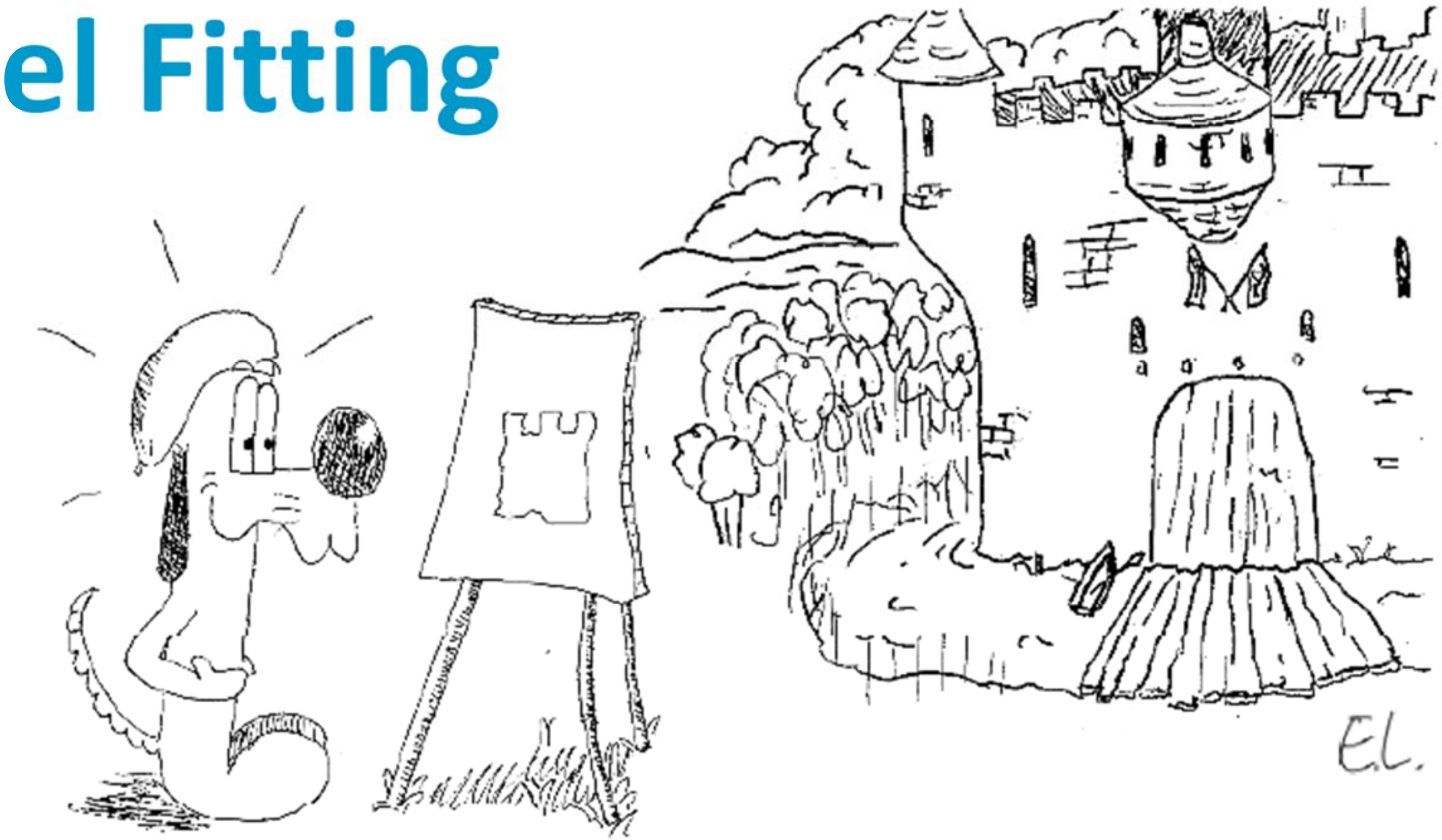
Model Fitting



1. What is model fitting ?
2. Linear Regression
3. Linear Regression with ℓ^1 norm
4. Heavy Tail

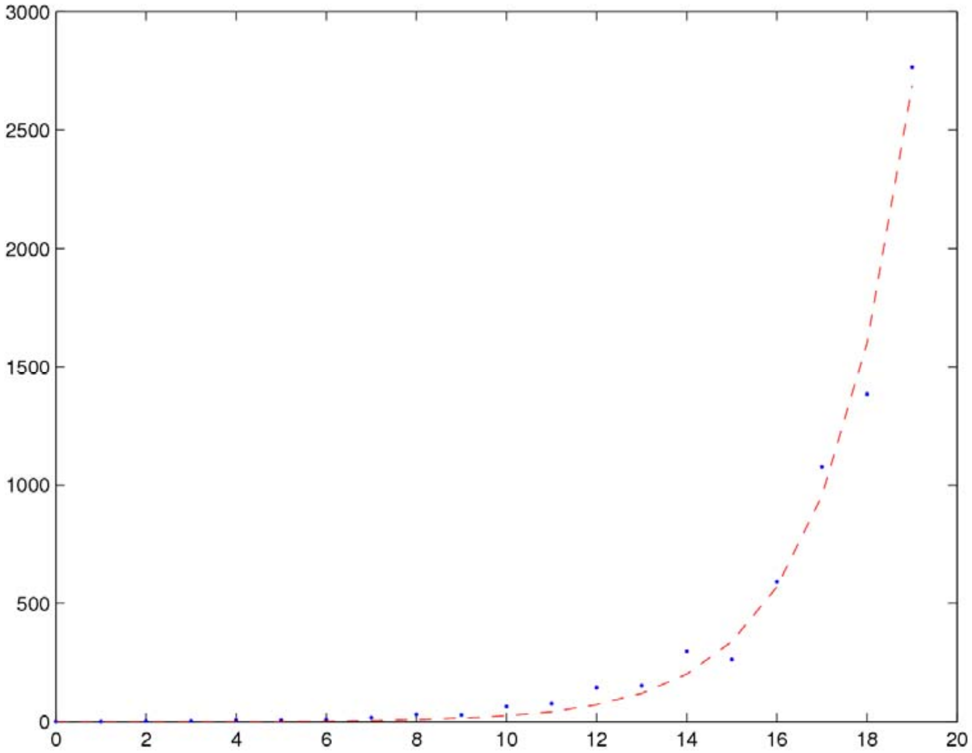
Virus Infection Data

- We would like to capture the growth of infected hosts

- Explanatory model: $\boldsymbol{y_i = f_i(\vec{\beta})}$
  - $\boldsymbol{y_i}$: collection of measured data
