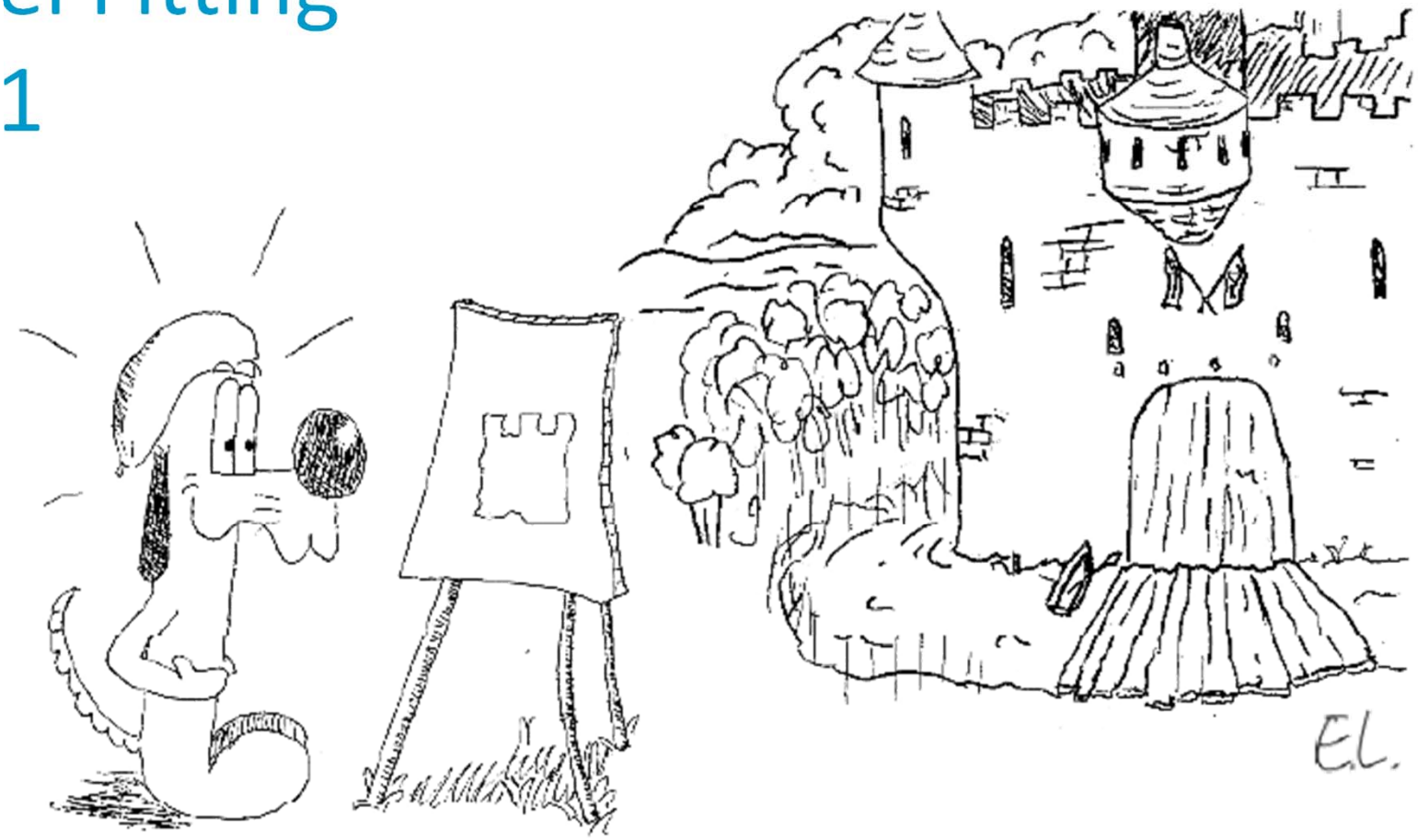
Model Fitting
Part 1



Jean-Yves Le Boudec

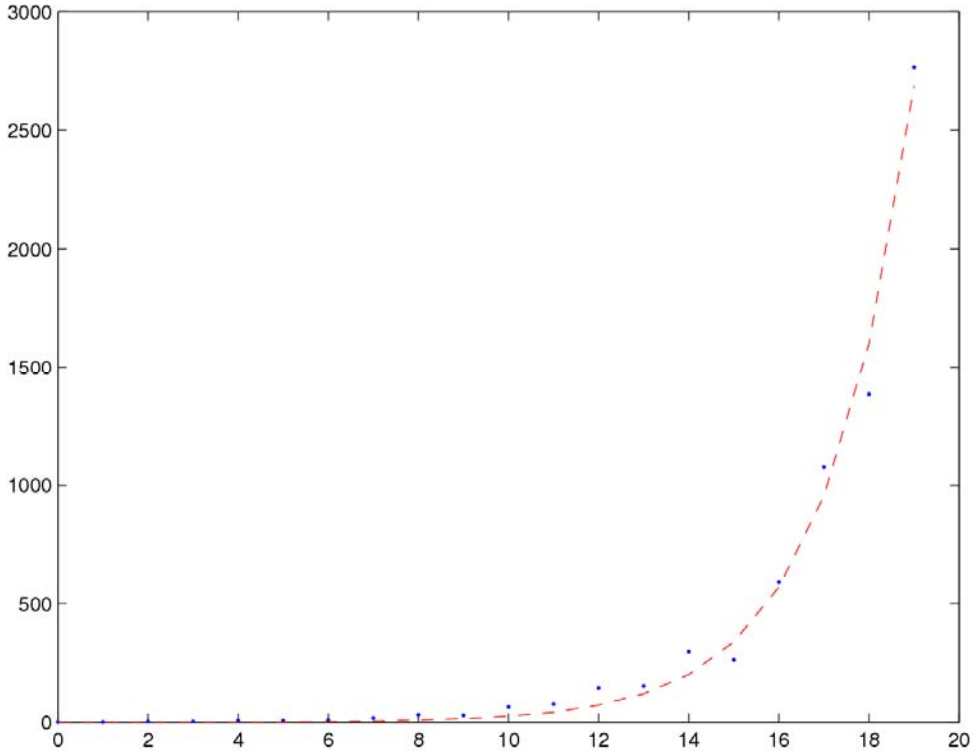
March 2019

Contents

Virus Infection Data

We would like to capture the growth of infected hosts

(explanatory model)

