Undirected network reconstruction - part 2

Wessel van Wieringen
w.n.van.wieringen@vu.nl

Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

Two-gene pathway

Two-gene pathway

Two-gene pathways comprise two genes, and ignore the possibility there may be more.

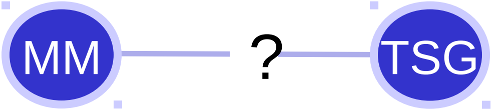
Cancer research example

Y_2 : gene expression measurements of a tumor suppressor gene

Y_1 : gene expression of a methylation marker

Question

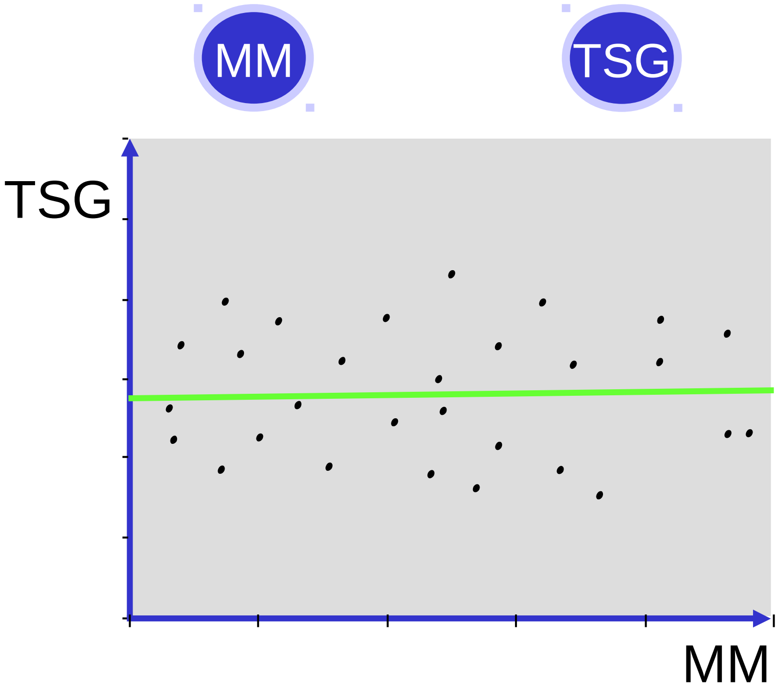
Does the methylation marker (MM) influence the expression of the tumor suppressor gene (TSG)?



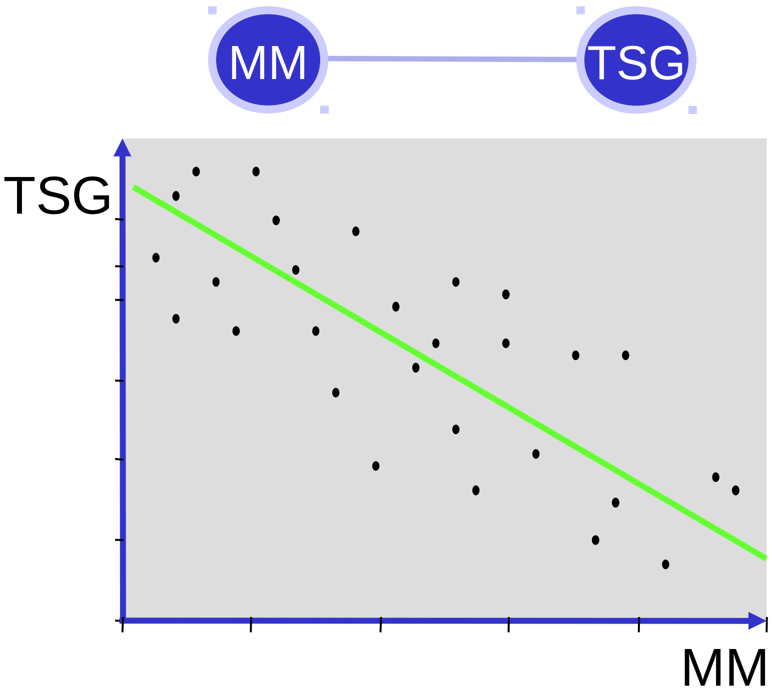
Two-gene pathway

TSG suppresses tumorigenesis. Ideally, its expression levels are high. If the expression levels of MM and TSG are dependent, we may aim to control those of TSG via MM.

Independence



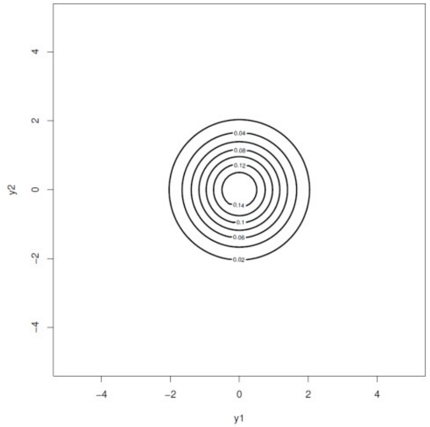
Dependence



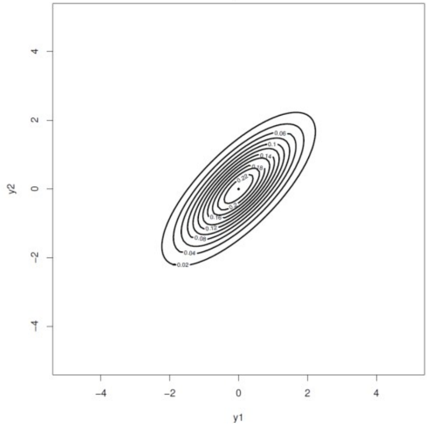
Correlation

Two-gene system

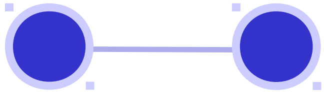
Calculate correlation between any two genes. If the correlation is large (in some sense), the two genes interact.



r = 0.027



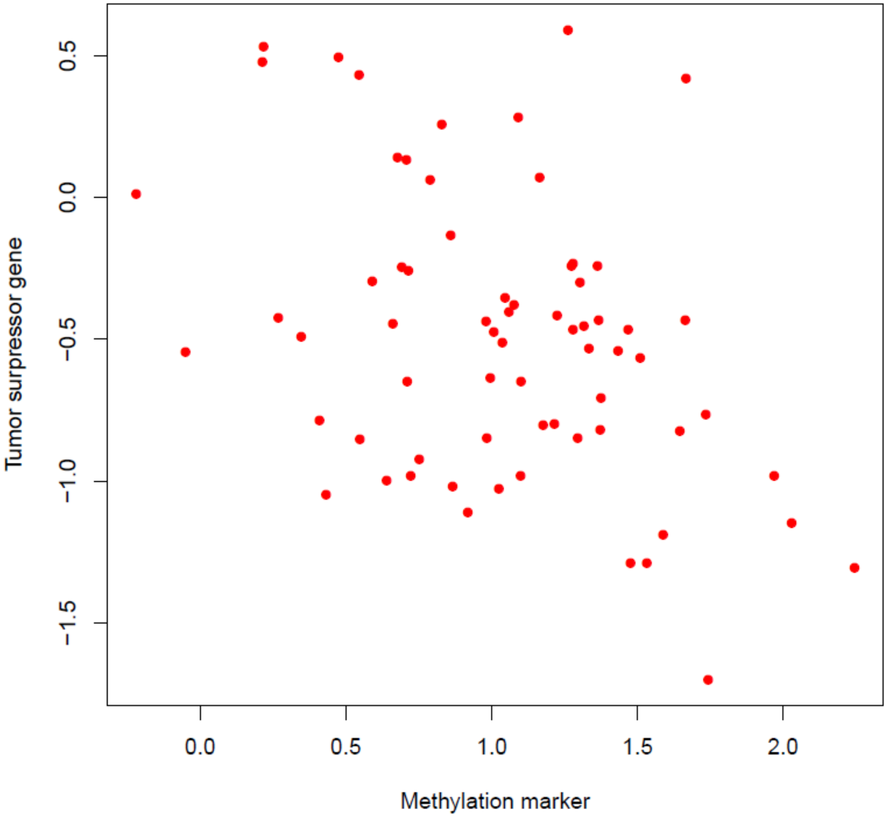
r = 0.693



Correlation

Cancer research example

Expression levels of the TSG vs. MM



Question: is there dependence between TSG and MM?

Correlation

Cancer research example

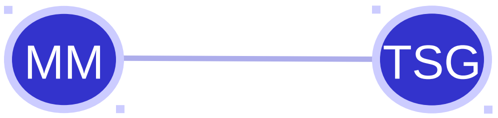
```
> cov(cbind(MM, TSG))
             MM         TSG
MM   0.23897377 -0.09787409
TSG -0.09787409  0.25388099

> cor(cbind(MM, TSG))
           MM       TSG
MM   1.000000 -0.397354
TSG -0.397354  1.000000

> rho <- cor(MM, TSG)
> T <- log((1+rho)/(1-rho))/2
> sd <- sqrt(1/(length(MM)-3))
> pvalue <- 2*pnorm(T, sd=sd)
> pvalue
[1] 0.0006984108
```

Conclusion

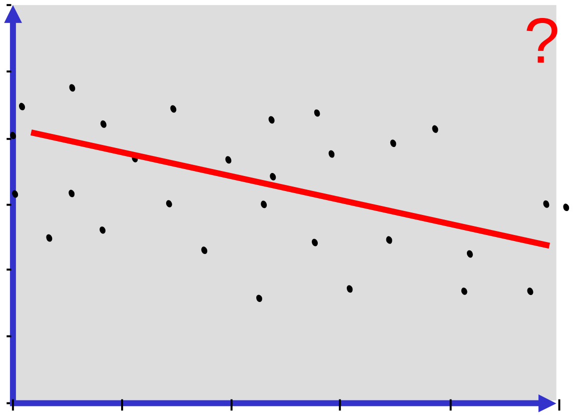
Significant association between of MM on TSG.



Note: the edge is undirected, the two variates are (causally) on a par.

Regression

Instead of using some measure to assess the dependence between two variables, one may explicitly model their relationship.



Regression analysis is a statistical method to estimate the relation among variables. E.g.:

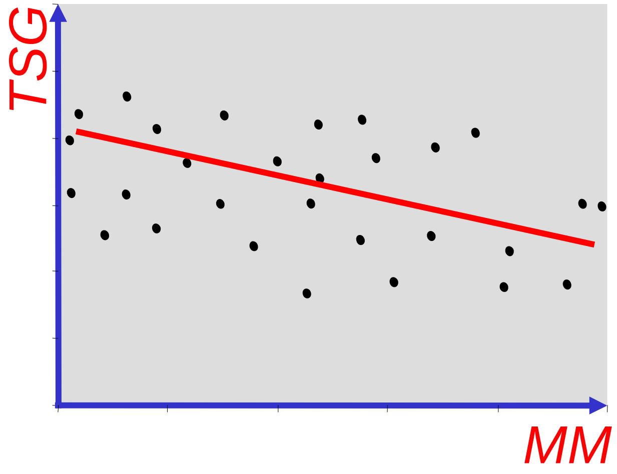
$$Y_{\mathrm{tsg}} = \mathrm{f}(Y_{\mathrm{mm}}) + \mathrm{error}$$

where f() is some function deemed appropriate. Commonly, f() is taken to be linear (as a first order approximation).

Two-gene pathway & regression

More formally, the simple linear regression model:

$$\underset{\text{TSG}}{\underline{Y_i}} = \beta_0 + \beta_1 \underset{\text{MM}}{\underline{X_i}} + \varepsilon_i$$



Some nomenclature:

$$\underbrace{Y_i}_{\text{response or dependent variable}} = \underbrace{\beta_0 + \beta_1}_{\text{regression parameter}} \underbrace{X_i}_{\text{covariate or explanatory variable}} + \underbrace{\varepsilon_i}_{\text{error} \approx \text{part of } Y \text{ not explained by the model}}$$

response or dependent variable

regression parameter

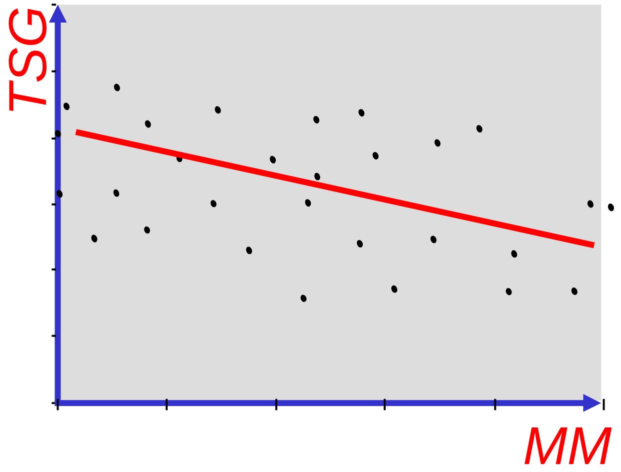
covariate or explanatory variable

error ≈ part of Y not explained by the model

Regression

More formally, the simple linear regression model:

$$\underset{\text{TSG}}{\underline{Y_i}} = \beta_0 + \beta_1 \underset{\text{MM}}{\underline{X_i}} + \varepsilon_i$$



with ε_i normally distributed with:

$$\begin{aligned} E(\varepsilon_i) &= 0 \\ \mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) &= \begin{cases} \sigma^2 & \text{if} \quad i_1 = i_2 \\ 0 & \text{if} \quad i_1 \neq i_2 \end{cases} \end{aligned}$$

In the above the unknown parameters are: β_0, β_1, σ^2.

Regression

Note

We write:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

while it is equivalent to write:

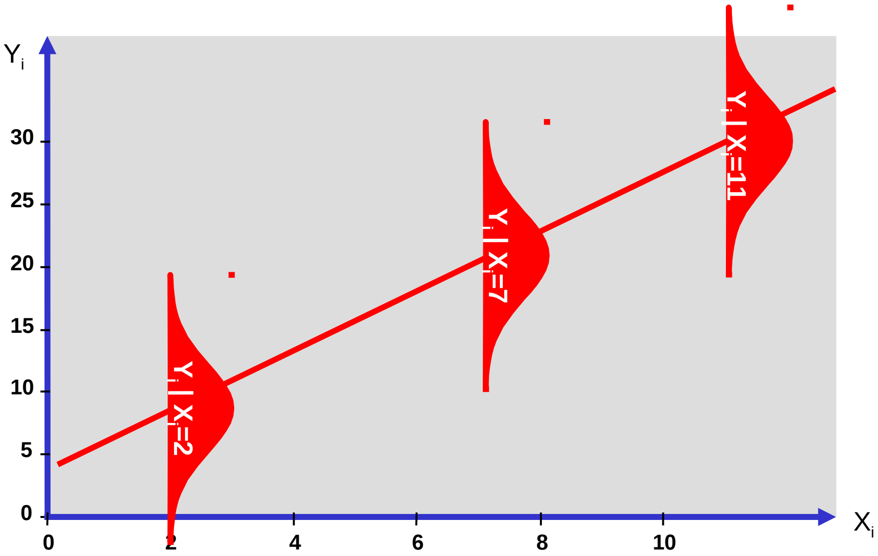
$$Y_i \,|\, X_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$

The latter explicitly assumes that the explanatory variable X_i is (temporarily) taken as non-random. It is to be read as: Y_i *conditional* on X_i is distributed as ….

