Topic3: Bivariate Data

Outline

Topic3: Bivariate Data

- Example: Male Olympic 100m sprints
- Bivariate Data
- Fitting a Least Squares Regression Model
- Step1: Construct a Scatter plot
- Step2: Fit the Least Square's Line (LSL)
- Step3: Construct a residual plot
- Step4: Calculate the correlation coefficient and the coefficient of determination
- Avoiding mistakes in Regression
- Linear regression with outliers

Example: Male Olympic 100m sprints

The Male 100 metres sprint race in the Olympics is one of the most prestigious events in Athletics. The reigning champion is often named ‘the fastest man in the world’, currently the Jamaican Usain Bolt.

Are the times getting faster in each Olympics? What would you predict the time for the next Olympics?

All the Medalists: **Men's 100-Meter Sprint**

One Race, Every Medalist Ever

Usain Bolt's 9.63 set an Olympic record in the 100. So how far ahead of every Olympic medalist is he?

By KEVIN QUEALY and GRAHAM ROBERTS

Bivariate Data

In most contexts we collect many pieces of information for each ‘individual’ in the data set and look at the relationships between the variables (eg age, height, weight and income for USyd students). This is called multivariate data.

- We just consider bivariate data, of the form: $\{(x_i, y_i)\}$ for $i = 1, 2, \ldots, n$.
- X is called the independent variable (or explanatory variable, predictor or regressor) and Y is called the dependent variable (or response variable). These are determined by the context of the data.

Fitting a Model

We are interested in fitting a model $Y = f(X) + error$, where the error is independent of the function $f(X)$ and follows a Normal curve (See Topic 6).

- Examples include $Y = \alpha + (X + \beta)^2 + \gamma + error$ (quadratic) or $Y = \alpha e^{\beta X}$ (exponential) or $Y = \alpha X^{\beta}$ (allometric).
- We will just consider the linear model: $Y = \alpha + \beta X + error$.
- We use the sample values $\{(x, y)\}$ to find an estimate of the model: $y = a + bx + residual$.
- Note that both the exponential and allometric models can be expressed as a linear model by taking logs of each side.

Fitting a Linear Regression (LSL)

Consider the following 5 steps:

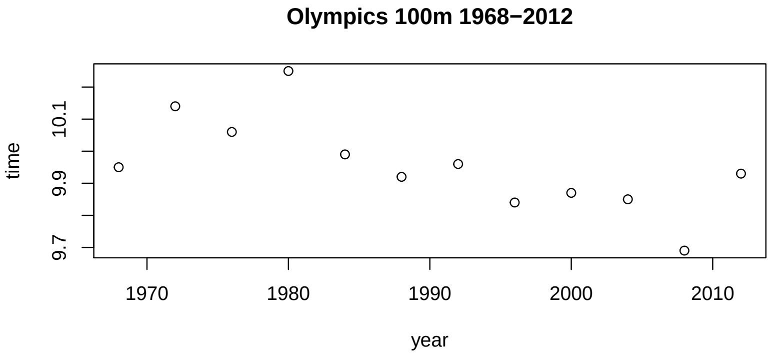
1. Construct a Scatterplot: y vs x.
2. If the plot looks linear, fit the Least Square's (Regression) Line (LSL): $\hat{y} = a + bx$.
3. Construct a residual plot: $res = y - \hat{y} = y - (a + bx)$ vs x
4. If the residual plot looks random, calculate the coefficient of determination (r^2) and correlation coefficient (r).
5. Work out predictions for y, by finding $\hat{y}$ for a certain given x.

Step1: Construct a Scatter plot

The 1st step is to construct a scatterplot of Y vs X. This is a graphical summary of the bivariate data.

This is 1st diagnostic: does the plot suggest a linear relationship between Y and X, or not?

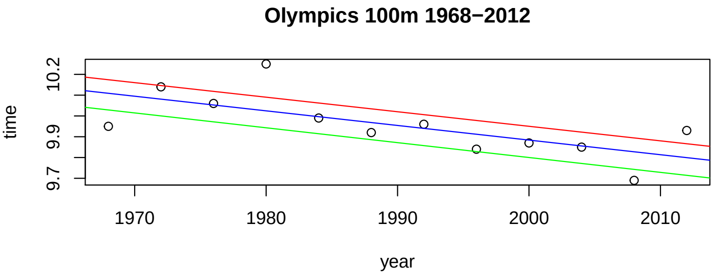
```r
## olympics <- read.csv("Olympics100m.csv",header=T)
year=c(olympics$Year)
time=c(olympics$Time)
plot(year,time, xlab="year", ylab="time",
     main="Olympics 100m 1968-2012")
```