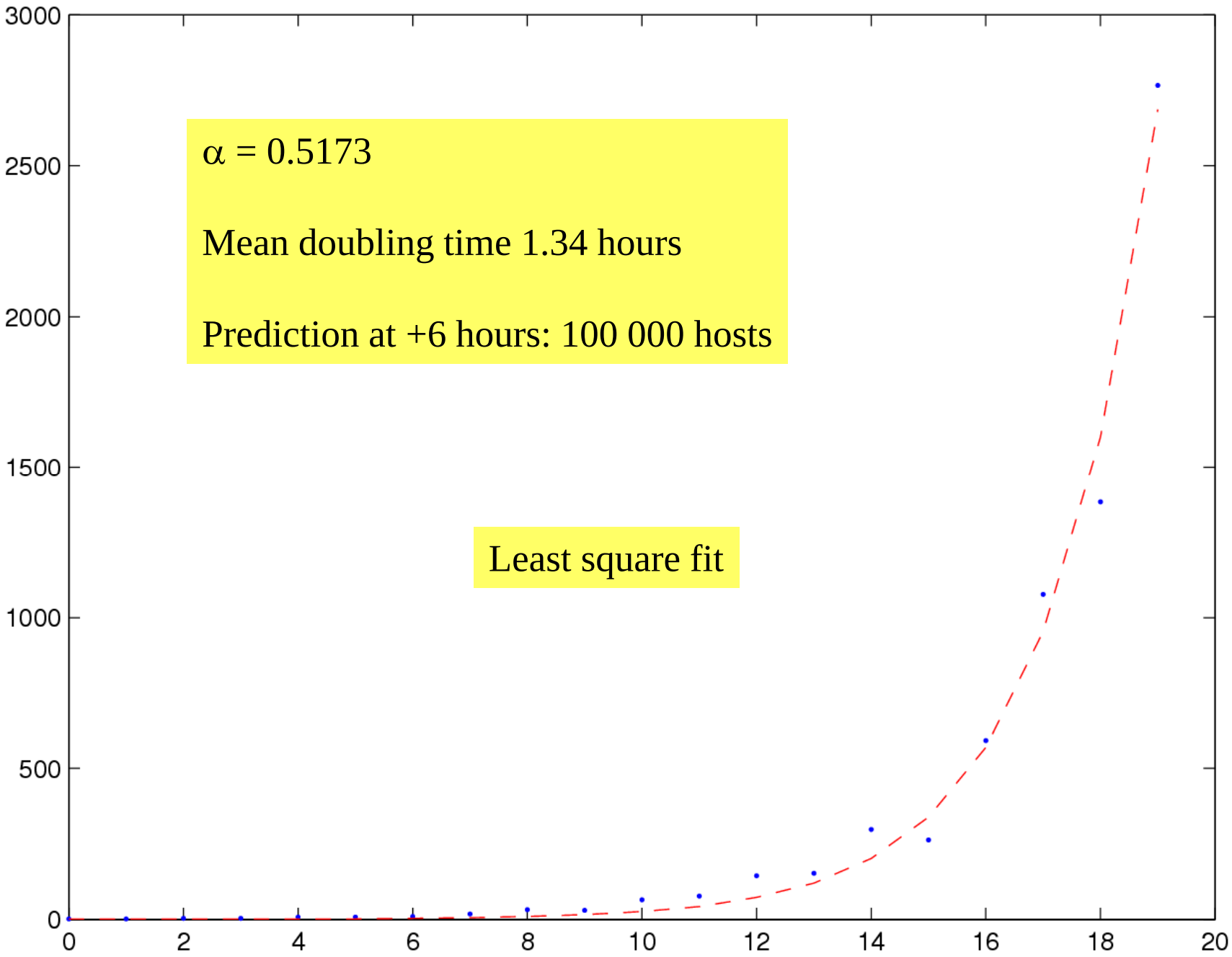


An exponential model seems appropriate

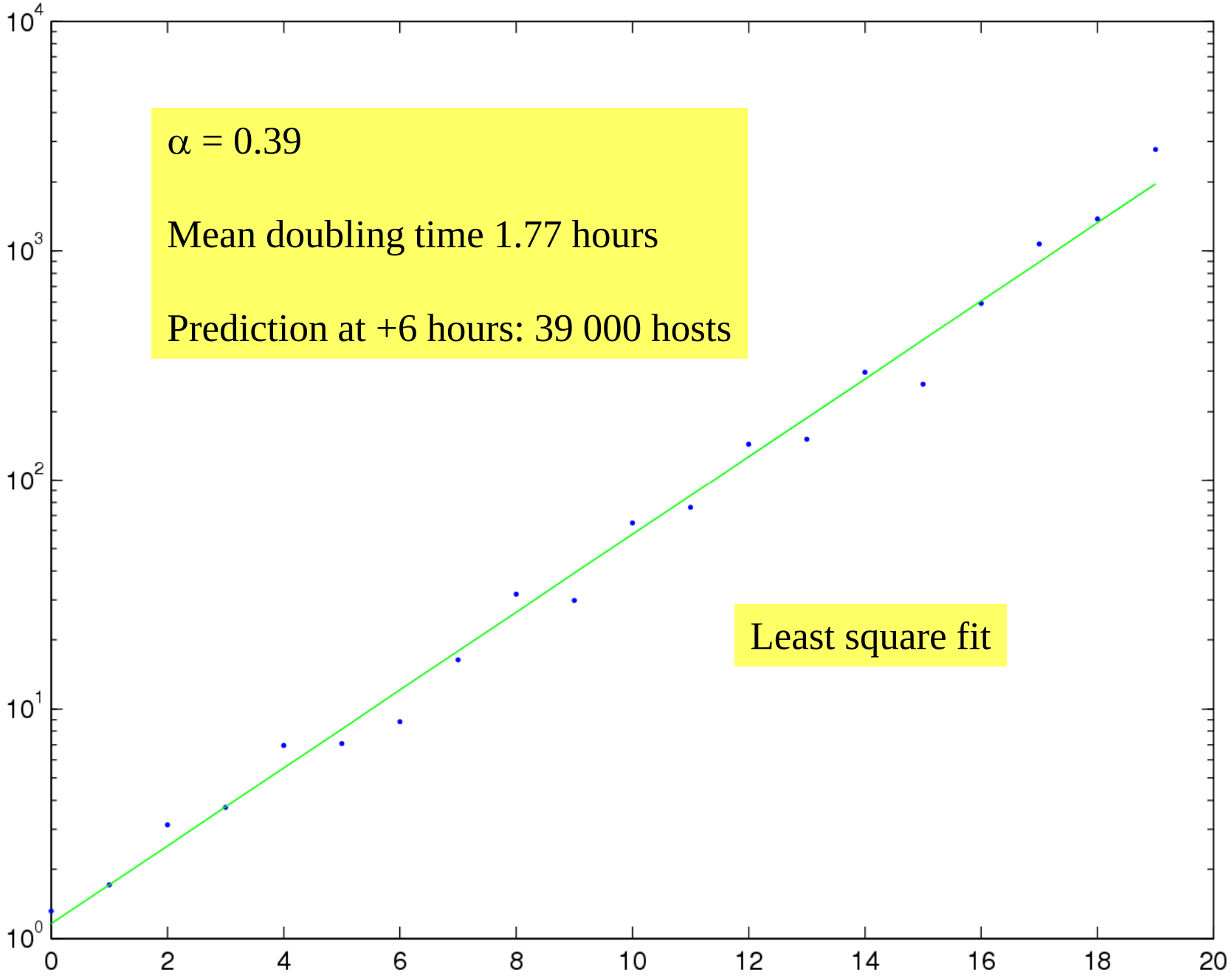
$$Y(t) = ae^{\alpha t}$$

How can we fit the model, in particular, what is the value of α ?

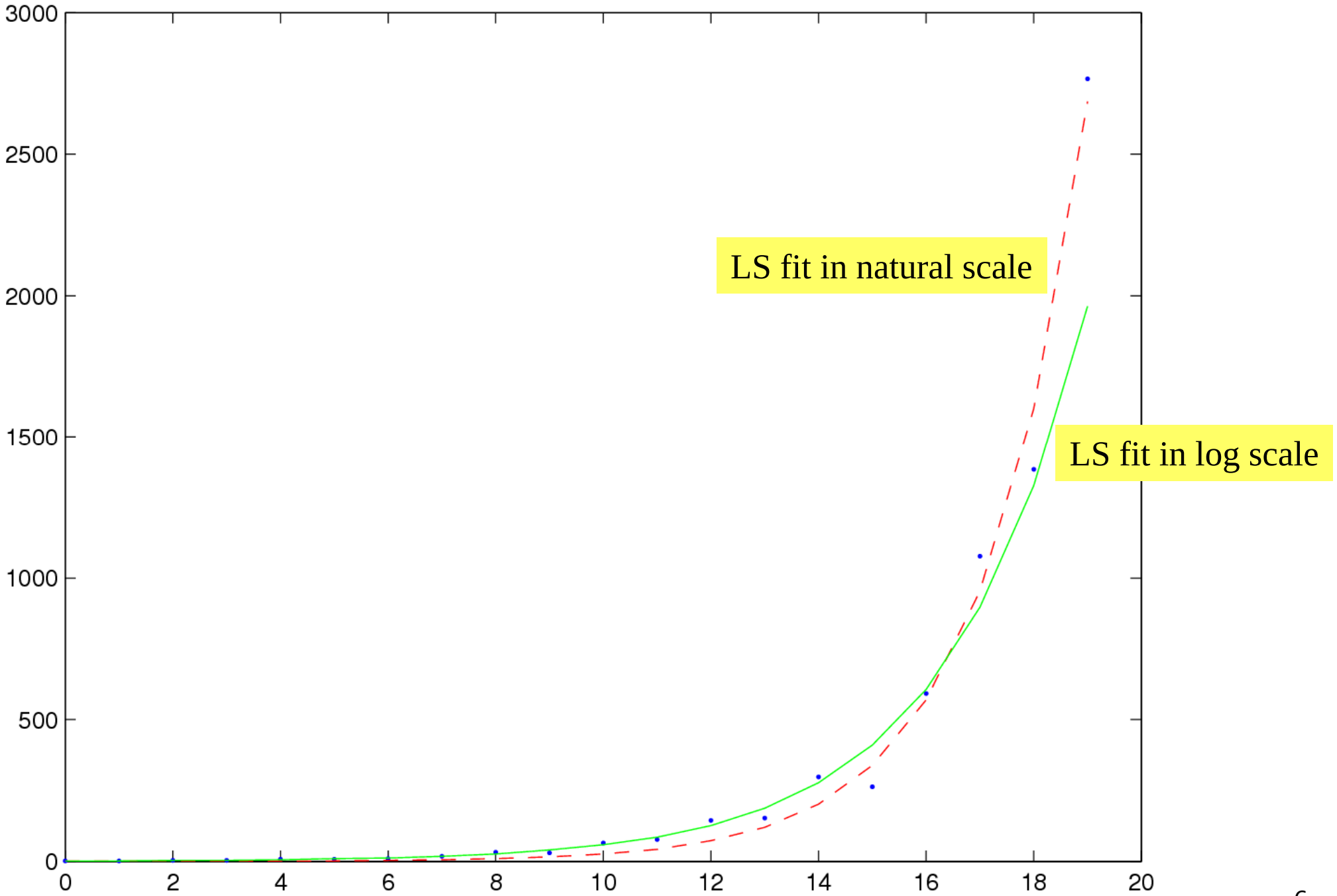
Least Square Fit of Virus Infection Data



Least Square Fit of Virus Infection Data In Log Scale

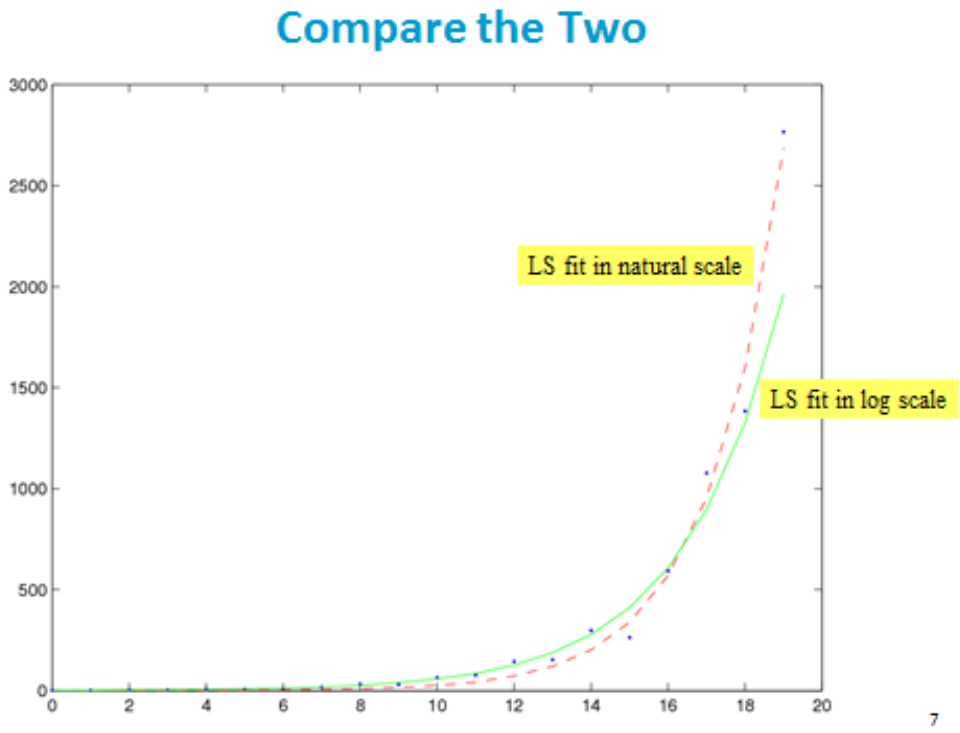


Compare the Two



Which scale should I use ?