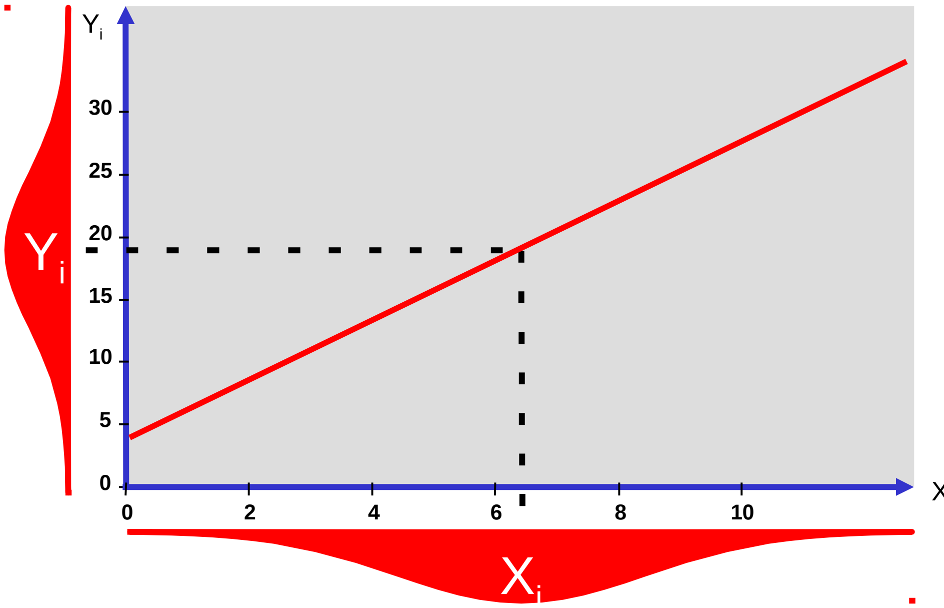
Regression

Conditional vs. marginal

The *conditional* distribution of Y_i on X_i



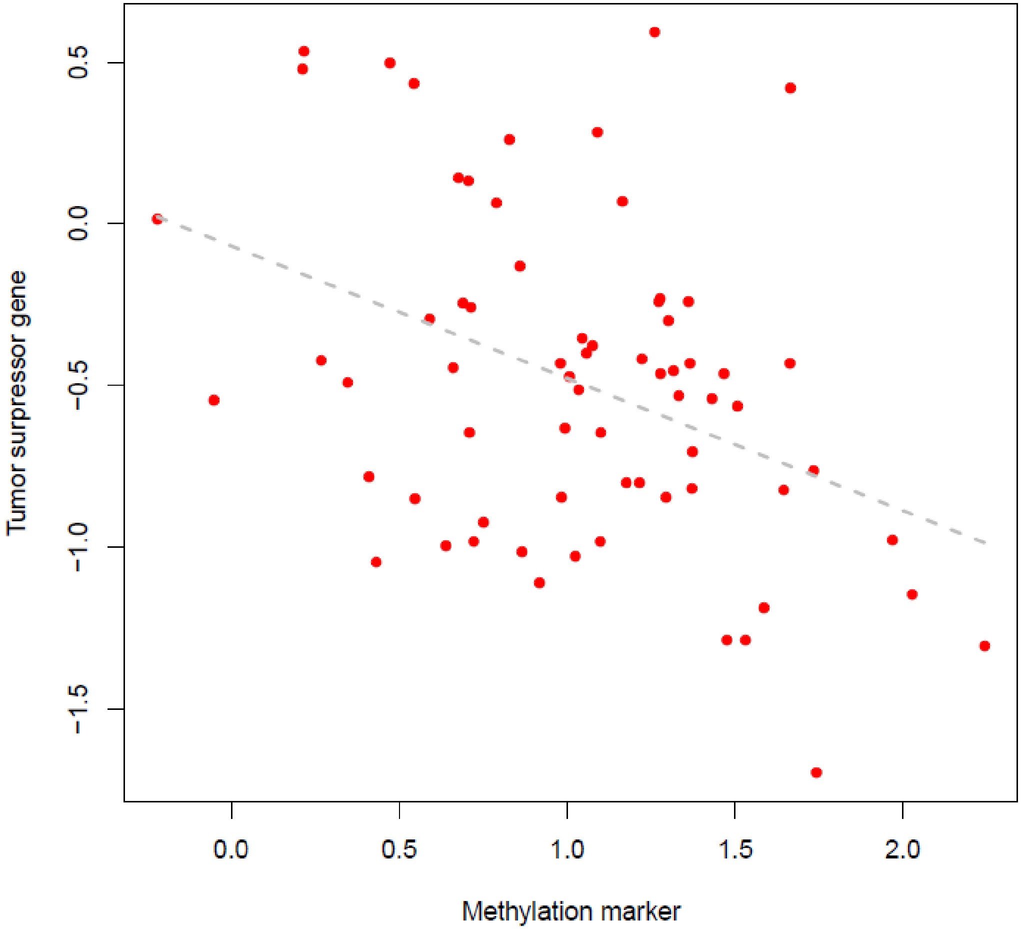
The unconditional (*marginal*) distribution of Y_i



Regression

Cancer research example

```
> plot(TSG ~ MM, ...)
> lines(regressionResults$fitted.values ~ MM, ...)
```



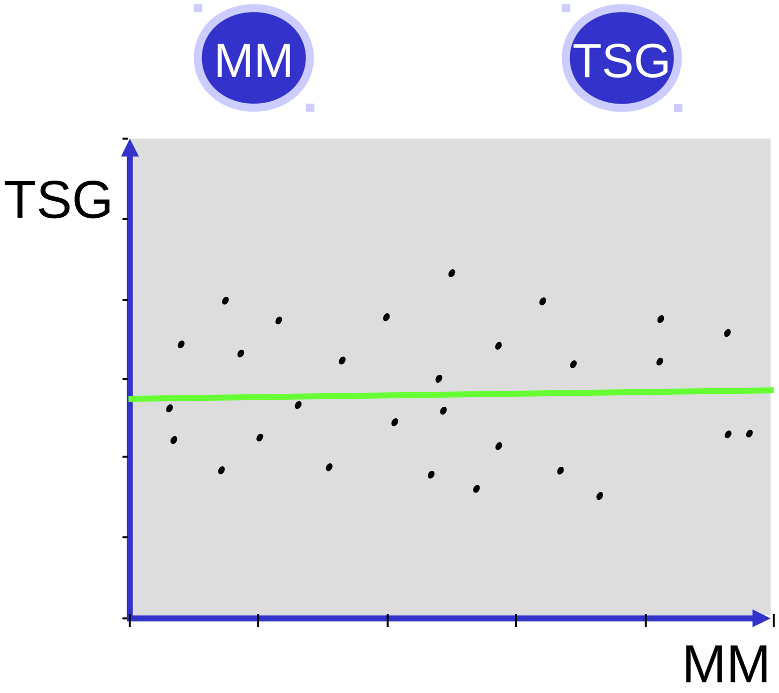
Conclusion
Significant effect of MM on TSG.

Thus, β ≠ 0. Hence, gene expression levels of MM and TSG are related.

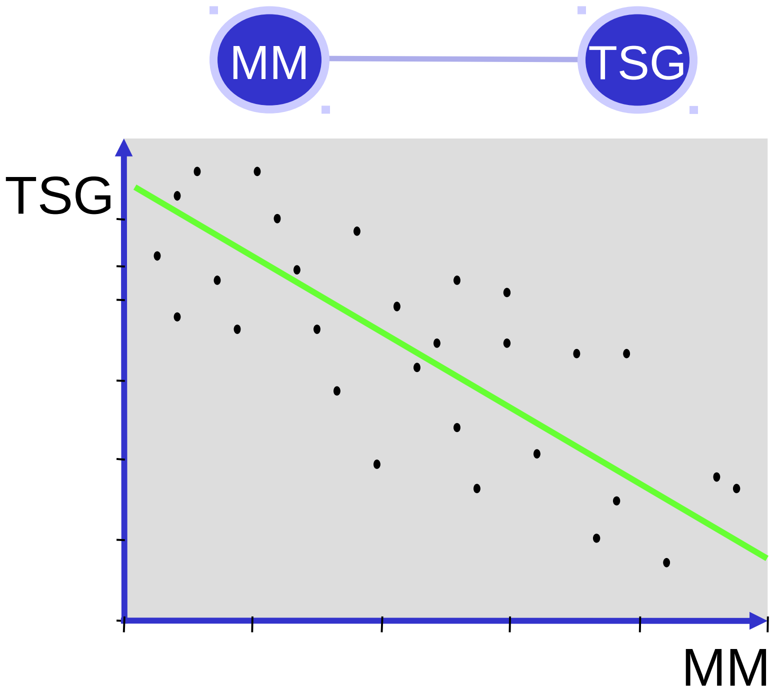
Correlation vs. regression

Cancer research example

In a two-gene pathway ρ=0 implies independence between its genes. What does β_1=0 say about independence?



Independence



Dependence

Correlation vs. regression

Assume $(Y_{MM}, Y_{TSG})^{\mathrm{T}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Consider the regression equations:

$$\begin{aligned} Y_{\mathrm{TSG}} &= \beta_{0,\mathrm{TSG}} + \beta_{1,\mathrm{MM}} Y_{\mathrm{MM}} + \varepsilon_{\mathrm{TSG}} \\ Y_{\mathrm{MM}} &= \beta_{0,\mathrm{MM}} + \beta_{1,\mathrm{TSG}} Y_{\mathrm{TSG}} + \varepsilon_{\mathrm{MM}} \end{aligned}$$

with:

$$\mathrm{Var}(\varepsilon_{\mathrm{MM}}) = \mathrm{Var}(Y_{\mathrm{MM}} \,|\, Y_{\mathrm{TSG}}) = \sigma^2_{\mathrm{MM}}$$
$$\mathrm{Var}(\varepsilon_{\mathrm{TSG}}) = \mathrm{Var}(Y_{\mathrm{TSG}} \,|\, Y_{\mathrm{MM}}) = \sigma^2_{\mathrm{TSG}}$$

Question

What is the relation between the *β*'s and *ρ*?

Correlation vs. regression

Then:

$$\beta_{\mathrm{MM}} = \frac{\sqrt{\mathrm{Var}(Y_{\mathrm{TSG}} \,|\, Y_{\mathrm{MM}})}}{\sqrt{\mathrm{Var}(Y_{\mathrm{MM}} \,|\, Y_{\mathrm{TSG}})}} \, \rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})$$

and:

$$\beta_{\mathrm{TSG}} = \frac{\sqrt{\mathrm{Var}(Y_{\mathrm{MM}} \,|\, Y_{\mathrm{TSG}})}}{\sqrt{\mathrm{Var}(Y_{\mathrm{TSG}} \,|\, Y_{\mathrm{MM}})}} \, \rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})$$

Rewritten this gives:

$$\begin{aligned} \beta_{\mathrm{MM}} &= \sigma_{\mathrm{TSG}} \, \sigma_{\mathrm{MM}}^{-1} \, \rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}}) \\ \beta_{\mathrm{TSG}} &= \sigma_{\mathrm{MM}} \, \sigma_{\mathrm{TSG}}^{-1} \, \rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}}) \end{aligned}$$

Hence, if *ρ*=0 so will the *β's* equal zero.

Correlation vs. regression

To validate this claim, simply condition on either Y_{mm} or Y_{tsg} in the bivariate normal distribution:

$$
\begin{aligned}
Y_{\mathrm{TSG}} \mid Y_{\mathrm{MM}} \\
&= \overbrace{\mu_{\mathrm{TSG}}}^{\mu_X} + \overbrace{\sigma_{\mathrm{TSG}}\sigma_{\mathrm{MM}}\rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})}^{\boldsymbol{\Sigma}_{XZ}} \, \overbrace{\sigma_{\mathrm{MM}}^{-2}}^{\boldsymbol{\Sigma}_{ZZ}^{-1}} \, (Y_{\mathrm{MM}} - \mu_{\mathrm{MM}}) \\
&= \beta_0 + \underbrace{\sigma_{\mathrm{TSG}}\sigma_{\mathrm{MM}}^{-1}\rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})}_{\beta_1} \, Y_{\mathrm{MM}}
\end{aligned}
$$

Correlation vs. regression

Note

The relation between *ρ* and the *β*'s can also be reversed.

From:

$$\begin{aligned} \beta_{\mathrm{MM}} &= \sigma_{\mathrm{TSG}}\,\sigma_{\mathrm{MM}}^{-1}\,\rho(Y_{\mathrm{MM}},Y_{\mathrm{TSG}}) \\ \beta_{\mathrm{TSG}} &= \sigma_{\mathrm{MM}}\,\sigma_{\mathrm{TSG}}^{-1}\,\rho(Y_{\mathrm{MM}},Y_{\mathrm{TSG}}) \end{aligned}$$

we obtain

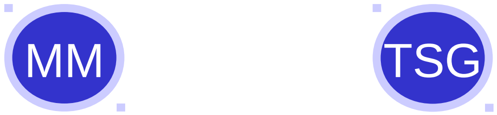
$$\begin{aligned} \rho(Y_{\mathrm{MM}},Y_{\mathrm{TSG}}) &= \mathrm{sign}(\beta_{\mathrm{MM}})\sqrt{\beta_{\mathrm{MM}}\,\beta_{\mathrm{TSG}}} \\ &= \mathrm{sign}(\beta_{\mathrm{TSG}})\sqrt{\beta_{\mathrm{MM}}\,\beta_{\mathrm{TSG}}} \end{aligned}$$

Thus, *ρ* and *β* are 1-1 related.

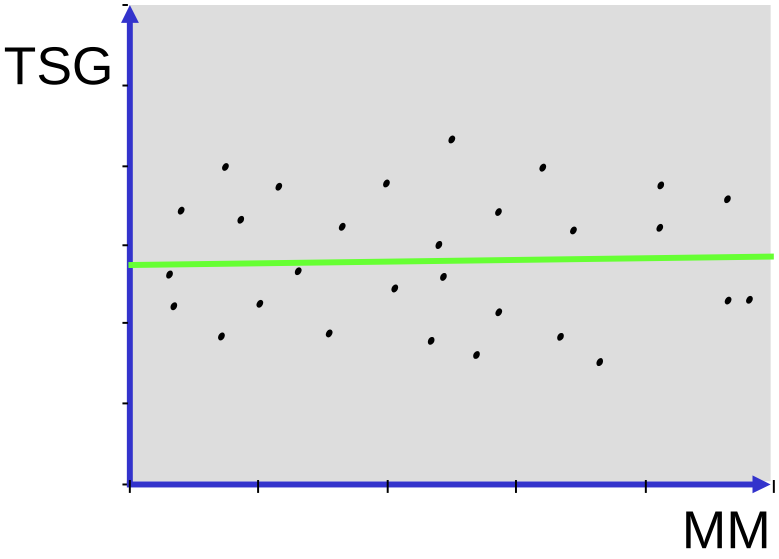
Correlation vs. regression

TSG & MM independent

Cond. indep. graph

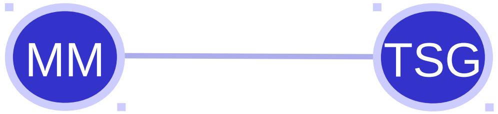


Data (ρ=0 ↔ β_1=0)

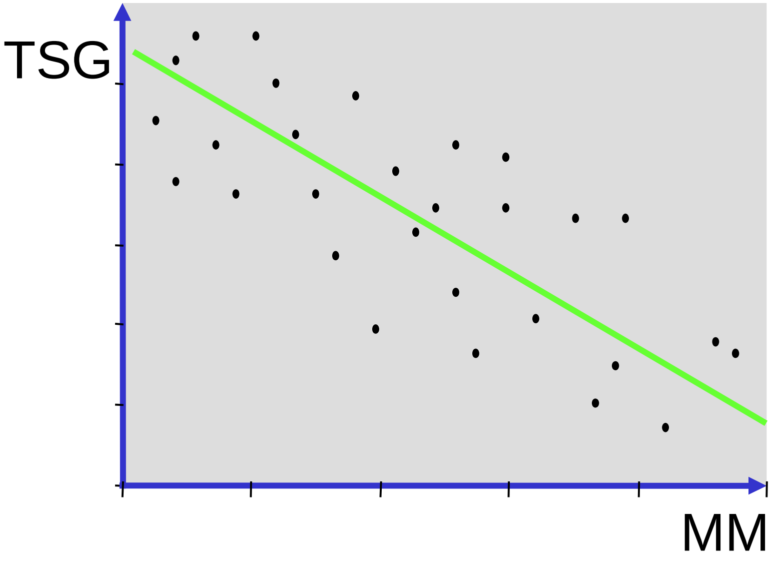


TSG & MM dependent

Cond. indep. graph



Data (ρ≠0 ↔ β_1≠0)

