

Lecture 6

Correlation & Regression

Theophanis Tsandilas

Memory experiment

Six participants are given a picture (20 cm x 30 cm) showing a ball in the air to view it for 3 to 18 seconds.

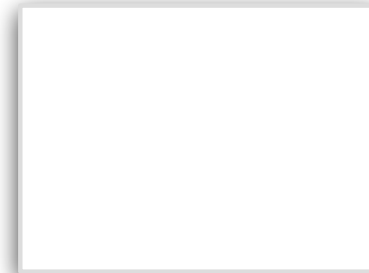


Memory experiment

Six participants are given a picture (20 cm x 30 cm) showing a ball in the air to view it for 3 to 18 seconds.



The following day, they are shown an empty frame (20 cm x 30 cm) and are asked to indicate the position of the ball in the original picture as precisely as possible.



The researcher measures the error: the distance between the position indicated by the participant and its true value.

Memory experiment: Results

The experimental results are as follows:

Participant	Presentation Time (sec)	Position Error (cm)
1	3	6.3
2	6	3.9
3	9	2.3
4	12	2.0
5	15	2.8
6	18	1.6

Memory experiment: Questions

Is there a relationship between the **Presentation Time** and the **Position Error**?

If yes, how could we quantify and describe this relationship?

Correlation & dependence

Is any statistical relationship between two random variables.

Example: The relationship between *Presentation Time* and *Position Error* in the previous example.

Causal relationships

There is a causality process: a **cause** and an **effect**.

Example: There is a causal relationship between *environmental temperature* and *electricity consumption*.

A drop in temperature causes households to consume more electricity for heating purposes.

Non-causal relationships

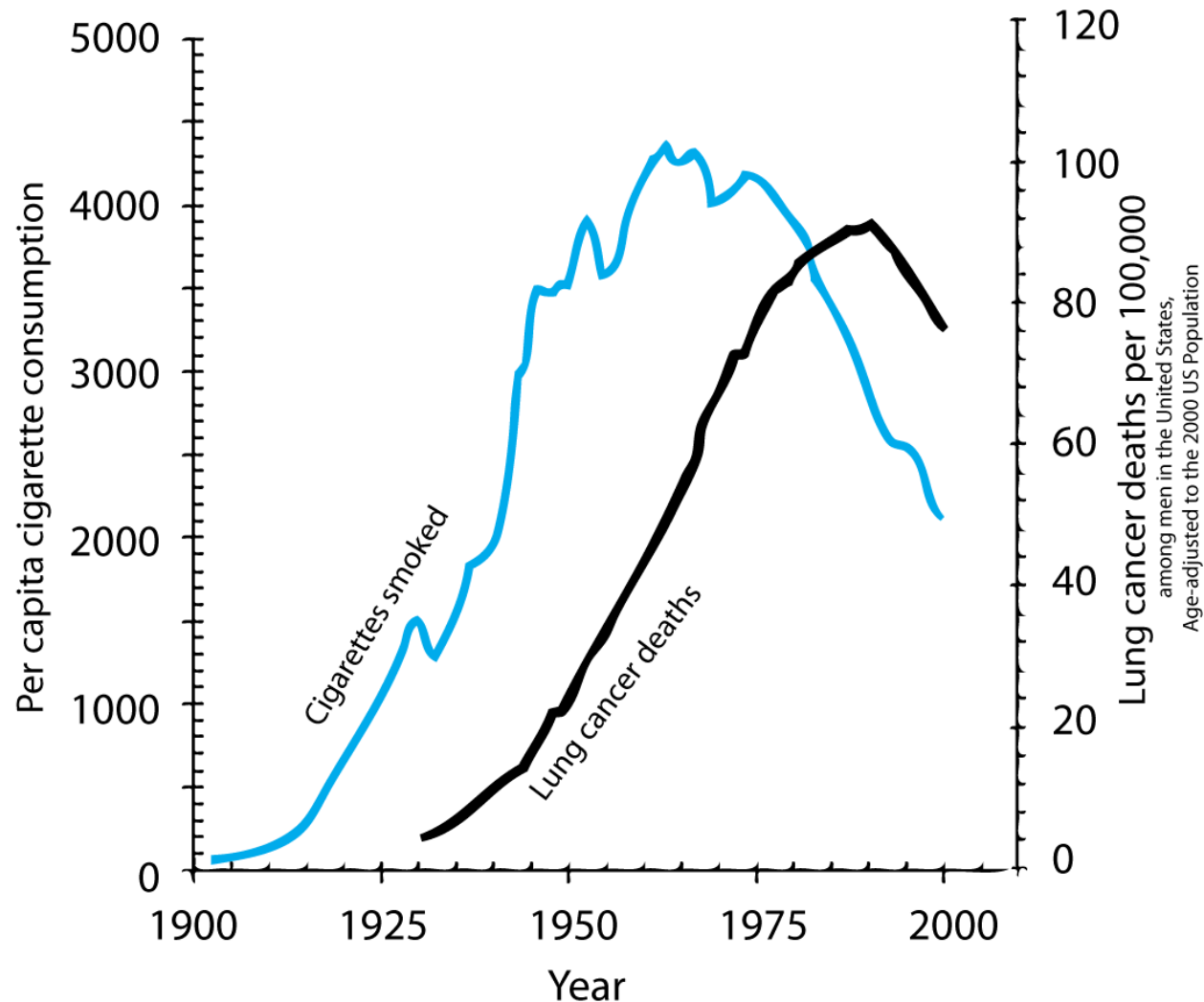
(example from wikipedia)

Sleeping with one's shoes on is strongly correlated with *waking up with a headache*.

Conclusion: sleeping with one's shoes on causes headache.

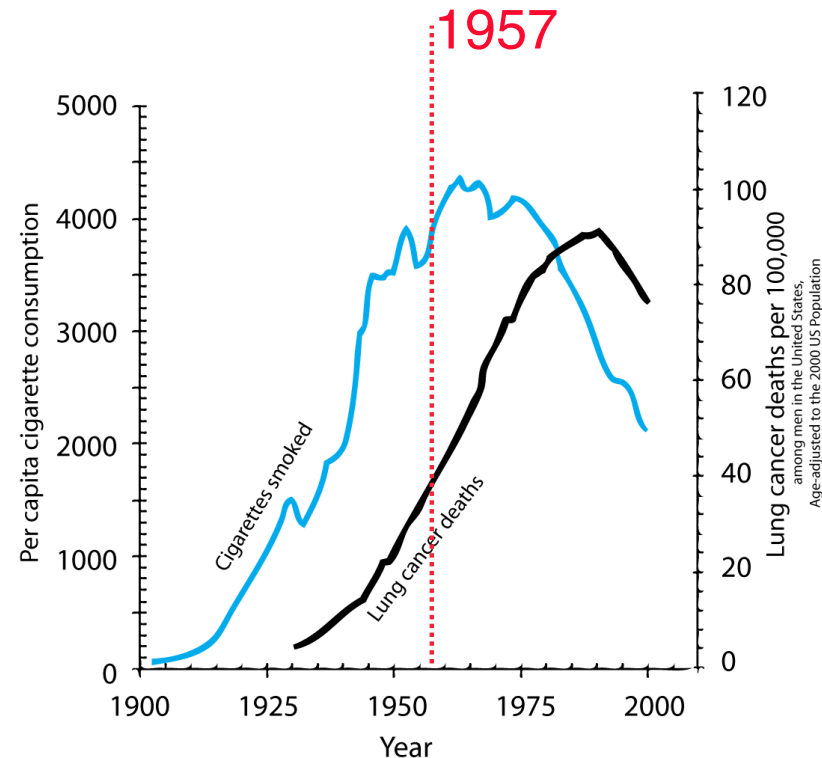
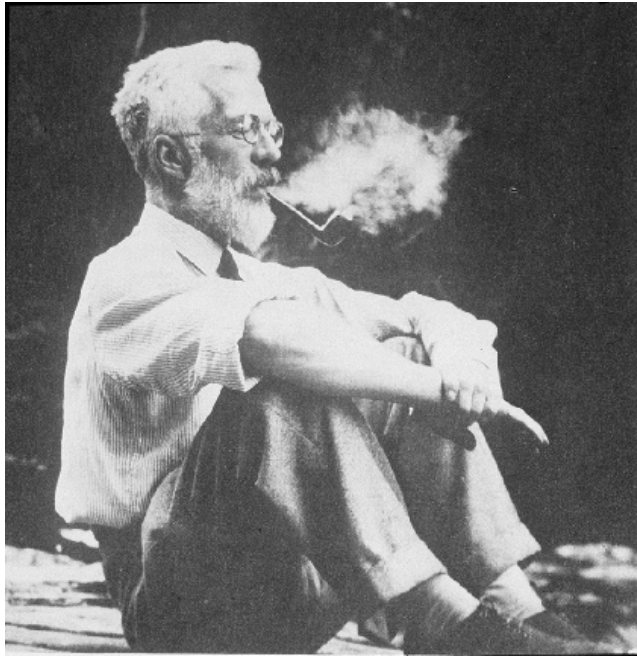
Correlation does NOT imply causation. A more likely explanation is that both effects are caused by a third factor such as *going to bed drunk*. Thus, the conclusion is wrong.

Smoking and lung cancer



https://en.wikipedia.org/wiki/File:Smoking_lung_cancer.png

Smoking and lung cancer



The theory that increased smoking is “the cause” of the change in apparent incidence of lung cancer is not even tenable in face of this contrast.

Ronald Fischer, 1957 (<https://www.york.ac.uk/depts/maths/histstat/fisher269.pdf>)

even the most prominent statisticians can be wrong about statistics!

Covariance

Measures the average tendency of two variables X and Y to change together (*covary*):

$$Cov_{X,Y} = \frac{\sum_{i=1}^N (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{N}$$

where our sample is $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$

Compare this equation with the equation measuring variance.

Estimating the population covariance

We usually want to estimate the **population** $\sigma_{X,Y}$ **covariance** from the sample:

$$\hat{\sigma}_{X,Y} = \frac{\sum_{i=1}^N (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{N - 1}$$

which gives an unbiased estimate.

Example in R: memory experiment

```
> time <- c(3, 6, 9, 12, 15, 18)
> error <- c(6.3, 3.9, 2.3, 2.0, 2.8, 1.6)
> cov(time, error)
[1] -8.13
```

Covariance vs. variance

Variance is a special case of covariance:

If the two variables are identical ($X = Y$), covariance reduces to variance.

Covariance and independence

If two variables are independent, their covariance will be zero.

The converse is not always true:

zero covariance does not imply independence!

Example

X	$Y = X^2$
-2	4
-1	1
0	0
1	1
2	4

Y is fully determined by X, but their covariance is 0!

```
> x<-c(-2,-1,0,1,2)
> y<-x^2
> cov(x,y)
[1] 0
```


Drawbacks of covariance measure

Hard to interpret:

It is scaled in units of the product of X and Y

See again our example:

```
> time <- c(3, 6, 9, 12, 15, 18)
> error <- c(6.3, 3.9, 2.3, 2.0, 2.8, 1.6)
> cov(time, error)
[1] -8.13
```

How can one interpret this covariance value?

Correlation coefficients

Can be regarded as **standardized (normalized) covariance**:

Scaled by the standard deviations of X and Y

Pearson correlation coefficient (or *Pearson's r*)

A measure of **linear correlation** between two random variables X and Y.

The population correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

covariance

standard deviation of X standard deviation of Y

And its sample estimate:

$$\hat{\rho}_{XY} = r_{XY} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Pearson correlation coefficient (or *Pearson's r*)

Note: It is also known as **Pearson's product-moment correlation**.

