Nonlinear regression

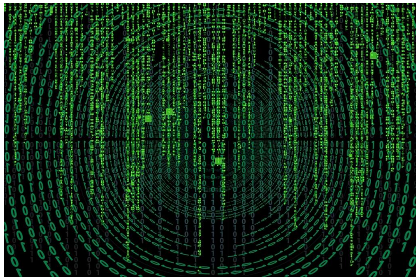
S. Benzekry

1. Fitting a model

Data

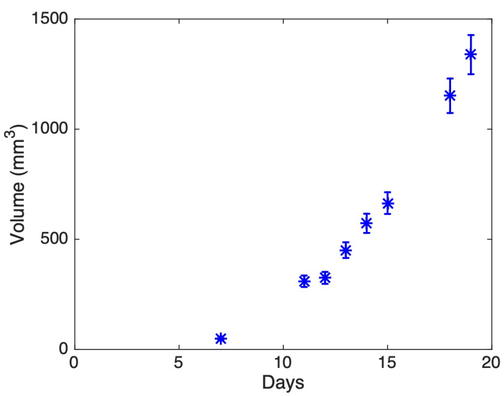
Theory

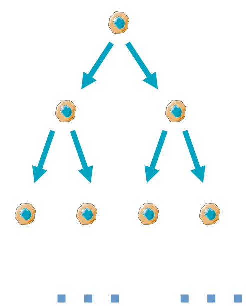
Mathematical model



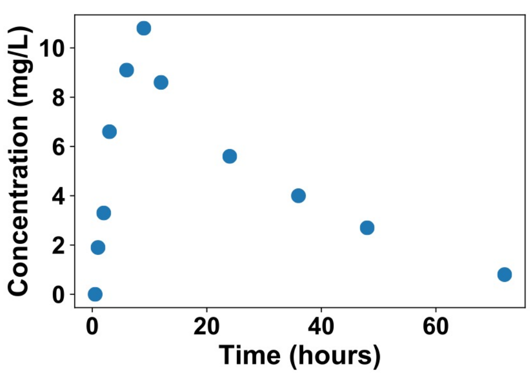
?

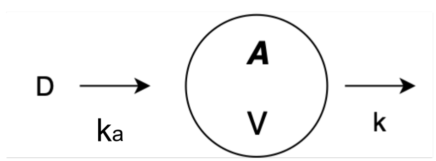
$$M : \begin{array}{ccc} \mathbb{R} \times \mathbb{R}^p & \to & \mathbb{R} \\ (t, \theta) & \mapsto & M(t, \theta) \end{array}$$





$$M(t, \theta) = e^{\theta t}$$



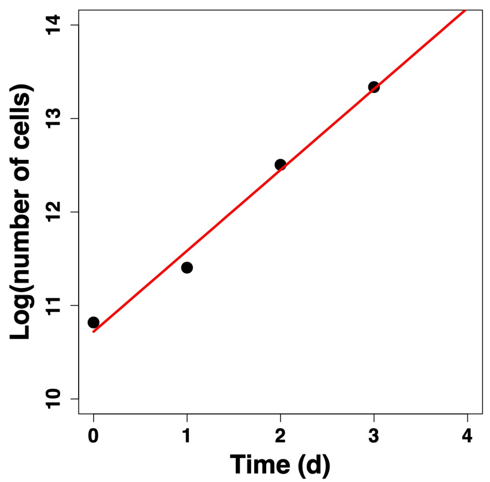


$$\begin{cases} \frac{dA_a}{dt} = -k_a A_a \\ \frac{dA}{dt} = k_a A_a - kA \\ A_a(t=0) = D, \quad (t=0) = 0 \end{cases}$$

$$C(t) = \frac{A(t)}{V}.$$

Linear regression

$$y = \theta_0 + \theta_1 t + \varepsilon$$



Question: what is the « best » linear approximation of y ?

$$\overset{2}{}$$

$$n \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} \longrightarrow$$

M rectangular

no solution

$$\Leftrightarrow y = M \cdot \theta$$

$\times M^T (\in M_{2,n})$

$$\Rightarrow \underbrace{M^T y}_{M_{2,n} \cdot M_{n,1}} = \underbrace{M^T M}_{M_{2,n} \cdot M_{n,2}} \cdot \theta \quad M_{2,1} \longrightarrow$$

$$M_{2,1} \qquad M_{2,2} \cdot M_{2,1}$$

one unique solution

(if the square matrix $M^T M$ is invertible)

$$\hat{\boldsymbol{\theta}} = \left(\boldsymbol{M}^T \boldsymbol{M}\right)^{-1} \boldsymbol{M}^T \boldsymbol{y}$$

Formalism

- **Observations**: *n* couples of points (t_j, y_j), with $y_j \in \mathbb{R}$ (or $\mathbb{R}^m$).

  We will denote $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$ and $t = (t_1, \cdots, t_n)$.