Step2: Fit the Least Square's Line (LSL)

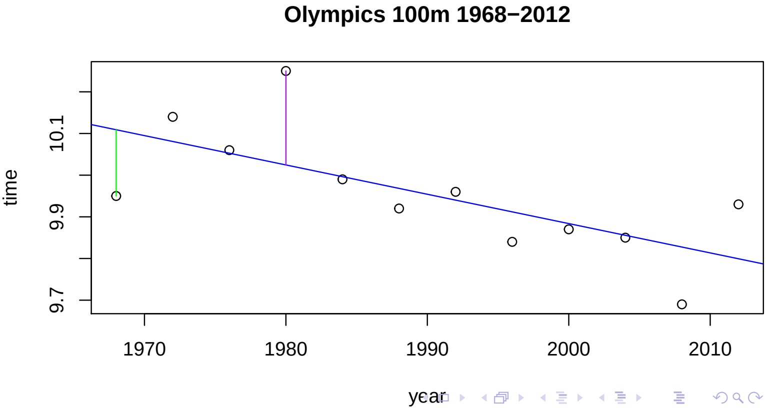
If the scatter plot looks linear, then we fit the Least Square's Regression Line (LSL).

- By 'eye', there are many possible lines, which could be drawn on the scatter plot – but which one is optimal?



- For each candidate line $f(x) = a + bx$ for all a and b, we focus on the resultant set of residuals: $res = e(a, b) = y - (a + bx)$.

  This is the gaps between the line and the actual points. For example, in the plot below the green residual is -0.1581 and the purple residual is 0.22603.



- We consider the 'best' line, to have the smallest residuals, determined by their sum of squared residuals

$$S(a,b) = \sum_{i=1}^{n} e(a,b)^2 = \sum_{i=1}^{n} (y - (a + bx))^2$$

- We minimise $S(a,b)$ by solving $\frac{\partial}{\partial a} = 0$ and $\frac{\partial}{\partial b} = 0$. This gives

$$\sum_{i=1}^{n} y_i - na - b \sum_{i=1}^{n} x_i = 0$$

and

$$\sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i - b \sum_{i=1}^{n} x_i^2 = 0$$

.

- As long as the x_i are not all equal, there is a unique solution for the intercept and the slope,

$$\boxed{a = \bar{y} - b\bar{x}}$$

and

$$\boxed{b = \frac{S_{xy}}{S_{xx}}}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = (n-1)s_x^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 = (n-1)s_y^2$$

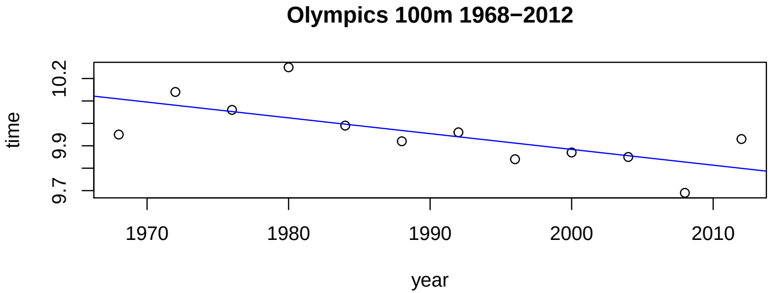
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)$$

and s_x^2 and s_y^2 are the variance of $\{x\}$ and $\{y\}$ respectively.

Note: The natural numerical summaries for bivariate data are: $\bar{x}, \bar{y}, s_x, s_y$, so the LSL is a combination of these summaries: $\hat{y} = a + bx$.

```
model = lm(time~year)
model$coeff

##  (Intercept)         year
## 23.957226107 -0.007036713
```

```r
plot(year,time, xlab="year", ylab="time",
     main="Olympics 100m 1968-2012")
abline(model$coeff, col="blue")
```



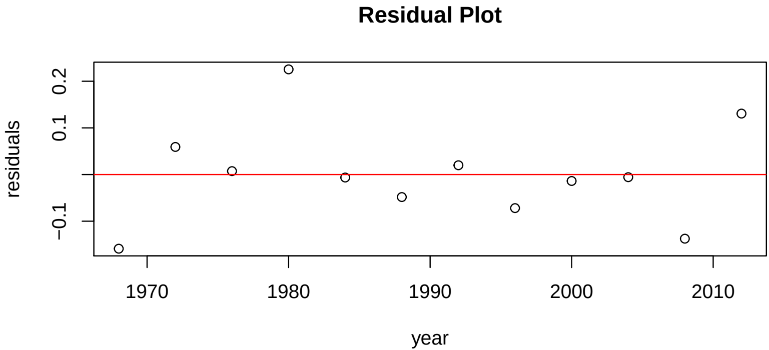
Step3: Construct a residual plot

A residual plot is a scatter plot of the residuals $res = e = y - (a + bx)$ vs x.