Example in R: memory experiment

```
> time <- c(3, 6, 9, 12, 15, 18)
> error <- c(6.3, 3.9, 2.3, 2.0, 2.8, 1.6)
> cov(time, error)/(sd(time)*sd(error))
[1] -0.8347951
> cor(time, error)
[1] -0.8347951
```

The correlation coefficient is always bounded by -1 and 1

Correlation coefficient's bounds

$$r_{XY} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \hat{\sigma}_Y}$$

The covariance gets its maximum value when $X = Y$, which means that X and Y are perfectly correlated.

But in this case:

$$\hat{\sigma}_{X,Y} = \hat{\sigma}_{X,X} = \hat{\sigma}_X^2 = \hat{\sigma}_X \hat{\sigma}_Y$$

Thus, the maximum correlation value is 1.

Interpretation

The Person correlation coefficient (r) has **no units**.

The closer the correlation is to 1 or -1, the stronger is the **linear relationship** between X and Y .

A value close to zero seems to suggest a weak or absent linear relationship between X and Y .

Interpretation

But:

Is a correlation of $r = .2$ twice as strong as a correlation of $r = .1$?

Unfortunately, we cannot argue that.

Cohen's [1988] conventions

Cohen's has proposed guidelines for interpretation of correlation values for the behavioral sciences:

r	<i>Effect Size</i>
.1	Small
.3	Medium
.5 or higher	High

Others have tried to apply similar recommendations for other domains.

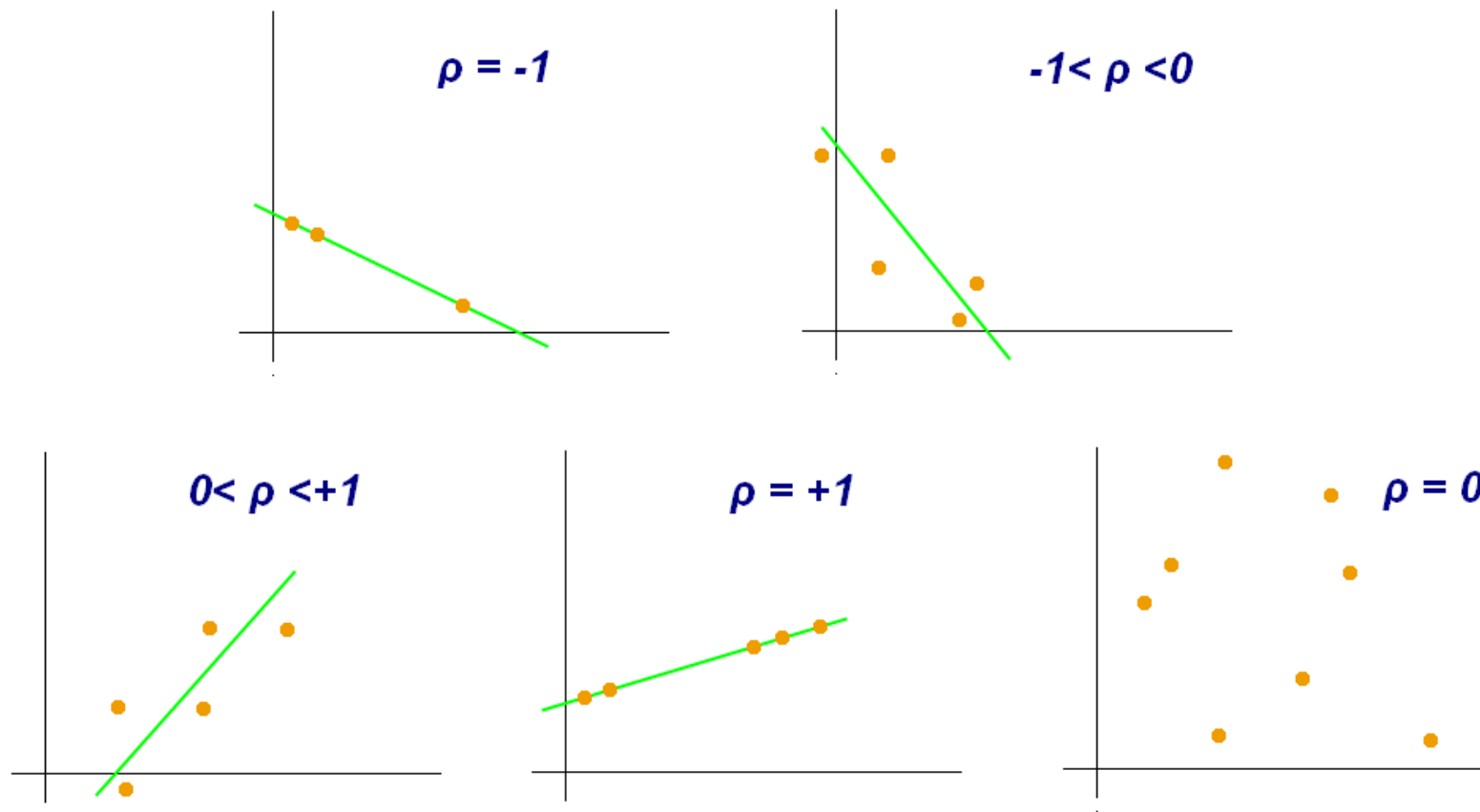
"Attaching such labels can be dangerous and is usually best avoided."
(Thomas Baguley)

Cohen's own advice for caution

"The terms 'small,' 'medium,' and 'large' are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation...."

Visually assessing correlation

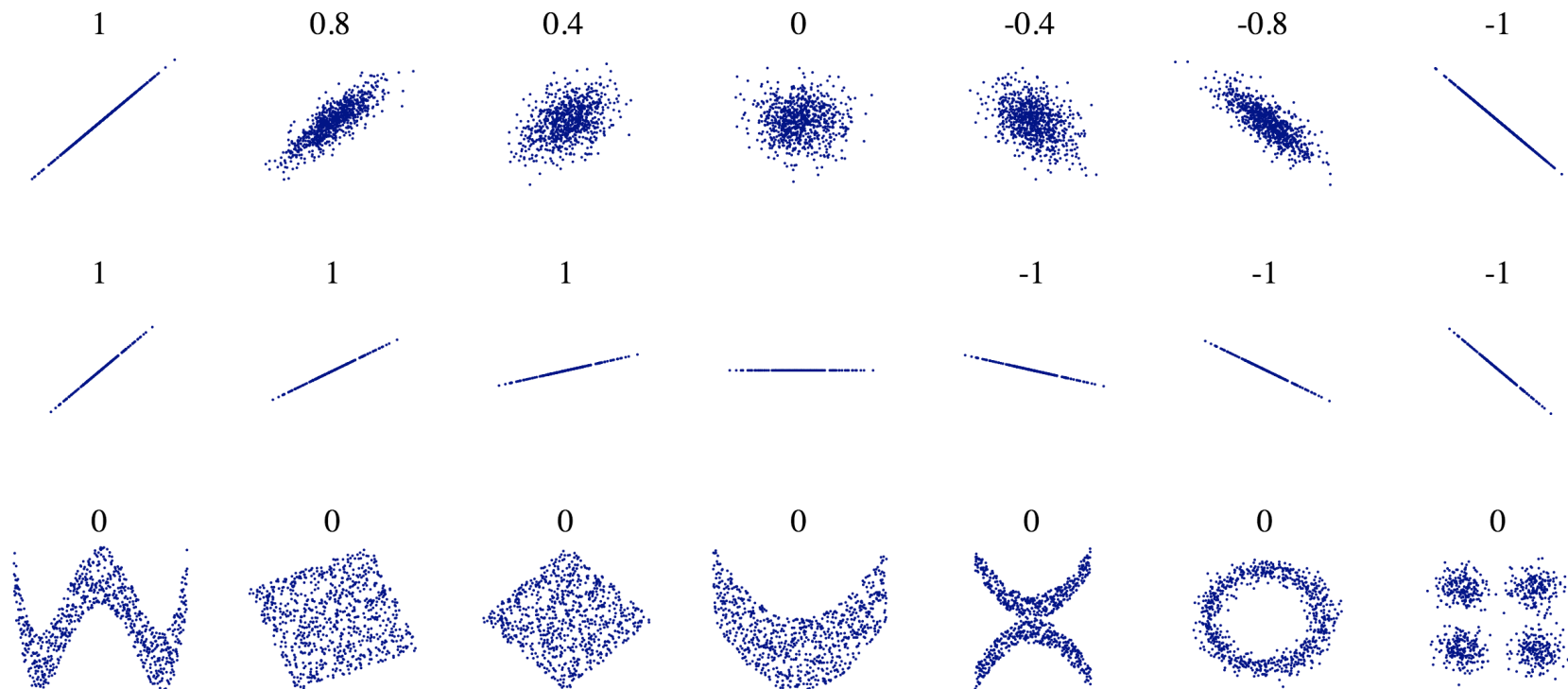
Examples of Pearson's correlation coefficients for different datasets (shown with scatterplots).



(from wikipedia)

Visually assessing correlation

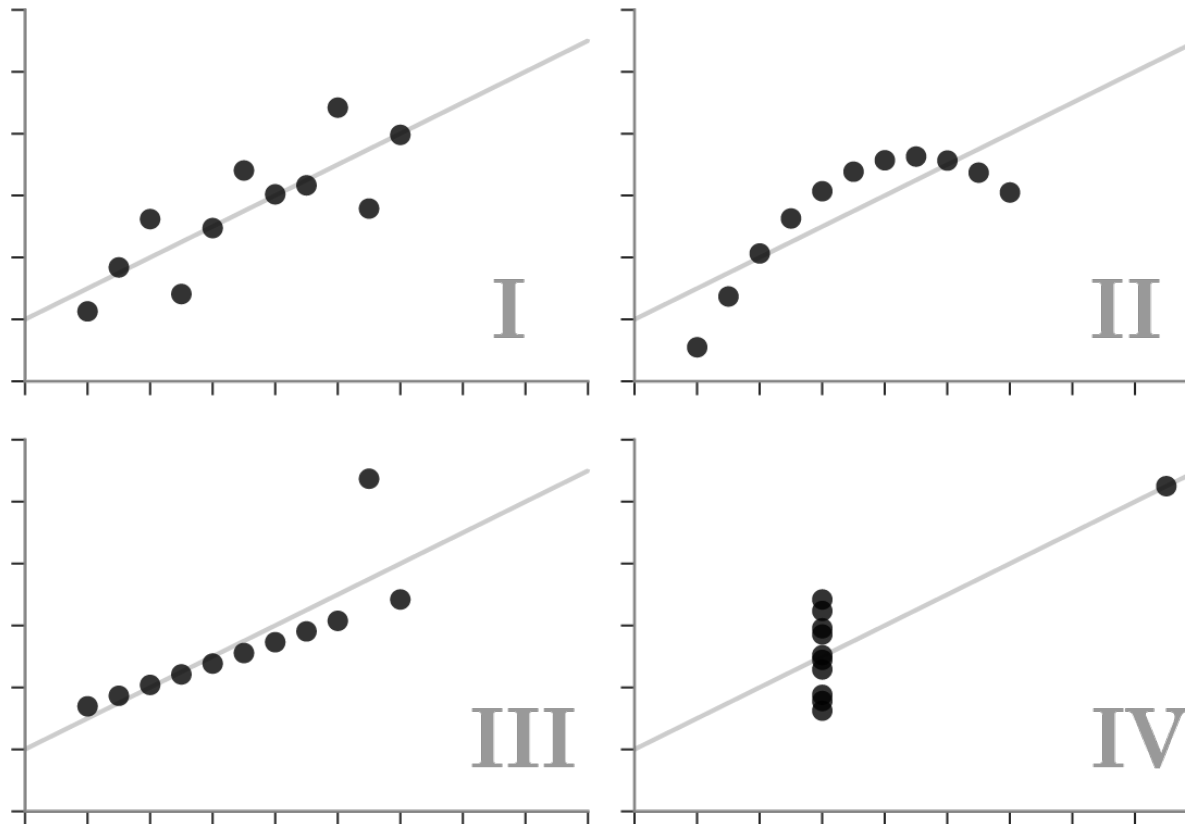
Examples of Pearson's correlation coefficients for different datasets (shown with scatterplots).