- **Structural model:** a function

$$M : \begin{array}{ccc} \mathbb{R} \times \mathbb{R}^p & \to & \mathbb{R} \\ (t, \theta) & \mapsto & M(t, \theta) \end{array}$$

- The (unknown) vector of **parameters** $\theta^* \in \mathbb{R}^p$

Goal = find θ^*

Statistical model

$$y_j = M(t_j; \theta^*) + e_j$$

- « True » parameter θ^*
- e_j = error = measurement error + structural error
- $(y_1, \cdots, y_n)$ are realizations of **random variables**

$$Y_j = M(t_j; \theta^*) + \varepsilon_j$$

Y_j, ε_j = r.v.

y_j, e_j = realizations

- $(y_1, \cdots, y_n)$ = **sample** with probability density function $p(y|\theta^*)$
- An **estimator** of θ^* is a random variable function of Y, denoted $\hat{\theta}$:

$$\hat{\theta} = h(Y_1, \cdots, Y_n)$$

Linear least-squares: statistical properties

$$Y = M\theta^* + \varepsilon$$

$$\hat{\theta}_{LS} = (M^T M)^{-1} M^T Y \;\Leftrightarrow\; \hat{\theta}_{LS} = \underset{\theta \in \mathbb{R}^p}{argmin} \parallel Y - M\theta \parallel^2$$

Proposition:

Assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, *then*

$$\hat{\theta}_{LS} \sim \mathcal{N}(\theta^*, \sigma^2 (M^T M)^{-1})$$

From this, standard errors and confidence intervals can be computed on the parameter estimates

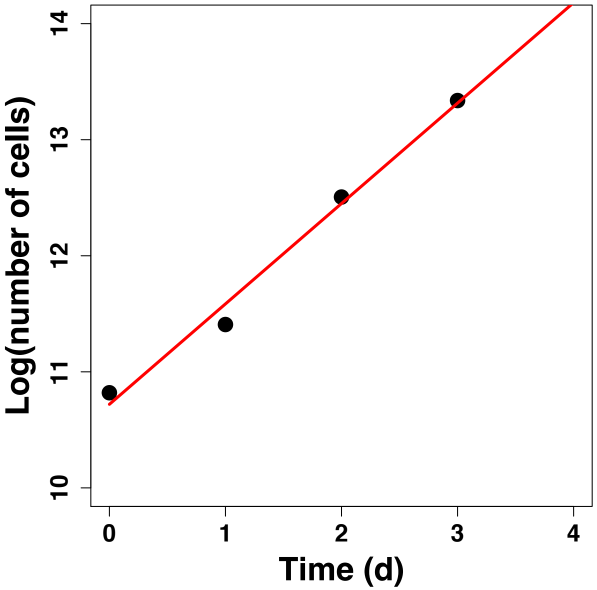
$$s^2 = \frac{1}{n-p} \parallel y - M\hat{\theta}_{LS} \parallel^2 \qquad se\left(\hat{\theta}_{LS,p}\right) = s\sqrt{(M^T M)^{-1}_{p,p}} \qquad IC_\alpha(\theta^*) = \hat{\theta}_{LS,p} \pm t^{\alpha/2}_{n-p} s \sqrt{(M^T M)^{-1}_{p,p}}$$

Example: tumor growth

$$nb_j \simeq N_0 e^{\lambda t_j} = N_0 e^{\frac{\log(2)}{DT} t_j}$$

$$\ln(nb_j) \simeq \log(N_0) + \lambda t_j$$

$$y = \theta_1 + \theta_2 t + \varepsilon$$



$$\hat{\lambda} = \widehat{\theta_2} = 0.865 \Longrightarrow \widehat{DT} = \frac{\log(2)}{\hat{\lambda}}\ 19.2 \text{ hours}$$

$$se(\widehat{\theta_2}) = 0.004, \mathrm{rse}(\widehat{\theta_2}) = 0.005$$

$$IC(DT) = (18.8, 19.7) \text{ hours}$$

Statistical test for the model parameters

$$\hat{\theta} \sim \mathcal{N}(\theta^*, \sigma^2 (M^T M)^{-1})$$

For $k = 1,2 \ldots, p$ $\quad t_k = \dfrac{\theta_k - \theta_k^*}{se_k}$ $\quad$ t-distribution with $n - p$ degrees of freedom

$\Longrightarrow$ t-test (Wald test)

H₀ : « $\theta_k = 0$ » versus H₁ : « $\theta_k \neq 0$ »

Under the null hypothesis, $t_{stat} = \dfrac{\hat{\beta}_k}{se_k}$ follows a t-distribution with $n - d$ degrees of freedom

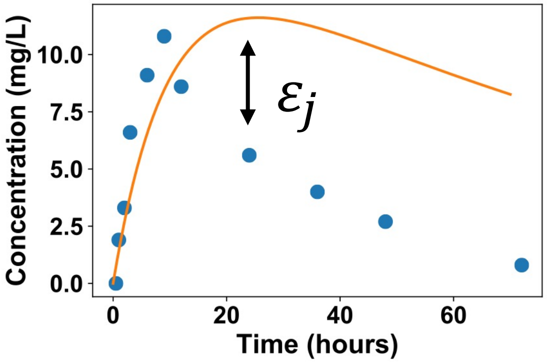
p-value:

$$\mathbb{P}(|t_{n-p}| \geq t_{stat}) = 2\left(1 - \mathbb{P}(t_{n-p} \leq t_{stat})\right)$$

Nonlinear regression: least-squares

$$Y = M(t;\theta^*) + \varepsilon$$

$$\hat{\theta}_{LS} = \underset{\theta \in \mathbb{R}^p}{argmin} \parallel Y - M(t;\theta) \parallel^2$$



Linearization: $M(t,\theta) = M(t,\theta^*) + J \cdot (\theta - \theta^*) + o(\theta - \theta^*), \quad J = D_\theta M(t,\theta^*)$

Proposition:

Assume $\varepsilon \sim \mathcal{N}(0,\sigma^2 I)$. Then, for large n, approximately

$$\hat{\theta}_{LS} \sim \mathcal{N}(\theta^*, \sigma^2 (J^T J)^{-1})$$

⇒ standard errors, confidence intervals

Confidence interval and prediction interval

$$Y = M(t;\theta) + \varepsilon$$

- Prediction at new time t_{new} $\qquad \hat{M}_{new} = M\left(t_{new}, \hat{\theta}\right)$

