right\}.$$

Examples to make relation linear

Sometimes Y is a function of multiple covariates which is not a linear combination like

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

If we have a rough idea about the relation as in Cobb-Douglas, we can transform appropriately to make the effects of the covariates add up in a linear manner.

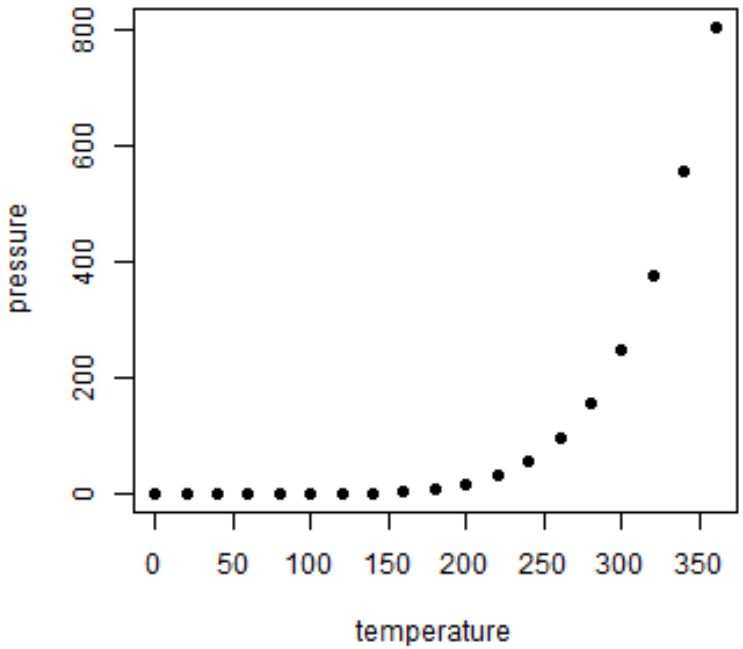
- $Y = \alpha x_1 \gamma^{\beta x_2} \quad \longrightarrow \quad log(Y) = log(\alpha) + log(x_1) + log(\gamma^{\beta}) x_2$

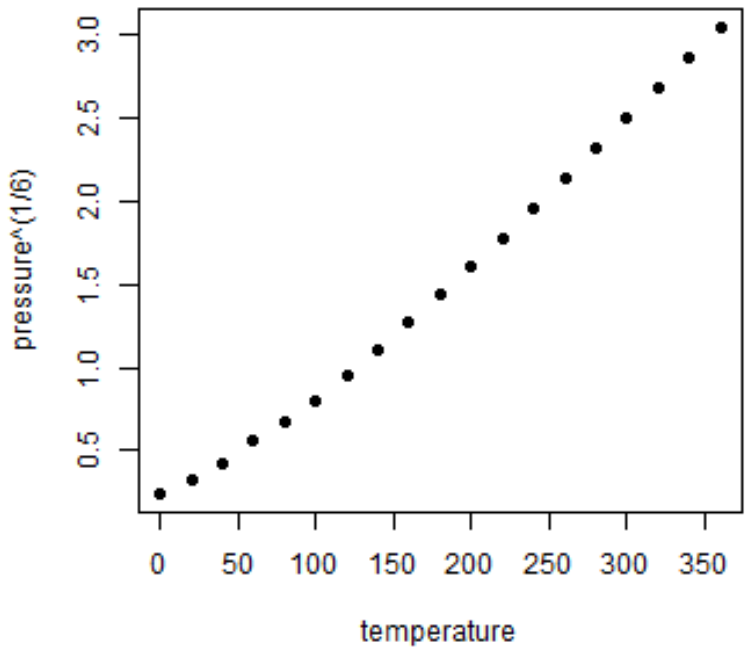
  Regress $log(Y)$ on $log(x_1)$ and x_2.
- $Y = e^{\alpha {x_1}^{\beta} {x_2}^{\gamma}} \rightarrow log(log(Y)) = log(\alpha) + \beta log(x_1) + \gamma log(x_2)$

  Regress $log(log(Y))$ on $log(x_1)$ and $log(x_2)$.