(from wikipedia)

Same stats, different graphs

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.

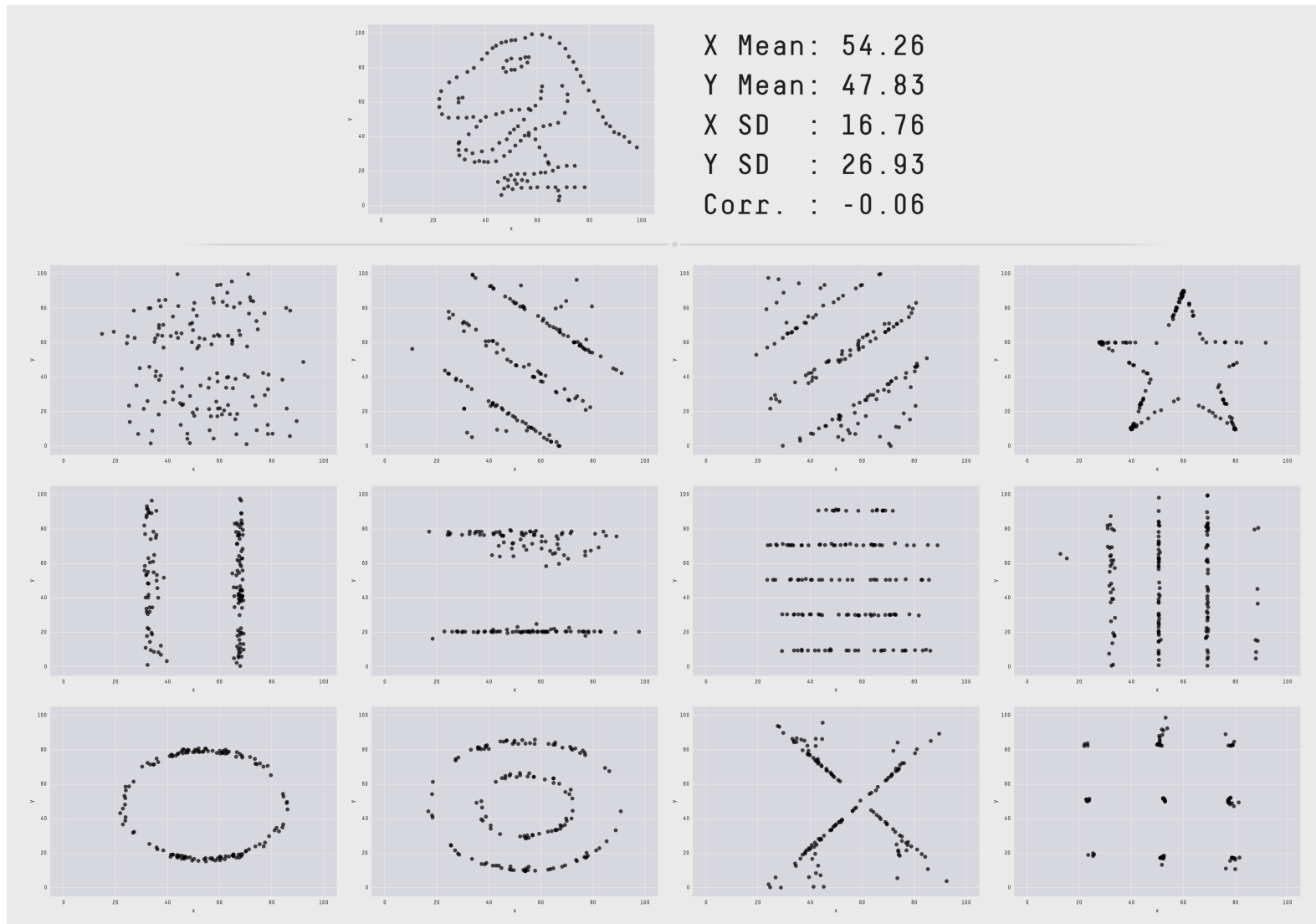


“...make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.” (Anscombe, 1973)

Same stats, different graphs

Matejka & Fitzmaurice (CHI 2017)

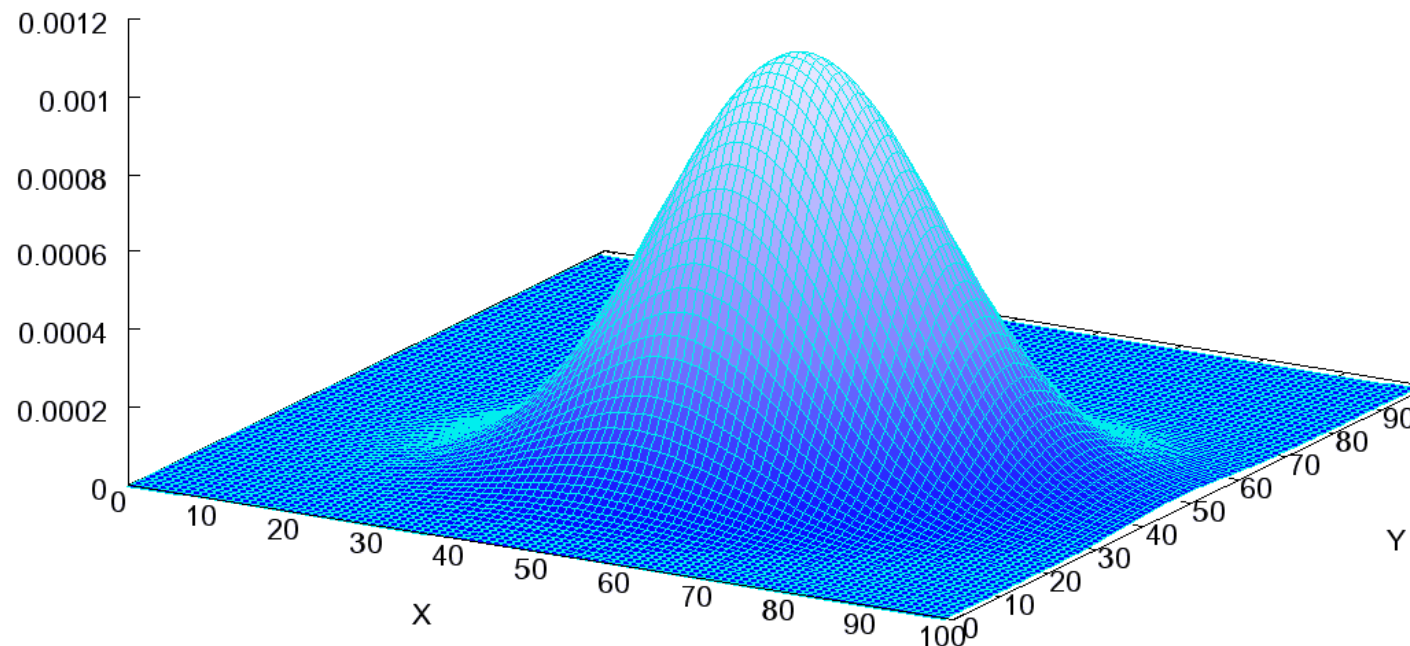
<https://www.autodeskresearch.com/publications/samestats>



Bivariate normal distribution

Extending the notion of normality to two dimensions

A bivariate normal joint density distribution



Bivariate normal distribution

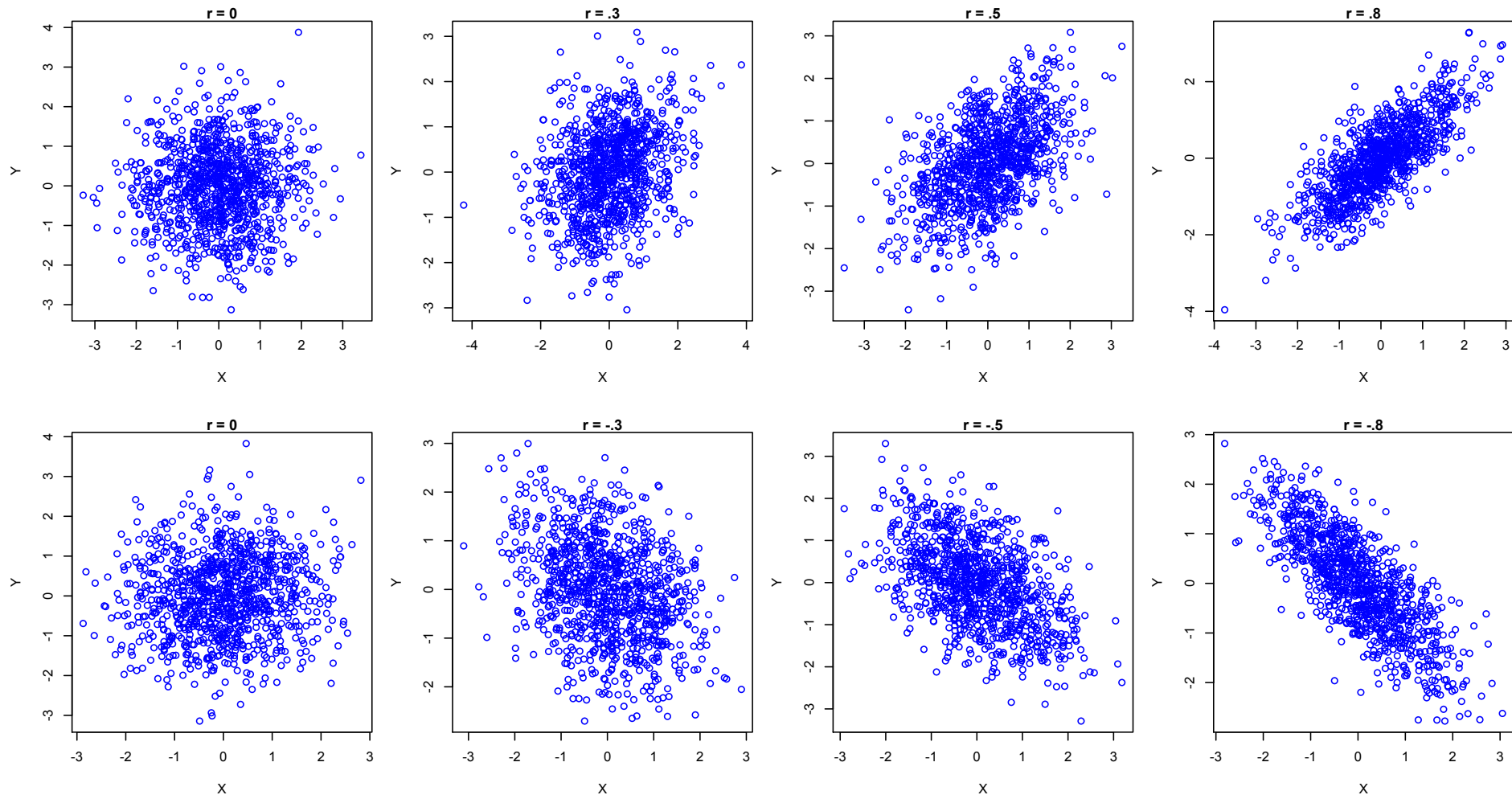
Two variables X and Y are said to be **bivariate normal** or (**jointly normal**):

If their linear combination $aX + bY$ has a normal distribution for all a and b .