- Uncertainty on parameter estimate $\hat{\theta}$ $\Longrightarrow$ confidence interval on $\hat{M}_{new}$

$$\hat{M}_{new} \sim \mathcal{N}\left(M_{new}, Var\left(\hat{M}_{new}\right)\right)$$

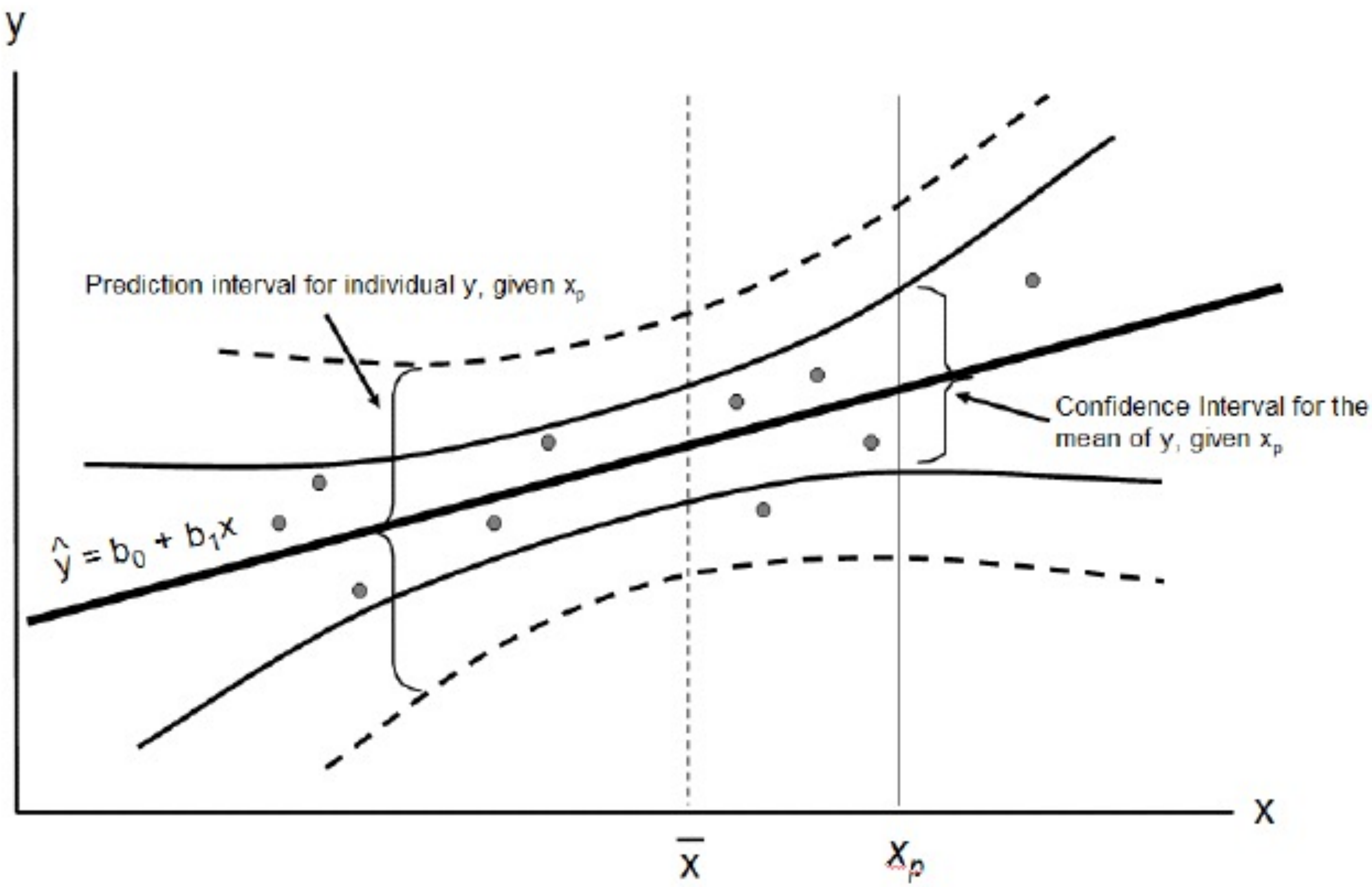
- Uncertainty on parameter estimate $\hat{\theta}$

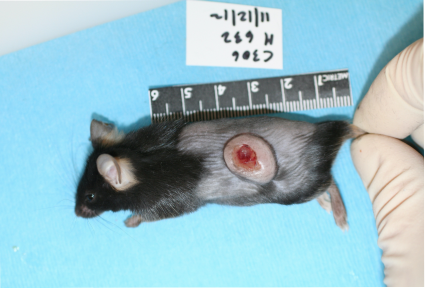
\+ uncertainty on observation ε (e.g. measurement error) $\Longrightarrow$ prediction interval on $\hat{M}_{new}$

$$y_{new} = M_{new} + \varepsilon$$

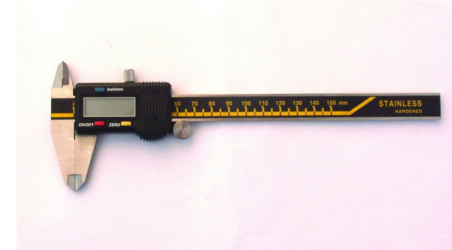
$$y_{new} \sim \mathcal{N}\left(\hat{M}_{new}, Var\left(\hat{M}_{new}\right) + \sigma^2 I\right)$$