  - $\boldsymbol{i}$: index of measurement
  - $\boldsymbol{f_i}$: array of functions
  - $\boldsymbol{\beta_i}$: parameter we would like to obtain

- An exponential model seems appropriate

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

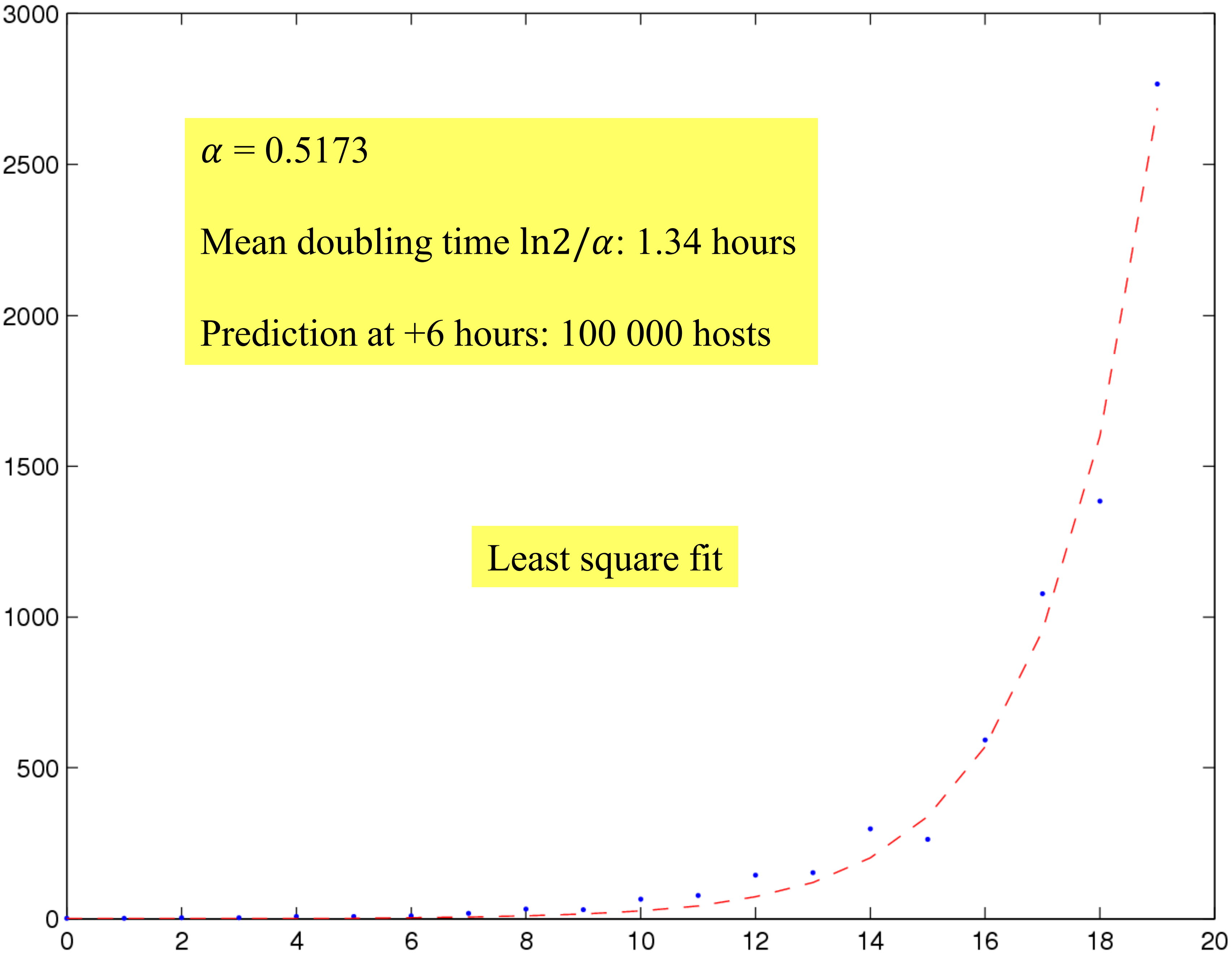


$$\boldsymbol{Y(t) = a e^{\alpha t}}$$

$$\vec{\beta} = (a, \alpha),\ f_i(\vec{\beta}) = f_i(a, \alpha) = a e^{\alpha t_i}$$