Attention: Two variables X and Y may be independently normal (i.e., when $a = 0$ or $b = 0$) **but not jointly normal.**

Bivariate normal distribution

Plotting bivariate normal datasets for positive (top) and negative (bottom) correlation values.



R code

```
# Requires to first install the mvtnorm library
library("mvtnorm")

# mean values for our X, Y variables
means <- c(0, 0)

# Four different covariance matrices that correspond to different
# positive correlation values: 0, .3, .5 and .8
cov.mat.1 <- matrix(c(1,0,0,1), nrow=2, ncol=2)
cov.mat.2 <- matrix(c(1,.3,.3,1), nrow=2, ncol=2)
cov.mat.3 <- matrix(c(1,.5,.5,1), nrow=2, ncol=2)
cov.mat.4 <- matrix(c(1,.8,.8,1), nrow=2, ncol=2)

# Random sampling (n = 1000) from bivariate normal distributions
xy.1 <- rmvnorm(1000, mean = means, sigma = cov.mat.1)
xy.2 <- rmvnorm(1000, mean = means, sigma = cov.mat.2)
xy.3 <- rmvnorm(1000, mean = means, sigma = cov.mat.3)
xy.4 <- rmvnorm(1000, mean = means, sigma = cov.mat.4)

# Plotting the results
par(mfrow=c(2,4), mar = c(4,4,1,1), pty='s', cex.main = 1.1)

plot(xy.1, ylab = "Y", xlab = "X", main = "r = 0", col = 'blue')
plot(xy.2, ylab = "Y", xlab = "X", main = "r = .3", col = 'blue')
plot(xy.3, ylab = "Y", xlab = "X", main = "r = .5", col = 'blue')
plot(xy.4, ylab = "Y", xlab = "X", main = "r = .8", col = 'blue')
```

R code

```
# Requires to first install the mvtnorm library  
library("mvtnorm")
```

```
# mean values for our X, Y variables  
means <- c(0, 0)
```

```
# Four different covariance matrices that correspond to different  
# positive correlation values: 0, .3, .5 and .8  
cov.mat.1 <- matrix(c(1,0,0,1), nrow=2, ncol=2)  
cov.mat.2 <- matrix(c(1,.3,.3,1), nrow=2, ncol=2)  
cov.mat.3 <- matrix(c(1,.5,.5,1), nrow=2, ncol=2)  
cov.mat.4 <- matrix(c(1,.8,.8,1), nrow=2, ncol=2)
```



	X	Y
X	var(X)	cov(X,Y)
Y	cov(Y,X)	var(Y)

Correlation and independence

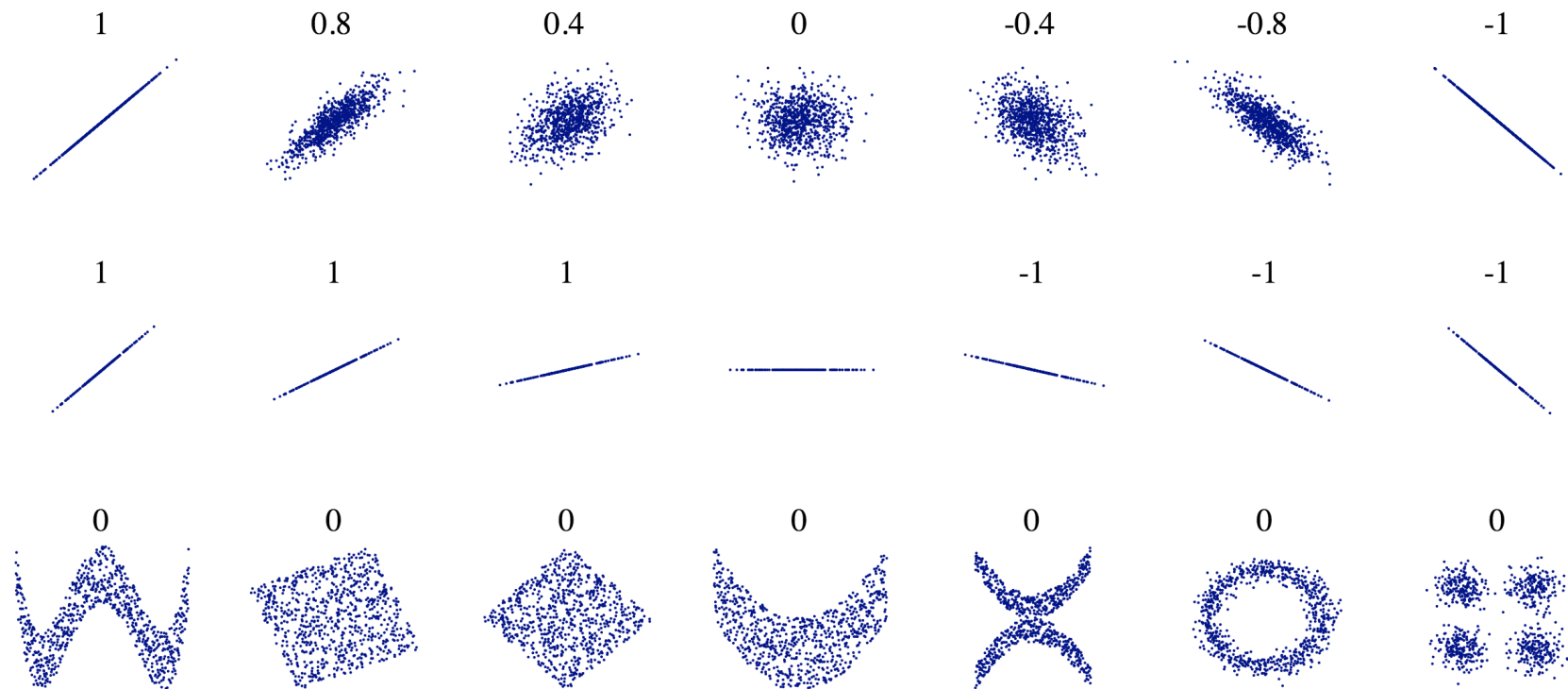
We said that a zero covariance (and a zero correlation) does not imply independence.

However:

If two variables X and Y **are bivariate normal** and **their correlation is zero** ($r_{XY} = 0$), then they are independent.

Correlation and independence

For which of these cases, are X and Y independent?



Correlation and statistical inference

How do we construct the confidence interval for a correlation coefficient?

Problem: The correlation coefficient is bounded between -1 and 1 . Unless the sample size is large, its sampling distribution cannot be assumed as normal.

Fisher's z transformation

Fisher (1921) showed that the following transformation:

$$z_r = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

leads to a normal distribution with a standard error that is approximately equal to $\frac{1}{\sqrt{N-3}}$

Thus, one can construct a CI for z_r and then apply the inverse transformation in order to calculate the CI for the correlation coefficient r .

Fisher's z transformation

What you need to know:

The transformation assumes that **X and Y are bivariate normal**. However, this assumption is often reasonable unless the data are severely skewed or contain extreme outliers.

Example: memory experiment

Provide an interval estimate of the correlation between Presentation Time and Position Error.

Participant	Presentation Time (sec)	Position Error (cm)
1	3	6.3
2	6	3.9
3	9	2.3
4	12	2.0
5	15	2.8
6	18	1.6

Example: R Code

```
> time  
[1] 3 6 9 12 15 18  
> error  
[1] 6.3 3.9 2.3 2.0 2.8 1.6  
> cor.test(time,error)
```

Pearson's product-moment correlation

```
data: time and error  
t = -3.0325, df = 4, p-value = 0.03868  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.98144189 -0.07204004  
sample estimates:  
cor  
-0.8347951
```

Example: conclusion

Pearson's correlation between *Time* and *Error* is

$$r_{Time,Error} = -.83, 95\% \text{ CI } [-.98, -.07]$$

Note that the CI is asymmetric and quite wide (N = 6).

Other correlation coefficients

The **phi (ϕ) coefficient**: For dichotomous variables.

Example: Correlation between gender (women, men) and employment (yes, no)

Spearman's rho: For discrete ordinal variables. Appropriate for monotonic relationships (whether linear or not). Less sensitive to extreme outliers.

Kendall's τ rank coefficient: Alternative to Spearman's rho and has some advantages.

Limitations of correlation coefficients

They make no distinction between *predictor* (X) and *outcome* (Y).

They simply quantify a relationship between two variables X and Y .

Not adequate for creating predictive **statistical models** that connect multiple random variables together.

Regression

Its goal is to ***model*** a relationship between two (or more) random variables

In the simplest case of two variables X and Y , the goal is to determine a function f such that $Y = f(X)$

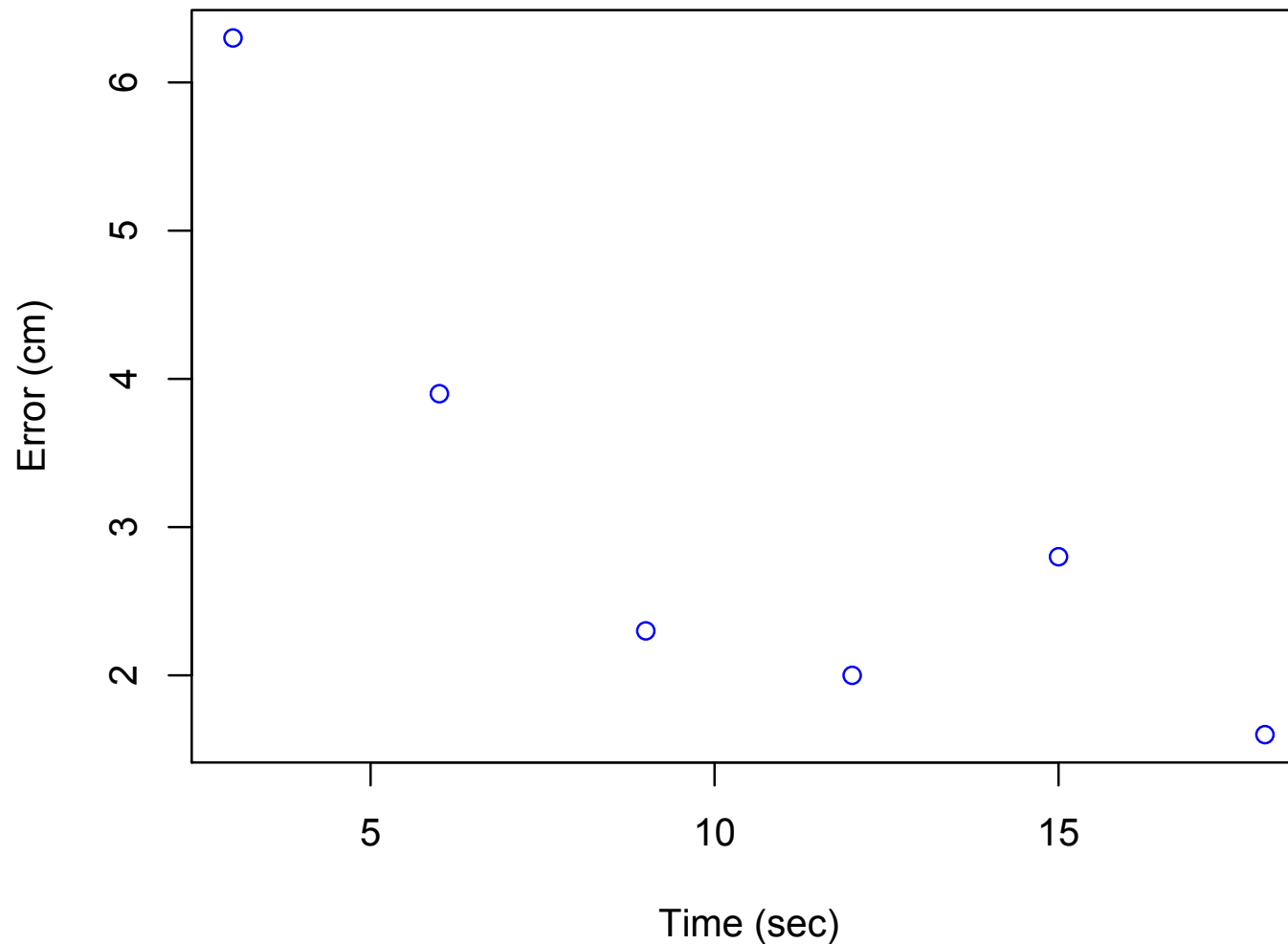
Example: memory experiment

Can we find a function $f: \text{Time} \rightarrow \text{Error}$ that describes the relationship between Presentation Time and Position Error?