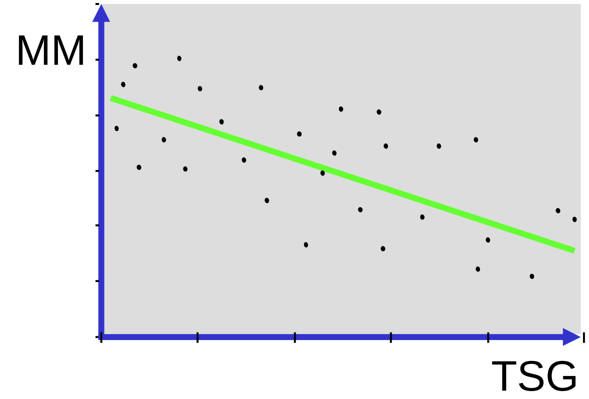


Correlation vs. regression

Undirected edges only

A closer look at:

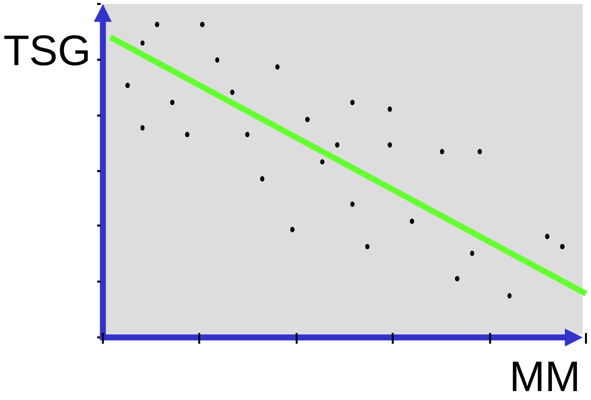
$$\begin{aligned}\beta_{\mathrm{MM}} &= \sigma_{\mathrm{TSG}}\,\sigma_{\mathrm{MM}}^{-1}\,\rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})\\ \beta_{\mathrm{TSG}} &= \sigma_{\mathrm{MM}}\,\sigma_{\mathrm{TSG}}^{-1}\,\rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}})\end{aligned}$$



The correlation is symmetric:

$$\rho(Y_{\mathrm{MM}}, Y_{\mathrm{TSG}}) = \rho(Y_{\mathrm{TSG}}, Y_{\mathrm{MM}})$$

and the variances are both positive.

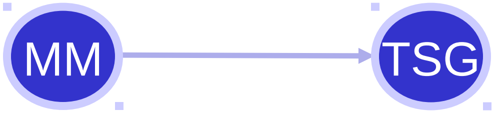


Hence, the signs of β_{MM} and β_{TSG} are identical.

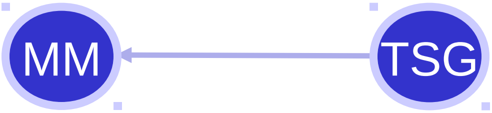
Correlation vs. regression

Undirected edges only

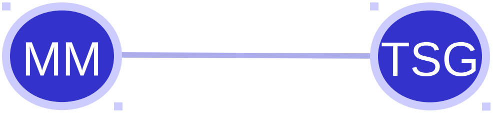
Due to the symmetry of ρ, it does not distinguish between:



or



Hence:



In regression analysis the random variables Y_{MM} and Y_{TSG} are not on equal footing. The equation:

$$Y_{\mathrm{TSG}} = \mathrm{f}(Y_{\mathrm{MM}}) + \mathrm{error}$$

suggests MM → TSG. However, the β's are one-to-one related. Consequently, also regression does not provide a clue about the direction of the relationship.

Interpretation pitfall

Eat chocolate, win the Nobel!

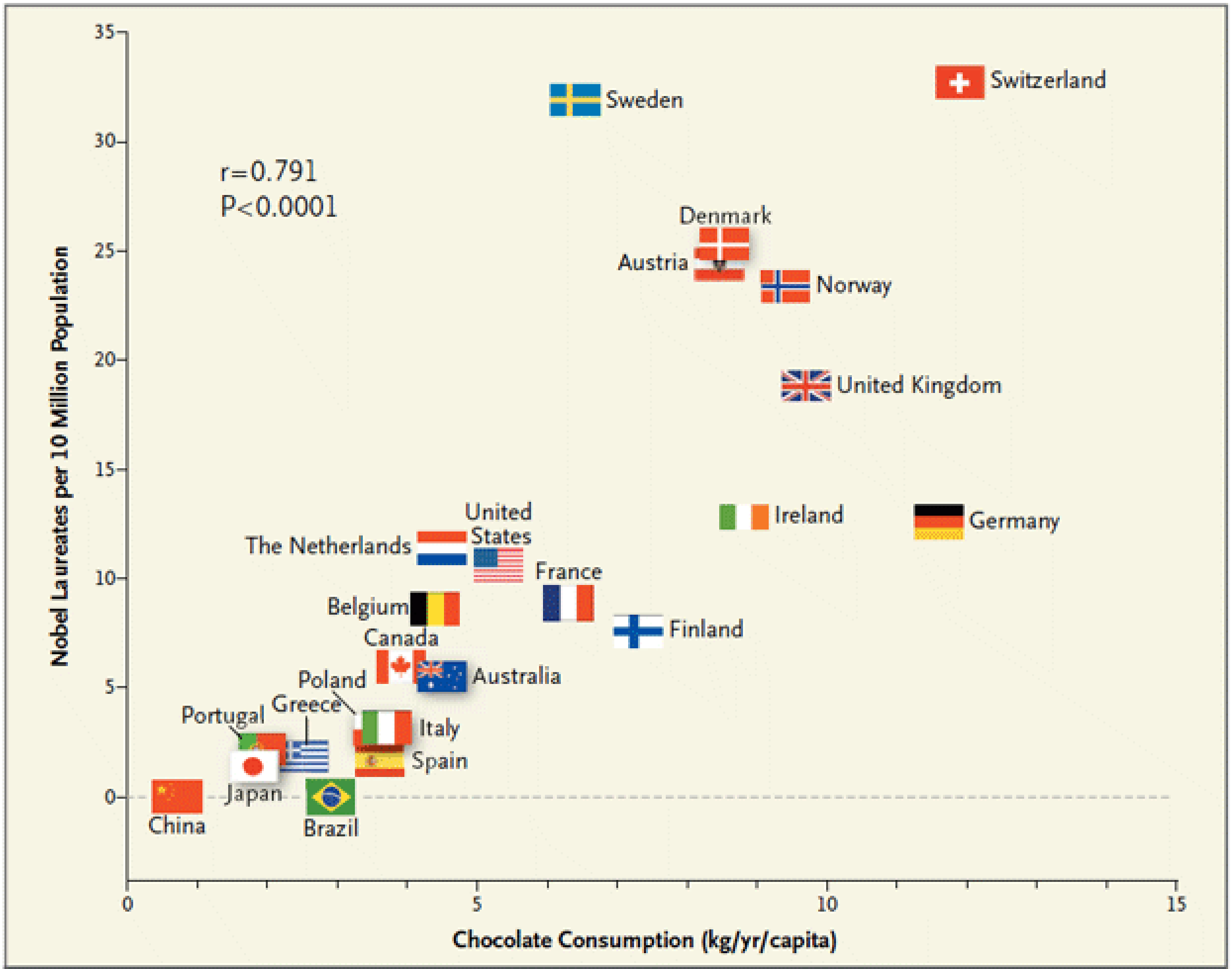


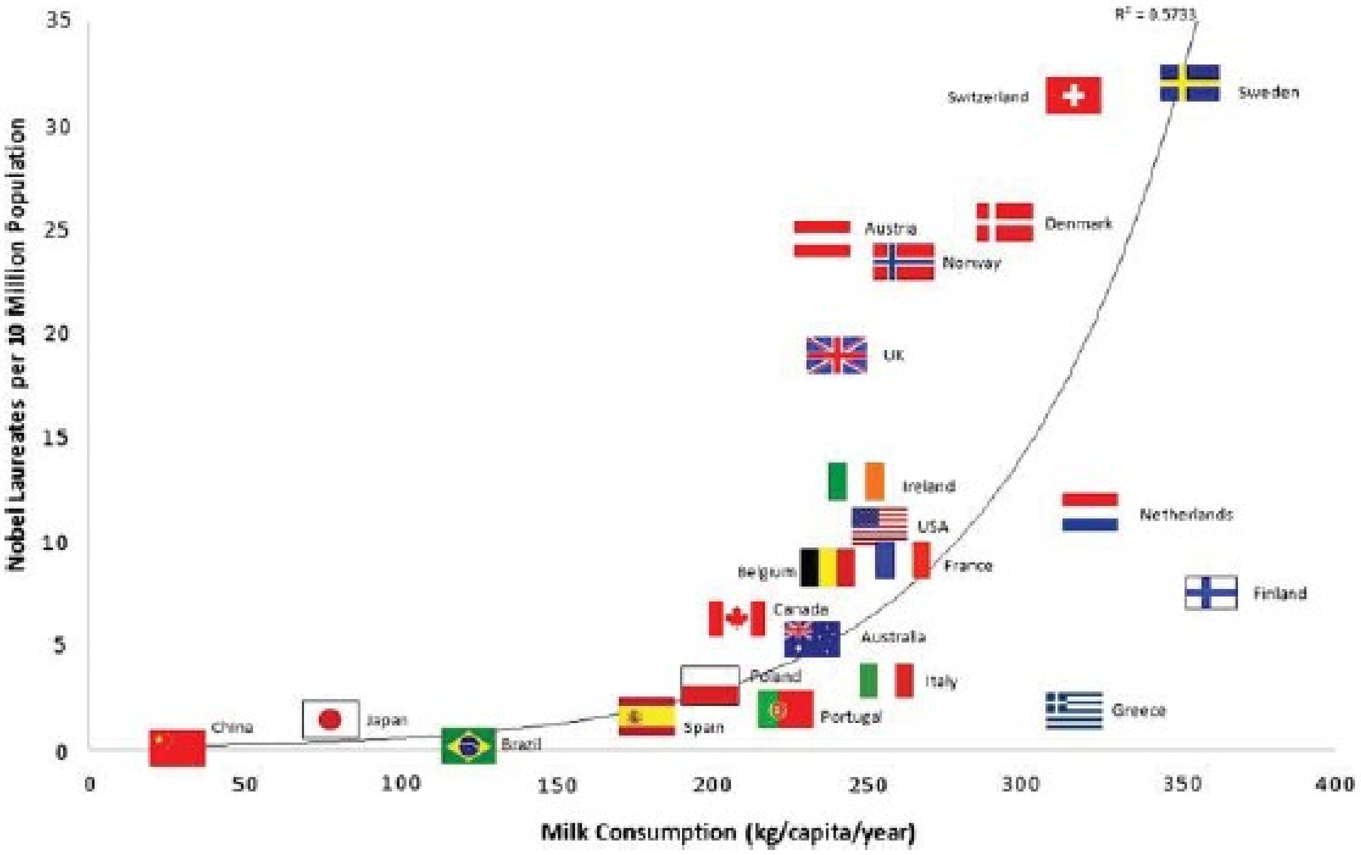
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

Messerli, 2012.

Interpretation pitfall

Even better: drink milk, win the Nobel!

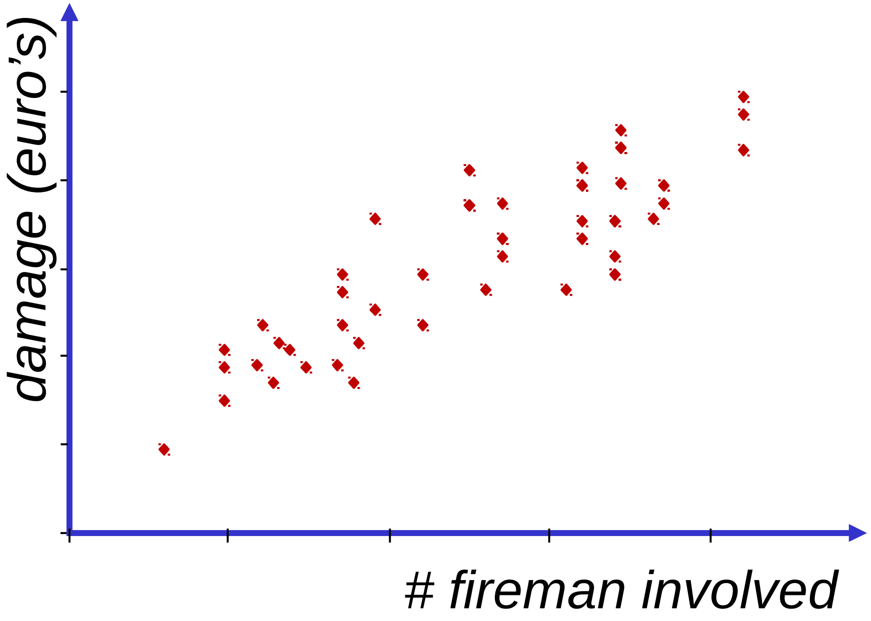


Best: drink chocolate-milk, win the Nobel?

Linthwaite, Fuller, 2013.

Interpretation pitfall

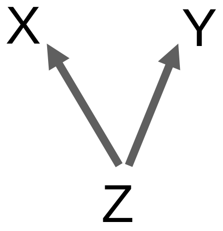
Does the involvement of more fireman result in more damage?



Possible interpretations of these data:

X → Y More firemen result in more damage.

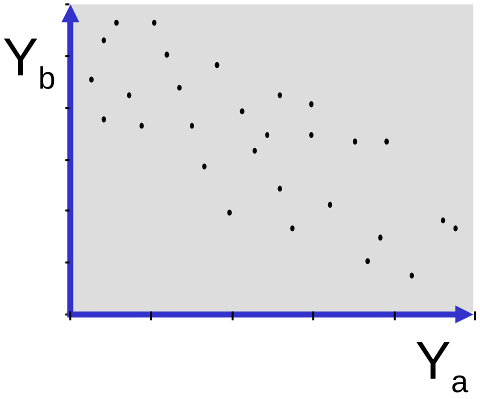
X ← Y More damage results in more firemen.



A bigger fire (Z) results in more firemen and more damage.

Interpretation pitfall

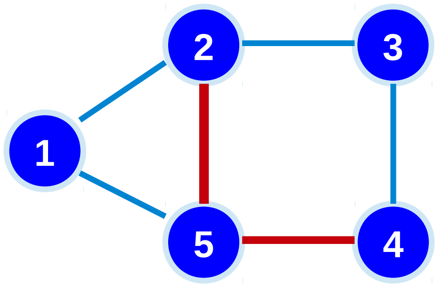
What to conclude about the relation between the activity levels of molecules A and B?

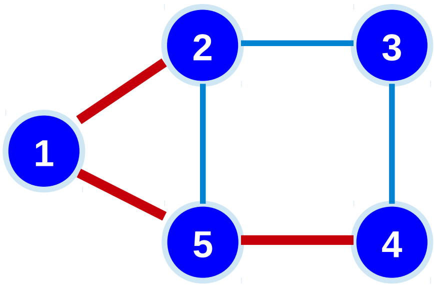


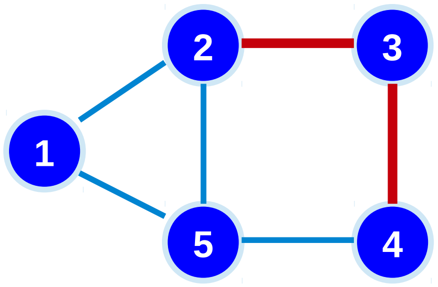
Question

Could others be responsible for observed (in)dependence?