Confidence interval vs prediction interval





Error models for tumor volume

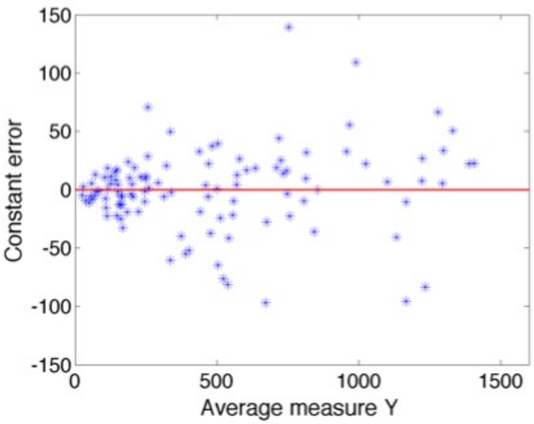


$$\varepsilon_j \ i.i.d \ \mathcal{N}(0, \sigma_j)$$

Constant

$$\sigma_j = \sigma, \forall j$$

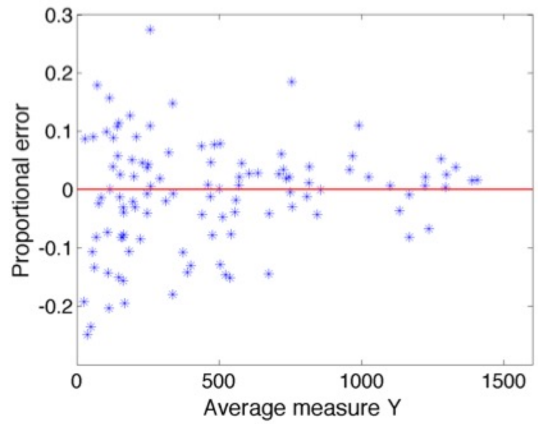
$$p = 0.004$$



Proportional

$$\sigma_j = \sigma M\left(t_j, \hat{\theta}\right)$$

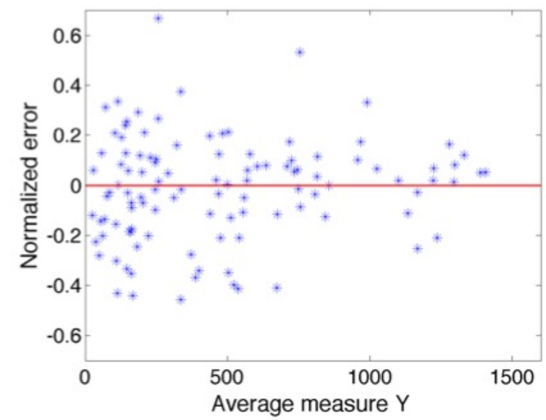
$$p = 0.083$$



Specific

$$\sigma_i = \begin{cases} \sigma M\left(t_j, \hat{\theta}\right)^{\alpha}, & M\left(t_j, \hat{\theta}\right) \geq V_m \\ \sigma V_m^{\alpha}, & M\left(t_j, \hat{\theta}\right) < V_m \end{cases}$$

$$p = 0.2$$



Note: combined error model = $\sigma_j = a + bM\left(t_j, \hat{\theta}\right)$

Nonlinear regression: Likelihood maximization

$$Y = M(t; \theta^*) + \varepsilon$$

The likelihood is defined by

$$L(\theta) = p(y_1, \cdots, y_n | \theta) = \prod_{j=1}^{n} p(y_j | \theta)$$

It is the probability to observe y if the parameter is θ.

The maximum likelihood estimator (MLE) is the value of θ that maximizes the likelihood

$$\hat{\theta}_{MV} = \underset{\theta}{argmax} L(\theta)$$

Asymptotic properties of the MLE

Proposition:

Under regularity assumptions on L*, when* $n \to +\infty$

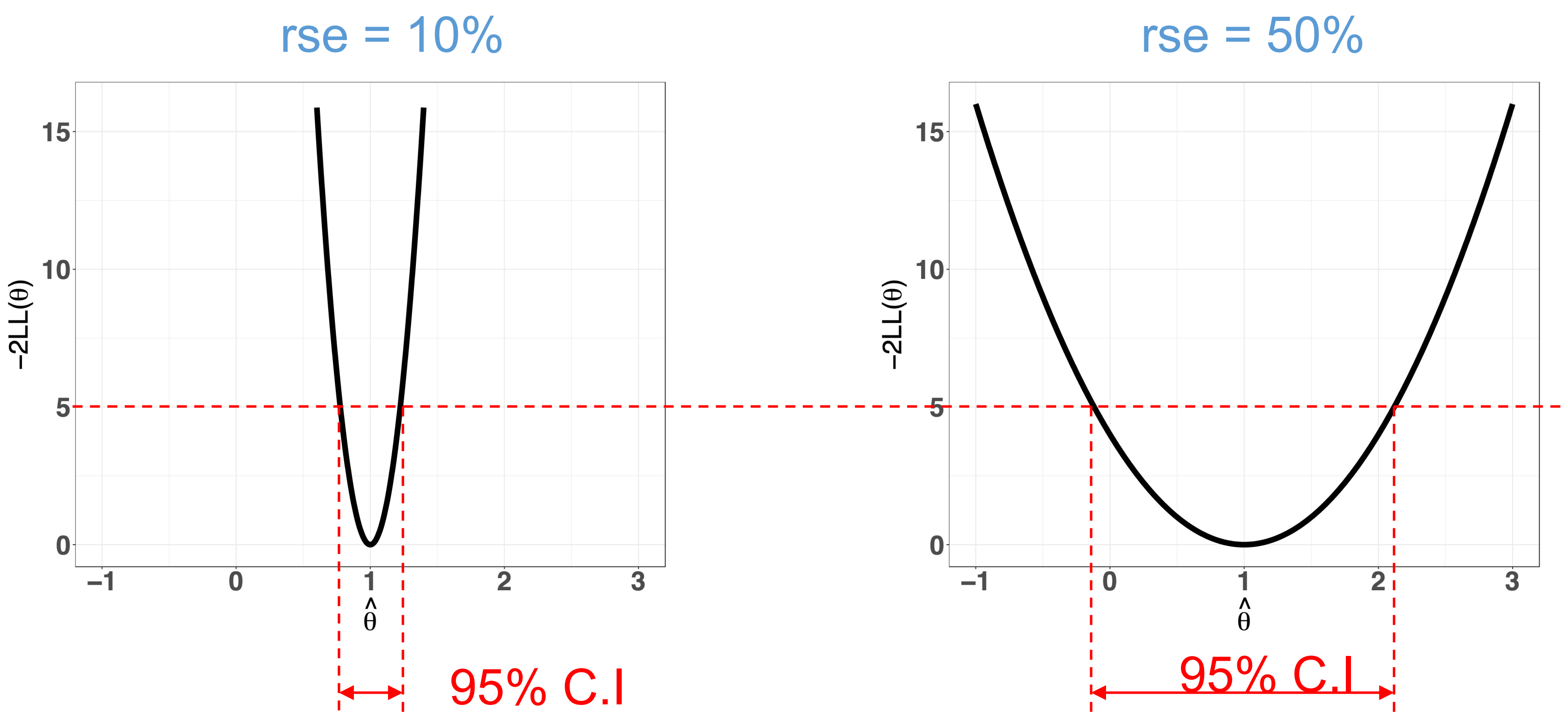
1. $\hat{\theta}_{MV} \longrightarrow \theta^*$ *(consistency)*
2. $\hat{\theta}_{MV}$ *is asymptotically of minimal variance (it reaches the Cramér-Rao bound):*

$$\sqrt{n}\left(\hat{\theta}_{MV} - \theta^*\right) \rightharpoonup \mathcal{N}\left(0, I_{\theta^*}^{-1}\right)$$

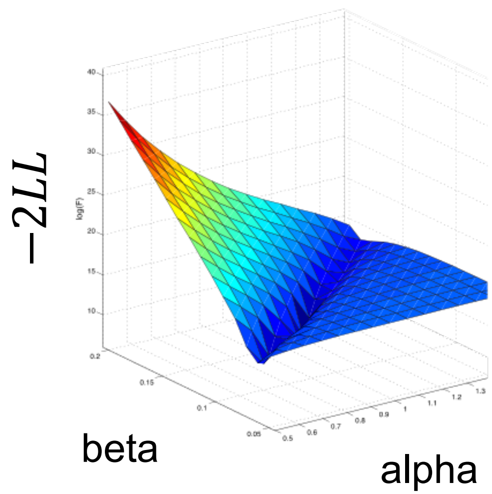
where I_{θ^*} *is the Fisher information matrix*

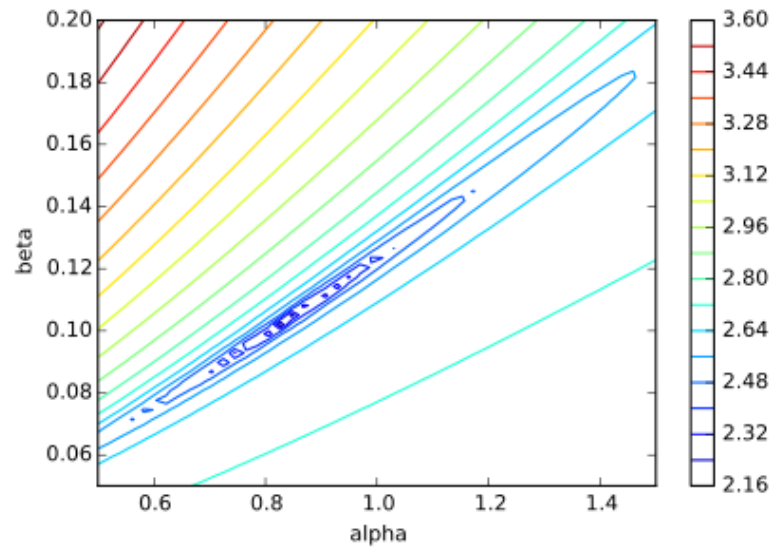
$$(I_{\theta^*})_{j,k} = \mathbb{E}\left[\left\{\frac{\partial \log(p(Y|\theta^*))}{\partial \theta_j}\right\}\left\{\frac{\partial \log(p(Y|\theta^*))}{\partial \theta_k}\right\}\right] = \mathbb{E}\left[-\left(\frac{\partial^2 \log\left(p(Y|\theta^*)\right)}{\partial \theta_j \partial \theta_k}\right)\right].$$

Precision of the estimates



Correlation between estimates





Correlation matrix of the estimates

	R.S.E.(%)			
alpha_pop	3.09	1		
beta_pop	5.68	**0.98574**	1	
b	23.8	-0.00055974	0.022018	1

	MIN	MAX	MAX/MIN
Eigen values	0.014	2	1.4e+2

small r.s.e on alpha and beta, but large correlation

MLE: normal errors

$$Y_j = M(t_j;\theta^*) + \varepsilon_j, \varepsilon_j \sim \mathcal{N}(0,\sigma)$$

$$p(y_j|\theta,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(y_j - M(t_j,\theta)\right)^2}{2\sigma^2}}, L(\theta,\sigma) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{-\frac{\| y - M(t,\theta) \|^2}{2\sigma^2}}$$