Participant	Presentation Time (sec)	Position Error (cm)
1	3	6.3
2	6	3.9
3	9	2.3
4	12	2.0
5	15	2.8
6	18	1.6

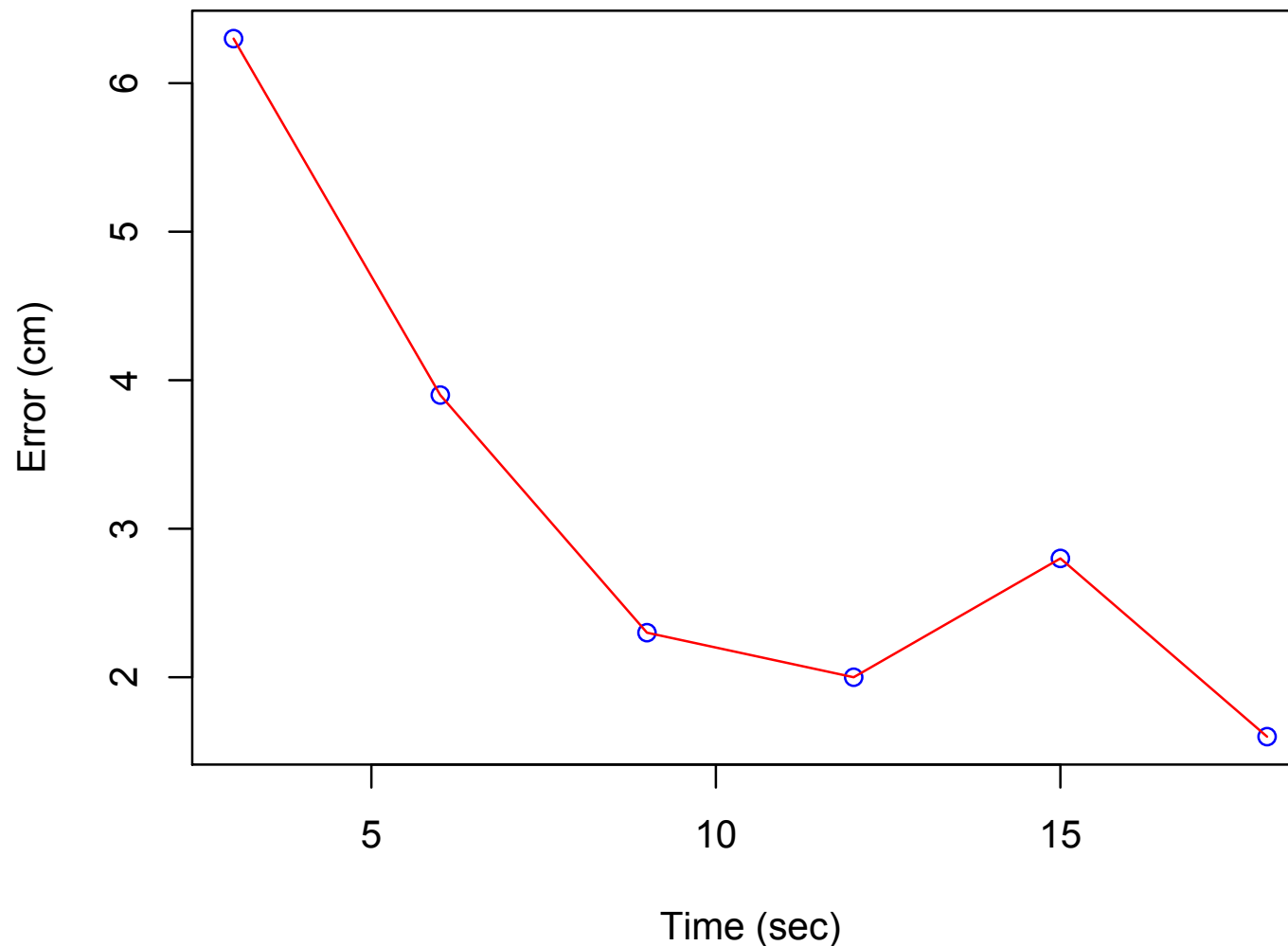
Example: memory experiment

Let's first plot our data.



Example: memory experiment

One might be tempted to choose the function that perfectly fits the sampled data!



Example: memory experiment

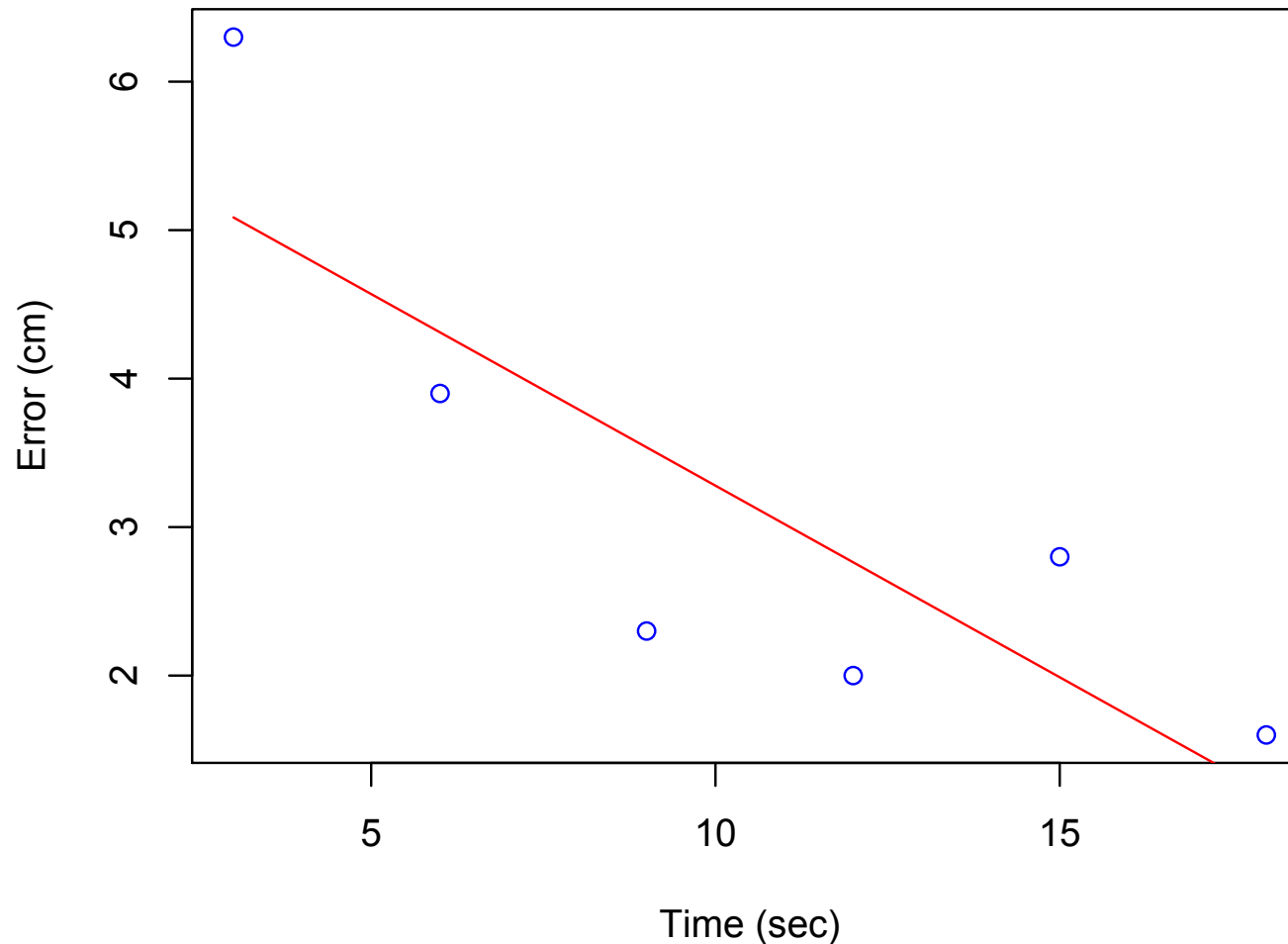
One might be tempted to choose the function that perfectly fits the sampled data!

This is generally a bad approach. Such a function will not help us understand the true relationship between X and Y and make inference about the population.

In the absence of previous theory about this relationship, it makes more sense to start with the simplest possible model.

Example: memory experiment

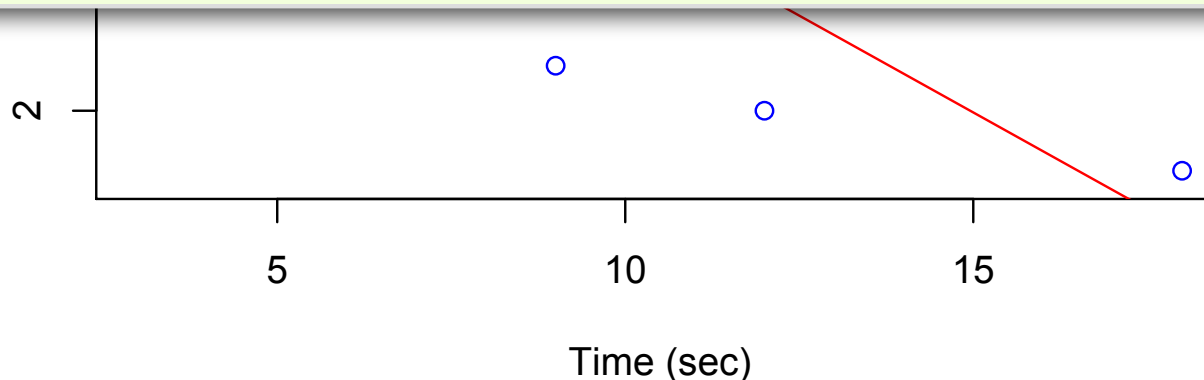
This is a simple linear function, and it is probably a more reasonable choice.



Example: memory experiment

This is a simple linear function, and it is probably a more reasonable choice.

Even a rough model of a relationship is informative. With enough data, more sophisticated models could be possibly derived. But simple models are often better than complex ones.