t_i: time of ith measurement

Least Square Fit of Virus Infection Data



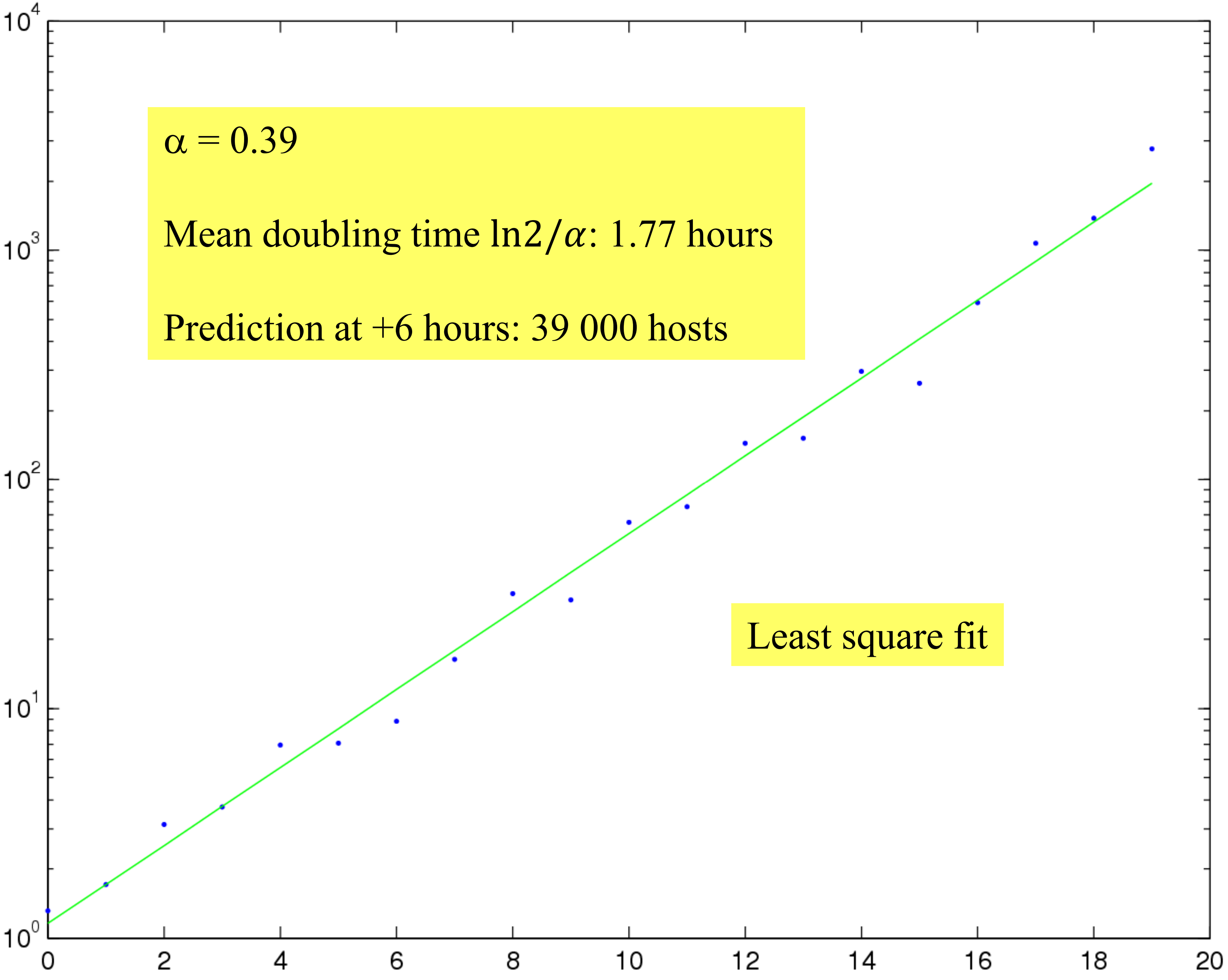
$\alpha = 0.5173$

Mean doubling time $\ln 2/\alpha$: 1.34 hours

Prediction at +6 hours: 100 000 hosts

Least square fit

Least Square Fit of Virus Infection Data In Log Scale



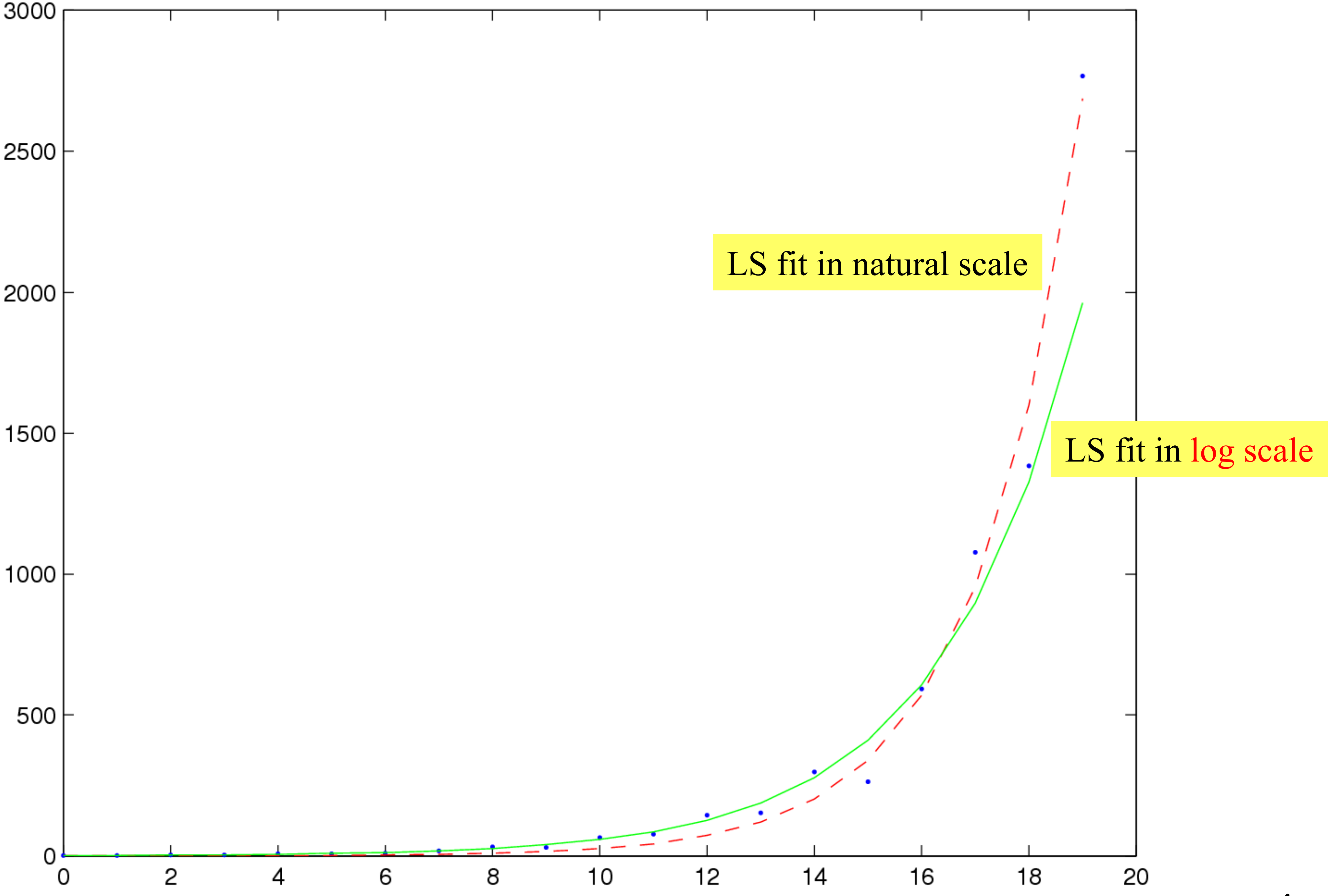
$\alpha = 0.39$

Mean doubling time $\ln 2/\alpha$: 1.77 hours

Prediction at +6 hours: 39 000 hosts

Least square fit

Compare the Two



Which Fitting Method should I use ?

- Which optimization criterion should I use?

- The answer is in a *statistical model.*
  - Model not only the interesting part, but also the **noise**