Maximize $L(\theta,\sigma)$ ⇔minimize $F(\theta,\sigma) = -\log(L(\theta,\sigma))$

$$F(\theta,\sigma) = n\log\left(\sigma\sqrt{2\pi}\right) + \frac{\| y - M(t,\theta) \|^2}{2\sigma^2}$$

$$\frac{\partial F}{\partial \sigma}\left(\hat{\theta},\hat{\sigma}\right) = 0 \Rightarrow \hat{\sigma} = \frac{1}{n} \| y - M(t,\hat{\theta}) \|^2$$

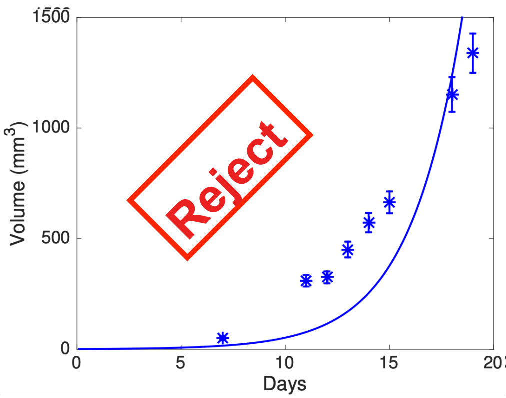
$$\Rightarrow \hat{\theta} = \underset{\theta}{argmin} \| y - M(t,\theta) \|^2$$

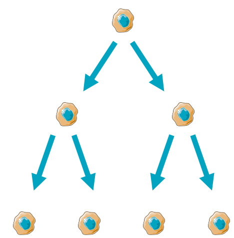
Maximum likelihood ⇔ Least-squares

Application: tumor growth

What are **minimal** biological processes able to recover the **kinetics** of (experimental) tumor growth?

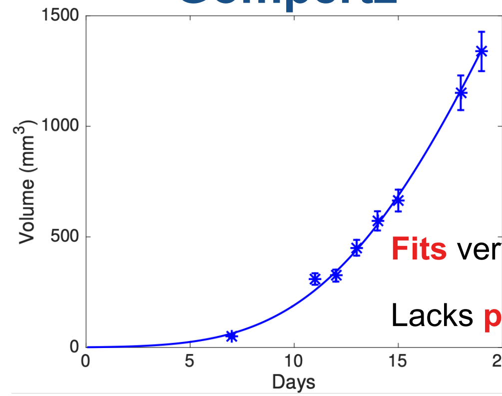
Exponential





$$\frac{dV}{dt} = aV$$

Gompertz

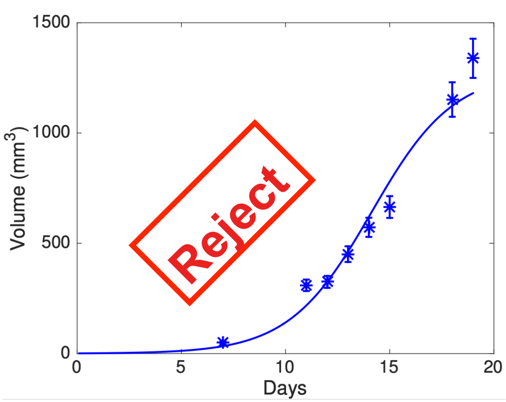


$$\frac{dV}{dt} = \alpha e^{-\beta t} V$$

Fits very well

Lacks **physiological** interpretation

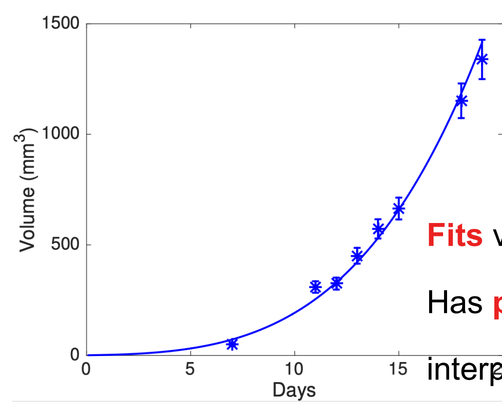
Logistic



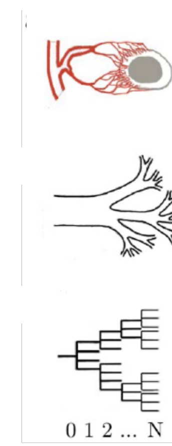
Competition

$$\frac{dV}{dt} = aV\left(1 - \frac{V}{K}\right)$$

Power law



$$\frac{dV}{dt} = \alpha V^{\gamma}$$



Fits very well

Has **physiological** interpretation

Benzekry et al., PLoS Comp Biol, 2014

Goodness of fit metrics

Sum of Squared Errors

$$SSE^i = \sum_{j=1}^{n^i} \left(\frac{V_j^i - V(t_j^i, \hat{\theta}^i)}{\hat{\sigma}_j^i} \right)^2$$

Akaike Information Criterion

$$AIC^i = -2l(\hat{\theta}^i) + 2p$$

↑ number of parameters

Model	SSE	AIC	RMSE	R2	p > 0.05	#
Power law	**0.164(0.0158 - 0.646)[1]**	**-18.4(-43.2 - 1.63)[1]**	**0.415(0.145 - 0.899)[1]**	**0.97(0.801 - 0.998)[1]**	100	2
Gompertz	0.176(0.019 - 0.613)[2]	-16.9(-48.2 - 1.1)[2]	0.433(0.156 - 0.875)[2]	0.971(0.828 - 0.997)[2]	100	2
Logistic	0.404(0.0869 - 0.85)[3]	-5.41(-18.4 - 3.88)[3]	0.665(0.331 - 1)[3]	0.908(0.712 - 0.989)[3]	100	2
Exponential	1.9(0.31 - 3.56)[4]	10.7(-5.38 - 23.1)[4]	1.4(0.595 - 1.95)[4]	0.69(0.454 - 0.944)[4]	15	1