Linear regression

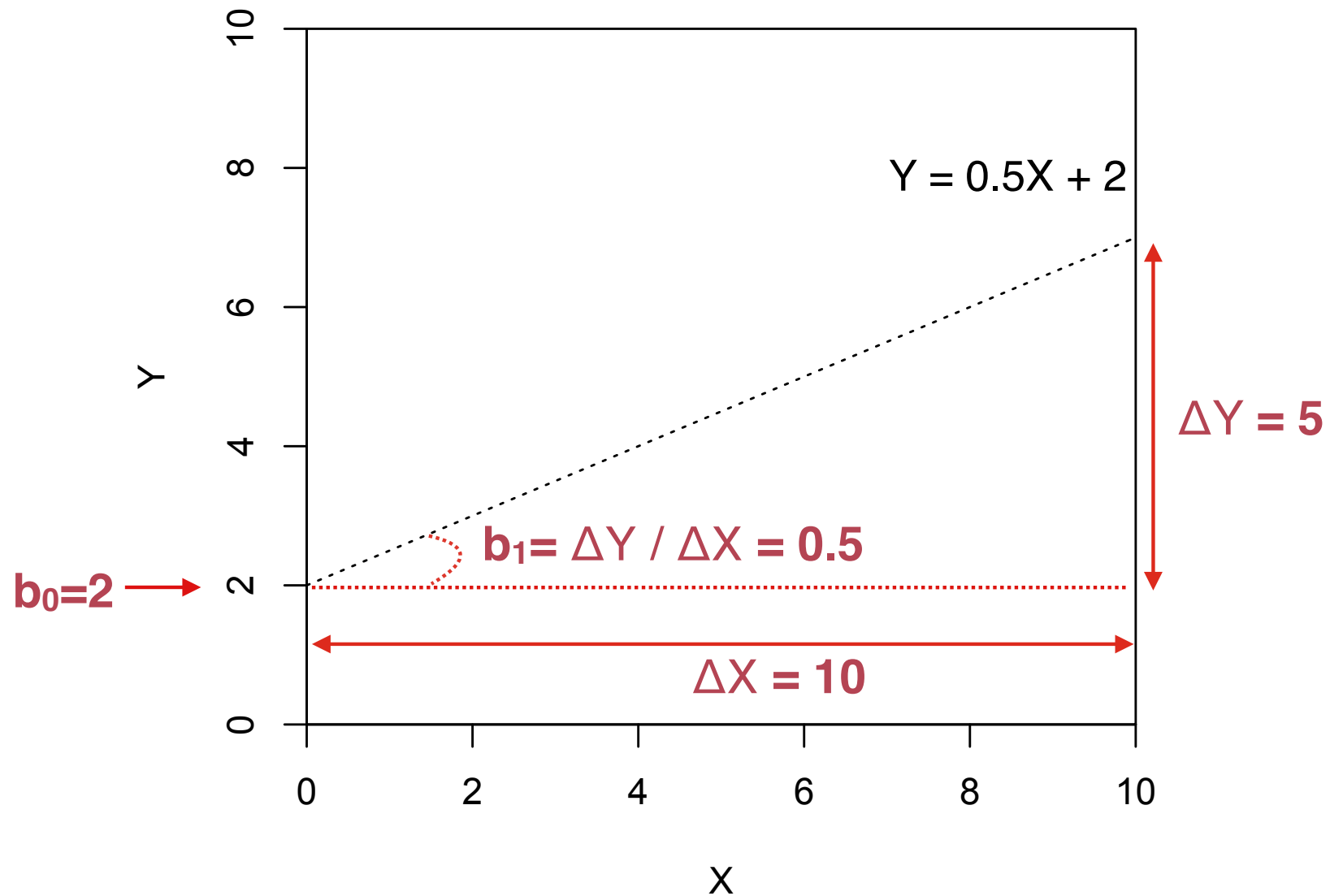
It is an approach for modeling a relationship between X and Y as a linear function:

$$Y = b_0 + b_1X$$

b_0 : known as the **intercept** (or constant)

b_1 : known as the **slope**

Linear functions



R code

Drawing the function:

```
curve(0.5*x+2, ylim = c(0,10), xlim = c(0,10))
```

To make it appear exactly as in the previous slide:

```
curve(0.5*x+2, ylim = c(0,10), xlim = c(0,10),  
      lty = 3, yaxs="i", xaxs="i", ylab="Y", xlab="X")
```

Force starting at (0, 0)

Regression terminology

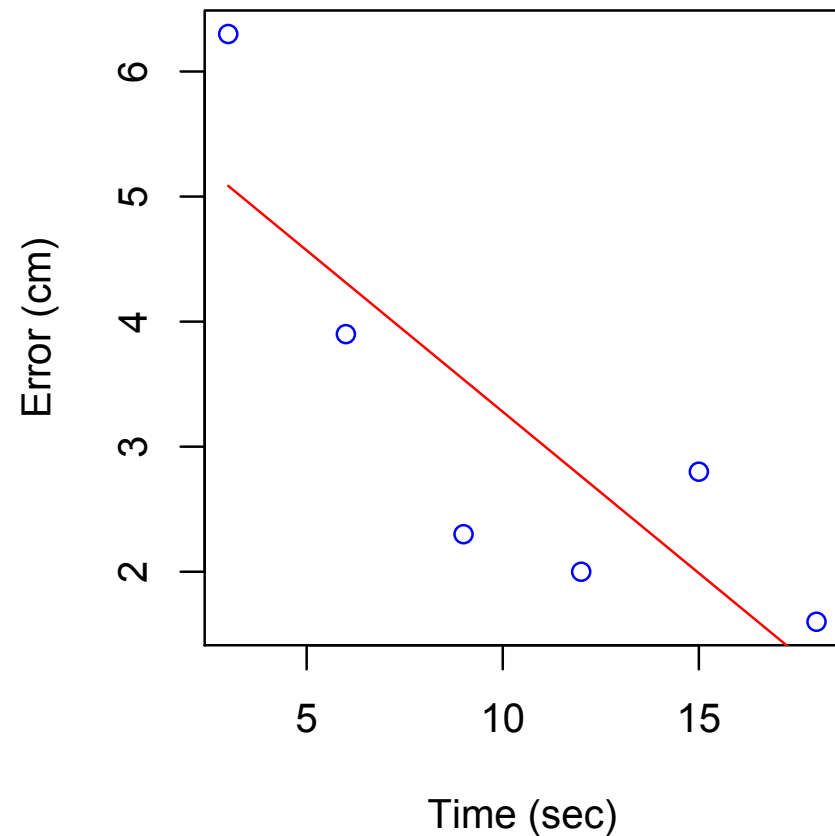
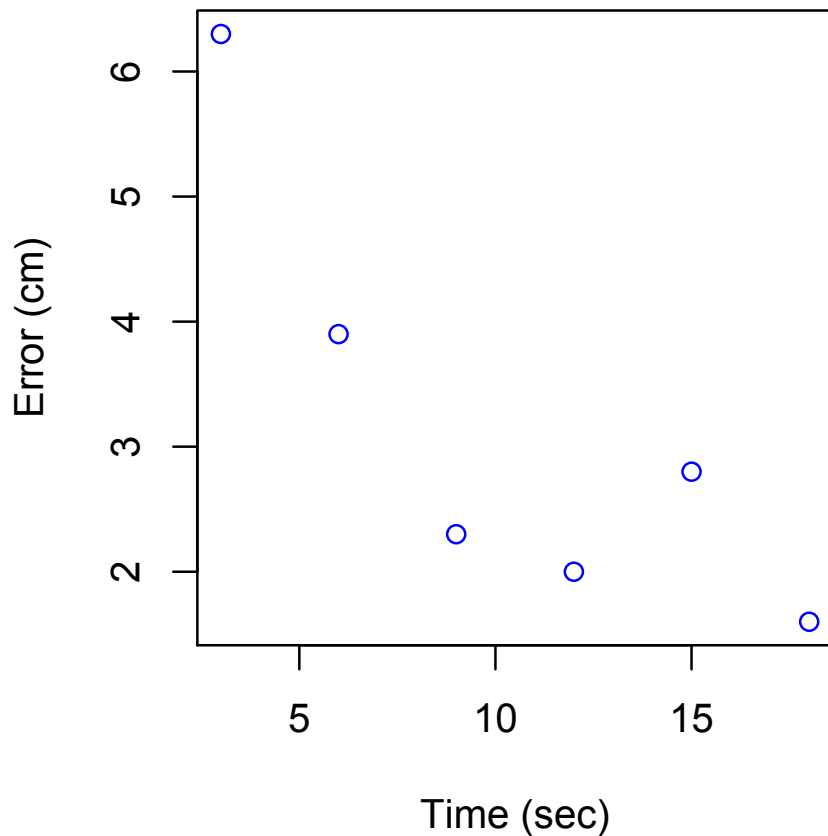
X is known as the *predictor*, and Y is known as the *outcome*

Other terms used for X : *independent variable*, *explanatory variable*, *regressor*, or *covariate*.

Other terms used for Y : *dependent variable*, *response*, *regressand*, *criterion*, or *measurement variable*.

Fitting a line

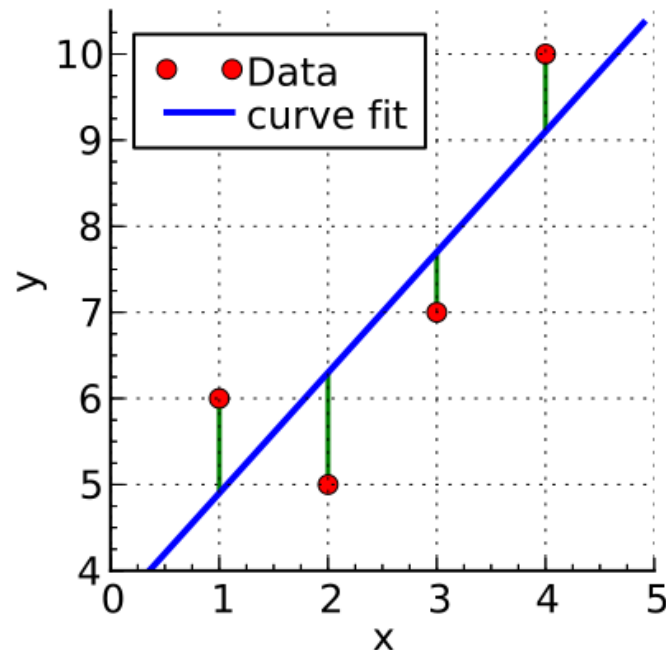
Problem: Giving a sample, how do we decide which line is the **best fit**?



Residuals (errors)

A residual is the difference between the observed value y_i (for a given point x_i) and its predicted outcome \hat{y}_i :

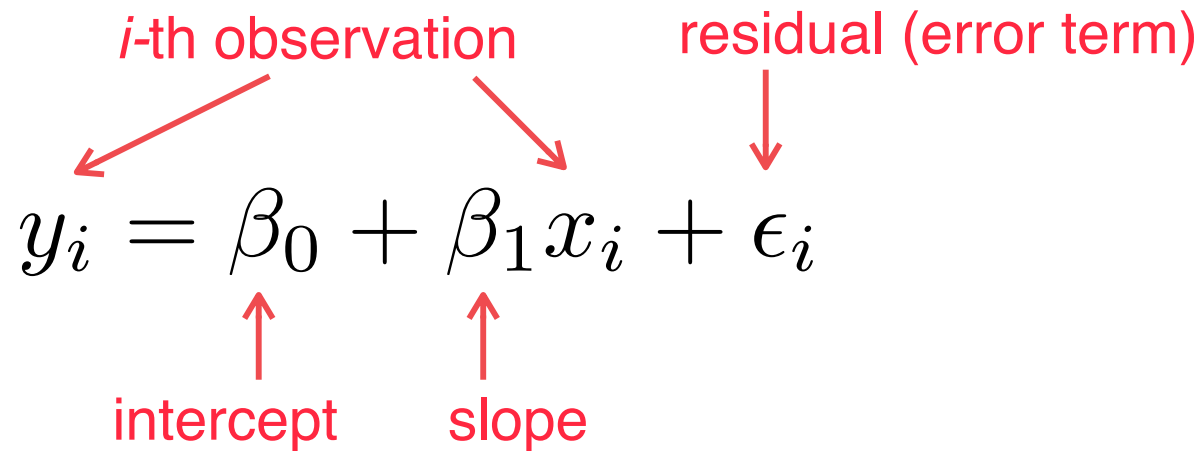
$$e_i = y_i - \hat{y}_i$$



Here, residuals are shown in green.