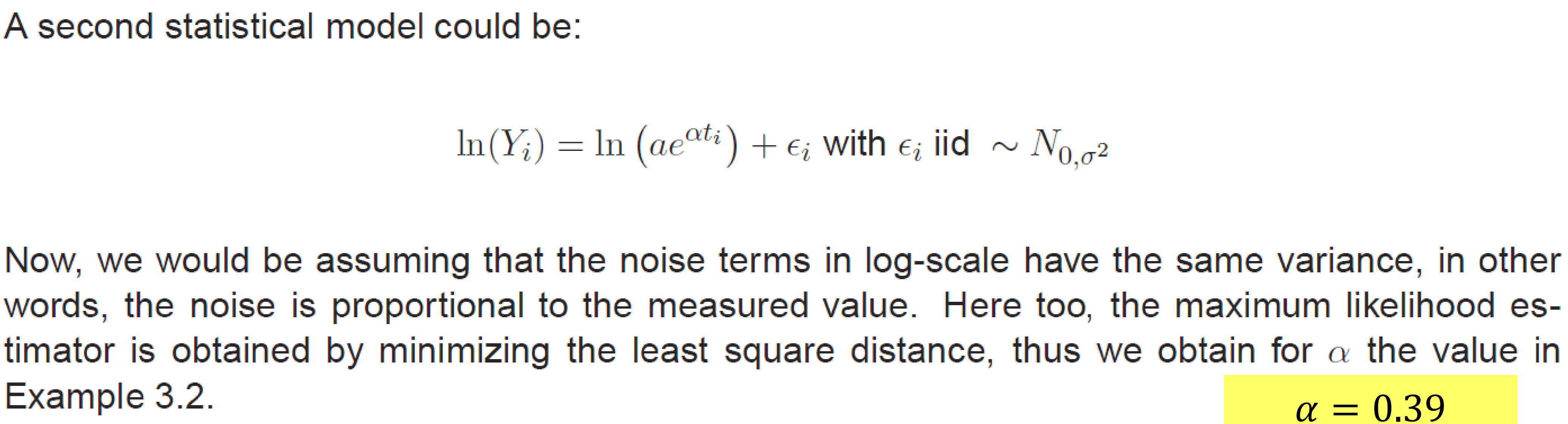
- For example

$$Y_i = a e^{\alpha t_i} + \epsilon_i \text{ with } \epsilon_i \text{ iid } \sim N_{0,\sigma^2}$$

The parameter is $\theta = (a, \alpha, \sigma)$.

We will see in Section 3.1.2 that the maximum likelihood estimator for this model is the one that minimizes the mean square distance. Thus, with this model, we obtain for α the value in Example 3.1.

$\alpha = 0.5137$
data in normal scale



A second statistical model could be:

$$\ln(Y_i) = \ln\left(ae^{\alpha t_i}\right) + \epsilon_i \text{ with } \epsilon_i \text{ iid } \sim N_{0,\sigma^2}$$