Root Mean Squared Errors

$$RMSE^i = \sqrt{\frac{1}{n-p} SSE^i}$$

R^2

$$R^{2,j} = 1 - \frac{\sum_j \left(V_j^i - V(t_j^i; \hat{\theta}^i) \right)^2}{\sum_j \left(V_j^i - \overline{V^i} \right)^2}$$

Parameter values and identifiability

Model	Par.	Unit	Median value (CV)	NSE (%) (CV)
Power law	α	$mm^{3(1-\gamma)} \cdot day^{-1}$	0.886 (30.8)	8.17 (52.5)
	γ	-	0.788 (7.56)	2.28 (58.6)
Gompertz	α_0	day^{-1}	1.68 (23.5)	6.11 (82.9)
	β	day^{-1}	0.0703 (28)	8.35 (92.9)
Logistic	a	day^{-1}	0.474 (13.3)	2.93 (23.3)
	K	mm^3	1.92e+03 (36.7)	15.8 (28.7)
Exponential	a	day^{-1}	0.356 (12.9)	2.53 (19.4)
Generalized logistic	a	$[day^{-1}]$	2555 (148)	2.36e+05 (137)
	K	$[mm^3]$	4378 (307)	165 (220)
	α	-	0.0001413 (199)	2.36e+05 (137)

NSE = Normalized Standard Error ⟷ practical identifiability

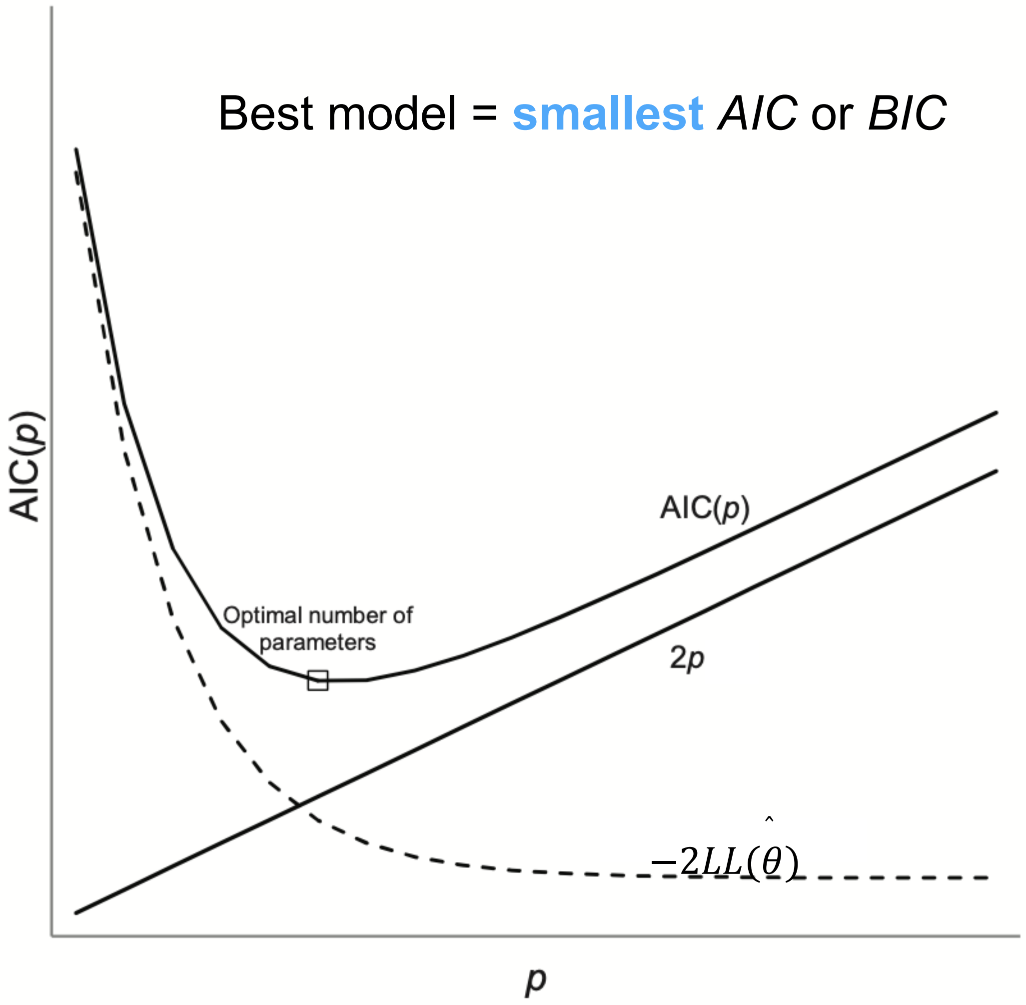
$$\hat{\theta} \sim \mathcal{N}\left(\theta^*, \hat{\sigma}^2 \left(J \cdot J^T\right)^{-1}\right)$$

$$se\left(\hat{\theta}^k\right) = \sqrt{\hat{\sigma}^2 \left(J \cdot J^T\right)_{k,k}}$$