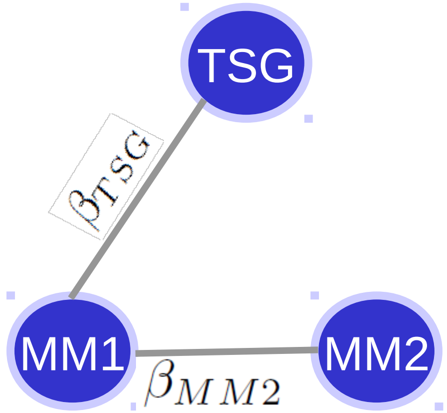
2 and 4 could be connected in many ways, e.g.:







Interpretation pitfall



Cancer research example

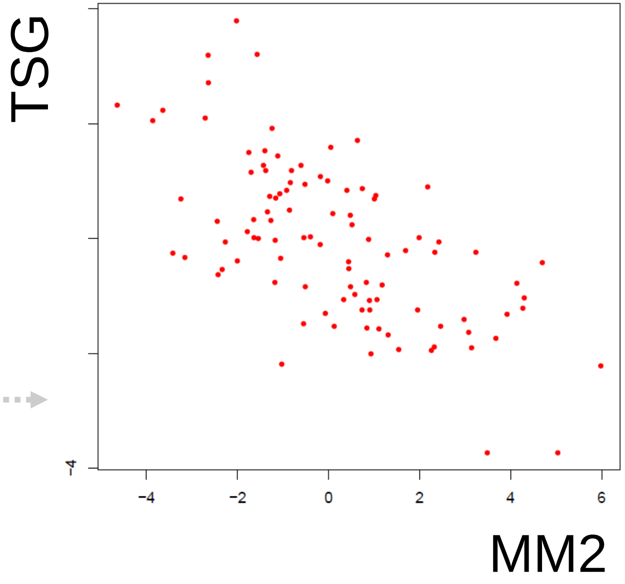
An alternative explanation by model:

$$\begin{aligned} TSG_i &= \beta_{TSG}\, MM1_i + \varepsilon_{TSG,i} \\ MM2_i &= \beta_{MM2}\, MM1_i + \varepsilon_{MM2,i} \end{aligned}$$

with

$$\begin{aligned} MM1_i &\sim \mathcal{N}(0, \sigma^2_{MM1}) \\ \varepsilon_{TSG,i} &\sim \mathcal{N}(0, \sigma^2_{TSG}) \\ \varepsilon_{MM2,i} &\sim \mathcal{N}(0, \sigma^2_{MM2}) \end{aligned}$$

Simulate

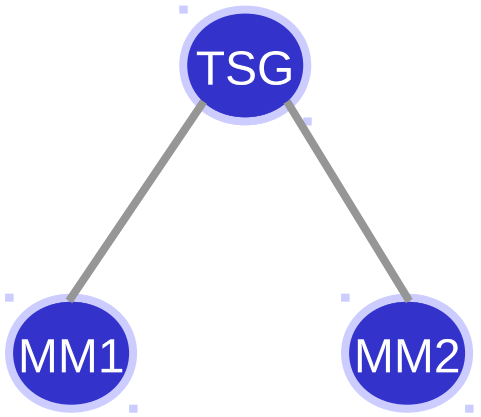


Even though there is no direct (causal) relationship between TSG and MM2 they may appear to be related.

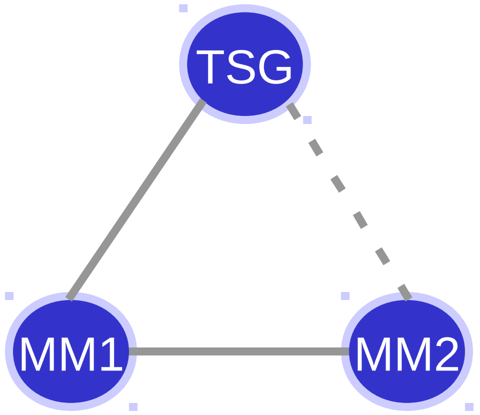
Interpretation pitfall

Cancer research example

The (independence) graph of the 3-gene pathway underlying the regression model:



An alternative graph that may explain the data equally well:



X ——— Y : X are Y cond. dependent

X - - - - Y : X and Y are correlated

Regression

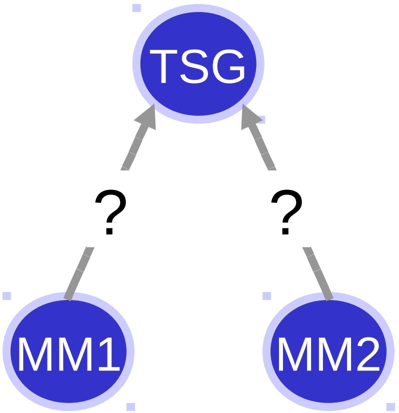
Regression

Cancer research example

Y : gene expression measurements of a tumor suppressor gene

X_1 : gene expression of methylation marker 1

X_2 : gene expression of methylation marker 2



Question

Do the methylation markers (MMs) influence the expression of the tumor suppressor gene (TSG)?

Revisited later.

Regression

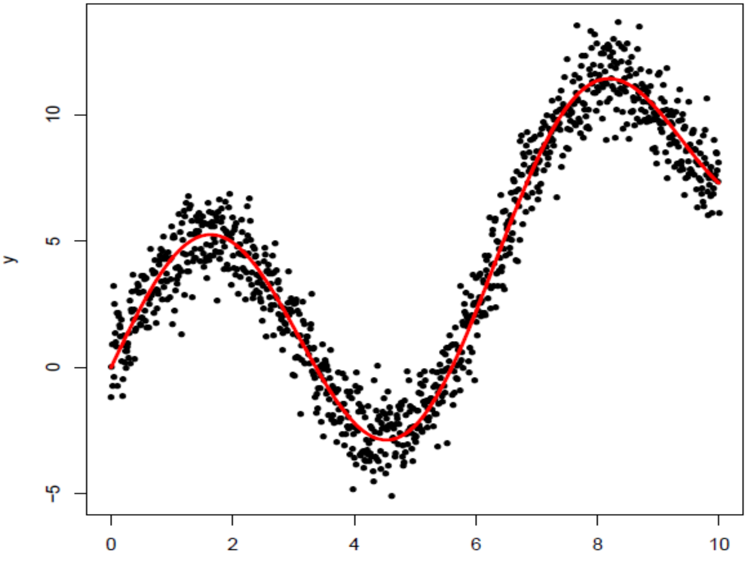
The simple linear regression model:

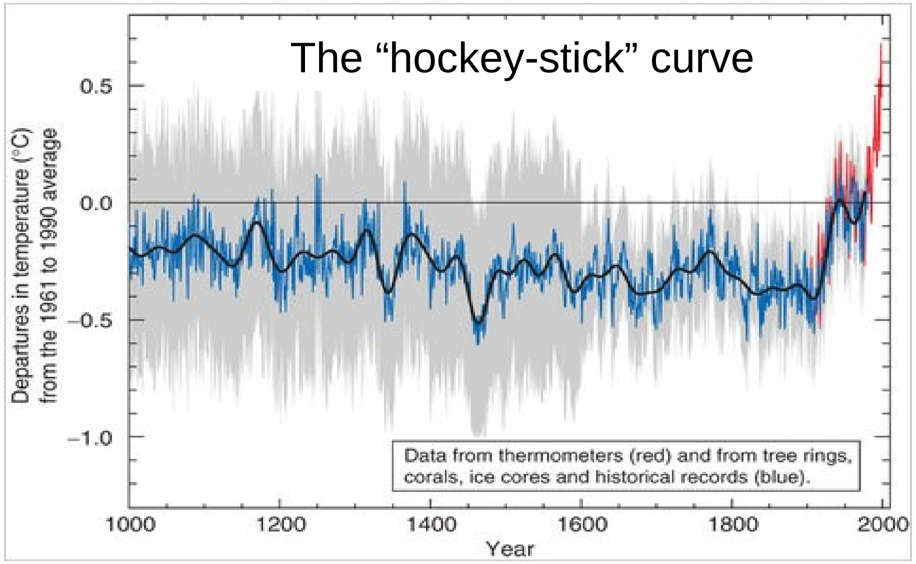
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

is *linear* in the regression parameters.
Hence, the following extensions are linear too:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$
$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$$

Examples of linear models



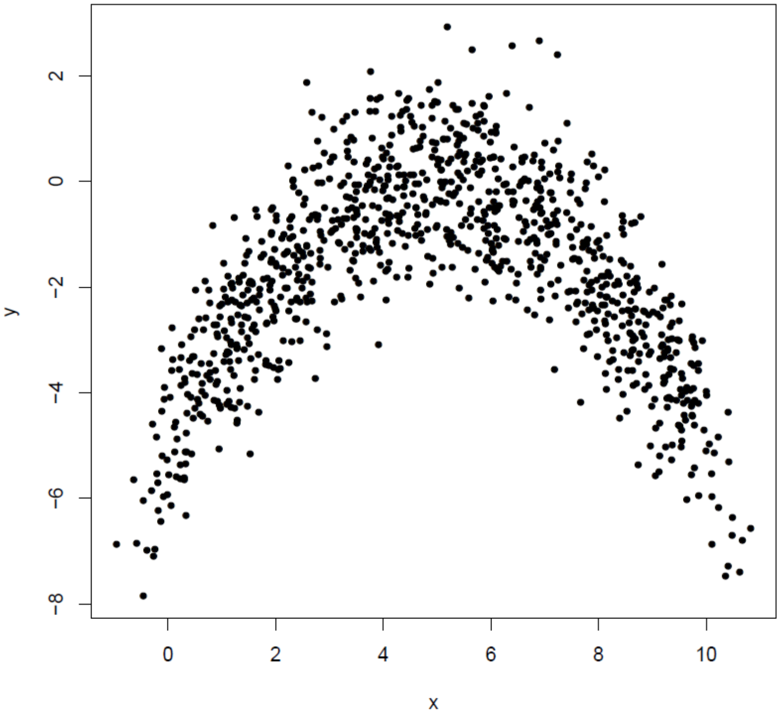

The “hockey-stick” curve

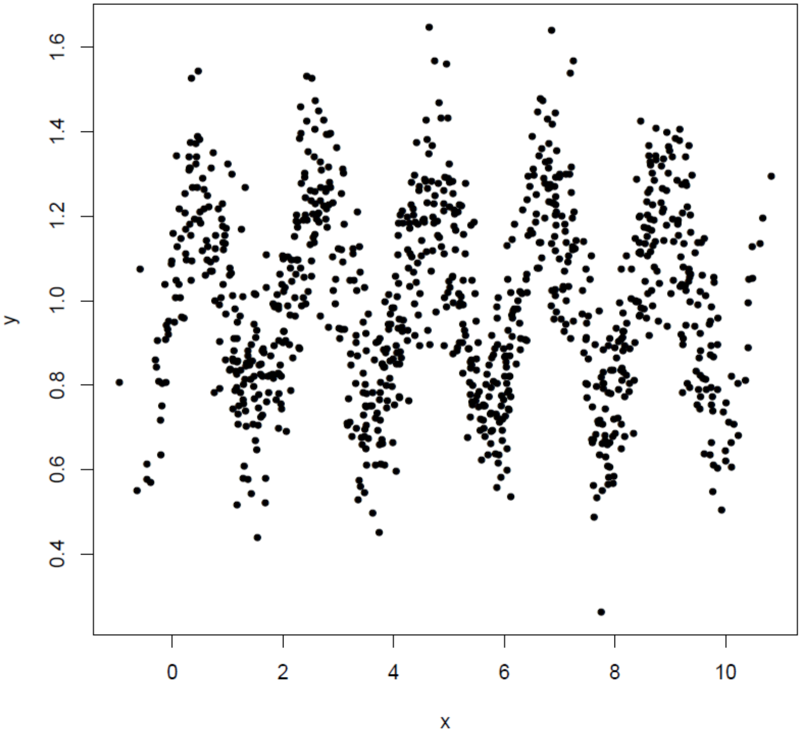
Regression: parameter estimation

Question

Can a quadratic relationship
be modelled by linear regression?

Can $Y = \beta_0 + \beta_1 \sin(\beta_2 X)$ be fitted by linear regression?





Regression

In general, the linear regression model is:

$$Y_i \quad = \quad \beta_0 + \beta_1 \, X_{i1} + \beta_2 \, X_{i2} + \ldots + \beta_{p-1} \, X_{i,p-1} + \varepsilon_i$$

e.g.:

$$Y_{i,\mathrm{TSG}} \quad = \quad \beta_0 + \beta_{\mathrm{MM1}} X_{i,\mathrm{MM1}} + \beta_{\mathrm{MM2}} X_{i,\mathrm{MM2}} + \varepsilon_i$$

with the distribution assumptions:

$$\begin{aligned} \varepsilon_i \quad & \sim \quad \mathcal{N}(0, \sigma^2) \\ E(\varepsilon_i) \quad & = \quad 0 \\ \mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) \quad & = \quad \begin{cases} \sigma^2 & \text{if} \quad i_1 = i_2 \\ 0 & \text{if} \quad i_1 \neq i_2 \end{cases} \end{aligned}$$

In the above the unknown parameters are: β_0, β_1, …, $\beta_{p\text{-}1}$, σ^2.

Regression

In matrix notation (simplifying notation):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n})$$

The (*n* x 1)-, (*p* x 1)-, (*n* x 1)-dimensional vectors with observations, parameters, and errors:

$$\mathbf{Y} = \left(\begin{array}{c} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{array} \right), \quad \boldsymbol{\beta} = \left(\begin{array}{l} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{array} \right), \quad \boldsymbol{\varepsilon} = \left(\begin{array}{c} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{array} \right).$$

Regression

The (*n* x *p*) *design matrix*:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n,p-1} \end{pmatrix}$$

E.g. in the tumor suppressor example:

```
                 intercept            MM1            MM2
sample 1                 1       -0.42796        0.26441
sample 2                 1        4.21648       -3.86460
sample 3                 1       -1.14688       -1.22544
sample 4                 1       -0.46377        0.12756
sample 5                 1        0.86248        1.16049
...                    ...            ...            ...
...                    ...            ...            ...
```

→ **X**

Regression

Question

Consider simple model for length in terms of sex:

$$Y_i = \beta_0 + \beta_1 \times \mathrm{SEX}_i + \varepsilon_i$$

Two design matrices:

	intercept	sex
sample 1	1	-1
sample 2	1	1
sample 3	1	-1
sample 4	1	-1
sample 5	1	1
...	...	...
...	...	...

	intercept	sex
sample 1	3	-2
sample 2	3	2
sample 3	3	-2
sample 4	3	-2
sample 5	3	2
...	...	...
...	...	...

What are the differences between resulting models?

Regression

The specifics of the design matrix depend on the model employed. E.g. consider the two equivalent models:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 \times \mathrm{SEX}_i + \varepsilon_i \\
Y_i &= \beta_1 \times \mathrm{FEMALE}_i + \beta_2 \times \mathrm{MALE}_i + \varepsilon_i
\end{aligned}
$$

with corresponding design matrices:

	female	male
sample 1	0	1
sample 2	1	0
sample 3	0	1
sample 4	0	1
sample 5	1	0
...	...	...
...	...	...

	intercept	sex
sample 1	1	-1
sample 2	1	1
sample 3	1	-1
sample 4	1	-1
sample 5	1	1
...	...	...
...	...	...