A. The natural scale because it is the simplest (principle of parsimony)
B. The log scale because it is better adapted to exponential growth
C. I don't know



What is model fitting ?

Given some data Y_i and a parametric model $f_i(\theta)$ we assume that :

$$Y_i = f_i(\theta) + \text{noise}$$

The model fitting problem is to estimate the unknown parameter θ

For example: the model is $Y_i = a\mathrm{e}^{\alpha t_i}$

or

$$\log Y_i = \log a + \alpha t_i$$

These two models are equivalent; the parameter is $\theta = (a, \alpha)$

The signal processing interpretation

Given a model $Y_i = f_i(\theta) + \mathbf{noise}$ and a *score function,* find θ that minimizes the score

Classical scores

- Least Square: $\mathbf{score} = \sum_i (\mathbf{noise}_i)^2$
- Weighted Least Square: $\mathbf{score} = \sum_i (w_i \times \mathbf{noise}_i)^2$
- ℓ^1 score: $\mathbf{score} = \sum_i |\mathbf{noise}_i|$
- Weighted ℓ^1 score: $\mathbf{score} = \sum_i w_i |\mathbf{noise}_i|$

Which one should we use ?

The statistical interpretation

The data Y_i is output by a simulator, with parameter θ, which we want to estimate

This forces us to make some assumption about the *noise* term; for example, we could consider any one of the following models

1. $Y_i = a\mathrm{e}^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
2. $Y_i = a\mathrm{e}^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
3. $\log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$

For each of these models, the parameter is now $\theta = (a, \alpha, \sigma)$

Estimating a statistical model: maximum likelihood

Principle: find θ that maximizes the pdf $f_Y(y|\theta)$, also called likelihood, where y is the data

This is a robust and consistent estimation method

Example: model 1 : $Y_i = a\mathrm{e}^{\alpha t_i} + \epsilon_i, \qquad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$

$$f_Y(y|a,\alpha,\sigma) = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^I} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\left(y_i - a\mathrm{e}^{\alpha t_i}\right)^2} \qquad (1)$$

the problem is to find (a, α, σ) that maximizes (1)

Least Square Score = Gaussian iid Noise

Assume model (homoscedasticity)

$$Y_i = f_i(\vec{\beta}) + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \epsilon_i \text{ iid } \sim N_{0,\sigma^2}$$

THEOREM 3.1.1 (Least Squares). *For the model in Eq.(3.5),*

1. *the maximum likelihood estimator of the parameter* $(\vec{\beta}, \sigma)$ *is given by:*

   (a) $\hat{\beta} = \arg\min_{\vec{\beta}} \sum_i \left(y_i - f_i(\vec{\beta})\right)^2$

   (b) $\hat{\sigma}^2 = \frac{1}{I} \sum_i \left(y_i - f_i(\hat{\beta})\right)^2$

i.e. the signal processing method with Least Squares score is the same as the statistical method with iid gaussian noise

Example: model 1 : $Y_i = a\mathrm{e}^{\alpha t_i} + \epsilon_i, \qquad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$

$$f_Y(y|a, \alpha, \sigma) = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^I} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\left(y_i - a\mathrm{e}^{\alpha t_i}\right)^2} \qquad (1)$$

the problem is to find (a, α, σ) that maximizes (1)

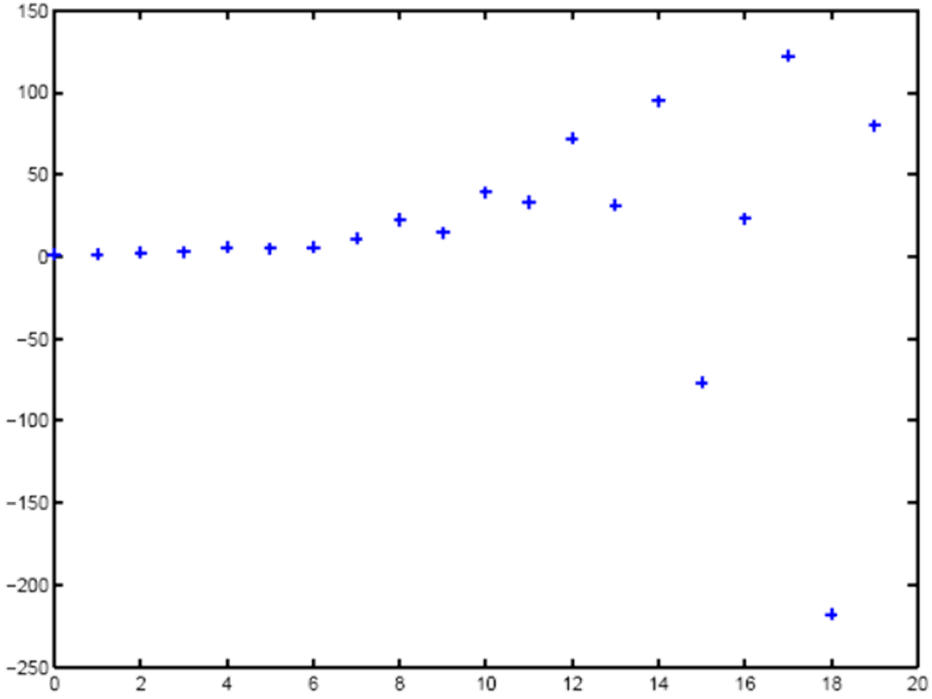
Thm 3.3.1⇒

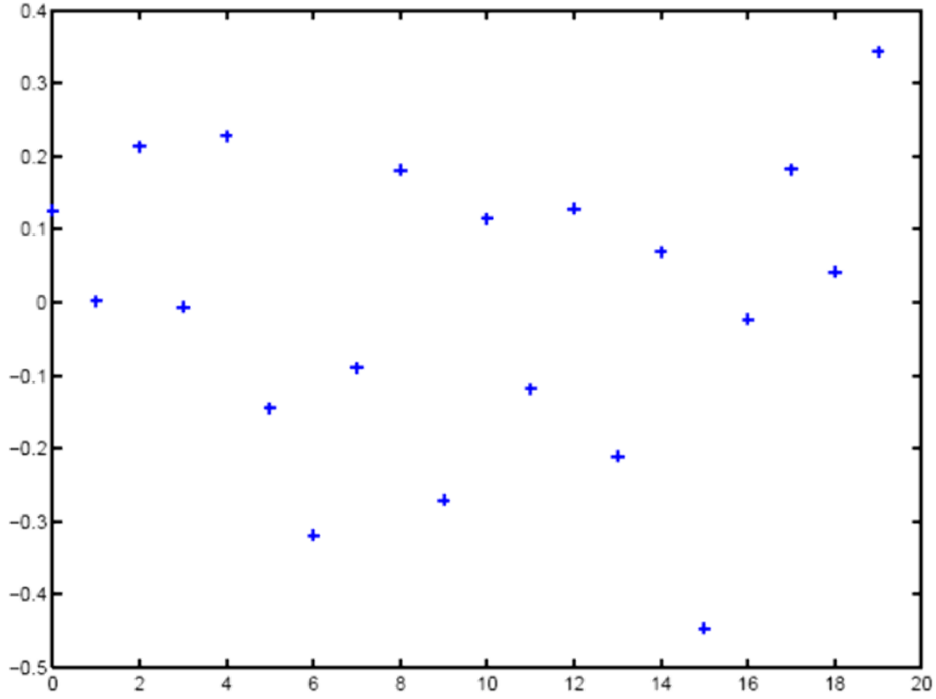
$$(\hat{a}, \hat{\alpha}) = \text{argmin}_{a,\alpha} \ \sum_{i=1}^{I}(y_i - ae^{\alpha t_i})^2$$

$$\hat{\sigma}^2 = \frac{1}{I}\sum_{i}^{I}\left(y_i - \hat{a}\, e^{\hat{\alpha} t_i}\right)^2$$

The residuals can help validate a fitting method

Residuals = estimate of noise terms (after estimation of parameter) – they should be consistent with the model





Model 1:

$$Y_i = ae^{\alpha t_i} + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2)$$

Model 3:

$$\log Y_i = \log a + \alpha t_i + \epsilon_i, \epsilon_i \sim iid\ N(0, \sigma^2)$$

Model 1 appears not to fit, residuals have a variance that grows with t_i, model 3 appears OK; we should prefer it for fitting this data set

LS score functions

Given a model to be fitted

$$Y_i = f_i(\theta) + \text{noise}$$

we have the equivalences:

Score	**Statistical model**
Least Square: $\text{score} = \sum_i (\text{noise}_i)^2$	$\text{noise} \sim \text{iid } N(0, \sigma^2)$
Weighted Least Square: $\text{score} = \sum_i (w_i \times \text{noise}_i)^2$	$\text{noise}_i \sim \text{iid } N\left(0, \frac{\sigma^2}{w_i^2}\right)$

i.e. WLS gives a weight inversely proportional to σ_i of noise

Which score corresponds to which model ?

A. A1 B2 C3
B. A1 B3 C2
C. A2 B1 C3
D. A2 B3 C1
E. A3 B1 C2
F. A3 B2 C1
G. I don't know

1. $Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
2. $Y_i = ae^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
3. $\log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$

$$A: \sum_i^I \left(Y_i - ae^{\alpha t_i}\right)^2 \qquad B: \sum_i^I \left(\frac{Y_i - ae^{\alpha t_i}}{ae^{\alpha t_i}}\right)^2$$

$$C: \sum_i^I (\log Y_i - \log a - \alpha t_i)^2$$

Compare models 2 and 3

1. $Y_i = ae^{\alpha t_i} + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
2. $Y_i = ae^{\alpha t_i}(1 + \epsilon_i), \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$
3. $\log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$

Assume model 2 and take log:

$$\log Y_i = \log a + \alpha t_i + \log(1 + \epsilon_i)$$
$$\approx \log a + \alpha t_i + \epsilon_i$$

So we expect models 2 and 3 to be approximately equivalent.