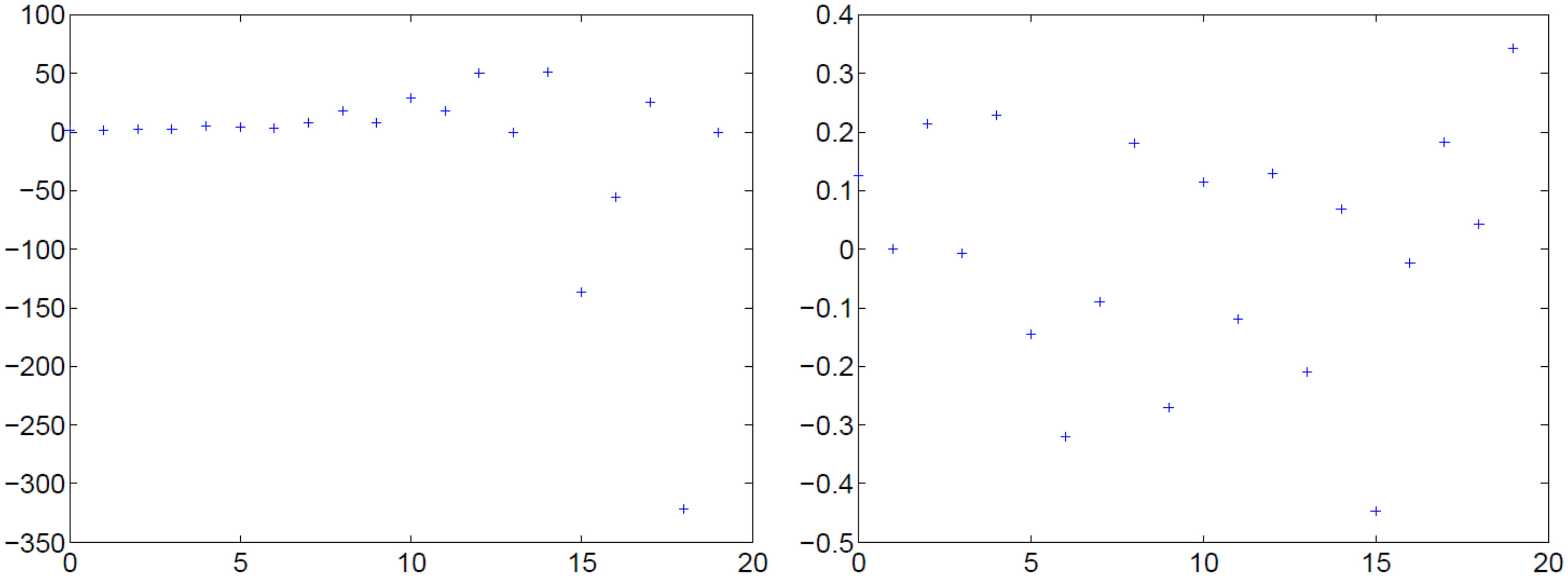
Now, we would be assuming that the noise terms in log-scale have the same variance, in other words, the noise is proportional to the measured value. Here too, the maximum likelihood estimator is obtained by minimizing the least square distance, thus we obtain for α the value in Example 3.2.

$\alpha = 0.39$
data in log scale

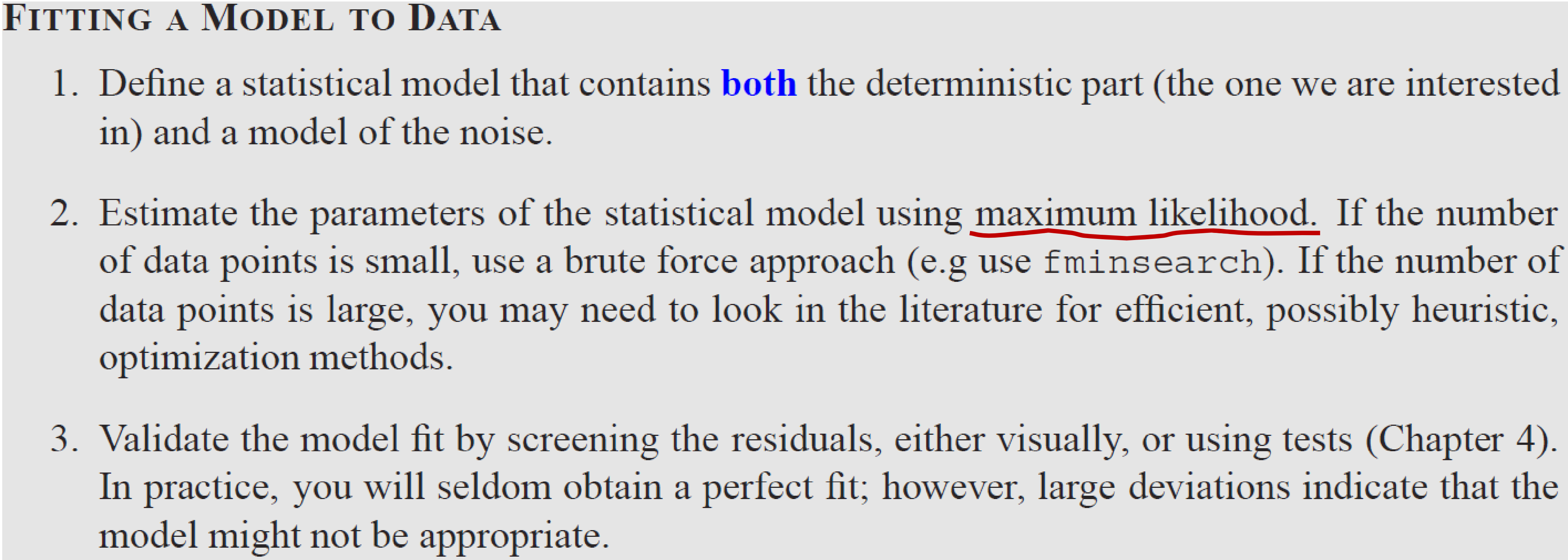
- How can I tell **which is correct** ?

Look at Residuals

We can validate either model by plotting the residuals:



We see clearly that the residual for the former model do not appear to be normally distributed, and the converse is true for the former model, which is the one we should adopt. Therefore, an acceptable fitting is obtained by minimizing least squares in log-scale.



FITTING A MODEL TO DATA

1. Define a statistical model that contains **both** the deterministic part (the one we are interested in) and a model of the noise.