Regression

The regression model thus is:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

To illustrate the notation simplification:

$$\begin{aligned} Y_i &= \mathbf{X}_{i*}\,\boldsymbol{\beta} + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1\,X_{i1} + \beta_2\,X_{i2} + \ldots + \beta_{p-1}\,X_{i,p-1} + \varepsilon_i \end{aligned}$$

where $\mathbf{X}_{i*}$ denotes the i-th row of the design matrix.

The distributional assumptions become:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}) &= [\mathbb{E}(\varepsilon_1), \mathbb{E}(\varepsilon_2), \ldots, \mathbb{E}(\varepsilon_n)]^\top \\ &= (0, 0, \ldots, 0)^\top = \mathbf{0}_{n \times 1} \end{aligned}$$

Regression

and (independence of samples):

$$\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) = \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{n \times n}$$

as

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_i) = \sigma^2$$
$$\mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = 0 \quad \text{if } i_1 \neq i_2$$

The expectation of the vector of observations:

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

as:

$$\begin{aligned}\mathbb{E}(Y_i) &= \mathbb{E}(\beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i) \\ &= \mathbb{E}(\beta_0) + \mathbb{E}(\beta_1 X_{i,1}) + \ldots + \mathbb{E}(\beta_{p-1} X_{i,p-1}) + \mathbb{E}(\varepsilon_i) \\ &= \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{p-1} X_{i,p-1}\end{aligned}$$

Regression

Model

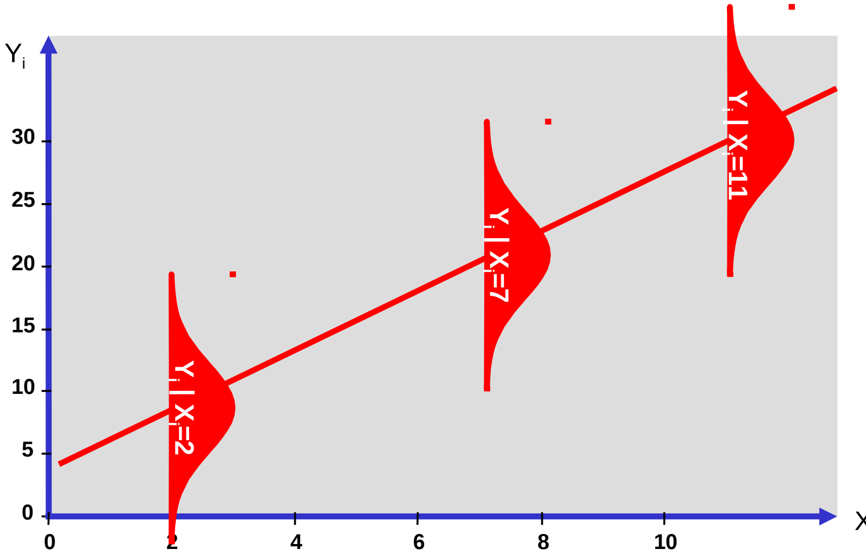
We write:

$$Y_i = \mathbf{X}_{i,*}\boldsymbol{\beta} + \varepsilon_i$$

while it is equivalent to write:

$$Y_i \,|\, \mathbf{X}_{i,*} \sim N(\mathbf{X}_{i,*}\boldsymbol{\beta}, \sigma^2)$$

The latter explicitly assumes that the explanatory variable **X** is (temporarily) taken as non-random. It is to be read as: **Y** *conditional* on **X** is distributed as ….



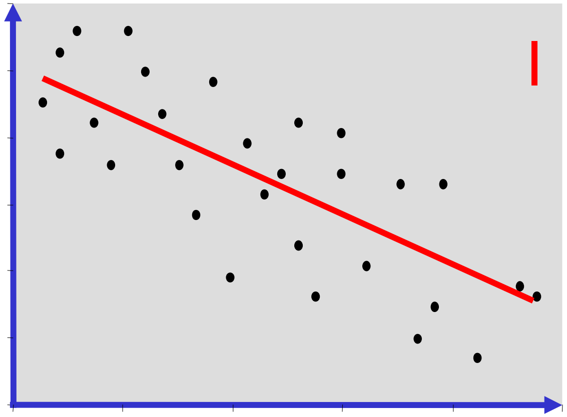
Regression

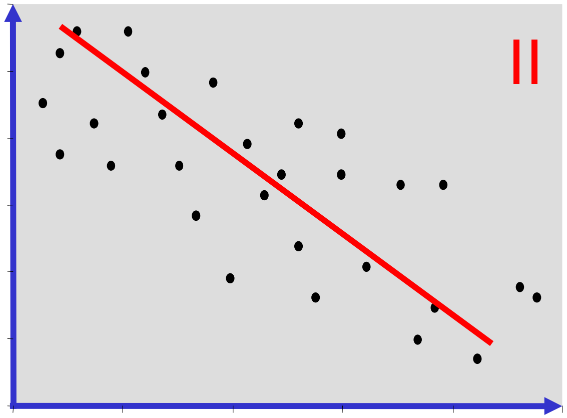
Parameter estimation

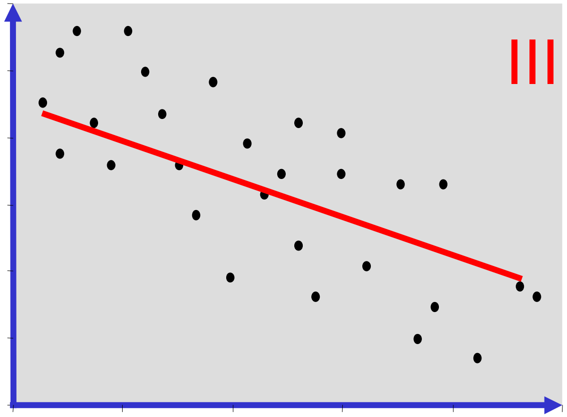
Regression: parameter estimation

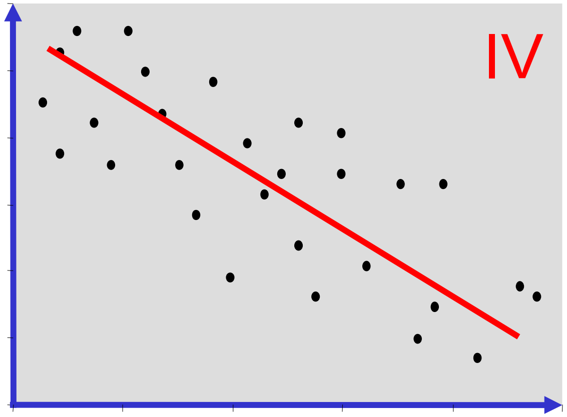
Question

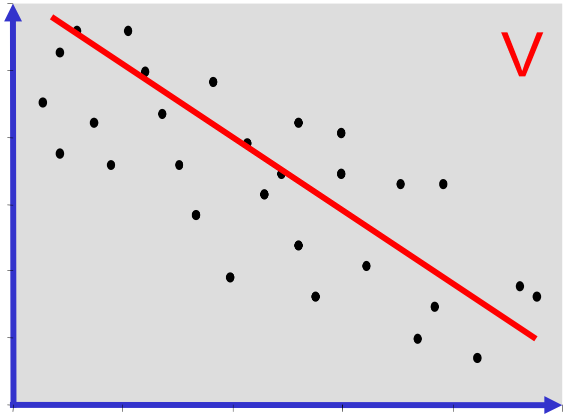
What is now the best model? Best in what sense?

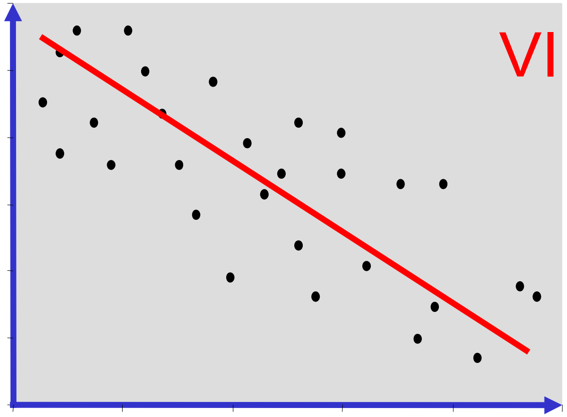










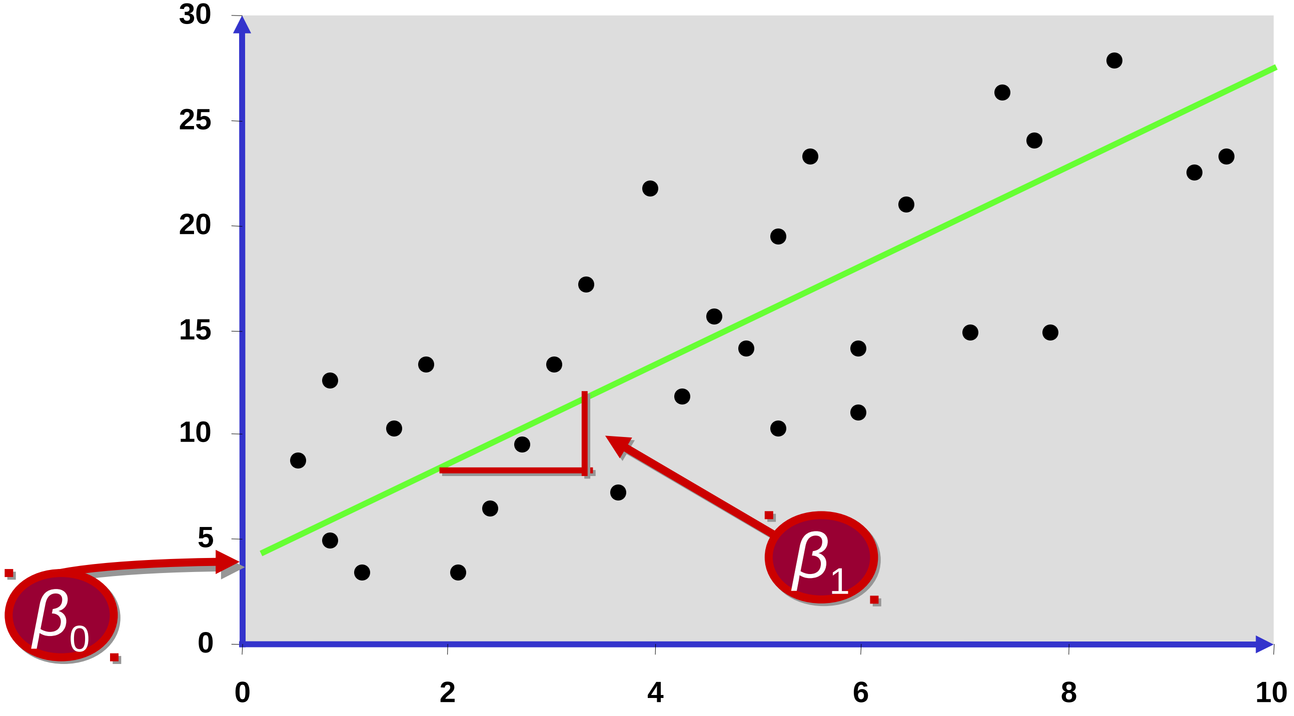


Regression: parameter estimation

We search a *linear* (= straight line) relation:

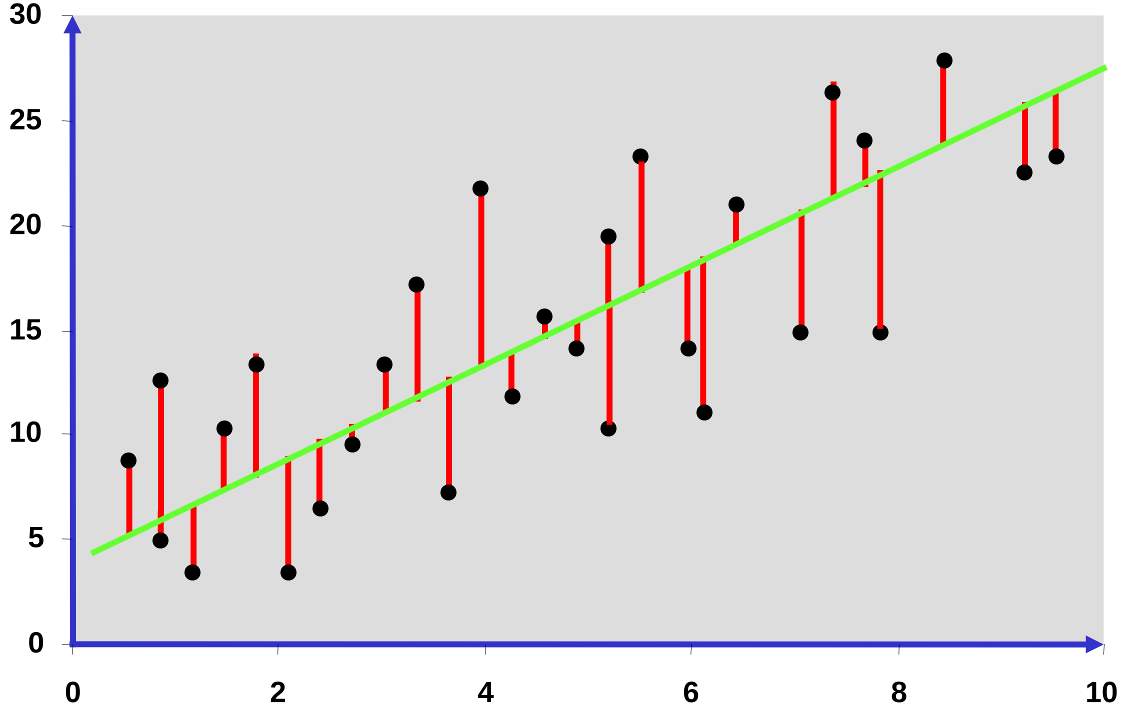
$$Y = \beta_0 + \beta_1 X.$$

How to choose β_0 and β_1?



Regression: parameter estimation

β_0 and β_1 are chosen such that the total quadratic distance of the observations to the regression line is minimal.



Regression: parameter estimation

Estimation

Use maximum likelihood. Hereto, note that:

$$\begin{aligned} Y_i &= \mathbf{X}_{i*}\,\boldsymbol{\beta} + \varepsilon_i \\ E(Y_i) &= \mathbf{X}_{i*}\,\boldsymbol{\beta} \end{aligned}$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and ε_{i_1}, ε_{i_2} independent if $i_1 \neq i_2$.