Take-Home Message

In order to fit a model you need to find an appropriate score function.

One easy way to determine one is to consider an explicit model of the noise, and verify it by looking at residuals

A common mistake is to assume that noise has same variance when it is obviously wrong (e.g. model 1)

Furthermore, we may be interested in finding confidence intervals for the estimated parameters.

2. Least Squares

Very commonly used

Efficient solution, lots of softwares

And gives confidence intervals

$$Y_i = f_i(\vec{\beta}) + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \epsilon_i \text{ iid } \sim N_{0,\sigma^2}$$

THEOREM 3.1 (Least Squares). *For the model in Eq.(3.5),*

1. *the maximum likelihood estimator of the parameter* $(\vec{\beta}, \sigma)$ *is given by:*

   (a) $\hat{\beta} = \arg\min_{\vec{\beta}} \sum_i \left(y_i - f_i(\vec{\beta})\right)^2$

   (b) $\hat{\sigma}^2 = \frac{1}{I} \sum_i \left(y_i - f_i(\hat{\beta})\right)^2$

Confidence Intervals

2. *Let K be the square matrix of second derivatives (assumed to exist), defined by*

$$K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}$$

If K is invertible and if the number I of data points is large, $\hat{\beta} - \vec{\beta}$ is approximately gaussian with 0 mean and covariance matrix K^{-1}.

Alternatively, for large I, an approximate confidence set at level γ for the jth component β_j of $\vec{\beta}$ is implicitly defined by

$$-2I \ln(\hat{\sigma}) + 2I \ln\left(\hat{\sigma}(\hat{\beta}_1, ..., \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} ... \hat{\beta}_p)\right) \geq \xi_1$$

where $\hat{\sigma}^2(\vec{\beta}) = \frac{1}{I} \sum_i \left(y_i - f_i(\vec{\beta})\right)^2$ and ξ_1 is the γ quantile of the χ^2 distribution with 1 degree of freedom (for example, for $\gamma = 0.95$, $\xi_1 = 3.92$).

Example: model 3

$$\log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$$

$f_i = a' + \alpha t_i$ with $\log a = a'$; let $\beta_1 = \alpha$ and $\beta_2 = a'$

$$\frac{\partial f_i}{\partial \beta_1} = t_i, \qquad \frac{\partial f_i}{\partial \beta_2} = 1$$

$$K_{1,1} = \frac{1}{\sigma^2}\sum t_i^2 \,, K_{1,2} = K_{2,1} = \frac{1}{\sigma^2}\sum t_i \,, K_{2,2} = \frac{I}{\sigma^2}$$

$f_i = a' + \alpha t_i$ with $\log a = a'$; let $\beta_1 = \alpha$ and $\beta_2 = a'$

$$\frac{\partial f_i}{\partial \beta_1} = t_i, \qquad \frac{\partial f_i}{\partial \beta_2} = 1$$

$$K_{1,1} = \frac{1}{\sigma^2}\sum t_i^2 \,, K_{1,2} = K_{2,1} = \frac{1}{\sigma^2}\sum t_i \,, K_{2,2} = \frac{I}{\sigma^2}$$

$$K = \frac{1}{\sigma^2}\begin{bmatrix} \sum_i t_i^2 & \sum_i t_i \\ \sum_i t_i & I \end{bmatrix}$$

Example: model 3

$$\log Y_i = \log a + \alpha t_i + \epsilon_i, \quad \epsilon_i \text{ iid} \sim N(0, \sigma^2)$$

2. Let K be the square matrix of second derivatives (assumed to exist), defined by

$$K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}$$

If K is invertible and if the number I of data points is large, $\hat{\beta} - \vec{\beta}$ is approximately gaussian with 0 mean and covariance matrix K^{-1}.

$$K = \frac{1}{\sigma^2} \begin{bmatrix} \sum_i t_i^2 & \sum_i t_i \\ \sum_i t_i & I \end{bmatrix} = \frac{I}{\sigma^2} \begin{bmatrix} s_t^2 + \bar{t}^2 & \bar{t} \\ \bar{t} & 1 \end{bmatrix}$$

Matlab calls K^{-1} the variance-covariance matrix

$$\text{with } \bar{t} = \frac{1}{I} \sum_i t_i \,, s_t^2 = \frac{1}{I} \sum_i (t_i - \bar{t}\,)^2 \quad \Rightarrow K^{-1} = \frac{\sigma^2}{I s_t^2} \begin{bmatrix} 1 & -\bar{t} \\ -\bar{t} & s_t^2 + \bar{t}^2 \end{bmatrix}$$

$$\Rightarrow \alpha - \hat{\alpha} \sim N\left(0, \frac{\sigma^2}{I s_t^2}\right), a - \hat{a} \sim N\left(0, \frac{\sigma^2}{I s_t^2} (\bar{t}^2 + s_t^2)\right)$$

An approximate 95% confidence interval for $\boldsymbol{\alpha}$ is (with $\hat{\sigma}^2 = \frac{1}{I}\sum_i (\log Y_i - \log \hat{a} - \hat{\alpha}\, t_i)^2$)

A. $\alpha \pm 1.96 \ \frac{\hat{\sigma}}{s_t}$

B. $\hat{\alpha} \pm 1.96 \ \frac{\hat{\sigma}}{\sqrt{I} s_t}$

C. $\hat{\alpha} \pm 1.96 \ \frac{\hat{\sigma}}{I s_t}$

D. None of the above

E. I don't know

Linear Regression

Model 3 is a special case of *linear regression*

By definition, a linear regression model (with least squares) means a model of the form

$$Y_i = f_i(\beta) + \epsilon_i, \qquad \epsilon_i \sim iid\ N(0, \sigma^2)$$

where f_i is *linear with respect to* β

For such models, there are closed form solutions in matrix form for the maximum likelihood estimation solution, including exact confidence intervals – see thm 3.3

DEFINITION 3.1 (Linear Regression Model).

$$Y_i = (X\vec{\beta})_i + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \epsilon_i \text{ iid } \sim N_{0,\sigma^2} \tag{3.8}$$

where the unknown parameter $\vec{\beta}$ is in $\mathbb{R}^p$ and X is a $I \times p$ matrix. The matrix X supposed to be known exactly in advance. We also assume that

H *X has rank p*

THEOREM 3.3 (Linear Regression). *Consider the model in Definition 3.1; let $\vec{y}$ be the $I \times 1$ column vector of the data.*

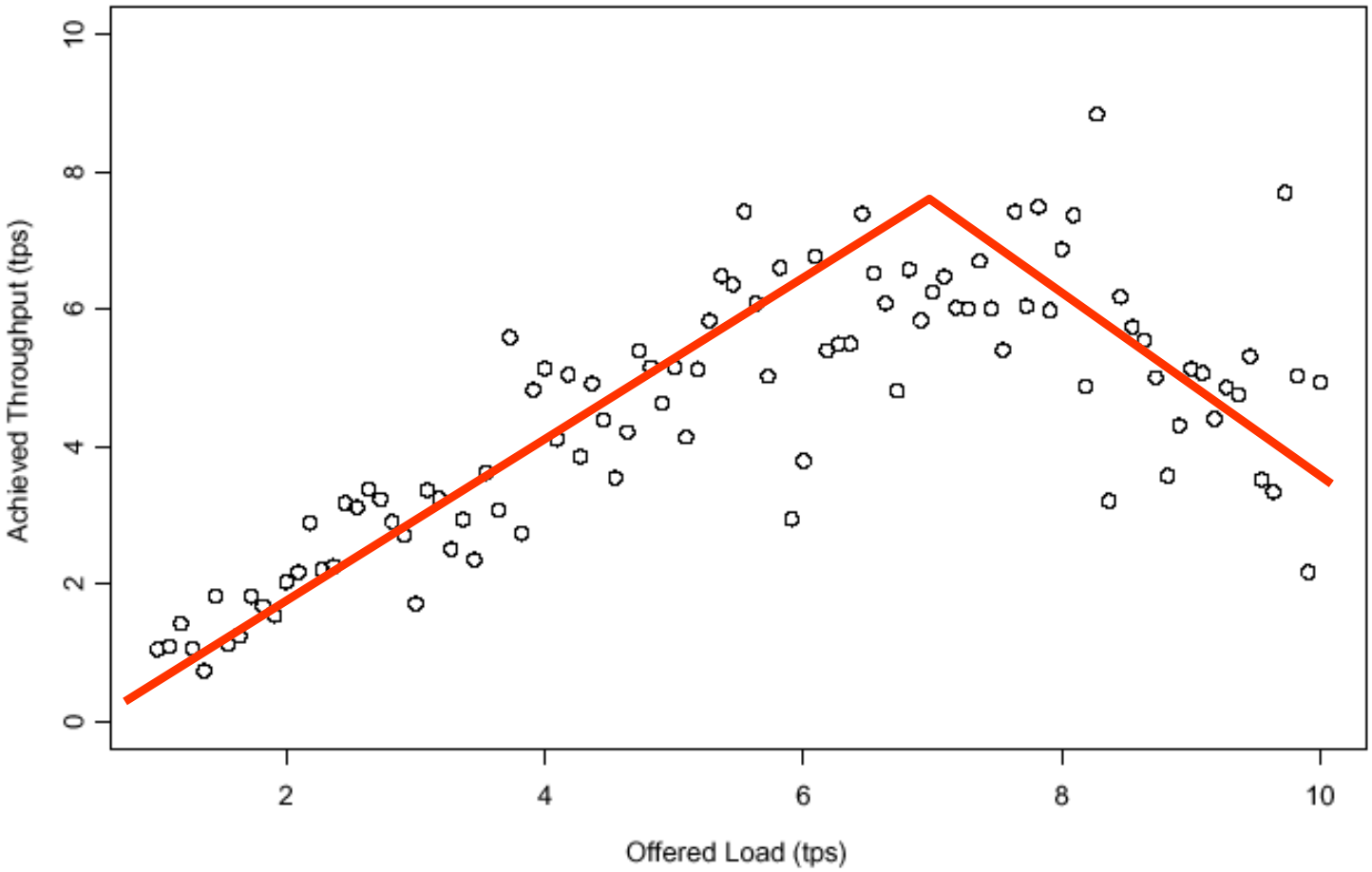
1. *The $p \times p$ matrix $(X^T X)$ is invertible*
2. *(Estimation) The maximum likelihood estimator of $\vec{\beta}$ is $\hat{\beta} = K\vec{y}$ with $K = (X^T X)^{-1} X^T$*
3. *(Standardized Residuals) Define the ith residual as $e_i = \left(\vec{y} - X\hat{\beta}\right)_i$. The residuals are zero-mean gaussian but are correlated, with covariance matrix $\sigma^2(Id_I - H)$, where $H = X(X^T X)^{-1} X^T$.*

   Let $s^2 = \frac{1}{I-p}\|e\|^2 = \frac{1}{I-p}\sum_i e_i^2$ (rescaled sum of squared residuals). s^2 is an unbiased estimator of σ^2.