It is a 2nd diagnostic: "Does the plot look random, or is there any pattern?"

- If the plot is random: then the LSL fit is good.
- If the plot shows a relationship between e and x, then the LSL is not adequate and we need to consider a more complex function or a transformation (eg $y = x^2$ or $y = log(x)$).

```r
residuals=model$res
plot(year,residuals,main="Residual Plot")
abline(h=0,col="red")
```



Step4: Calculate the correlation coefficient and the coefficient of determination

We have calculated the 'best' line, but is it a good line? How strong is the linear relationship between Y and X?

A fit is considered good when the set of data points is close to the LSL, so that the residuals are small. So we consider the relationship between s_{res}^2 (the variance of the residuals) and s_y^2 (the variance of y).

Pearson's correlation coefficient

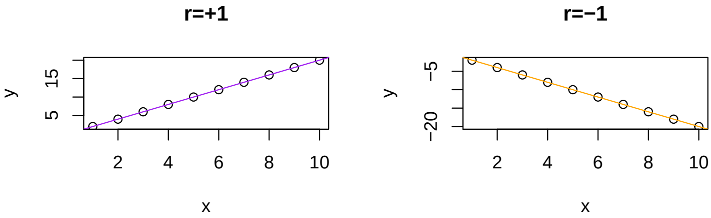
The correlation coefficient is

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}}$$

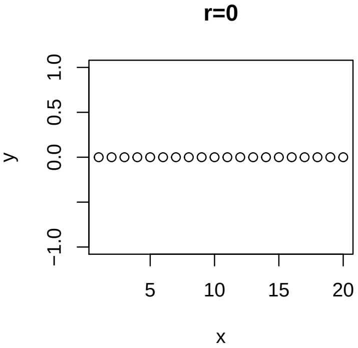
Properties:

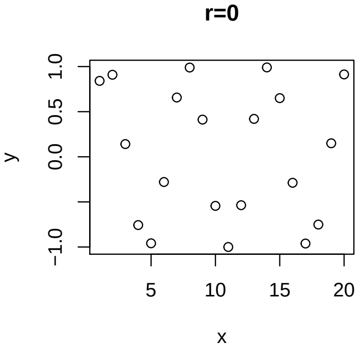
- r is symmetric in x and y.
- $-1 \leq r \leq 1$
  r is a numerical summary which indicates the strength of the linear assocation between y and x.

- $r = \pm 1$
  This corresponds to a perfect linear correlation, with all the data points lying on the LSL.

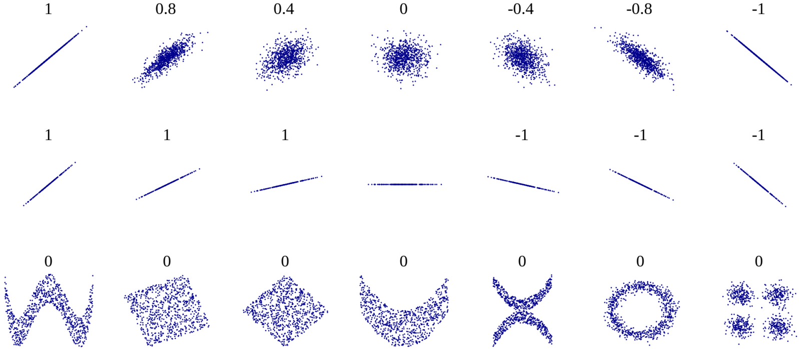


- $r = 0$
  This indicates no linear correlation, for example a line with zero slope, or a random scatter, or a non linear relationship.





Examples of Correlation Coefficients



▸ Link

Relationship between Correlation Coefficient and Slope

There is an interesting relationship between r and b

$$b = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} = r\frac{\sqrt{S_{yy}/(n-1)}}{\sqrt{S_{xx}/(n-1)}} = r\frac{s_y}{s_x}$$

Hence:

- The sign of r reflects the trend (slope) of the data.
- r is unaffected by a change of scale or origin.