Information criteria

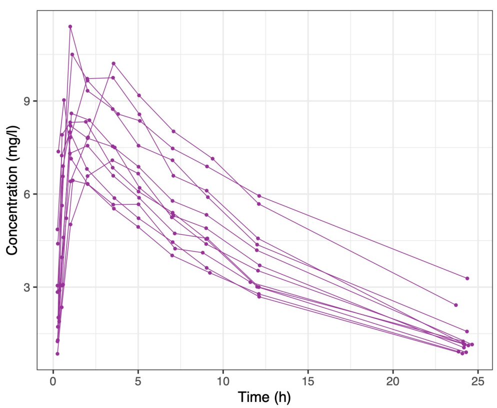
$AIC = -2LL(\hat{\theta}) + 2p$ $BIC = -2LL(\hat{\theta}) + \log(n)p$

Best model = **smallest** *AIC* or *BIC*



Population modeling: the two-steps approach

Population data



Individual fits

$$Y^{1} = M(t;\theta^{1}) + \varepsilon$$

$$Y^{2} = M(t;\theta^{2}) + \varepsilon$$

$$\vdots$$

$$Y^{N} = M(t;\theta^{N}) + \varepsilon$$

$$\hat{\theta}^{1}, \dots, \hat{\theta}^{N}$$

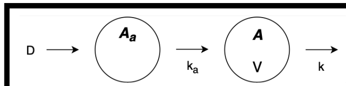
$$\hat{\theta}_{pop} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}^{i}$$

$$\hat{\Omega} = VCov\left(\hat{\theta}^{i}\right)$$

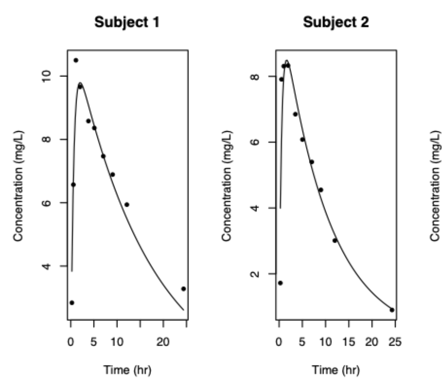
$$\mathcal{N}\left(\hat{\theta}_{pop}, \hat{\Omega}\right)$$

Population model

Individual structural model

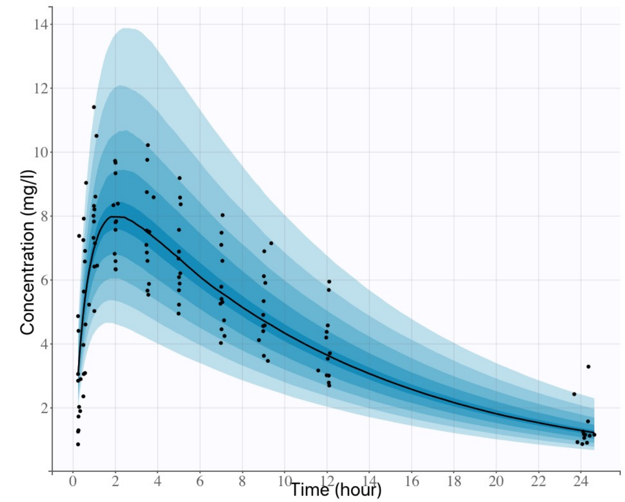


$$\begin{cases} \frac{dA_a}{dt} = -k_a A_a \\ \frac{dA}{dt} = k_a A_a - kA \\ A_a(t=0) = D, \quad (t=0) = 0 \end{cases} \qquad C(t) = \frac{A(t)}{V}.$$

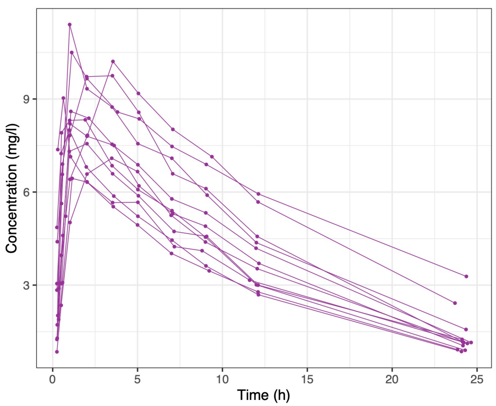


Population modeling: mixed-effects approach

Population fit (MLE)



Population data



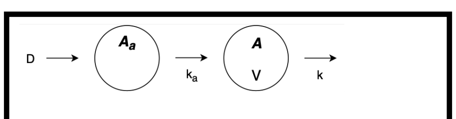
Population model

$$\theta^i = \theta_{pop} + \eta^i, \eta^i \sim \mathcal{N}(0, \Omega)$$

fixed effects

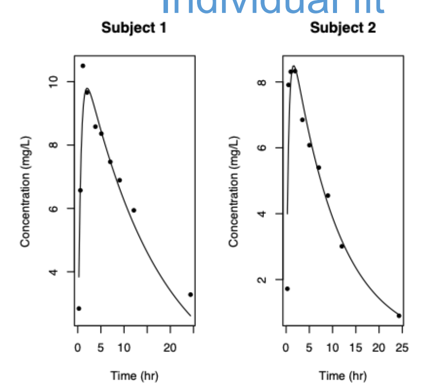
random effects

Individual structural model



$$\begin{cases} \frac{dA_a}{dt} = -k_a A_a \\ \frac{dA}{dt} = k_a A_a - kA \\ A_a(t=0) = D, \quad (t=0) = 0 \end{cases} \qquad C(t) = \frac{A(t)}{V}.$$

Individual fit



References

- Course « Statistics in Action with R » by Marc Lavielle

  http://sia.webpopix.org/index.html

- Seber, G. A., & Wild, C. J. (2003). Nonlinear regression. Hoboken (NJ): Wiley-Interscience.