   The standardized residuals defined by $r_i := \frac{e_i}{s\sqrt{1-H_{i,i}}}$ have unit variance and $r_i \sim t_{I-p-1}$. This can be used to test the model by checking that r_i are approximately normal with unit variance.
4. *(Confidence Intervals) Let $G = \left(X^T X\right)^{-1} = KK^T$; the distribution of $\hat{\beta}$ is gaussian with mean $\vec{\beta}$ and covariance matrix $\sigma^2 G$, and $\hat{\beta}$ is independent of e.*

   In particular, assume we want a confidence interval for a (non-random) linear combination of the parameters $\gamma = \sum_{j=1}^p u_j \beta_j$; $\hat{\gamma} = \sum_j u_j \hat{\beta}_j$ is our estimator of γ. Let $g = \sum_{j,k} u_j G_{j,k} u_k = \sum_k \left(\sum_j u_j K_{j,k}\right)^2$ (g is called the variance bias). Then $\frac{\hat{\gamma}-\gamma}{\sqrt{g}s} \sim t_{I-p}$. This can be used to obtain a confidence interval for γ.

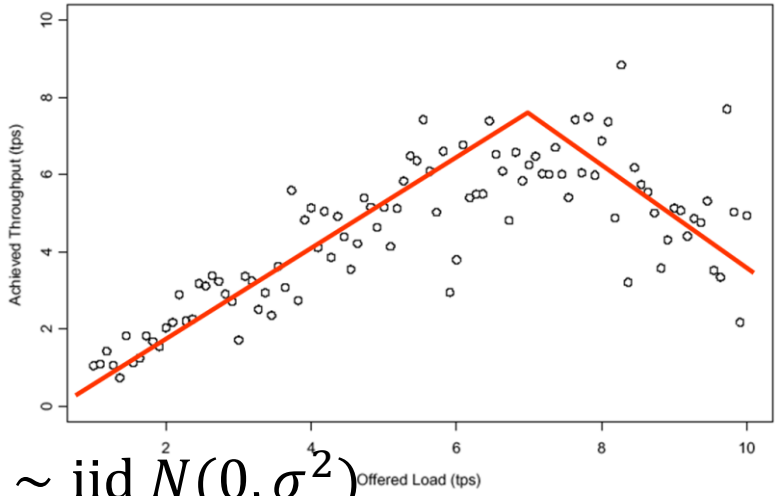
Example: Joe's shop



$$Y_i = (a + bx_i)1_{x_i \le \xi} + (c + dx_i)1_{x_i > \xi} + \epsilon_i, \epsilon_i \sim \text{iid } N(0, \sigma^2)$$

with $a + b\xi = c + d\xi$

Is this a linear regression model ?

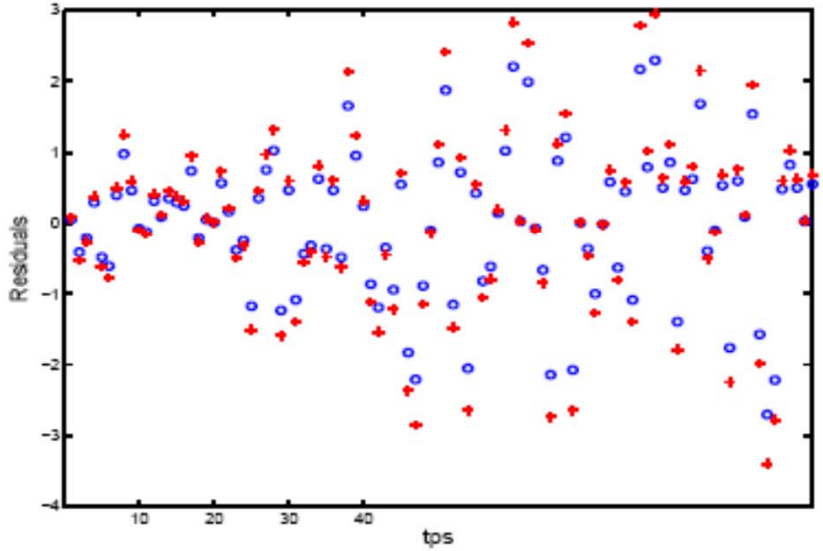


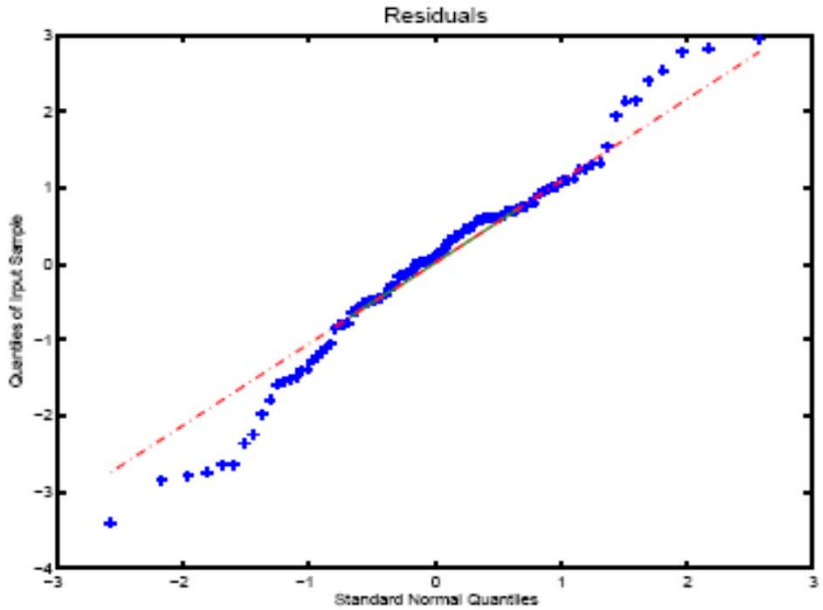
$$Y_i = (a + bx_i)1_{x_i \leq \xi} + (c + dx_i)1_{x_i > \xi} + \epsilon_i, \epsilon_i \sim \text{iid } N(0, \sigma^2)$$
with $a + b\xi = c + d\xi$

A. Yes
B. Yes if we assume ξ is known
C. No, even if ξ is not known
D. I don't know

Validate the Assumptions with Residuals

We also computed the residuals e_i (crosses) and standardized residuals r_i (circles). There is little difference between both types of residuals. They appear reasonably normal, but one might criticize the model in that the variance appears smaller for smaller values of x. The normal qqplot of the residuals also shows approximate normality (the qqplot of standardized residuals is similar and is not shown).





Non-linear case: use the likelihood function

2. *Let K be the square matrix of second derivatives (assumed to exist), defined by*

$$K_{j,k} = \frac{1}{\sigma^2} \sum_i \frac{\partial f_i}{\partial \beta_j} \frac{\partial f_i}{\partial \beta_k}$$

If K is invertible and if the number I of data points is large, $\hat{\beta} - \vec{\beta}$ is approximately gaussian with 0 mean and covariance matrix K^{-1}.

Alternatively, for large I, an approximate confidence set at level γ for the jth component β_j of $\vec{\beta}$ is implicitly defined by

$$-2I \ln(\hat{\sigma}) + 2I \ln\left(\hat{\sigma}(\hat{\beta}_1, ..., \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} ... \hat{\beta}_p)\right) \geq \xi_1$$

where $\hat{\sigma}^2(\vec{\beta}) = \frac{1}{I} \sum_i \left(y_i - f_i(\vec{\beta})\right)^2$ and ξ_1 is the γ quantile of the χ^2 distribution with 1 degree of freedom (for example, for $\gamma = 0.95$, $\xi_1 = 3.92$).

EXAMPLE 3.8: JOE'S SHOP - BEYOND THE LINEAR CASE - ESTIMATION OF ξ. In Example 3.6 we assumed that the value ξ after which there is congestion collapse is known in advance. Now we relax this assumption. Our model is now the same as Eq.(3.9), except that ξ is also now a parameter to be estimated.

To do this, we apply maximum likelihood estimation. We have to maximize the log-likelihood $l_{\vec{y}}(a, b, d, \xi, \sigma)$, where $\vec{y}$, the data, is fixed. For a fixed ξ, we know the value of (a, b, d, σ) that achieves the maximum, as we have a linear regression model. We plot the value of this maximum versus ξ (Figure 3.2) and numerically find the maximum. It is for $\xi = 77$.

To find a confidence interval, we use the asymptotic result in Theorem B.3.1. It says that a 95% confidence interval is obtained by solving $l(\hat{\xi}) - l(\xi) \leq 1.9207$, which gives $\xi \in [73, 80]$.