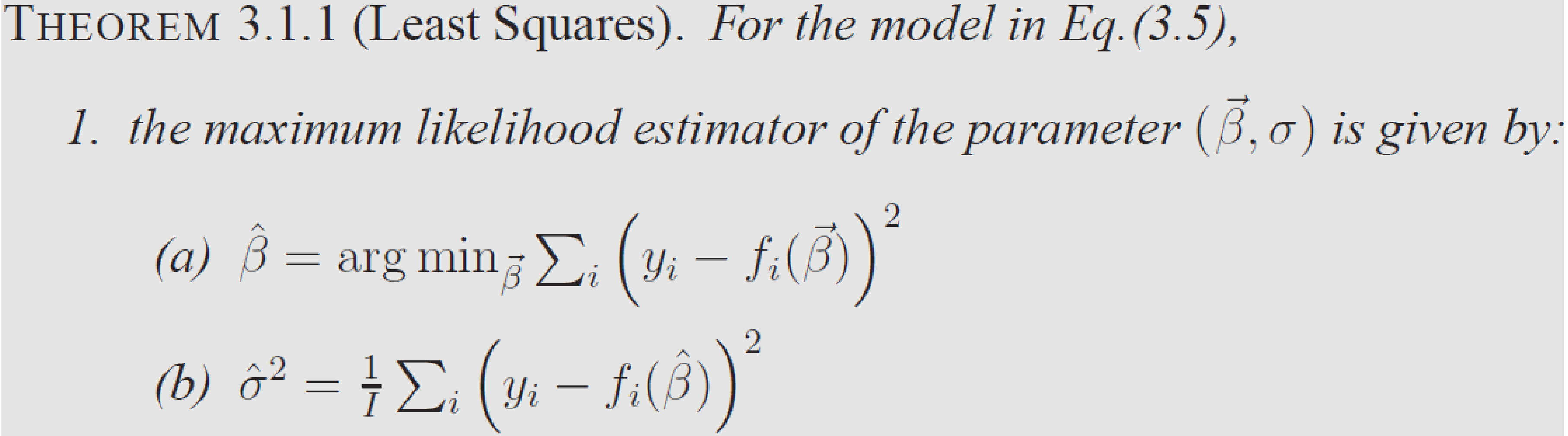
2. Estimate the parameters of the statistical model using maximum likelihood. If the number of data points is small, use a brute force approach (e.g use `fminsearch`). If the number of data points is large, you may need to look in the literature for efficient, possibly heuristic, optimization methods.

3. Validate the model fit by screening the residuals, either visually, or using tests (Chapter 4). In practice, you will seldom obtain a perfect fit; however, large deviations indicate that the model might not be appropriate.

Least Square Fit

- General model (homoscedasticity: **unknown same finite variance**)

$$Y_i = f_i(\vec{\beta}) + \epsilon_i \text{ for } i = 1, \ldots, I \text{ with } \underbrace{\epsilon_i \text{ iid } \sim N_{0,\sigma^2}}_{\textbf{This must be satisfied}} \qquad (3.5)$$



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

- The theorem says:

$$\text{minimize least squares} \overset{\text{equivalent}}{\Longleftrightarrow} \text{compute } \textbf{MLE} \text{ for this model}$$

- This is how we computed the estimates for the virus example.

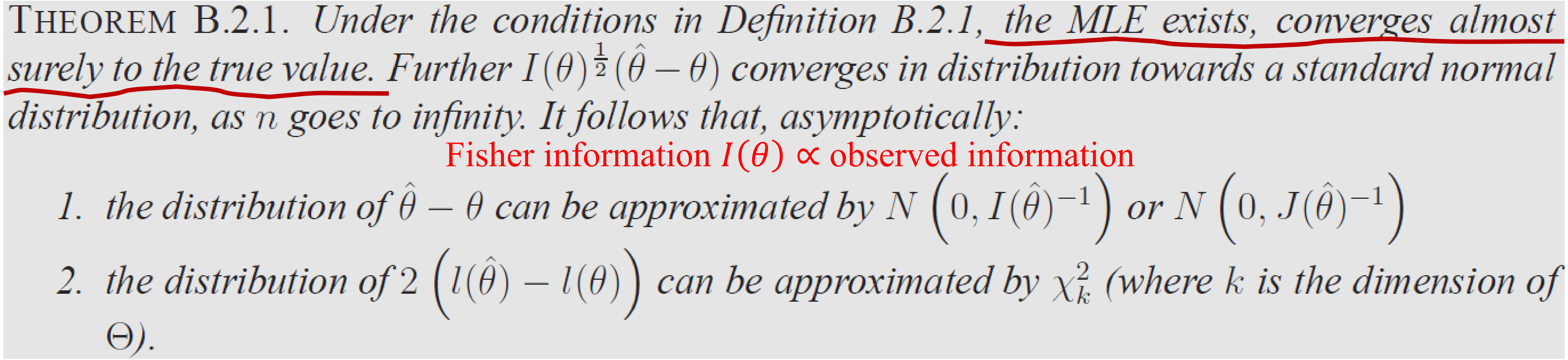
Maximum Likelihood Estimator

- What is **MLE**? Denote by $\boldsymbol{f(x)}$ the **pdf** (probability density function) of the **random variable** in the model. The MLE of the model is

The value of $\boldsymbol{\theta = \hat{\theta}}$ which maximizes $\boldsymbol{f(x_1, \dots, x_n|\theta)}$
(1) $\boldsymbol{\theta}$: the parameters to be estimated
(2) $\boldsymbol{\vec{x} = (x_1, \dots, x_n)}$: the available data

- That is because we have the following asymptotic convergence for MLE.