The coefficient of determination

The coefficient of determination is the proportion of variability of y explained by x for a model, or in our context, the proportion of variability explained by the linear regression.

$$\boxed{r^2 = \frac{s_y^2 - s_{res}^2}{s_y^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}}$$

Properties:

- $0 \leq r^2 \leq 1$
- $r^2 = 1$

  This arises when $res_i = 0 \ \ \forall i$ and $s_{res}^2 = 0$, ie all of the variability of the model is associated with the linear regression. All points lie on the LSL.

- $r^2 \approx 1$
  This arises when $res_i = 0$ is small compared to s_y^2, ie most of the variability of the model is associated with the linear regression.
- $r^2 = 0$
  This arises when $s_{res}^2 = s_y^2$, ie none the variability of the model is associated with the linear regression.
- $r^2 \approx 0$
  This arises when $s_{res}^2 \approx s_y^2$, ie almost none the variability of the model is associated with the linear regression.
- Note that r^2 can be small and still indicate that the model is correct (ie a model where there is naturally low association between X and Y).

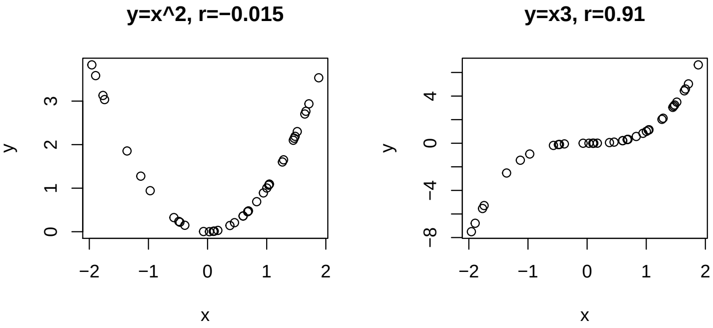
```
cor(year,time)

## [1] -0.6912573

cor(year,time)^2

## [1] 0.4778366
```

Hence, the linear associotion between year and time for the Olympics 100m sprint is -0.7 (fairly high). 48% of the variation in times is explained by the variation in years.

Avoiding mistakes in Regression

- Correlation does not imply causation
  A high value of r does not necessarily imply a causal relationship between X and Y. For example, December temperature and consumer spending).

- Causation does not imply (linear) correlation



y=x^2, r=−0.015

y=x3, r=0.91

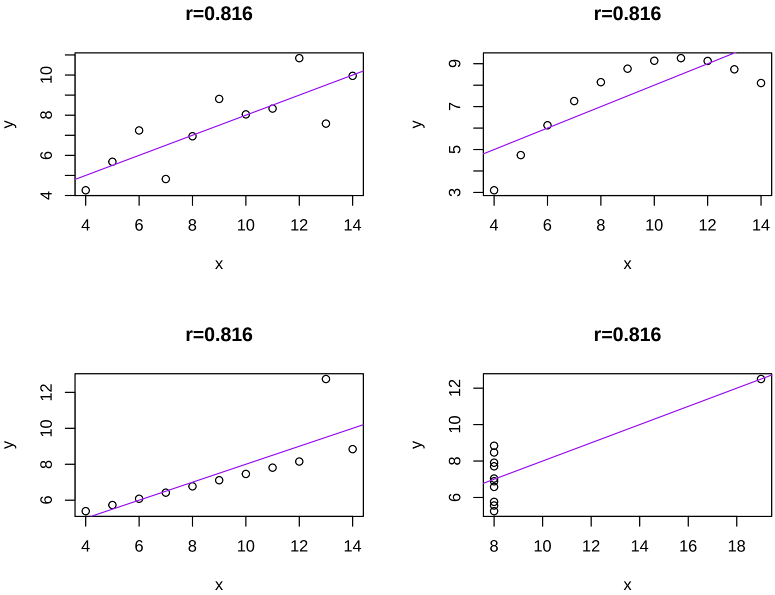
Avoiding mistakes in Regression

- The same value of r can correspond to very different models.