One may thus reformulate the model as:

$$Y_i \,|\, \mathbf{X}_{i*} \quad \sim \quad \mathcal{N}(\mathbf{X}_{i*}\boldsymbol{\beta}, \sigma^2)$$

Normality gives:

$$P(Y_i = y_i) \quad = \quad \frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp[-(y_i - \mathbf{X}_{i*}\,\boldsymbol{\beta})^2/2\sigma^2]$$

Regression: parameter estimation

Estimation

Using the independence of the samples, the likelihood is:

$$P(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp[-(y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2/(2\sigma^2)]$$

with log-likelihood:

$$\log[P(\mathbf{Y} = y)] = -n \log(\sqrt{2\pi}\sigma) - \tfrac{1}{2\sigma^2} \textstyle\sum_{i=1}^{n} (y_i - \mathbf{X}_{i,*}\boldsymbol{\beta})^2$$

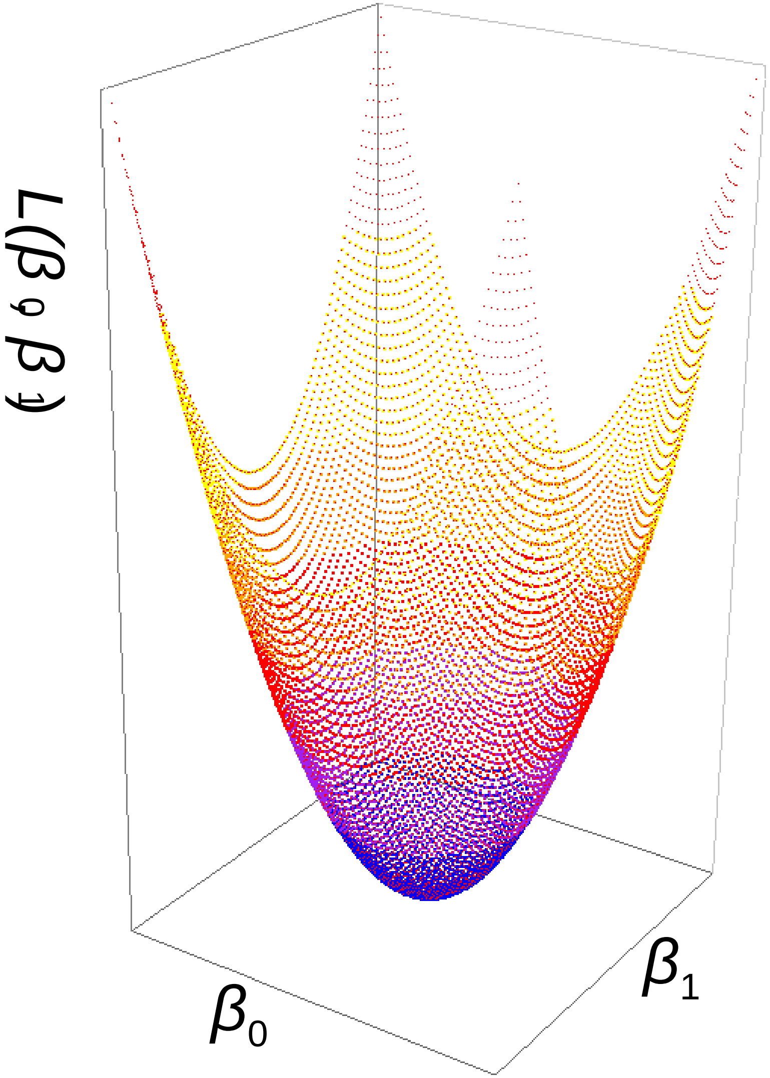
This is quadratic in the parameters (a parabola):

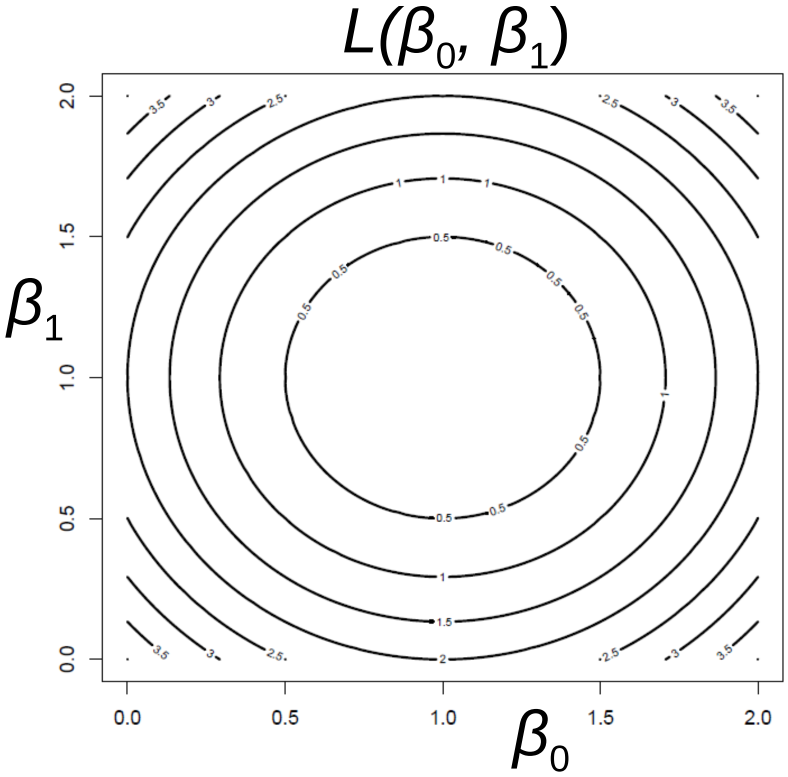
$$c_0 + c_1\,\beta_0 + c_2\,\beta_1 + c_3\,\beta_0^2 + c_4\,\beta_1^2 + c_5\,\beta_0\,\beta_1$$

where the c_k depend on **X** and **Y**.

Regression: parameter estimation

Plots of loss function vs. parameters (2d and 3d)

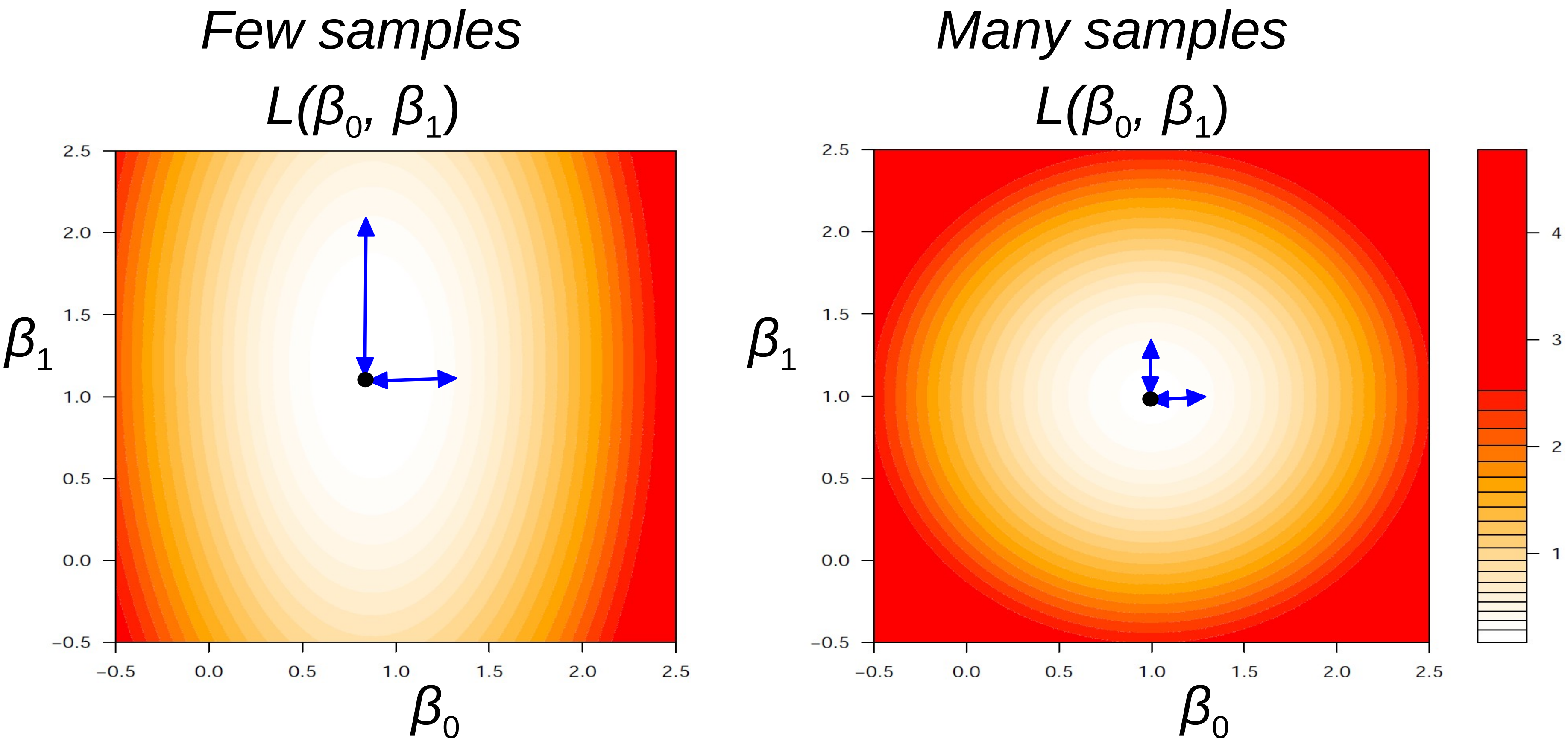




Regression: parameter estimation

Effect of sample size

Larger sample sizes yield better located (≈ less biased) and clearer (≈ lower variance) optima.



Few samples

Many samples

Regression: parameter estimation

Equate the loglikelihood's 1st order derivative to zero:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \; = \; \mathbf{X}^T \mathbf{Y}$$

Solving for $\boldsymbol{\beta}$ yields:

$$\hat{\boldsymbol{\beta}} \; = \; (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

For the ML estimation of σ², solve:

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (Y_i - \mathbf{X}_{i,*} \boldsymbol{\beta})^2 = 0$$

This yields:

$$\hat{\sigma}^2_{\mathrm{ML}} \; = \; \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_{i,*} \boldsymbol{\beta})^2$$

This estimator is however biased! For an unbiased estimator divide by *n*-*p* instead of *n*.

Regression: parameter estimation

Example (numerical)

Consider an experiment in which expression levels of a 3-gene pathway have been measured. The resulting data are:

	gene 1	gene 2	gene 3
sample 1	0.622	0.934	-1.915
sample 2	1.001	1.341	-2.140
sample 3	-0.468	-1.180	-0.088
sample 4	1.752	0.058	0.478

Wish to fit: $\mathbf{Y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbf{Y}$ gene 1. Then:

$$\underbrace{\begin{array}{r} \text{gene 1} \\ 0.622 \\ 1.001 \\ -0.468 \\ 1.752 \end{array}}_{\mathbf{Y}} \qquad\qquad \underbrace{\begin{array}{rrr} \text{intercept} & \text{gene 2} & \text{gene 3} \\ 1 & 0.934 & -1.915 \\ 1 & 1.341 & -2.140 \\ 1 & -1.180 & -0.088 \\ 1 & 0.058 & 0.478 \end{array}}_{\mathbf{X}}$$

Regression: parameter estimation

Example (numerical)

To evaluate $\hat{\beta} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y}$

calculate its constituents:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} 4.000 & 1.153 & -3.665 \\ 1.153 & 4.066 & -4.527 \\ -3.665 & -4.527 & 8.483 \end{pmatrix} \qquad \mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} 2.907 \\ 2.577 \\ -2.455 \end{pmatrix}$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{pmatrix} 0.494 & 0.240 & 0.341 \\ 0.240 & 0.722 & 0.489 \\ 0.341 & 0.489 & 0.526 \end{pmatrix}$$

and obtain:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} 1.215 \\ 1.359 \\ 0.961 \end{pmatrix}$$

Regression: parameter estimation

Example (numerical)

For σ² evaluate:

$$\hat{\sigma}^2_{\mathrm{ML}} \;=\; \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})^2$$

E.g.:

$$\mathbf{X}_{1,*}\hat{\boldsymbol{\beta}} \;=\; \begin{pmatrix} 1 & 0.934 & -1.915 \end{pmatrix} \begin{pmatrix} 1.215 \\ 1.359 \\ 0.961 \end{pmatrix}$$

This yields s² = 2.292 * 10⁻⁴.

The fitted model thus is:

$$\begin{aligned} Y_i &= \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2} + \varepsilon_i \\ &= 1.215 + 1.359 X_{i,1} + 0.961 X_{i,2} + \varepsilon_i \end{aligned}$$

with $\varepsilon_i \sim \mathcal{N}(0, 2.292 \times 10^{-4})$

Regression: parameter estimation

Fits

From the fitted model obtain the fits (the observation as expected by the model, i.e. the regression line)*:*

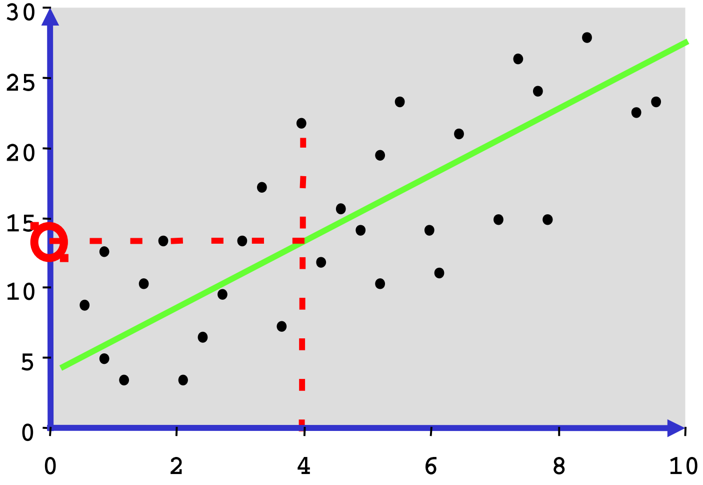
$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$$

as the error is best predicted by its mean, which is zero.

The fit of an individual observation is:

$$\hat{Y}_i \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \hat{\beta}_2 X_{i,2}$$

For novel data (X's) this formula may be used for *prediction*.



Regression: parameter estimation

Estimate behaviour

The estimates are unbiased:

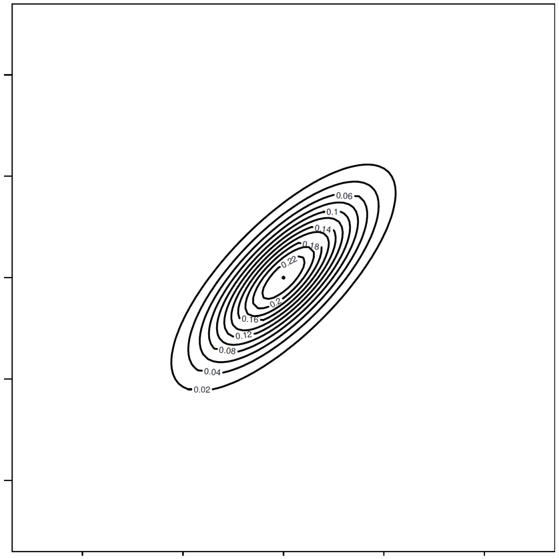
$$E(\hat{\beta}) \quad = \quad \beta$$

with variance (derivation in SM):

$$\mathrm{Var}(\hat{\beta}) \quad = \quad \sigma^2 \, (\mathbf{X}^T\mathbf{X})^{-1}$$

In particular:

$$\hat{\beta} \quad \sim \quad \mathcal{N}(\beta, \sigma^2 \, [\mathbf{X}^T\mathbf{X}]^{-1})$$



Note

Variance of the estimate mainly depends on design matrix.
In controlled experiments X is chosen s.t. the variance of the estimates is minimal.

Regression: parameter estimation

Estimate behaviour vs. design

Consider two experimental designs:

Orthogonal design

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Covariance matrix of estimates of $\boldsymbol{\beta}$ is diagonal.

Non-orthogonal design

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Covariance matrix of estimates of $\boldsymbol{\beta}$ is not diagonal.

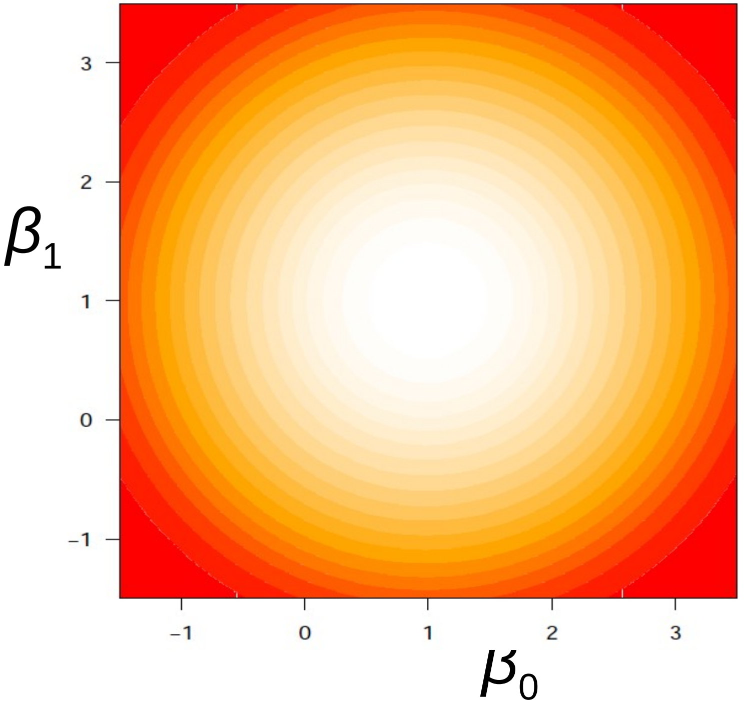
Regression: parameter estimation

Estimate behaviour vs. design

The orthogonality of the covariates determines the shape of the parabola.