More non-linear examples

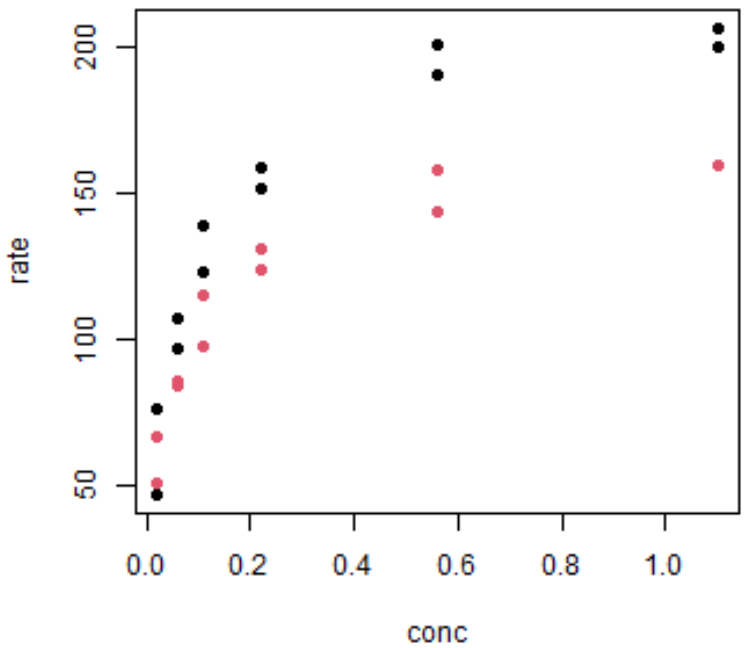
R dataset: pressure

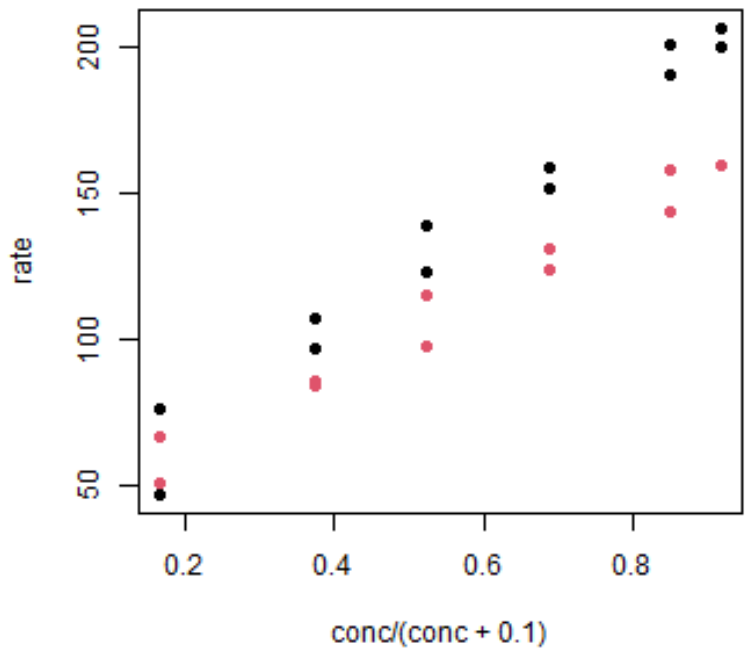




More non-linear examples

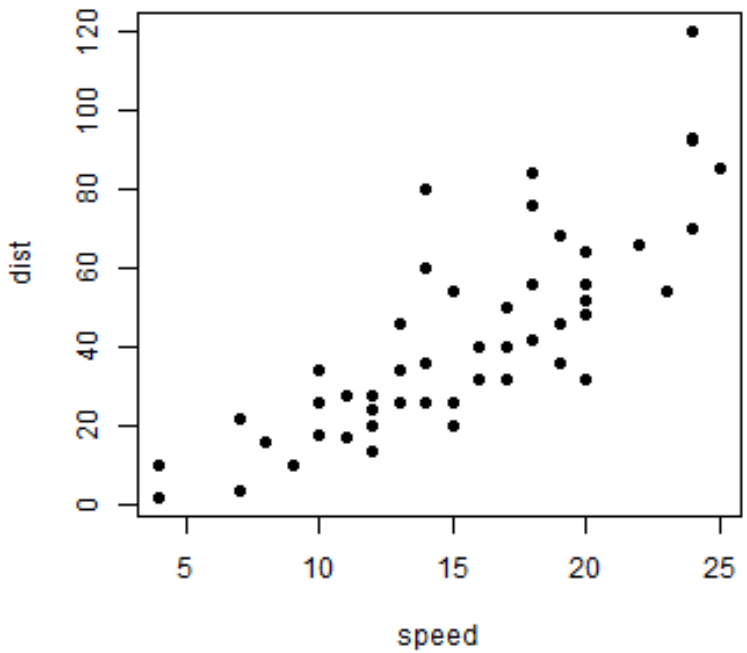
R dataset: Puromycin (black: treated, red: untreated)

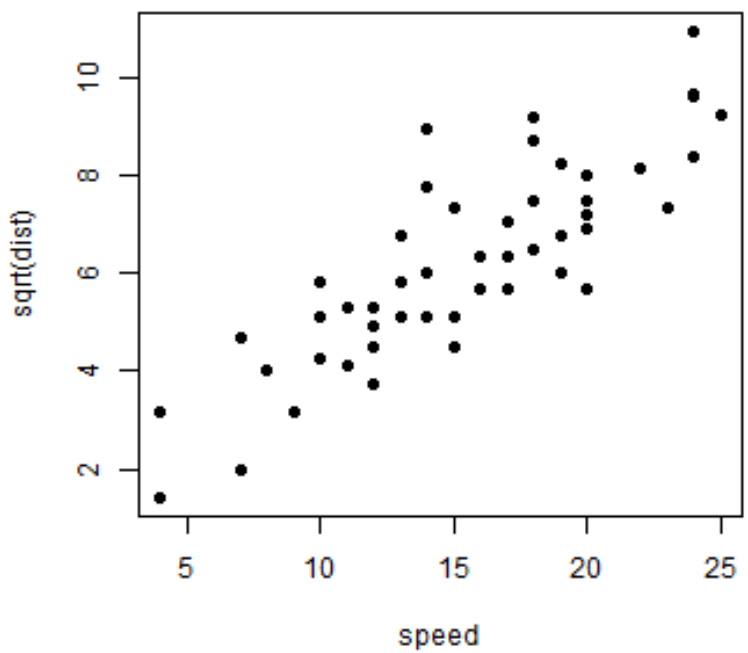




More non-linear examples

R dataset: cars



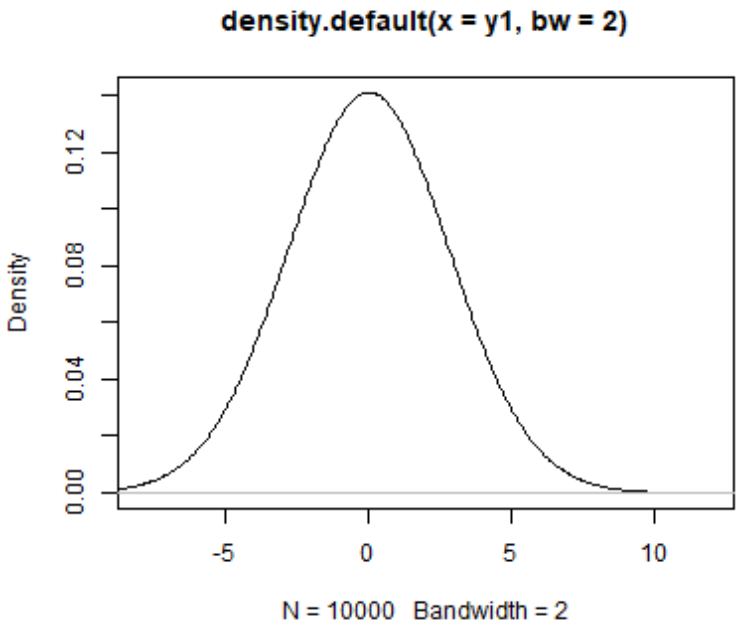


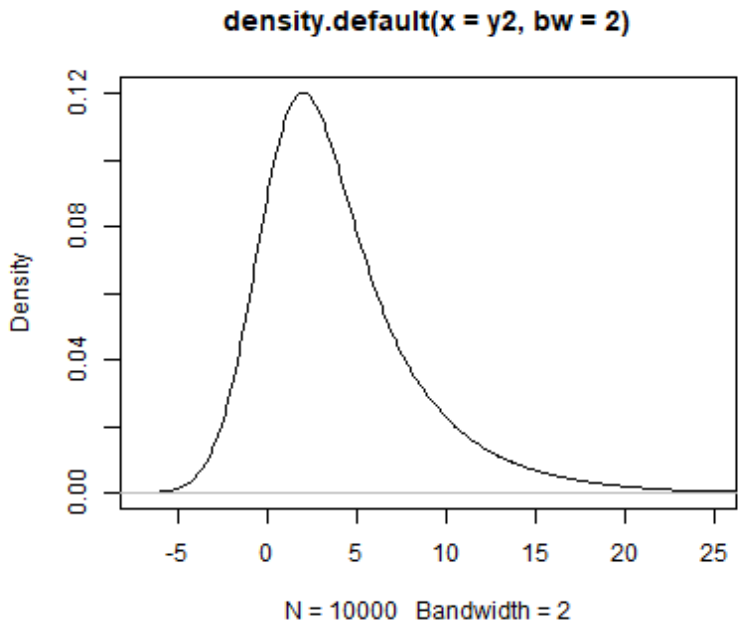
R data example with model fitting

Linear model fitting on cars data set and Puromycin data set.