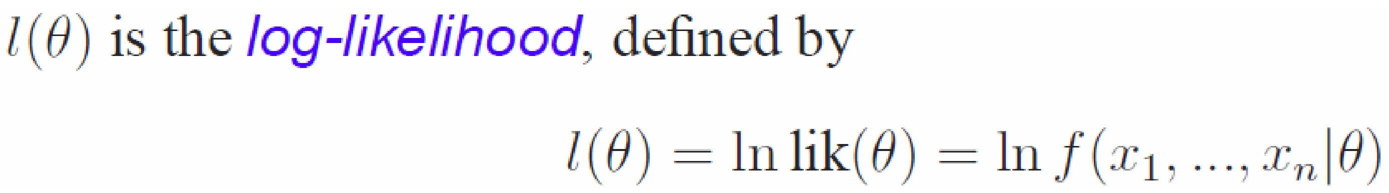
as the number of samples $\boldsymbol{n}$ goes to infinity



THEOREM B.2.1. *Under the conditions in Definition B.2.1, the MLE exists, converges almost surely to the true value. Further $I(\theta)^{\frac{1}{2}}(\hat{\theta} - \theta)$ converges in distribution towards a standard normal distribution, as n goes to infinity. It follows that, asymptotically:*

Fisher information $I(\theta) \propto$ observed information

1. *the distribution of $\hat{\theta} - \theta$ can be approximated by $N\left(0, I(\hat{\theta})^{-1}\right)$ or $N\left(0, J(\hat{\theta})^{-1}\right)$*
2. *the distribution of $2\left(l(\hat{\theta}) - l(\theta)\right)$ can be approximated by χ^2_k (where k is the dimension of Θ).*



$l(\theta)$ is the *log-likelihood*, defined by

$$l(\theta) = \ln \mathrm{lik}(\theta) = \ln f(x_1, \dots, x_n|\theta)$$

Maximum Likelihood Estimator

- **MLE** is ***dramatically*** simplified for one special case!

The value of $\theta = \hat{\theta}$ which maximizes $f(x_1, \dots, x_n|\theta)$
(1) θ: the parameters to be estimated
(2) $\vec{x} = (x_1, \dots, x_n)$: the available data

Maximum joint density at $\vec{x}$
over condition space θ:
Under which θ, $\vec{x}$ is most likely?

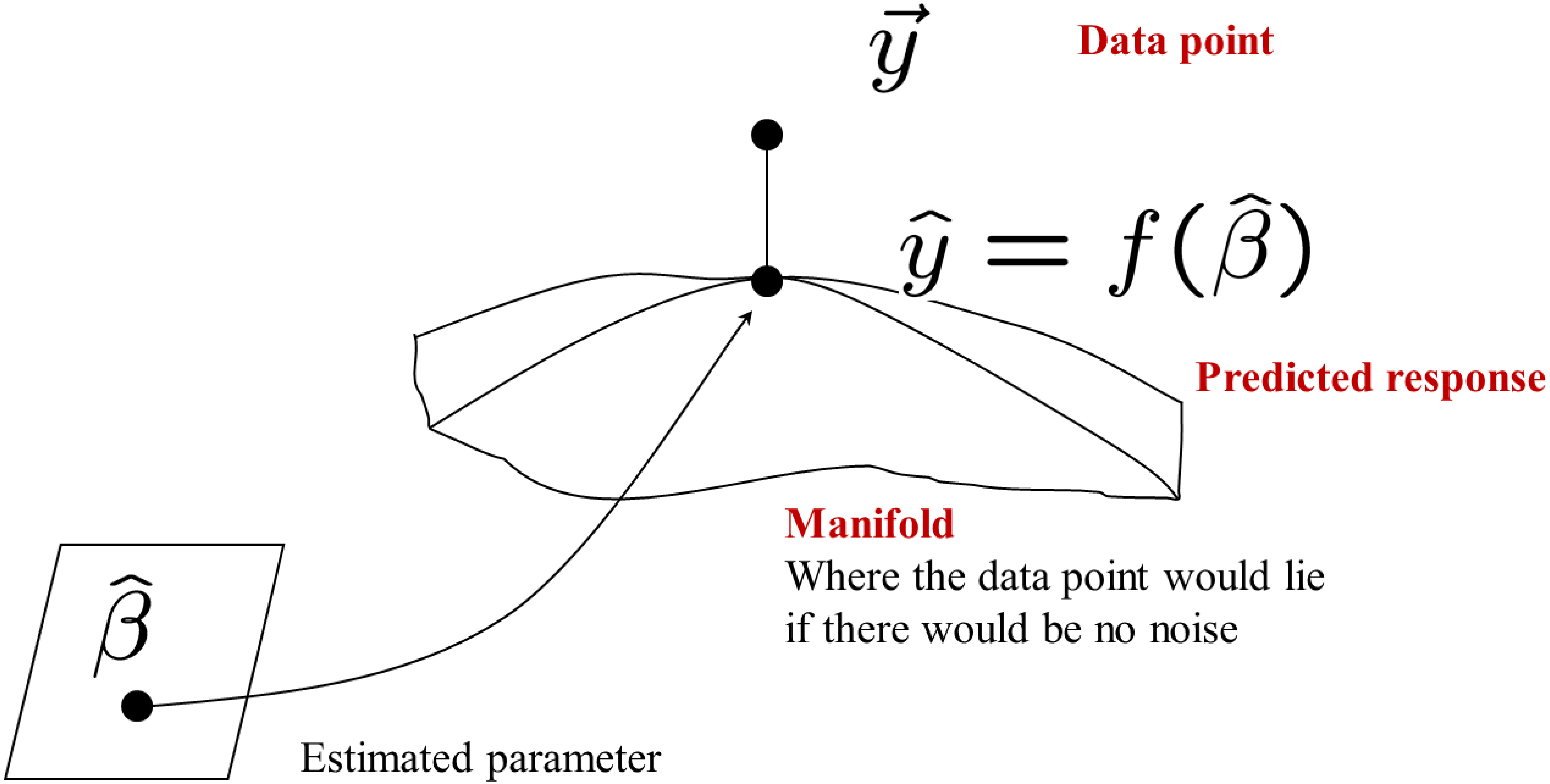
- For the Gaussian iid noise case, we have:

$$\begin{array}{c} n = I \\ \vec{x} = (\epsilon_1, \dots, \epsilon_n) \\ \theta = (\vec{\beta}, \sigma) \end{array} \Rightarrow f\left(\epsilon_1, \dots, \epsilon_n \middle| (\vec{\beta}, \sigma)\right) = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left(y_i - f_i(\vec{\beta})\right)^2}$$

$\forall\sigma$, the above expression is maximized over β when $\sum_{i=1}^{I}\left(y_i - f_i(\vec{\beta})\right)^2$ is minimized.
This completes the proof of Theorem 3.1.1.