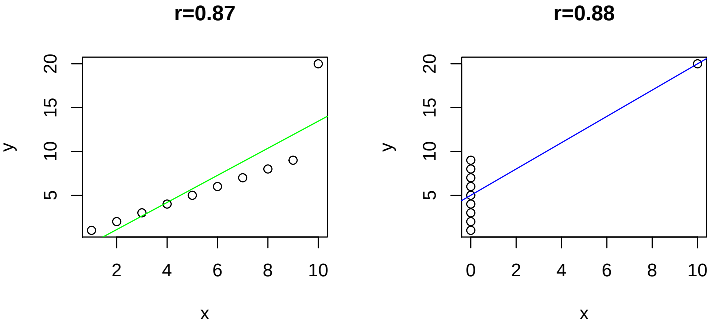
  The following data sets, called 'Ansombe's Quartet' were constructed by Francis Anscombe in 1973. They all have $\bar{x} = 9$, $s_x^2 = 11$, $\bar{y} = 7.5$, $s_y^2 = 4.127$, $r = 0.816$ and linear regression line $y = 3 + 0.5x$. But look how different they look!

▸ Anscombes Quartet

▸ Law Grad salaries


r=0.816
y
x
4
6
8
10
12
14
r=0.816
y
x
3
5
7
9
4
6
8
10
12
14
r=0.816
y
x
6
8
10
12
4
6
8
10
12
14
r=0.816
y
x
6
8
10
12
8
10
12
14
16
18

- Even one outlier can distort the model.
  It's vital to draw a scatter plot before considering r, because of the high influence of outliers.

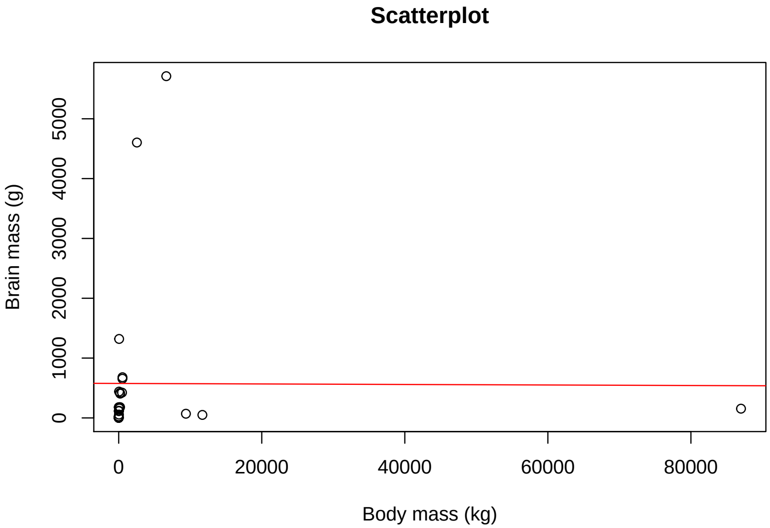


Example: Animals dataset: body vs brain weight

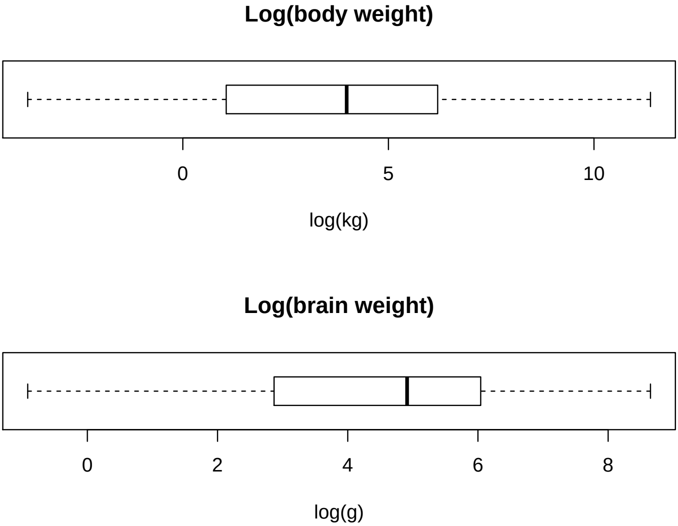
Body weight in kg and brain weight in g

```
library(MASS) ## The location of the package
head(Animals)

##                      body brain
## Mountain beaver      1.35   8.1
## Cow                465.00 423.0
## Grey wolf           36.33 119.5
## Goat                27.66 115.0
## Guinea pig           1.04   5.5
## Dipliodocus      11700.00  50.0

names(Animals)

## [1] "body"  "brain"
```