Please look at the file 'Examples of variable transformation in cars and Puromycin dataset.R' on canvas

Transformation for non-normal or skewed data

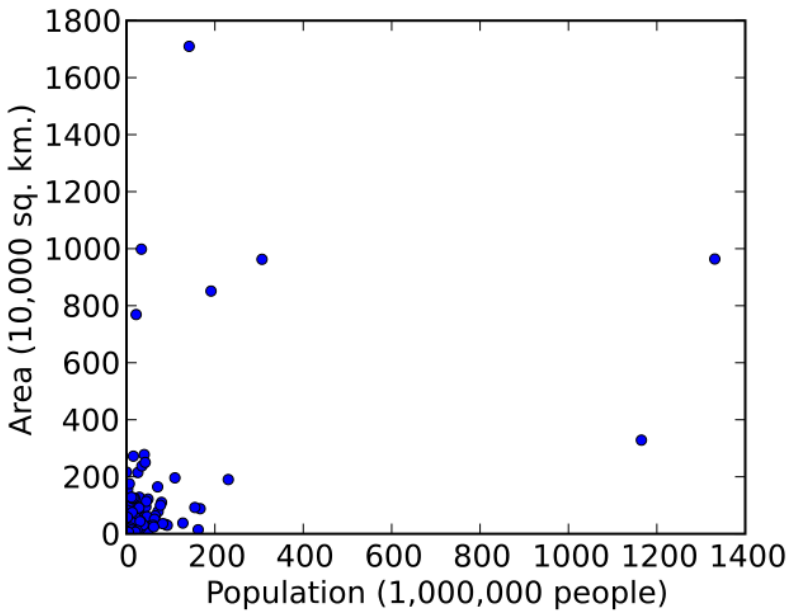


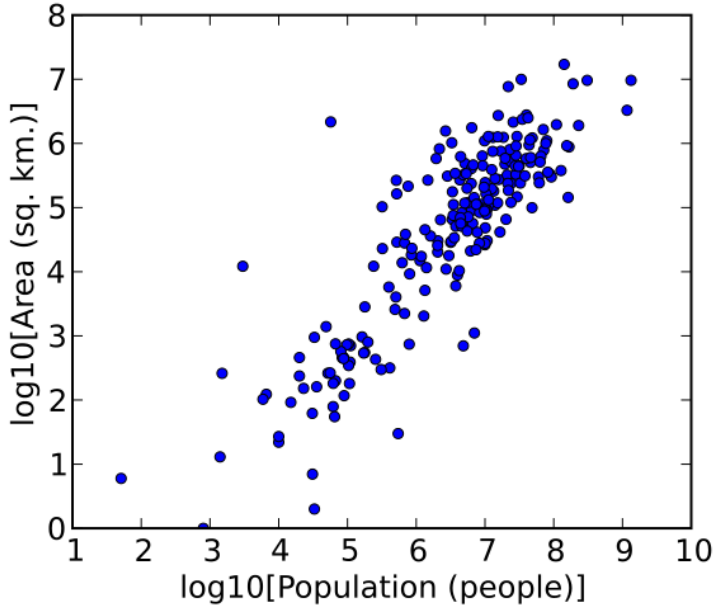


For skewed data, all values are not represented fairly.
The smaller values have high frequency and the larger values have low frequency.
The symmetric distribution on the left has a better represented and less chance of being affected by outliers.
For the bell shaped distribution on the left the mean, median and mode are roughly equal.

Transformation for skewed data (example)

The following scatter plot shows area vs population for several geographic regions.
The raw data shows that most regions have a small population and a small area. However there are some areas with large area and small population and the small area with large population.
Both the axes are skewed positively. The plot on the right shows the scatter plot after applying *log*10 on both axes.





[4]

[4] wikimedia.org

Summary till now

- Real data does not always conform to linear model assumptions.
- Transformation is required to make non-linear relation linear for simple linear regression.
- Transformation needs to be done with caution:
  - it should be invertible in most cases
  - it should not overfit the training data
  - it should not be too complex.
- Transformation also helps in handling skewed data to avoid long tails and outliers.

Breakout rooms. Discuss the papers you read.

Main introductory papers:

- Osborne, 2002, *Notes on the use of Data Transformations*
- S. Manikandan, 2010, *Data Transformations*

Recommended for further understanding (papers on canvas):

- Additive Effects: A comparison of linear model fits before and after transformation.
- Transformation bias: A paper discussing bias induced by unnecessary transformation.
- Log transformation: A discussion on use of log transformation for normality assumption.
- Rusina 2011: An application of Box-Cox transformation

Box-Cox

G. E. P. Box and D. R. Cox, *An Analysis of Transformations*

Goal: Find suitable transformation to perform linear regression.

$$f(y) = \beta_0 + \beta_1 x + \epsilon$$

$$f(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

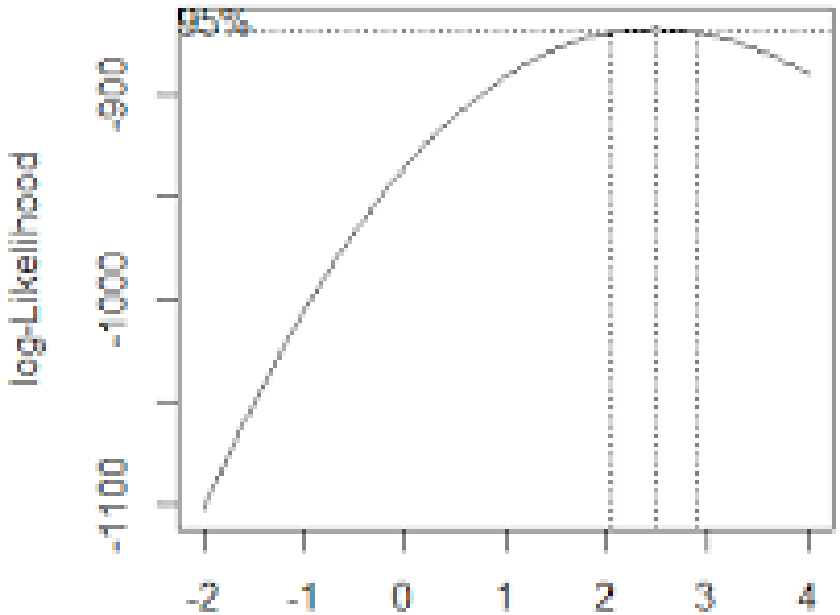
Estimate λ by maximum likelihood.
Regress $y^{(\lambda)}$ on x.