Other common notation

It's common to express a linear regression model as follows:



The diagram shows the linear regression equation $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with red arrows pointing to each term from descriptive text labels. The label '*i*-th observation' has two arrows pointing to y_i and x_i . The label 'intercept' has an arrow pointing to β_0 . The label 'slope' has an arrow pointing to β_1 . The label 'residual (error term)' has an arrow pointing to ϵ_i .

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

i-th observation residual (error term)

intercept slope

Least squares

The least squares criterion for fitting a line minimizes **the squares of the residuals** or better, it minimizes **the sum of squares of the residuals**:

$$SS_{residual} = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

A different way to think about it:

The least squares criterion **minimizes the variance** of the residuals (vertical errors).

Solution to the least squares criterion

Slope:

$$b_1 = \frac{\sum_{i=1}^N (x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{\sum_{i=1}^N (x_i - \hat{\mu}_X)^2} = \frac{Cov(X, Y)}{Var(X)} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

Intercept:

$$b_0 = \hat{\mu}_Y - b_1 \hat{\mu}_X$$

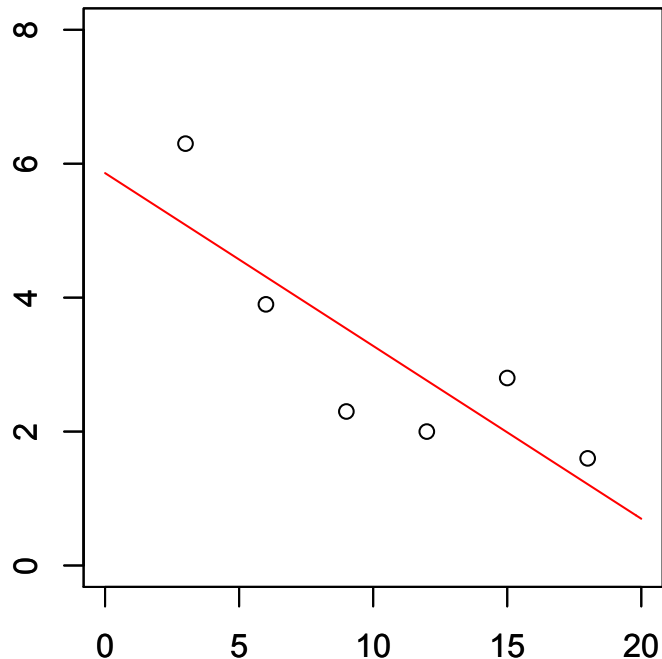
Example: memory experiment

Participant	Presentation Time (sec)	Position Error (cm)
1	3	6.3
2	6	3.9
3	9	2.3
4	12	2.0
5	15	2.8
6	18	1.6

```
> time <- c(3,6,9,12,15,18)
> error <- c(6.3,3.9,2.3,2.0,2.8,1.6)
> b1 <- cov(time,error)/var(time)
> b0 <- mean(error) - b1*mean(time)
> cat("b0 =", b0, "b1 =", b1, "\n")
b0 = 5.86 b1 = -0.2580952
```

Example: memory experiment

Result: The equation for the best-fitting straight line is: $Y = 5.86 - 0.2581X$



```
> plot(time,error, xlim = c(0, 20), ylim = c(0,8))  
> par(new=TRUE)  
> curve(b1*x+b0, xlim = c(0,20), ylim = c(0,8), col = "red", ylab=NA, xlab=NA)
```

Is this the end of the story?

Not really!

Finding the best-fitting line for a sample of X, Y is like finding the best point estimate for a parameter such as the mean.

We need to quantify the quality of the **fit of our model**.

We also need to express the uncertainty about our parameter estimates for the slope (b_0) and the intercept (b_1)

Variance of the error vs. total variance

The variance of the error (residuals) is as follows:

$$\hat{\sigma}_{residual}^2 = \frac{SS_{residual}}{N - 1} = \frac{\sum_{i=1}^N e_i^2}{N - 1} = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N - 1}$$

The overall variance of Y is as follows:

$$\hat{\sigma}_Y^2 = \frac{SS_{total}}{N - 1} = \frac{\sum_{i=1}^N (y_i - \hat{\mu}_Y)^2}{N - 1}$$

Unexplained variance

The following ratio is commonly use to express the **proportion of unexplained variance** in the sample:

$$\frac{\hat{\sigma}_{residual}^2}{\hat{\sigma}_Y^2} = \frac{SS_{residual}}{SS_{total}}$$

This quantity gives the variance of the error as a percentage of the total variance in the outcome Y . Thus, it is used as a measure of **the lack of fit** of our model.

The R-square (R^2) measure

The explained variance is commonly known as the *R-square*:

$$R^2 = \frac{SS_{total} - SS_{residual}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

Example: memory experiment

Best-fitting line: $Y = 5.86 - 0.2581X$

Participant	Presentation Time (sec)	Position Error (cm)	Predicted Position Error (cm)	Residual
1	3	6.3	5.0857	1.2143
2	6	3.9	4.3114	-0.4114
3	9	2.3	3.5371	-1.2371
4	12	2.0	2.7629	-0.7629
5	15	2.8	1.9886	0.8114
6	18	1.6	1.2143	0.3857

Example: memory experiment

```
# Example of how to derive R square for a linear regression  
# where X = time and Y = position error
```

```
time <- c(3,6,9,12,15,18)  
poserror <- c(6.3,3.9,2.3,2.0,2.8,1.6)
```

```
# Derive the slope and the intercept with least squares
```

```
b1 <- cov(time,poserror)/var(time)  
b0 <- mean(poserror) - b1*mean(time)
```

```
# Define the regression function
```

```
f <- function(x){b0 + b1*x}
```

```
# Find the predicted position error values
```

```
poserror.predicted <- f(time)
```

```
# Find the residuals
```

```
residuals <- poserror - poserror.predicted
```

```
# Calculate R-squared
```

```
R2 <- 1 - var(residuals) / var(poserror)
```

```
cat("R squared =", R2)
```

Example: memory experiment

Result: The equation for the best-fitting straight line is:
 $Y = 5.86 - 0.2581X$

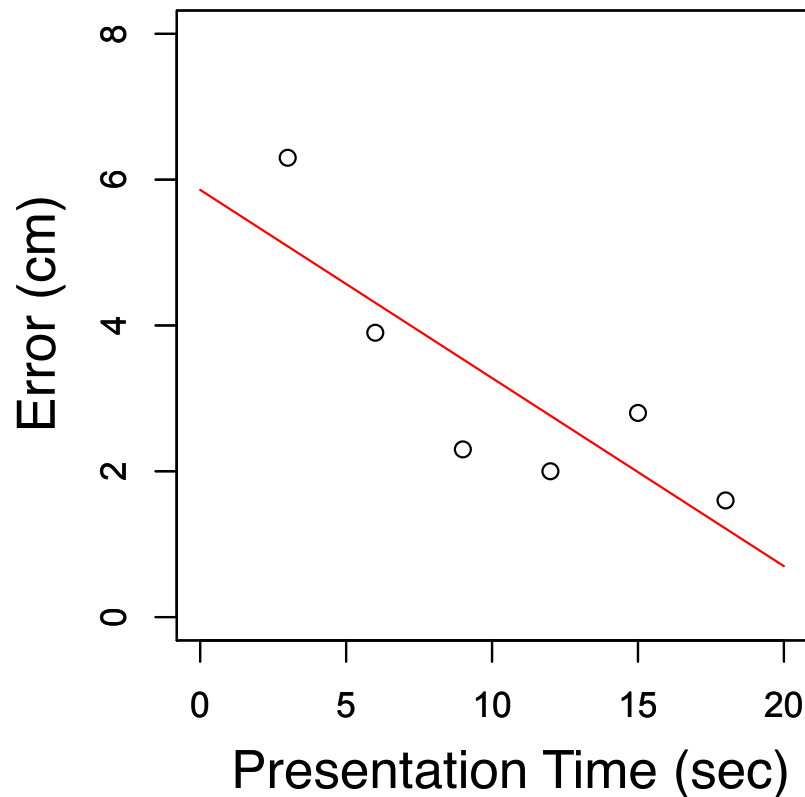
The explained sample variance is **$R^2 = .697$**

This means that the predictor (Time Presentation) explains 69.7% of the total sample variance in a simple linear model.

Things to consider (1)

For this particular scenario, how reasonable do you think a linear model is?

What happens if presentation time increases further (> 20 sec)?



Things to consider (2)

The above statistical model considers that the values for the predictor X (e.g., *presentation times*) are measured or controlled with no error.

This is normally a reasonable assumption for most experimental designs. However, it may not be always the case.

Things to consider (3)

As the sample size becomes small, finding a good fit by chance becomes easier.

For $N = 2$, the line always produces a perfect fit.

Adjusted R²

The R-square measure tends to overestimate the fit when the sample size N is low with respect to the number q of predictors.

Thus, the adjusted R-squared measure is often used:

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - q - 1}$$

For a simple linear regression with one predictor: $q = 1$

For our example:

$$adj. R \text{ squared} = 1 - (1 - .697) * (6 - 1) / (6 - 2) = .621$$

Correlation vs. R square

The Pearson correlation r between X and Y coincides with R : $R^2 = r^2$

```
> cor(time, error)^2  
[1] 0.6968829
```

The R square is more generic and is also used for regression analysis with multiple predictors.

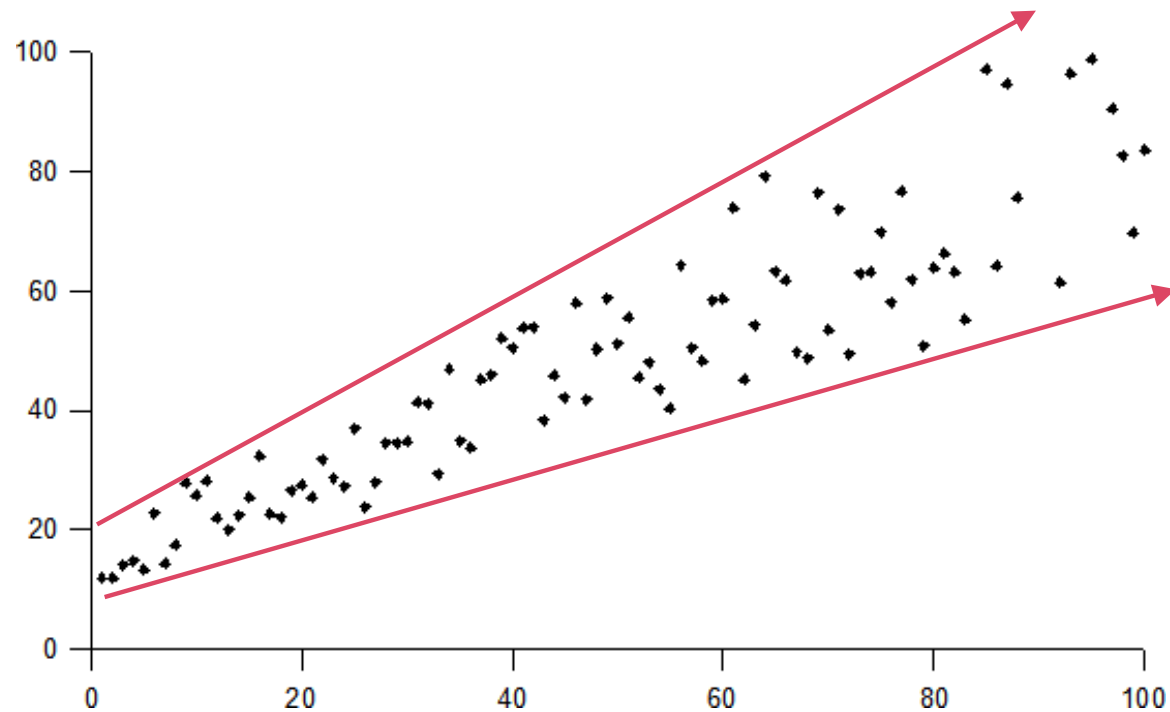
Statistical inference

We can construct confidence intervals (and t tests) for the slope (b_1) and the intercept (b_0) with the following assumptions for the **residuals (errors)**:

They are **independently** sampled from a **normal distribution** with a **constant variance**

Common assumption violations

Unequal variance (or *heteroscedasticity*)



Statistical inference

We can construct confidence intervals (and t tests) for the slope (b_1) and the intercept (b_0) with the following assumptions for the **residuals (errors)**:

They are **independently** sampled from a **normal distribution** with a **constant variance**

Explaining the full method for deriving such confidence intervals is out of the scope of this course.