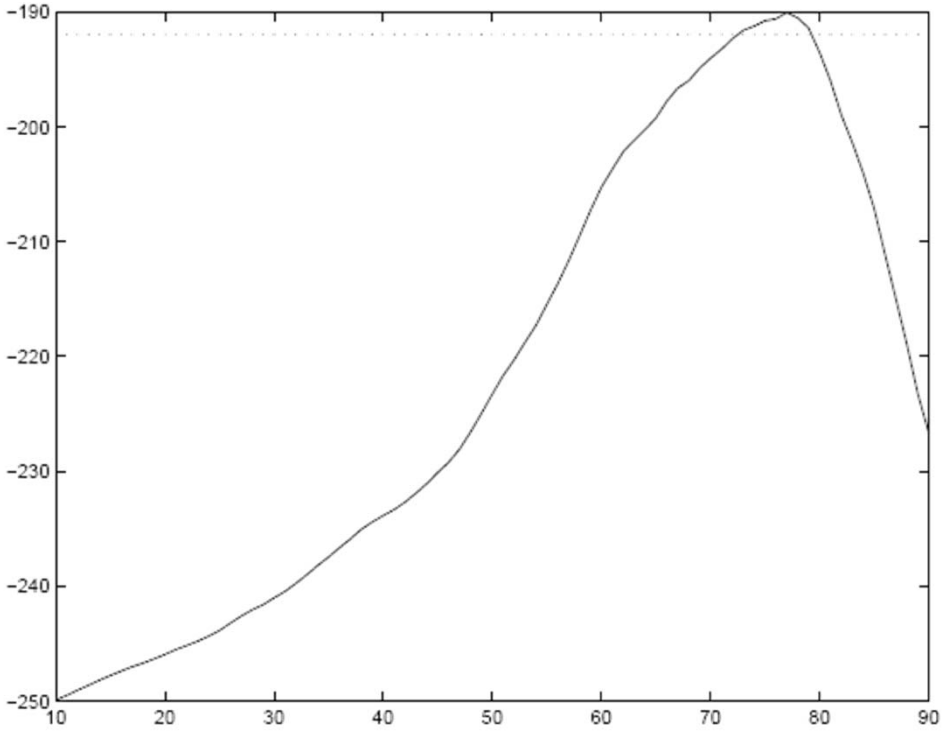


Figure 3.2: Log likelihood for Joes' shop as a function of ξ.

3. ℓ^1- Norm Minimization

Laplace Noise

Recall that the pdf of the $\exp(\lambda)$ distribution is $f(x) = \lambda e^{-\lambda x} 1_{x \geq 0}$ and its CDF is $F(c) = 1 - e^{-\lambda c}$

Consider the following sampling method:

Draw $Y \sim \exp(\lambda)$

With probability 0.5, let $X = Y$,
with probability 0.5 let $X = -Y$

The output X is called Laplace Noise

What is its pdf ?

The PDF of Laplace noise is …

A. $f\ (x)\ =\ \frac{\lambda}{2}\ e^{\lambda x}\ +\frac{\lambda}{2}\ e^{-\lambda x}$

B. $f(x) = \frac{\lambda}{2} e^{-\lambda|x|}$

C. $f(x) = \frac{\lambda}{2} e^{-\lambda|x|} + \frac{\lambda}{2} e^{-\lambda x}$

D. $f(x) = \frac{\lambda}{2} e^{-\lambda|x|} + \frac{\lambda}{2} e^{\lambda|x|}$

E. I don't know

ℓ^1-Norm Minimization = Laplace Noise

Assume model (homoscedasticity)

$$Y_i = f_i(\vec{\beta}) + \epsilon_i \text{ with } \epsilon_i \text{iid} \sim \text{Laplace}(\lambda)$$

THEOREM 3.2 (Least Deviation). *For the model in Eq.(3.7), the maximum likelihood estimator of the parameter* $(\vec{\beta}, \lambda)$ *is given by:*

1. $\hat{\beta} = \arg\min_{\vec{\beta}} \sum_i \left| y_i - f_i(\vec{\beta}) \right|$
2. $\frac{1}{\hat{\lambda}} = \frac{1}{I} \sum_i \left| y_i - f_i(\hat{\beta}) \right|$

i.e. the signal processing method with ℓ^1 score is the same as the statistical method with iid Laplace noise.

The ℓ^1 norm of a sequence $z = (z_1, ..., z_n)$ is $\|z\|_1 = \sum_{i=1}^{n} |z_i|$

LS and ℓ^1 score functions

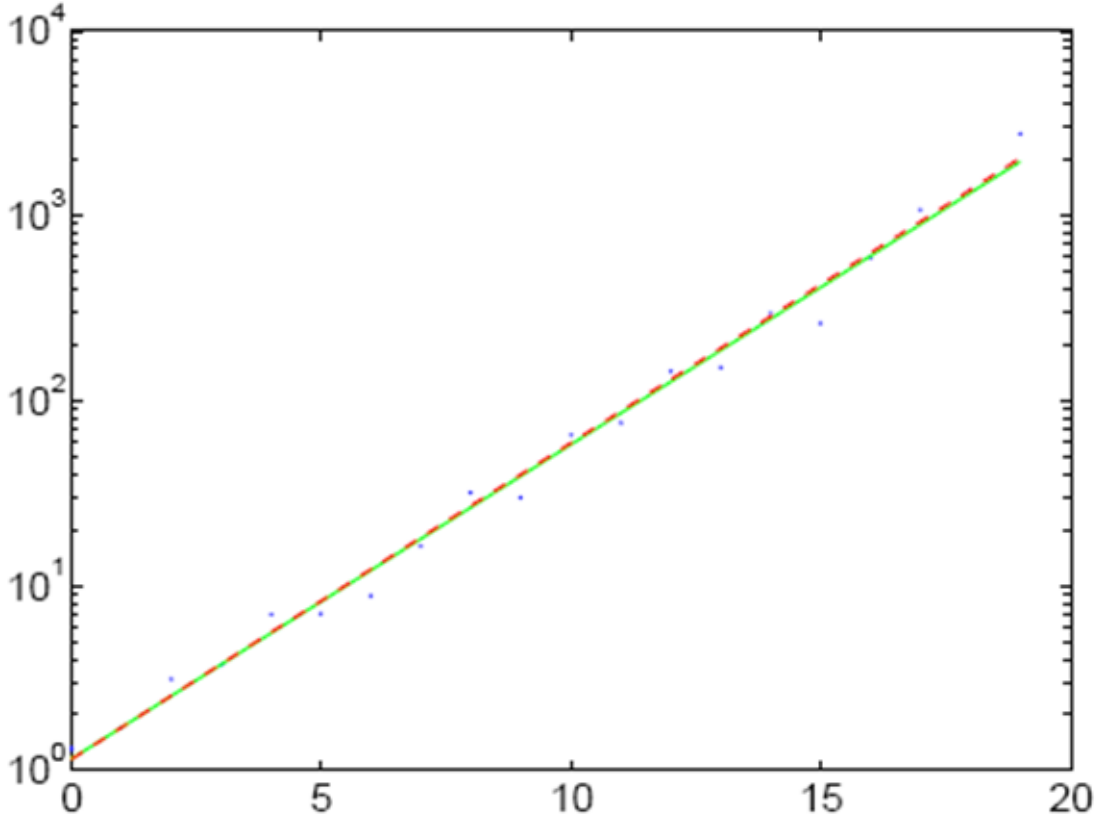
Given a model to be fitted

$$Y_i = f_i(\theta) + \text{noise}$$

we have the equivalences:

Score	**Statistical model**
Least Square: $\text{score} = \sum_i (\text{noise}_i)^2$	$\text{noise} \sim \text{iid } N(0, \sigma^2)$
Weighted Least Square: $\text{score} = \sum_i (w_i \times \text{noise}_i)^2$	$\text{noise}_i \sim \text{iid } N\left(0, \frac{\sigma^2}{w_i^2}\right)$
ℓ^1 score $= \sum_i \lvert\text{noise}_i\rvert$	$\text{noise} \sim \text{iid Laplace}(\lambda)$
ℓ^1 score $= \sum_i w_i \lvert\text{noise}_i\rvert$	$\text{noise}_i \sim \text{iid Laplace}(\lambda w_i)$

LS versus ℓ^1- norm minimization



Virus propagation example: both fits give the same result in log-scale

ℓ^1- norm minimization is more robust to « Outliers »

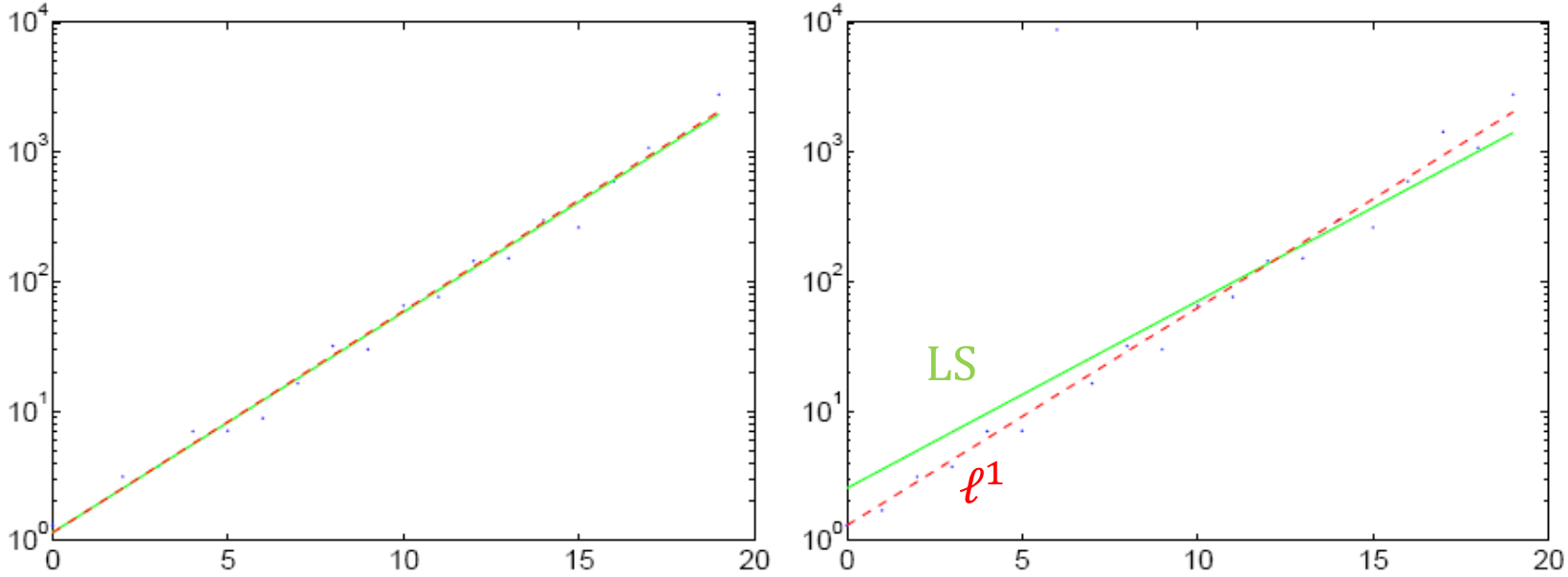


Figure 3.1: Fitting an exponential growth model to the data in Example 3.1, showing the fits obtained with least square (plain) and with ℓ^1 norm minimization (dashed) . First panel: original data; both fits are the same; Second panel: data corrupted by one outlier; the fit with ℓ^1 norm minimization is not affected, whereas the least square fit is.