Ensure $y > 0$!!!

Box-Cox

$$f(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

Estimate λ by maximum likelihood.
The following plot is generated by R show the log-likelihood as a function of λ. The three dotted vertical lines show the maximizer λ and the 95% range in which the mazimizer λ will lie.



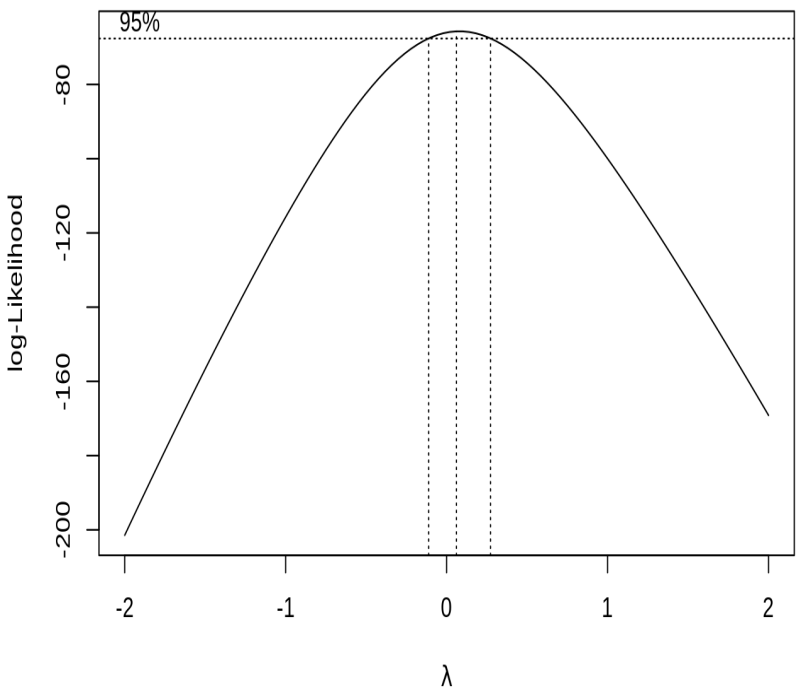
The maximizing λ from this plot is 2.5 and lies inside [2, 3] with probability 0.95.

Regress $\dfrac{y^{2.5} - 1}{2.5}$ on x.

Box-Cox

$$f(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

Estimate λ by maximum likelihood.



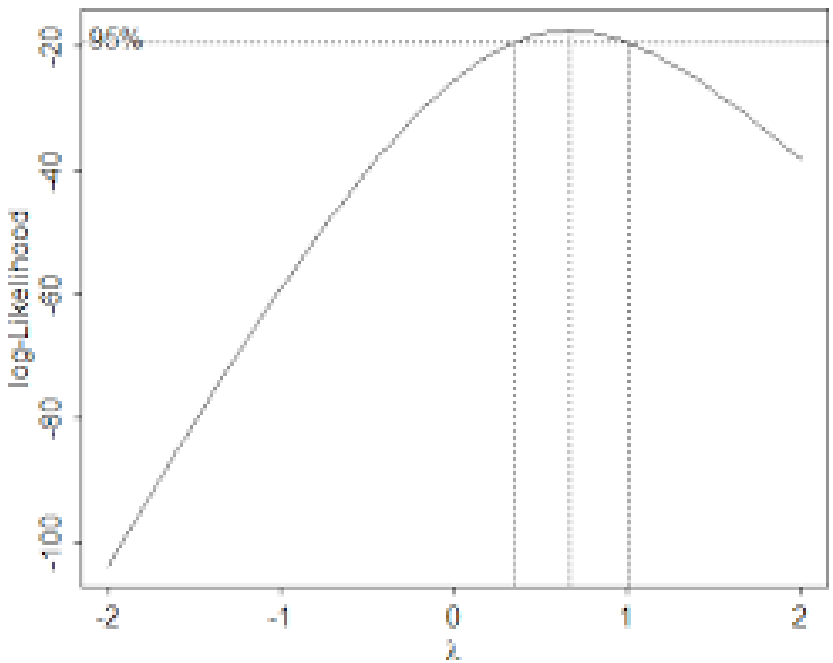
The maximizing λ from this plot is 0.1 and lies inside [-0.12, 0.3] with probability 0.95. We choose $\lambda = 0$.

Regress $log(y)$ on x.

Box-Cox

$$f(y) = y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ log(y), & \lambda = 0 \end{cases}$$

Estimate λ by maximum likelihood.



The maximizing λ from this plot is 0.66 and lies inside [0.33, 1] with probability 0.95.

Regress $\dfrac{y^{0.66} - 1}{0.66}$ on x.