R code

But we can use *R*'s linear model function to conduct a complete linear regression.

```
> lm(error ~ time)|
```

Call:

```
lm(formula = error ~ time)
```

Coefficients:

(Intercept)	time
5.8600	-0.2581

R code

```
> model <- lm(error ~ time)
```

```
> summary(model)
```

Call:

lm(formula = error ~ time)

Residuals:

1	2	3	4	5	6
1.2143	-0.4114	-1.2371	-0.7629	0.8114	0.3857

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.86000	0.99436	5.893	0.00415	**
time	-0.25810	0.08511	-3.033	0.03868	*

Signif. codes:

0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.068 on 4 degrees of freedom

Multiple R-squared: 0.6969,

Adjusted R-squared: 0.6211

F-statistic: 9.196 on 1 and 4 DF, p-value: 0.03868

R code

Getting the 95% CIs of the intercept and the slope:

```
> model <- lm(error ~ time)
> confint(model)
```

	2.5 %	97.5 %
(Intercept)	3.0992265	8.62077346
time	-0.4943956	-0.02179484

Results summary:

$b_0 = 5.86, 95\% \text{ CI } [3.10, 8.62]$

$b_1 = -0.26, 95\% \text{ CI } [-0.49, -0.02]$

Fitting non-linear relationships with linear regression

There may be theoretical reasons to a non-linear function, such as:

$$Y = b_0 + b_1 \ln(X), \text{ or}$$

$$\ln(Y) = b_0 + b_1 X, \text{ or}$$

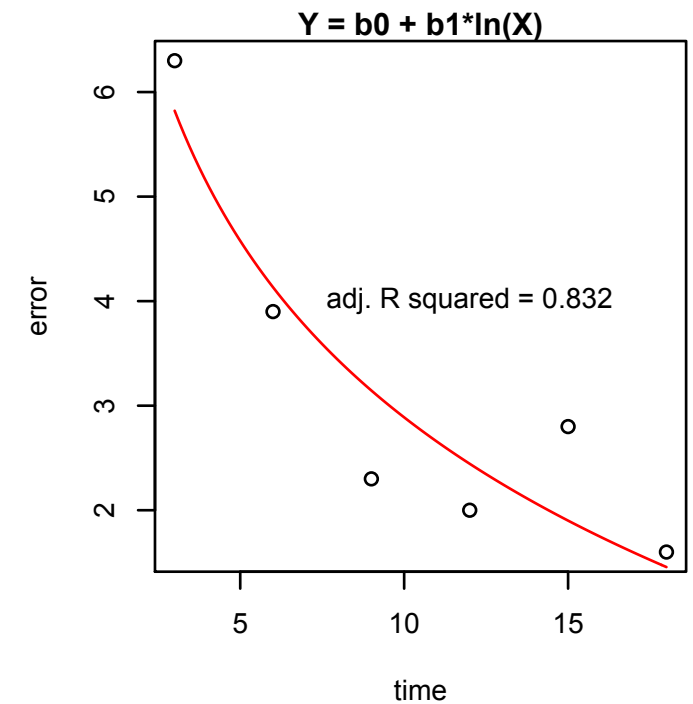
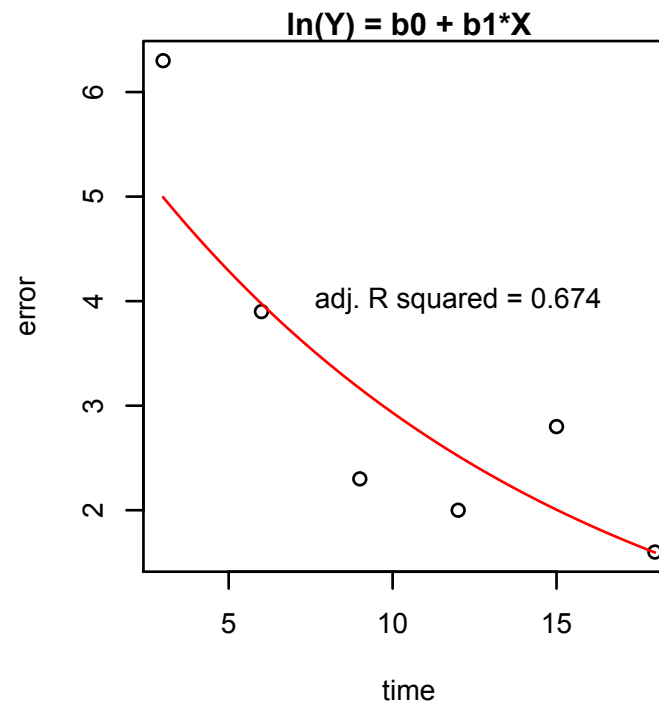
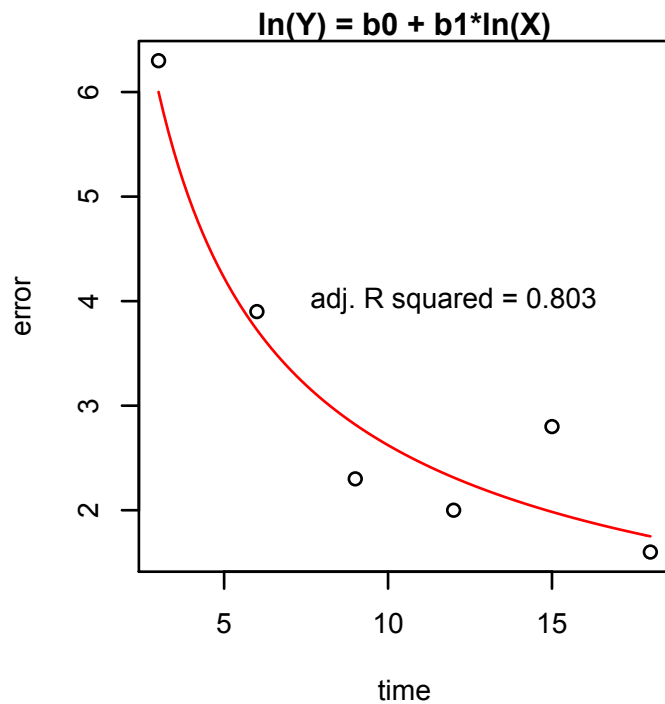
This assumes a lognormal distribution of errors as well as constant variances.

$$\ln(Y) = b_0 + b_1 \ln(X)$$

All these can be handled as linear regressions. However, all assumptions apply now to the residuals (errors) of the transformed variables.

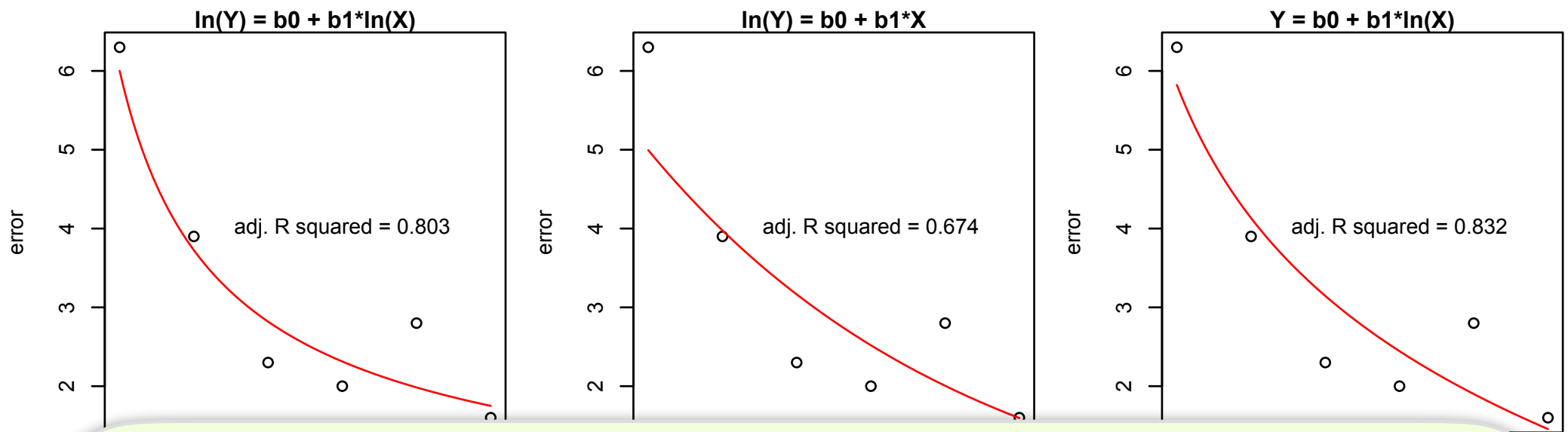
Example: memory experiment

Alternative models.



Example: memory experiment

Alternative models.

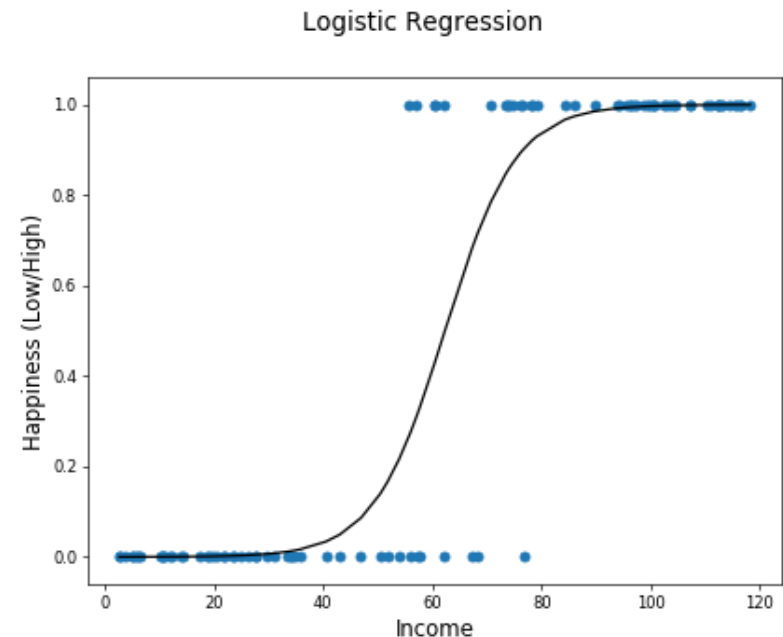


Logarithms may have a theoretical justification for variables that are bounded by zero. The last model results in the best fit, but this does not mean that it is the best one. The first model is an attractive alternative given that the distance error cannot be negative.

Logistic regression

The outcome Y is a binary (Bernoulli) variable with parameter $p = P(Y=1)$, that is, the probability that $Y = 1$. The model is reduced to a linear function as follows:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_i$$



Commonly used for binary classification.

Multiple linear regression

More than one predictors (independent variables):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$



residual (error term)

Generalized Linear Model

A generalization of the linear regression model that unifies a wide range of regression models: linear regression, logistic regression, Poisson regression, etc.

R provides an implementation through its *glm* function.

It's out of the scope of this class.