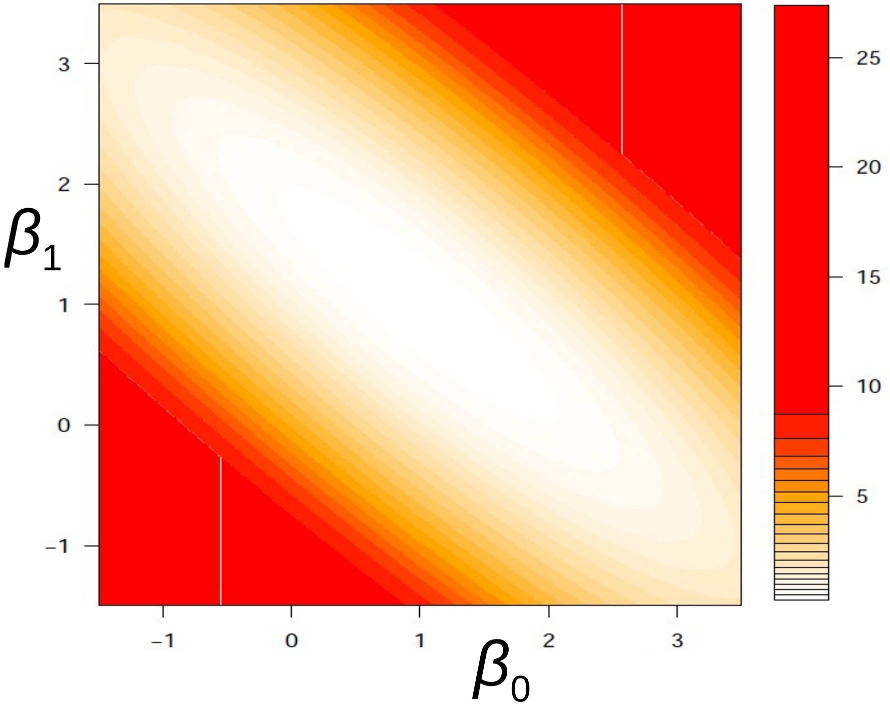
Orthogonal

$L(\beta_0, \beta_1)$



Non-orthogonal

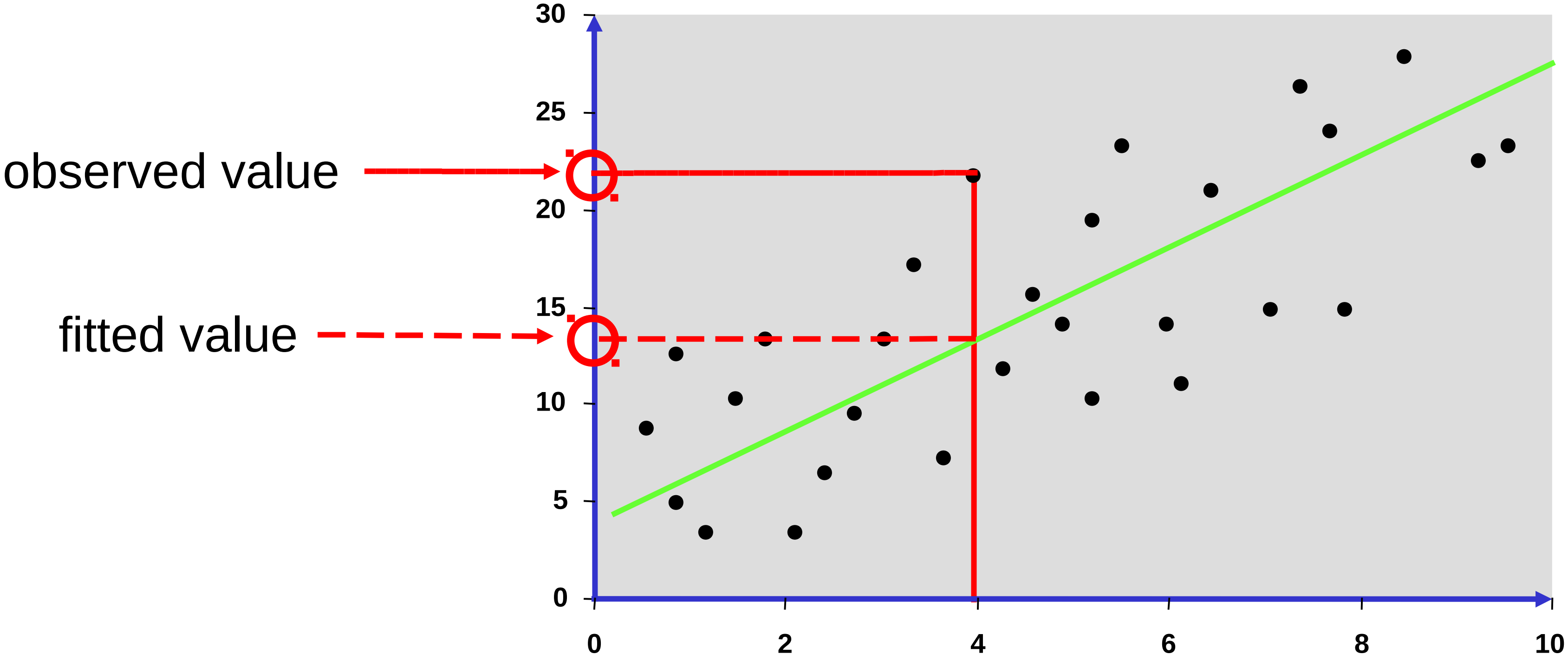
$L(\beta_0, \beta_1)$



Regression: parameter estimation

Residuals

Residual is the deviation between observation and model.



observed value

fitted value

Residual = observed value – fitted value:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}}$$

Regression: parameter estimation

Residual variance

Simply the variance of the residuals:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})}_{\textit{residuals}}{}^2$$

It is thus the variance of Y corrected for X. Or, the variance in Y not attributable to X. It is also denoted as: $\widehat{\mathrm{Var}}(\mathbf{Y} \mid \mathbf{X})$.

Ideally, this is small compared to $\widehat{\mathrm{Var}}(\mathbf{Y})$ as that would imply that the model is a good description of the data.

Regression: hypothesis testing

Testing

The variance of the estimate of $\boldsymbol{\beta}$ can now directly be obtained from:

$$\mathrm{Var}(\hat{\beta}) \quad = \quad \sigma^2 \, (\mathbf{X}^T \mathbf{X})^{-1}$$

its constituents are on previous slides.

This variance is used for testing ($H_0 : \beta_j = 0$), and the construction of confidence intervals, e.g.:

$$P\left\{\beta_1 \in \hat{\beta}_1 \pm 1.96 \sqrt{s^2[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}]_{1,1}}\right\} \approx 0.95$$

Regression: hypothesis testing

Testing

For each parameter we test the null hypothesis:

$$H_0 : \beta_j = 0$$

To evaluate this hypothesis we note that:

$$\frac{\hat{\beta}_j - \beta}{\hat{\sigma}_{\hat{\beta}_j}} \quad \sim \quad t_{n-p}$$

where:

$$\hat{\sigma}_{\hat{\beta}_j} \quad = \quad \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$$

Regression: hypothesis testing

Example (numerical)

```
R output of regression
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.21544    0.02126   57.18   0.0111 *
X[, 2]       1.35873    0.02572   52.83   0.0120 *
X[, 3]       0.96082    0.02195   43.77   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03026 on 1 degrees of freedom
Multiple R-squared: 0.9996,     Adjusted R-squared: 0.9989
F-statistic:  1400 on 2 and 1 DF,  p-value: 0.01889
```

T-statistics and p-values.

Note this uses the unbiased (rather than the ML) estimate of the error variance.

Regression: coefficient of determination

Define the *coefficient of determination*:

$$\begin{aligned} R^2(\mathbf{Y}, \mathbf{X}) &= \rho^2(\mathbf{Y}, \hat{\mathbf{Y}}) \\ &= \rho^2(\mathbf{Y}, \mathbf{X}\hat{\beta}) \end{aligned}$$

the squared correlation coefficient between ***Y*** and the columns of ***X***. Note: R^2 in [0,1].

An alternative interpretation of the R^2 comes from the sum of squares of the observation:

$$SYY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Regression: coefficient of determination

We may then write:

$$\begin{aligned} R^2 &= \frac{SYY - RSS}{SYY} = \frac{SYY/(n-1) - RSS/(n-1)}{SYY/(n-1)} \\ &= \frac{s_Y^2 - s_{\hat{\varepsilon}}^2}{s_Y^2} \end{aligned}$$

where:

$$\begin{aligned} s_{\hat{\varepsilon}}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = RSS/(n-1) \end{aligned}$$

The “percentage of explained variation” in ***Y*** by ***X***.

Regression: coefficient of determination

Example (numerical)

```
R output of regression
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.21544    0.02126   57.18   0.0111 *
X[, 2]       1.35873    0.02572   52.83   0.0120 *
X[, 3]       0.96082    0.02195   43.77   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03026 on 1 degrees of freedom
Multiple R-squared: 0.9996,     Adjusted R-squared: 0.9989
F-statistic:  1400 on 2 and 1 DF,  p-value: 0.01889
```

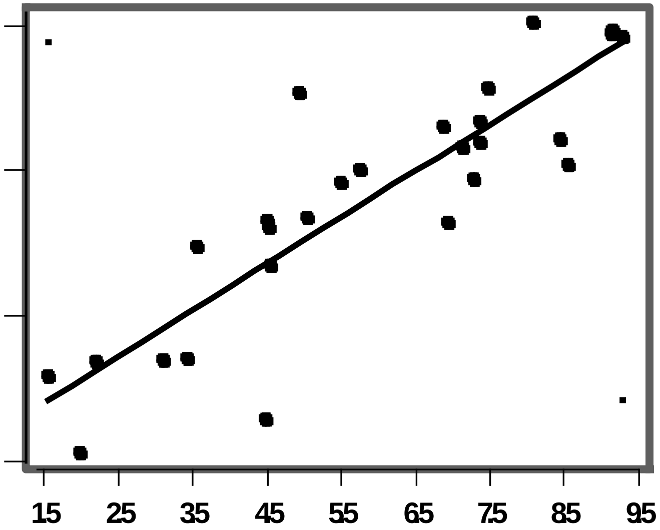
R^2: *coefficient of determination*.

Indicates the explanatory power of the model.

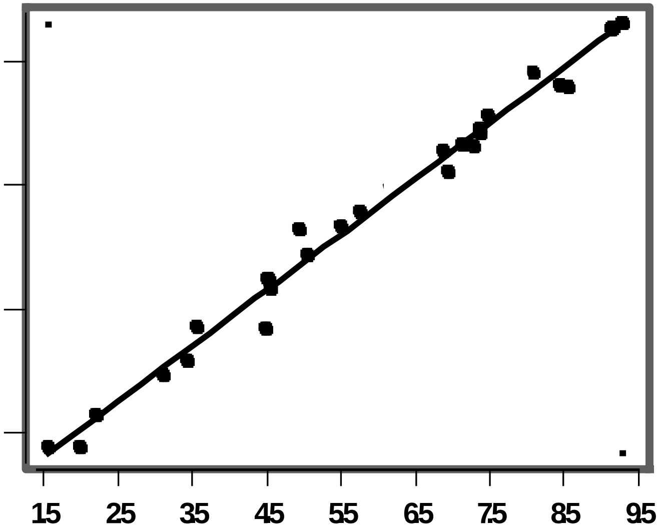
Regression: coefficient of determination

R^2 is the percentage of the variation in the measurements that is explained by the regression model.

$R^2 = 76.3\%$



$R^2 = 98.7\%$



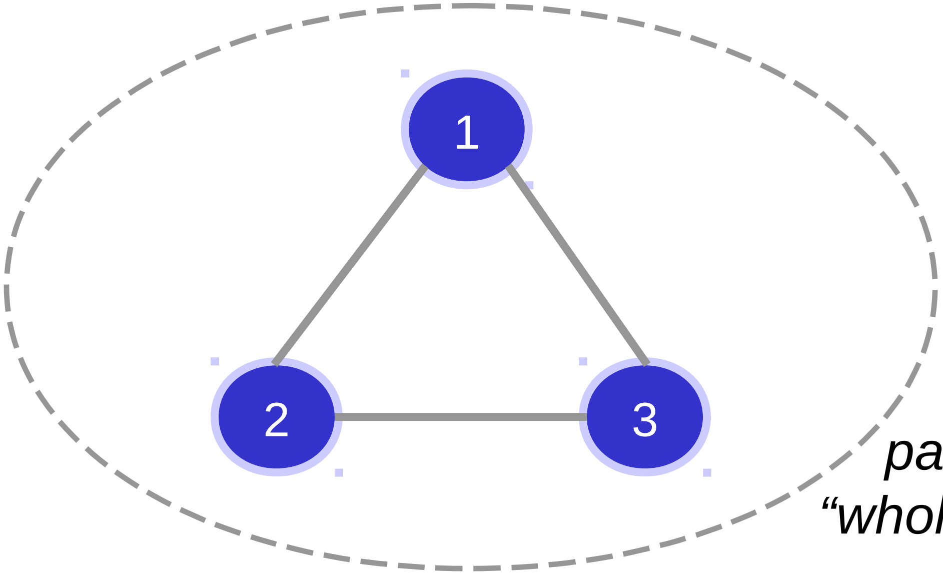
Large R^2 (> 80%): almost all variation in Y is explained by X. Hence, we can make precise predictions.

Small R^2: a substantial part of the variation in Y is explained by other factors.

Multi-gene pathway & regression

Multi-gene pathway & regression

Multi-gene pathways comprise of more than two genes, and assume no gene “lives” outside the pathway.



pathway =
“whole universe”

Two methods:

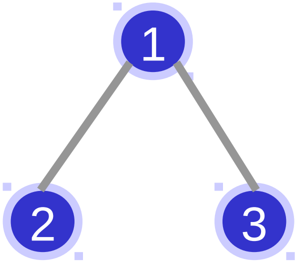
- Regression
- Correlation

Multi-gene pathway & regression

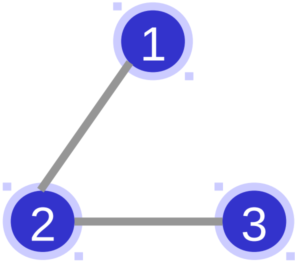
Regression method

Regress the expression data of each gene on that of all other genes.