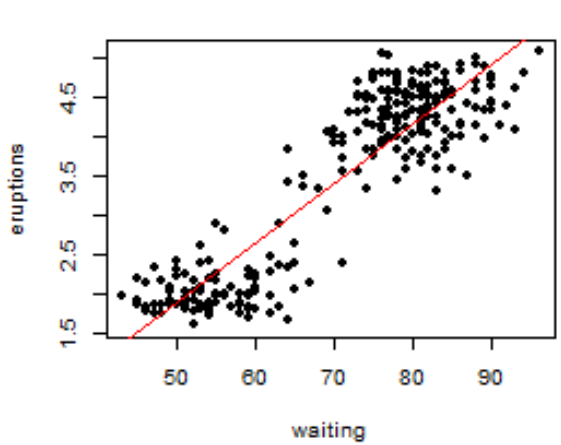
Box-Cox Demo

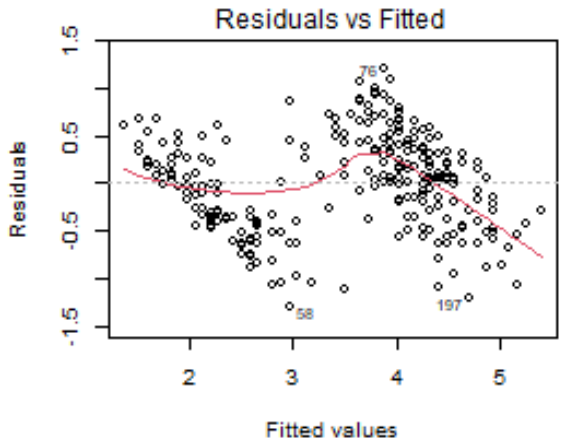
Rstudio

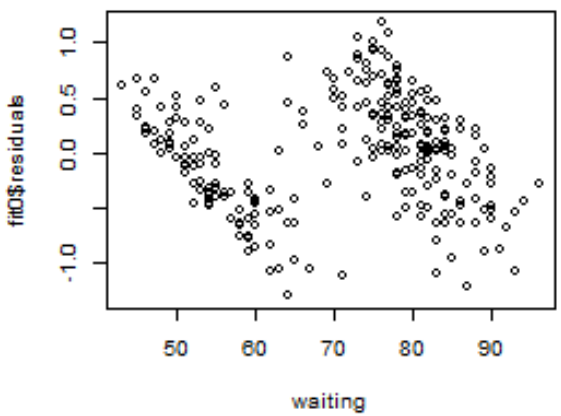
Variable manipulation

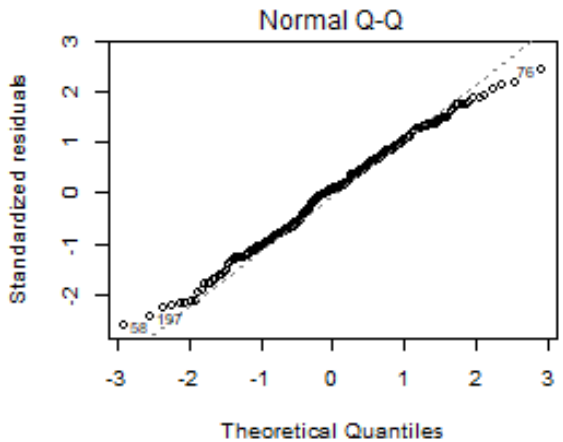
R dataset: faithful

```
fit0 <- lm(eruptions ~ waiting)
```







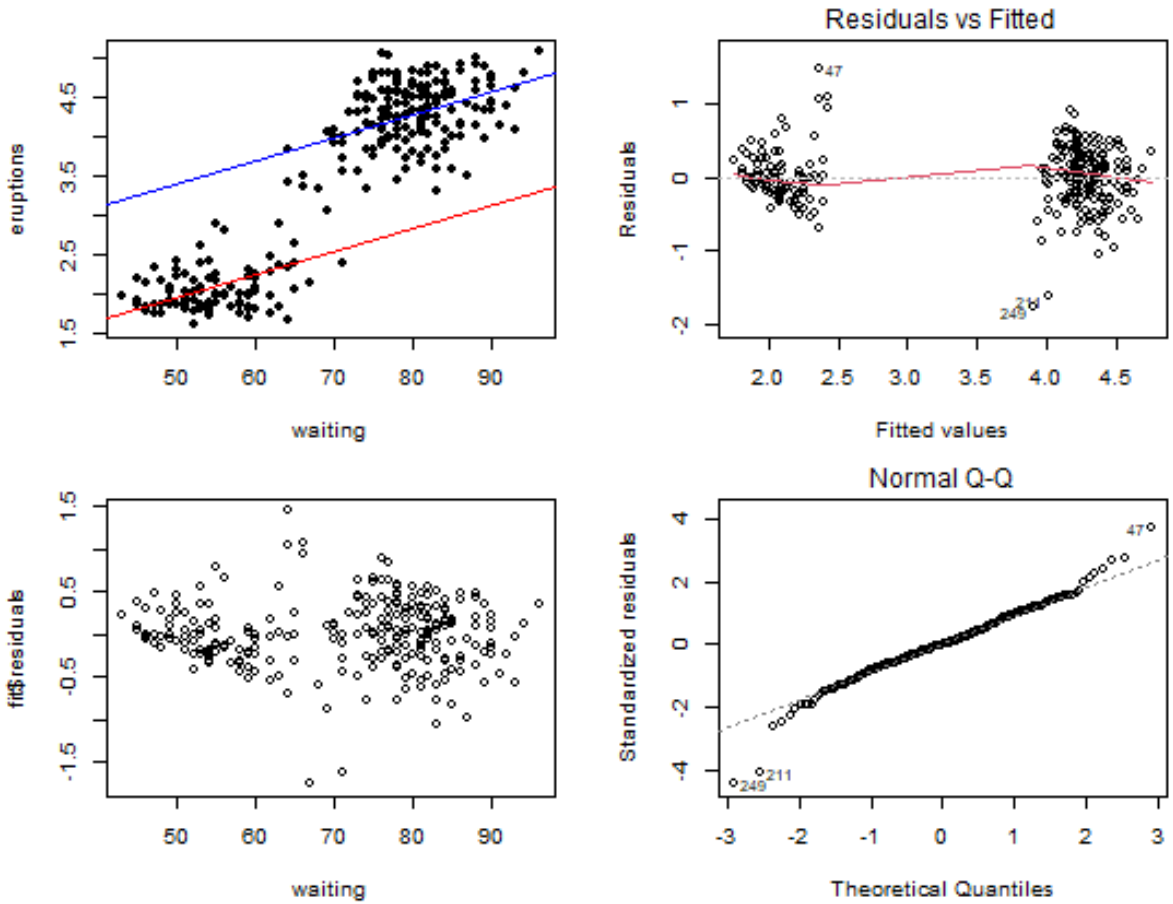


The residuals show a linear trend.

Variable manipulation

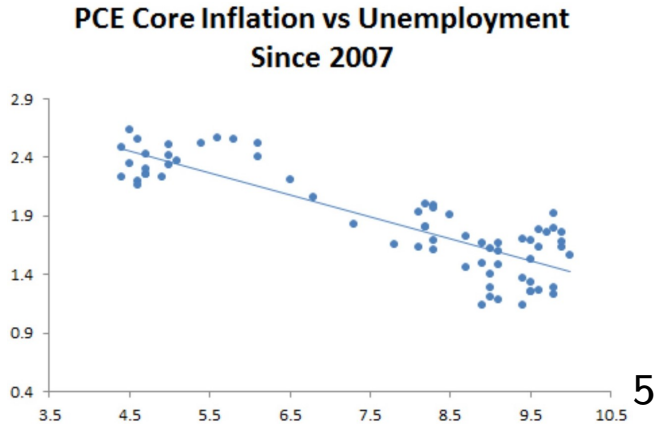
R dataset: faithful

```
fit <- lm(eruptions ~ waiting + factor(waiting>66))
```



Alternate models for non-linearity

Sometimes data is too complicated that transformation does not help.