	Least Square		ℓ^1 norm minimization	
	rate	prediction	rate	prediction
no outlier	0.3914	30300	0.3938	32300
with one outlier	0.3325	14500	0.3868	30500

To understand the difference between LS and ℓ^1 consider the simple example $y_i = m + \text{noise}$ i.e. fit a cloud of data to a single value

Least Square

Model:

$$y_i = m + \epsilon_i, \epsilon_i \text{ iid} \sim N(0, \sigma^2)$$

What is *m* ?

Confidence interval ?

L1 Norm Minimization

Model:

$$y_i = m + \epsilon_i, \epsilon_i \text{ iid} \sim \text{Laplace}(\lambda)$$

What is *m* ?

Confidence interval ?

What is the LS fit of $\boldsymbol{m}$ in $\mathbf{y}_i = \boldsymbol{m} + \mathbf{noise}$?

A. $m =$ median of y_i

B. $m = \bar{y}$ where $\bar{y} =$ mean of y_i

C. $m = \bar{y} + \frac{\hat{\sigma}^2}{\bar{y}}$ where $\hat{\sigma}^2 = \frac{1}{I}\sum_i^I (y_i - \bar{y})^2$

D. $m = \bar{y} + \frac{\hat{\sigma}^2}{\bar{y}}$ where $\hat{\sigma}^2 = \frac{1}{I-1}\sum_i^I (y_i - \bar{y})^2$

E. I don't know

What is the ℓ^1 fit of $\boldsymbol{m}$ in $\mathbf{y}_i = \boldsymbol{m} + \mathbf{noise}$?

A. $m = \bar{y}$ where $\bar{y} =$ mean of y_i

B. $m =$ median of y_i

C. $m = \bar{y} + \frac{\hat{\sigma}^2}{\bar{y}}$ where $\hat{\sigma}^2 = \frac{1}{I}\sum_{i=1}^{I}(y_i - \bar{y})^2$

D. $m = \bar{y} + \frac{\hat{\sigma}^2}{\bar{y}}$ where $\hat{\sigma}^2 = \frac{1}{I-1}\sum_{i=1}^{I}(y_i - \bar{y})^2$

E. I don't know

Linear Regression with ℓ^1 norm minimization

= ℓ^1 norm minimization + linear dependency on parameter

More robust than LS linear regression

Less traditional

DEFINITION 3.2 (Linear Regression Model with Laplace Noise).

$$Y_i = (X\vec{\beta})_i + \epsilon_i \text{ for } i = 1, \ldots, I \textit{ with } \epsilon_i \textit{ iid } \sim \textit{ Laplace } (\lambda)$$

This is convex programming

THEOREM 3.3.1. *Consider the model in Definition 3.2.1; let $\vec{y}$ be the $I \times 1$ column vector of the data. The maximum likelihood estimator of $\vec{\beta}$ is obtained by solving the linear program:*

$$\begin{aligned} \textit{minimize} \quad & \sum_{i=1}^{I} u_i \\ \textit{over} \quad & \vec{\beta} \in \mathbb{R}^p, u \in \mathbb{R}^I \\ \textit{subject to the constraints} \quad & u_i \geq y_i - \left(X\vec{\beta}\right)_i \\ & u_i \geq -y_i + \left(X\vec{\beta}\right)_i \end{aligned}$$

The maximum likelihood of the noise parameter λ is $\left(\frac{1}{I}\sum_{i=1}^{I}\left|y_i - \left(X\vec{\beta}\right)_i\right|\right)^{-1}$.

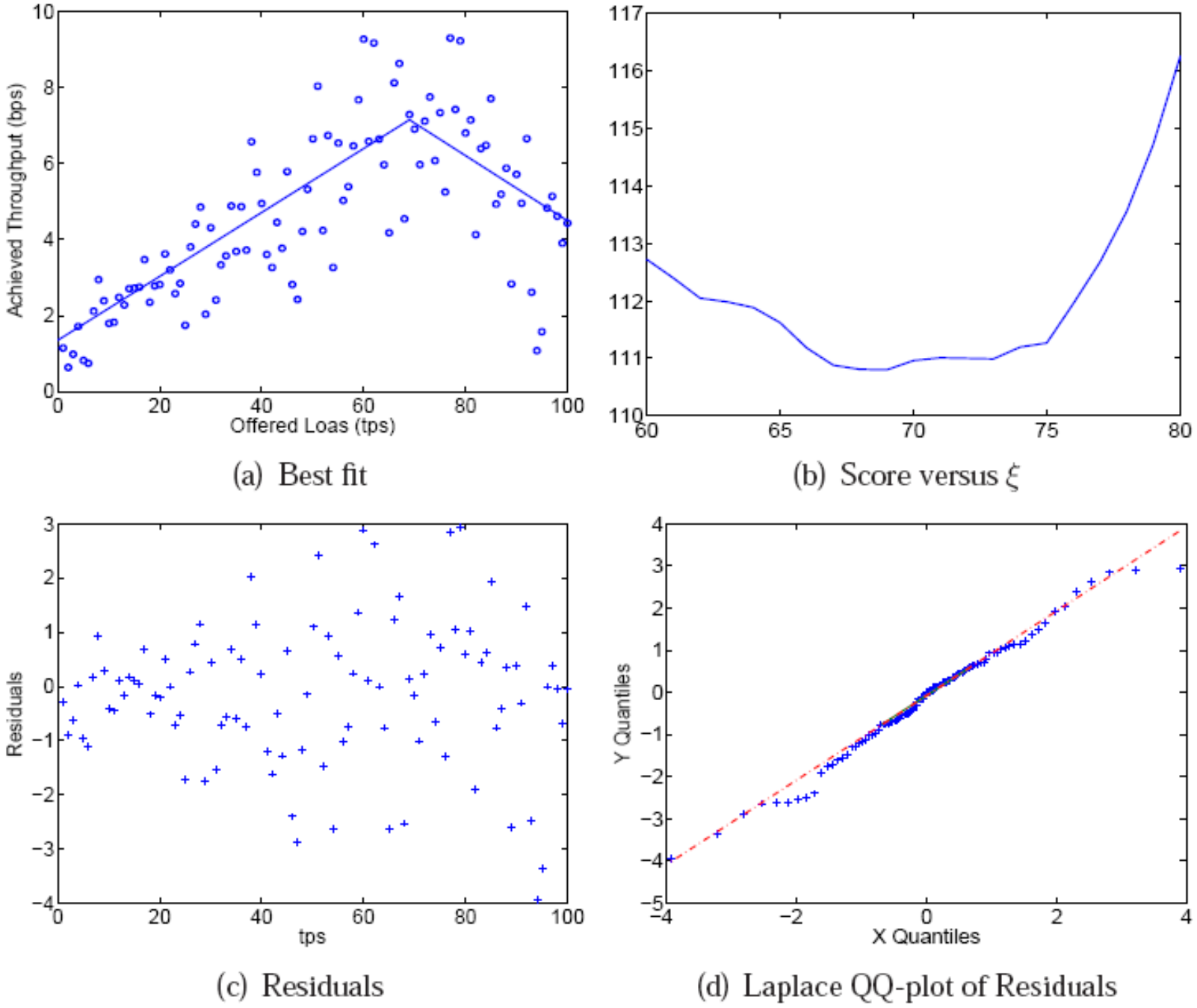
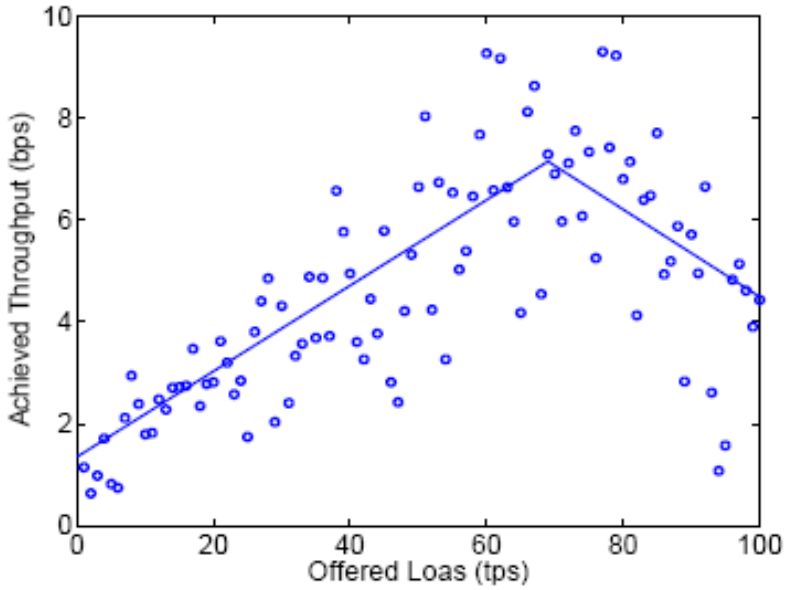


Figure 3.3: Modelling congestion collapse in Joe's shop with a piecewise linear function and ℓ^1 norm minimization of the errors.