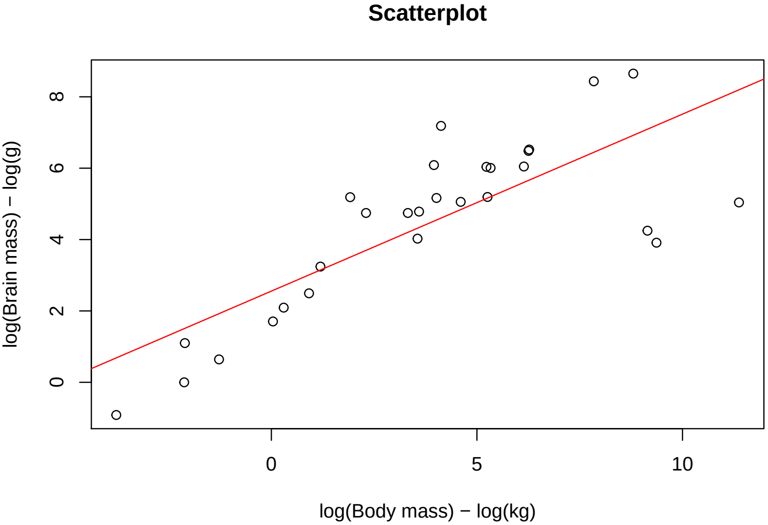
Outliers galore!



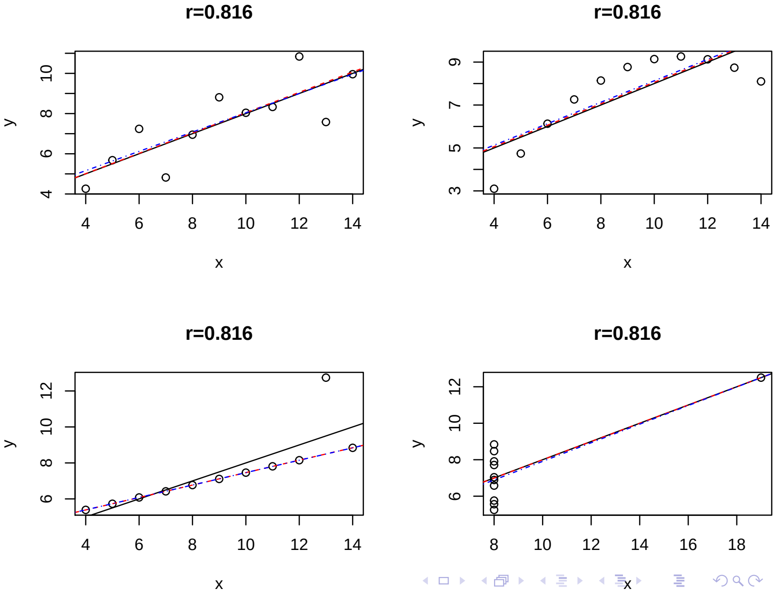
Log transformation



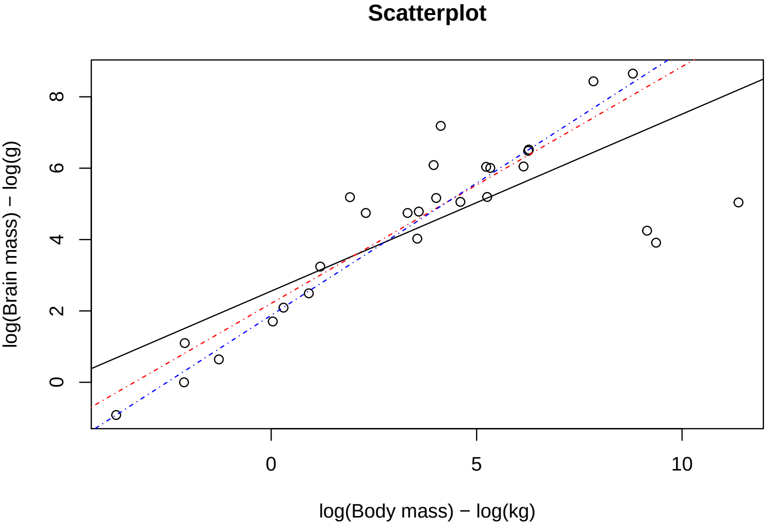
Log transformation linear regression



Robust regression



Robust regression



In R

The red line is MM regression

```
library(MASS)    # library
plot(x, y)
red_line = rlm(y~x)   # r[obust]lm()
abline(red_line, col="red")
```

The blue line is quantile regression

```
library(quantreg)    # library
plot(x, y)
blue_line = rq(y~x)   # r[egression]q[uantile]()
abline(blue_line, col="blue")
```