[5]

Then one can resort to some flexible modelling methods. Below are some examples. These will not be covered in this class.

- Nonlinear programming.
- Generalized additive models.
- Random forest or tree regression.
- Neural networks.
- Other non-parametric methods: kernel regression.

[5] economistsview.typepad.com

Takeaways from today's class

1. Linear model fitting and residual analysis.
2. The importance of transformation and variable manipulation in model fitting.
3. Identifying non-linear relations and apply required transformation: for linearity and unskewed data.
4. Apply Box-Cox method to obtain appropriate transformation.

Logistic Regression

Goal: Classification or probability estimation.
Data structure: $Y \in \{0, 1\}$.

Target: Express $P(Y = 1|x)$ in terms of X's.

If we had

$$P(Y = 1|x) = \beta_0 + \beta_1 x + \epsilon,$$

no guarantee that $P(Y = 1|x) \in [0, 1]$.

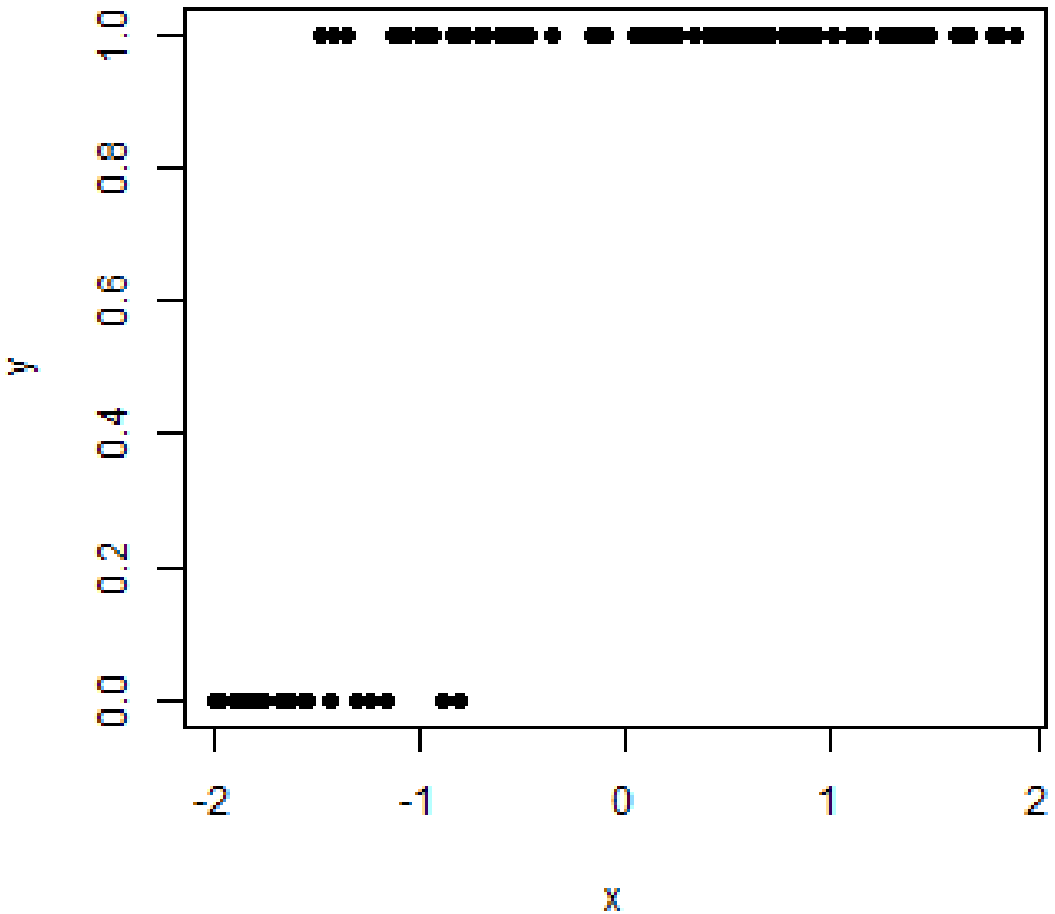
Solution:

$$log\left(\frac{P(Y = 1|x)}{1 - P(Y = 1|x)}\right) = \beta_0 + \beta_1 x + \epsilon.$$

Estimate $\widehat{P(Y = 1|x)} = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$.