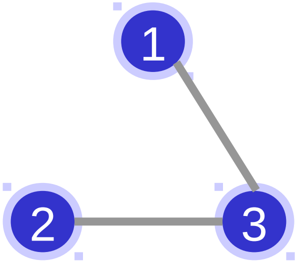
$$Y_1 = b_{01} + b_{21}Y_2 + b_{31}Y_3 + e_1$$



$$Y_2 = b_{02} + b_{12}Y_1 + b_{32}Y_3 + e_2$$



$$Y_3 = b_{03} + b_{13}Y_1 + b_{23}Y_2 + e_3$$



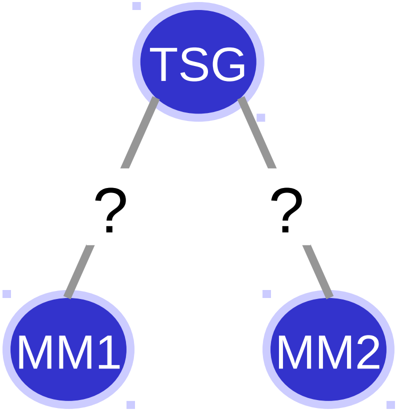
Multi-gene pathway & regression

Cancer research example

Y : gene expression measurements of a tumor suppressor gene

X_1 : gene expression of methylation marker 1

X_2 : gene expression of methylation marker 2

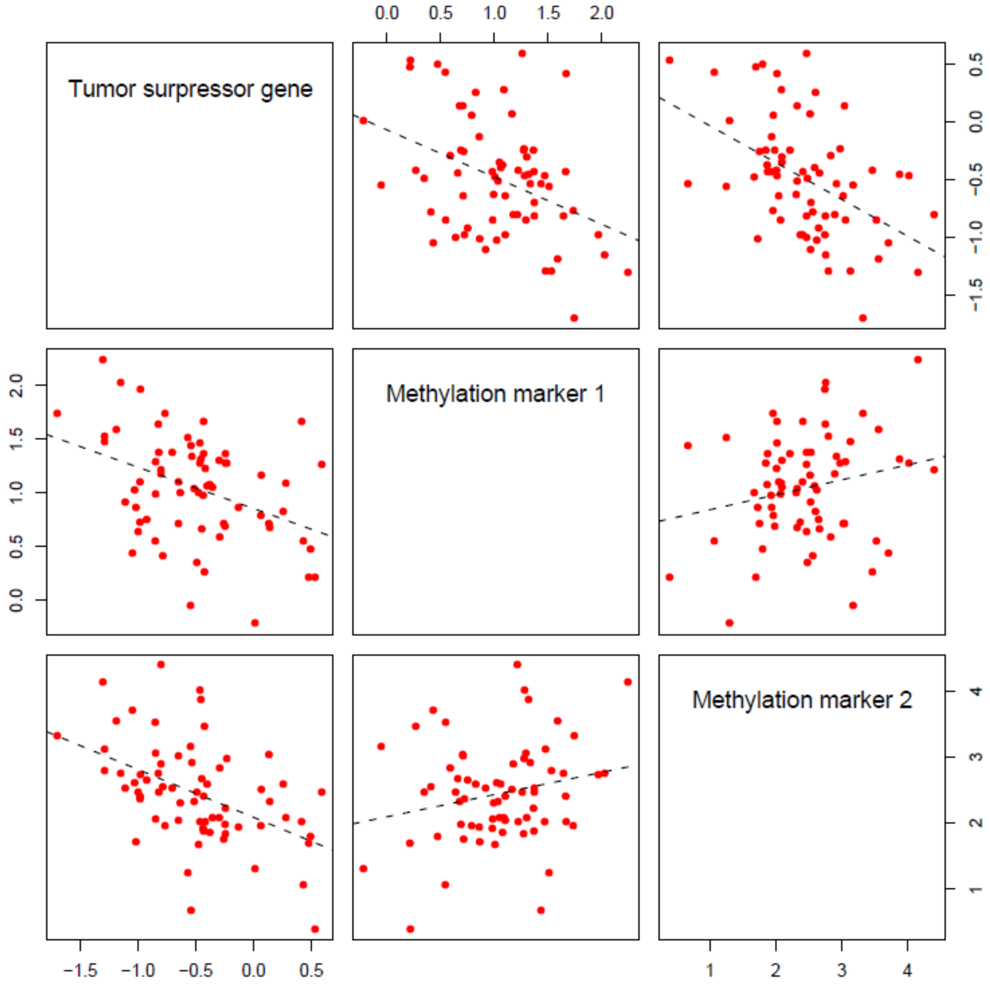


Question

Do the methylation markers (MMs) influence the expression of the tumor suppressor gene (TSG)?

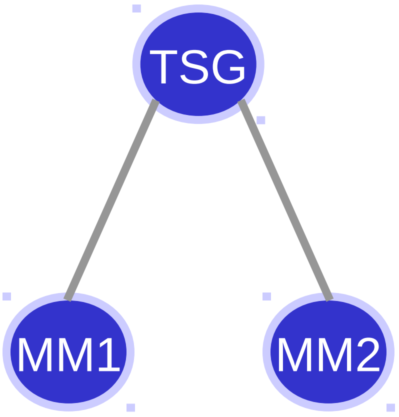
Multi-gene pathway & regression

```
# generate all pairwise scatterplots
> pairs(cbind(TSG, MM1, MM2))
```



Multi-gene pathway & regression

```
# perform multiple regression analysis
> regressionResults <- lm(TSG ~ MM1 + MM2)
> summary(regressionResults)
```

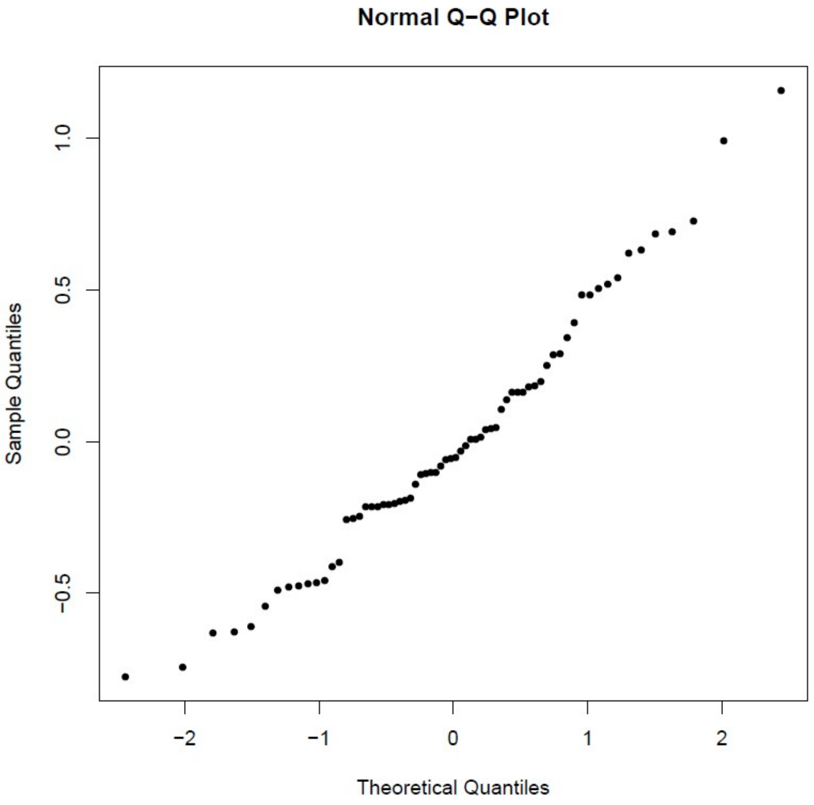


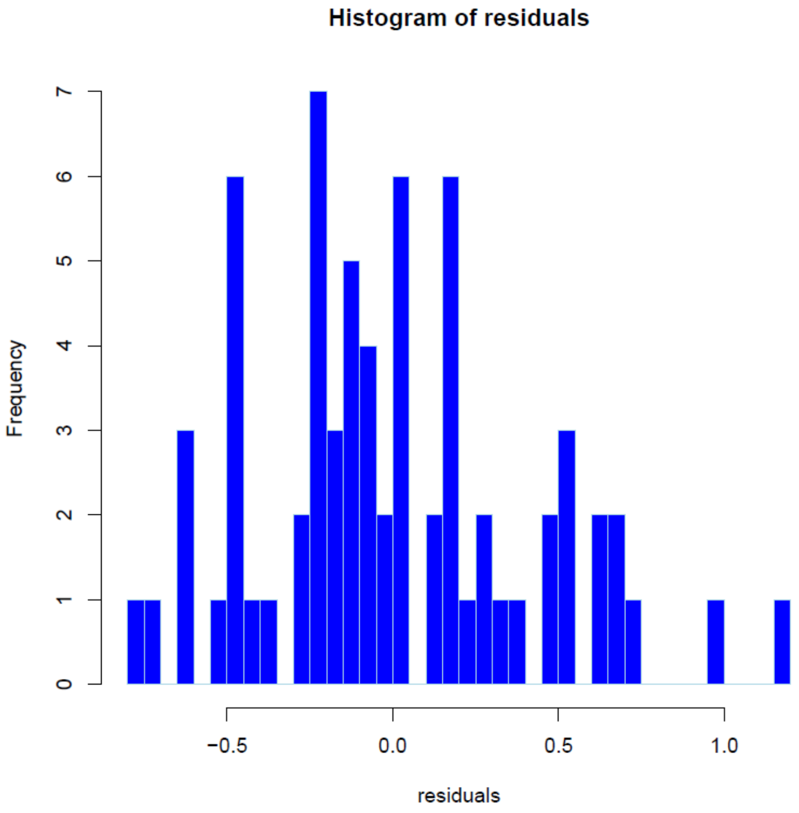
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.51023    0.18969   2.690 0.009077 **
MM1         -0.31679    0.10784  -2.938 0.004573 **
MM2         -0.27524    0.06941  -3.965 0.000185 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4212 on 65 degrees of freedom
Multiple R-squared: 0.3219,     Adjusted R-squared: 0.3011
F-statistic: 15.43 on 2 and 65 DF,  p-value: 3.286e-06
```

Multi-gene pathway & regression

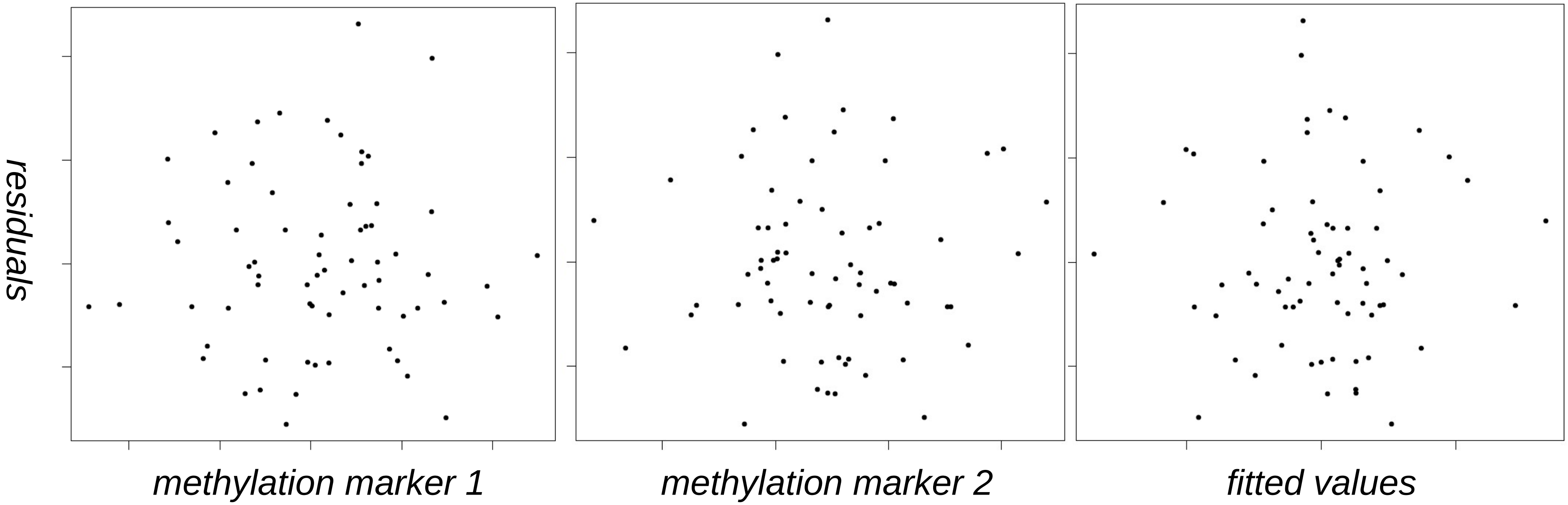
Check distributional assumption





Multi-gene pathway & regression

Check for other irregularities



Regression: parameter estimation

Model:

$$Y_{\mathrm{TSG}} = \hat{\beta}_{\mathrm{MM1}} X_{\mathrm{MM1}} + \hat{\beta}_{\mathrm{MM2}} X_{\mathrm{MM2}} + \mathrm{error}$$

The change in the response due to one in a covariate:

$$\partial Y_{\mathrm{TSG}} / \partial X_{\mathrm{MM1}} = \hat{\beta}_{\mathrm{MM1}}$$

Put differently:

$$\begin{aligned} \Delta Y_{\mathrm{TSG}} &= Y_{\mathrm{TSG},2} - Y_{\mathrm{TSG},1} \\ &= \hat{\beta}_{\mathrm{MM1}} \Delta X_{\mathrm{MM1}} = \hat{\beta}_{\mathrm{MM1}} (X_{\mathrm{TSG},2} - X_{\mathrm{TSG},1}) \end{aligned}$$

Suppose there is an optimal response value $Y_{\mathrm{TSG,ideal}}$. Then, set:

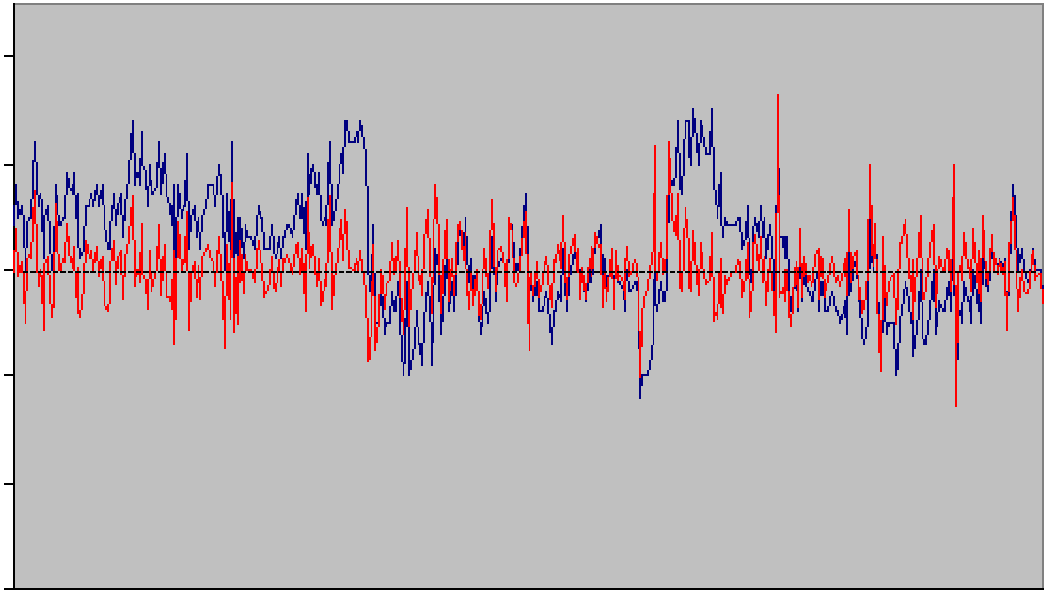
$$X_{\mathrm{TSG,new}} = (Y_{\mathrm{TSG,ideal}} - Y_{\mathrm{TSG,current}}) / \hat{\beta}_{\mathrm{MM1}} + X_{\mathrm{TSG,current}}$$

Substitute in the model: $Y_{\mathrm{TSG,new}} = Y_{\mathrm{TSG,ideal}} + \mathrm{error}$

Regression: parameter estimation

Hence, we can steer response to ideal value. But:
→ error causes deviations from $Y_{\mathrm{TSG,ideal}}$,
→ other methylation marker will not be constant.

Repeated application may yield cellular control:



Ideal value

▬ uncorrected expression level

▬ corrected expression level