Confidence Intervals

No closed form

We can use an approximate solution based on simulation ("Bootstrap")



a	1.32 ± 0.675
b	0.0791 ± 0.0149
c	11.7 ± 3.24
d	-0.0685 ± 0.0398

4. Bootstrap : Computation of Confidence Interval

We have a parametric model with parameters β, λ and want to find a confidence interval $[U, V]$ for, say, β_1 ; let $\hat{\beta}_1$ be our estimator of β_1

Recall that U, V is a function of the data (hence is random); we want to have $P_{\beta,\lambda}(U \leq \beta_1 \leq V) \geq 0.95$ for any β, λ

Assume someone tells us the true value of β, λ; in theory, we can compute the quantiles of the distribution of $\hat{\beta}_1$; let

$U_{\beta,\lambda} = 2.5\%$ quantile of $\hat{\beta}_1$ and $V_{\beta,\lambda} = 97.5\%$ quantile of $\hat{\beta}_1$

We have

$$P_{\beta,\lambda}\left(U_{\beta,\lambda} \leq \hat{\beta}_1 \leq V_{\beta,\lambda}\right) = 0.95$$

We have computed an estimator $\hat{\beta}, \hat{\lambda}$; we take as approximate confidence interval $U = U_{\hat{\beta},\hat{\lambda}}$ and $V = V_{\hat{\beta},\hat{\lambda}}$

Parametric Bootstrap Computation of CI (cont'd)

Let $U_{\beta,\lambda} = 2.5\%$ quantile of $\hat{\beta}_1$ and $V_{\beta,\lambda} = 97.5\%$ quantile of $\hat{\beta}_1$

We take as approximate confidence interval $U = U_{\hat{\beta},\hat{\lambda}}$ and $V = V_{\hat{\beta},\hat{\lambda}}$

How can we practically compute $U_{\beta,\lambda}$, $V_{\beta,\lambda}$?

By Monte-Carlo simulation:

Assume you know β, λ

Do $r = 1{:}\,R$

simulate Y_i , $i = 1 \ldots I$ from the model

obtain an estimate $\hat{\beta}_i^r$

end

$U_{\beta,\lambda} = 2.5\%$ quantile of $\hat{\beta}_i^r$, $V_{\beta,\lambda} = 97.5\%$ quantile of $\hat{\beta}_i^r$

e.g. with $R = 999, U_{\beta,\lambda} = \hat{\beta}_i^{(25)}$ and $V_{\beta,\lambda} = \hat{\beta}_i^{(975)}$

Do the above with $(\beta, \lambda) = (\hat{\beta}, \hat{\lambda})$ – this is called *Parametric Bootstrap*

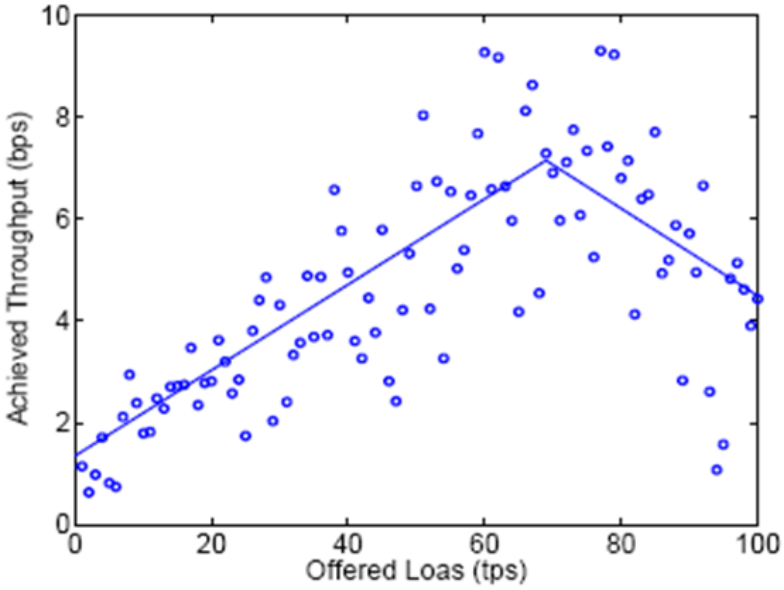
Example

We have obtained $\hat{\lambda} = 1.0055$

For $\mathrm{r} = 1{:}999$ we draw I residuals $\sim \mathrm{Laplace}(1.0055)$ and simulate n artificial data points $y_i^r, i = 1 \ldots I$ from the fitted model

For each replay experiment we estimate a, b, c, d and obtain $R = 999$ estimates $\hat{a}^r, \hat{b}^r, \hat{c}^r, \hat{d}^r$

A confidence interval for d is $\left[\hat{d}^{(25)}, \hat{d}^{(975)}\right]$



a	1.49 ± 0.601
b	0.079 ± 0.0143
c	11.2 ± 2.96
d	-0.062 ± 0.0368

Bootstrap with Resample From Residuals

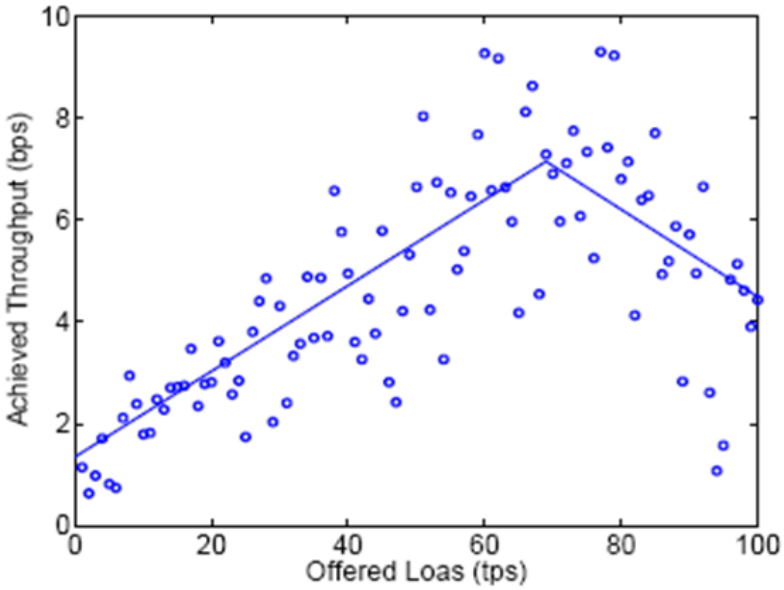
Instead of simulating the noise from the model (by drawing from Laplace(λ) assume we draw I noise samples from the residuals *with replacement*

This method is called *Bootstrap with Resampling from Residuals*

Algorithm 2 The Bootstrap with Re-Sampling From Residuals. The goal is to compute a confidence interval for some function $\varphi(\vec{\beta})$ of the parameter of the model in Definition 3.2.1. r_0 is the algorithm's accuracy parameter.

1: $R = \lceil 2\, r_0/(1-\gamma) \rceil - 1$ ▷ For example $r_0 = 25$, $\gamma = 0.95$, $R = 999$
2: estimate $\vec{\beta}$ using Theorem 3.3.1; obtain $\hat{\beta}$
3: compute the residuals $e_i = y_i - \left(X\hat{\beta}\right)_i$
4: **for** $r = 1 : R$ **do** ▷ Re-sample from residuals
5: draw I numbers with replacement from the list $(e_1, ..., e_I)$ and call them $E_1^r, ..., E_I^r$
6: generate the bootstrap replicate $Y_1^r, ..., Y_I^r$ from the estimated model:
7: $Y_i^r = \left(X\hat{\beta}\right)_i + E_i^r$ for $i = 1...I$
8: re-estimate $\vec{\beta}$, using Y_i^r as data, using Theorem 3.3.1; obtain $\vec{\beta}^r$
9: **end for**
10: $\left(\varphi_{(1)}, ..., \varphi_{(R)}\right) = \text{sort}\left(\varphi(\vec{\beta}^1), ..., \varphi(\vec{\beta}^R)\right)$
11: confidence interval for $\varphi(\vec{\beta})$ is $[\varphi_{(r_0)}\ ;\ \varphi_{(R+1-r_0)}]$

$\xi = 69$



CI with Parametric Bootstrap, 999 replays

a	1.49 ± 0.601
b	0.079 ± 0.0143
c	11.2 ± 2.96
d	-0.062 ± 0.0368

CI with Bootstrap, Resampling from Residuals, 999 replays

a	1.32 ± 0.675
b	0.0791 ± 0.0149
c	11.7 ± 3.24
d	-0.0685 ± 0.0398

Other Uses of the Bootstrap

The idea of the bootstrap is to use the data itself as estimate of the unknown distribution of the data

Can be used to obtain confidence or predictions intervals in a very simple way --- tends to underestimate

Example: CI for the mean of a data set y_i

the model is $Y_i \sim \text{iid } F()$ and the problem is to estimate the mean of $F(\quad)$

with the bootstrap we replace the unknown $F()$ by the empirical distribution of the data itself, i.e. the distribution that puts probability $\frac{1}{I}$ at every y_i , $i = 1 \ldots I$

Which algorithm is a correct implementation of the bootstrap for a 95% CI of the mean of $y_1, \ldots, y_I$?

A. A
B. B
C. Both
D. None
E. I don't know

Algorithm A

for $r = 1{:}\,999$
  for $n = 1{:}\,I$
    draw one integer
      $K \sim \text{unif}\{1, \ldots, I\}$
    $x_{n,r} = y_K$
  end
end
for $r = 1{:}\,999$
  $m_r = \text{mean}(x_{1,r}, \ldots, x_{I,r})$
end
CI $= [m_{(25)}, m_{(975)}]$

Algorithm B

for $r = 1{:}\,999$
  draw a random permutation
    σ of $\{1, \ldots, I\}$
    $x_{n,r} = y_{\sigma(n)}, n = 1 \ldots I$
  end
end
for $r = 1{:}\,999$
  $m_r = \text{mean}(x_{1,r}, \ldots, x_{I,r})$
end
CI $= [m_{(25)}, m_{(975)}]$

Example: Compiler Options, CI for mean