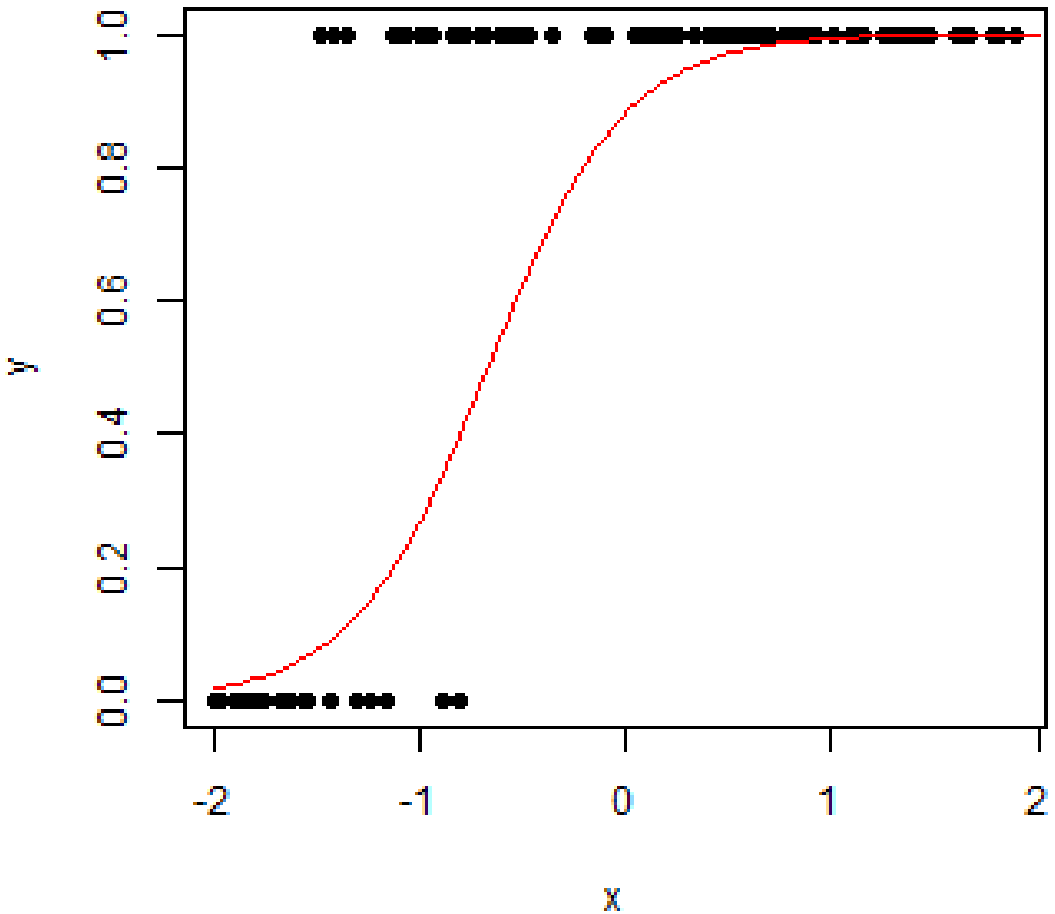
Logistic Regression

$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$



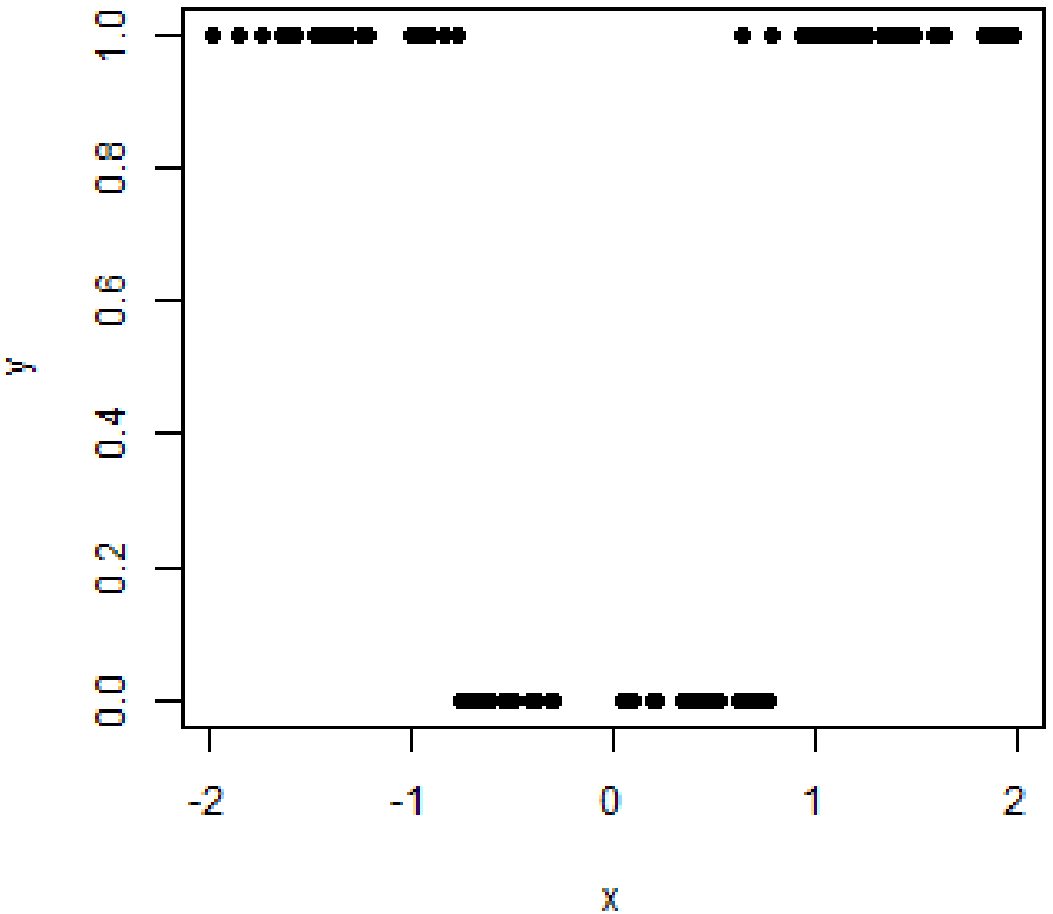
Logistic Regression

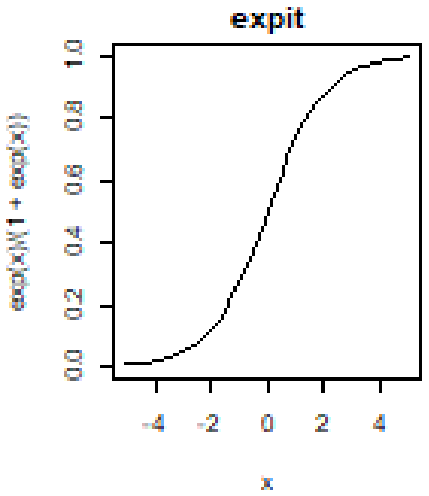
$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$



Logistic Regression

$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

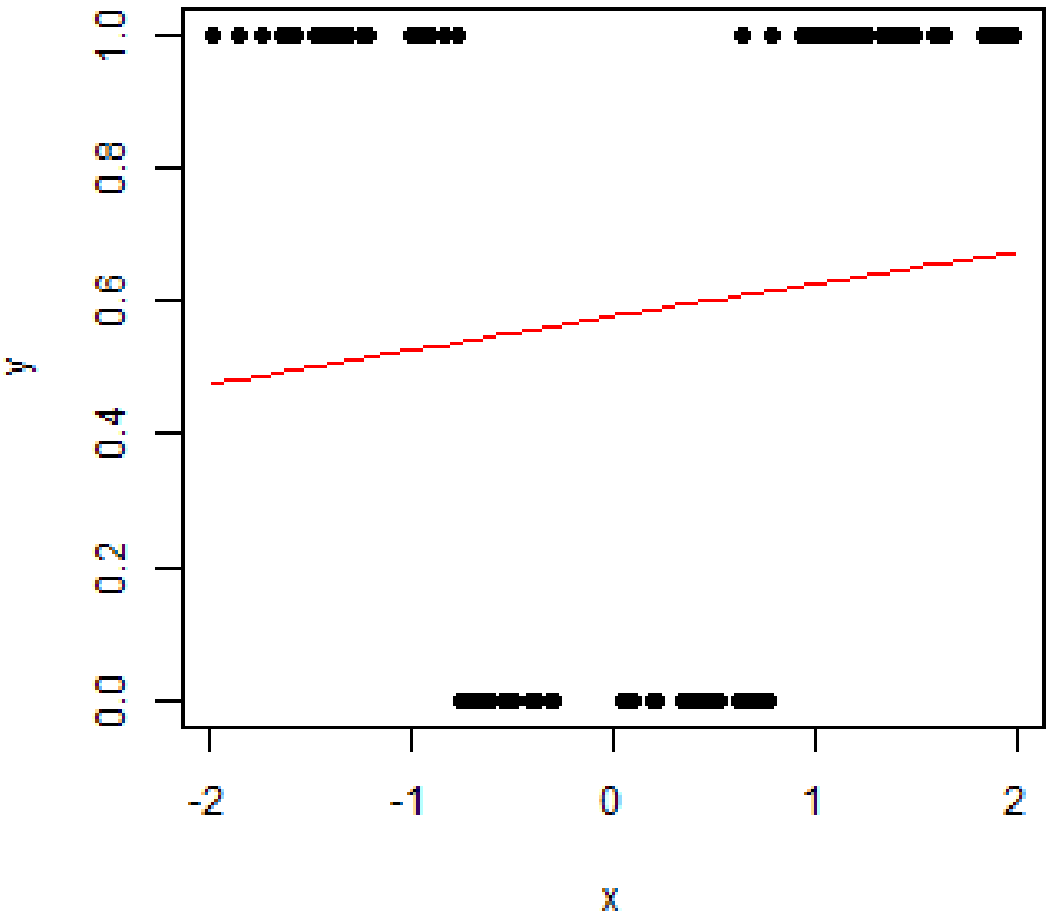


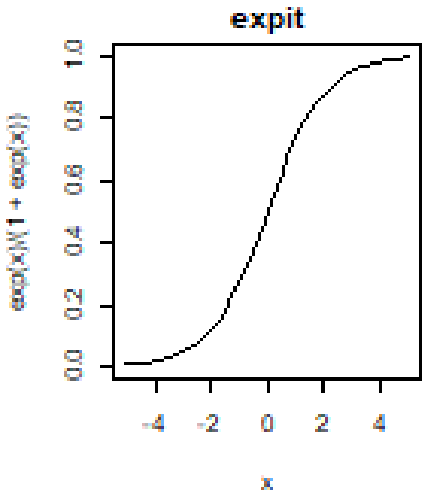


Transform x to x^2.

Logistic Regression

$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

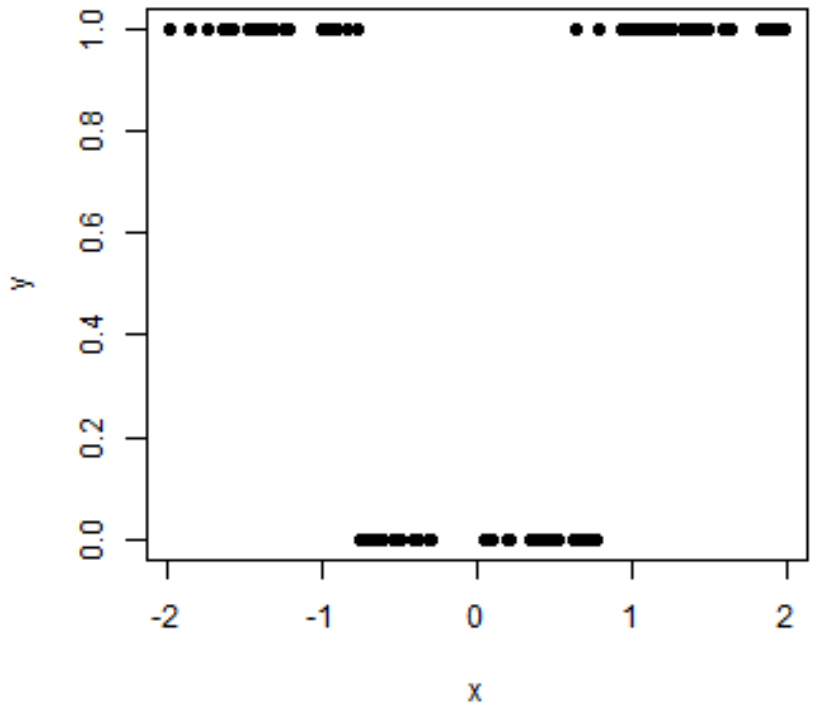


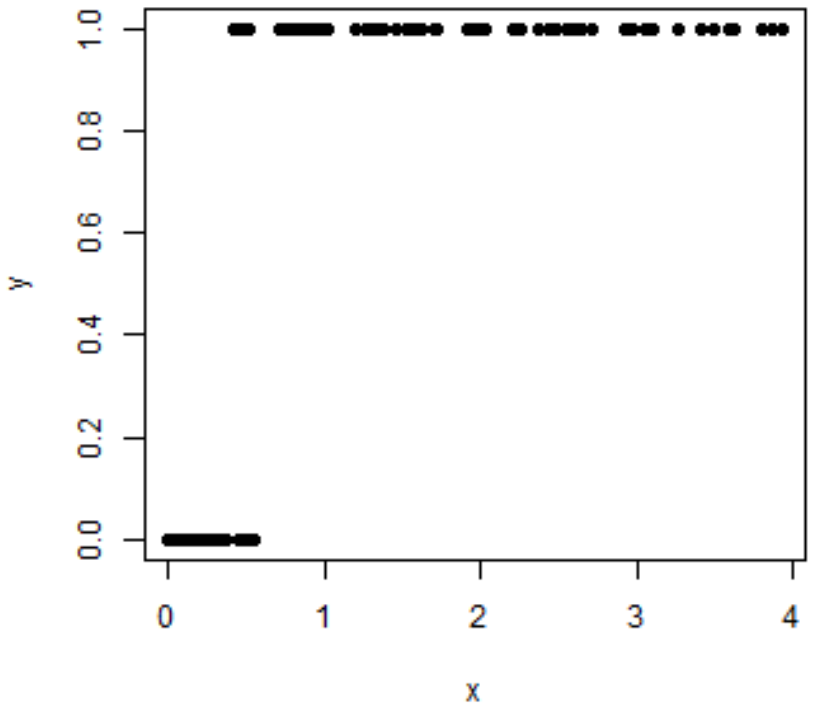


Logistic Regression

$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Transformation: x to x^2.





Any transformation of covariates, that roughly separates 0's and 1's by a single hyper-plane.

Logistic Regression

$$P(Y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Transformation: x to x^2.