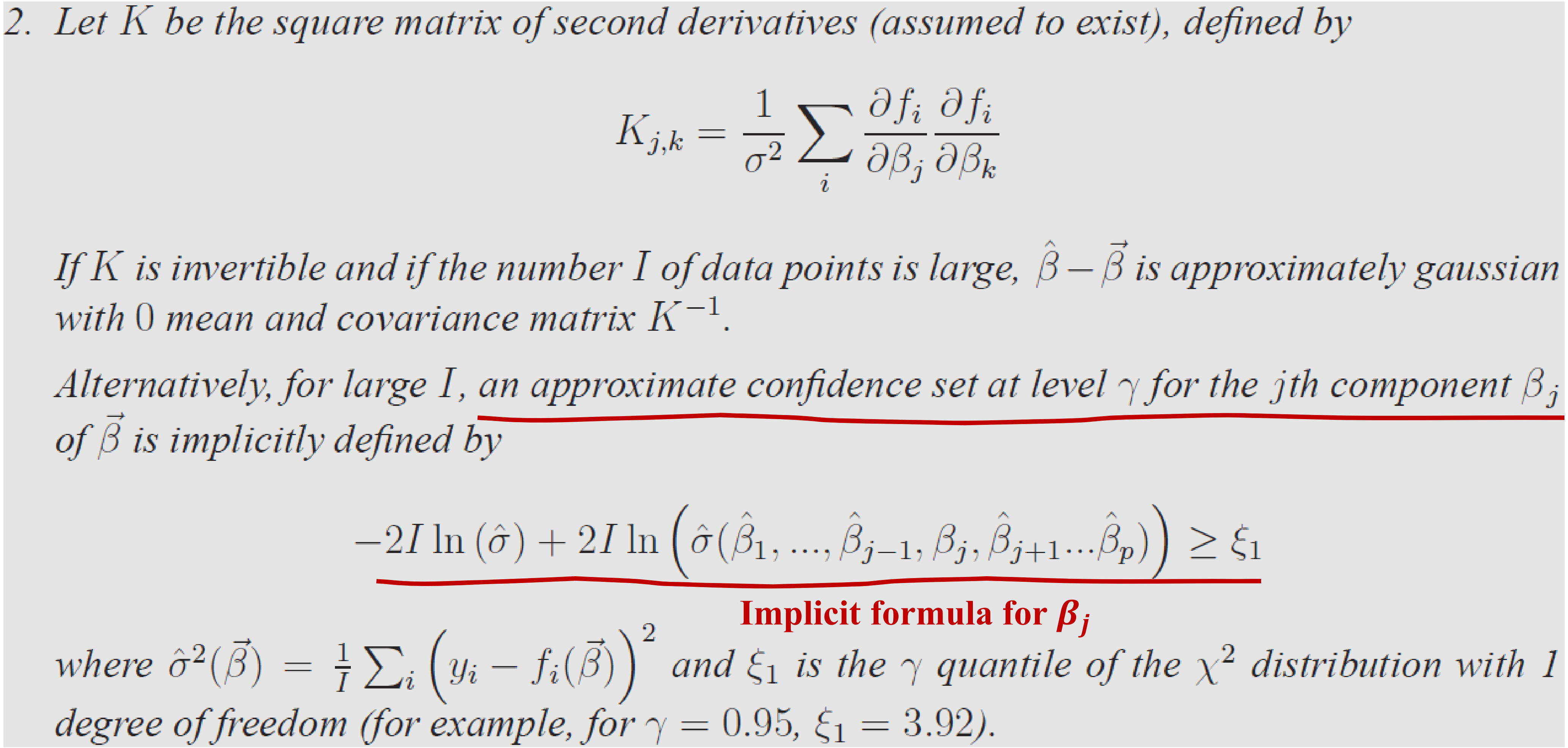
Maximum Likelihood Estimator ➔ Least Square Fit (or Mean Square Error)

Least Square as Projection


$\vec{y}$
Data point
$\hat{y} = f(\hat{\beta})$
Predicted response
Manifold
Where the data point would lie
if there would be no noise
$\hat{\beta}$
Estimated parameter

Least square projection (i.e., finding closest point) is robust to Gaussian noise.

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

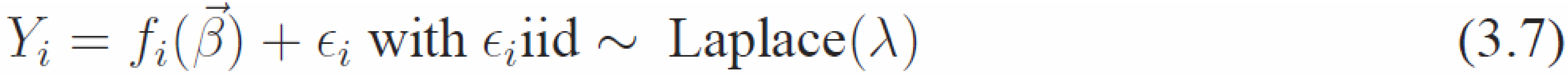
$$\underline{-2I \ln(\hat{\sigma}) + 2I \ln\left(\hat{\sigma}(\hat{\beta}_1, ..., \hat{\beta}_{j-1}, \beta_j, \hat{\beta}_{j+1} ... \hat{\beta}_p)\right) \geq \xi_1}$$

Implicit formula for $\boldsymbol{\beta_j}$

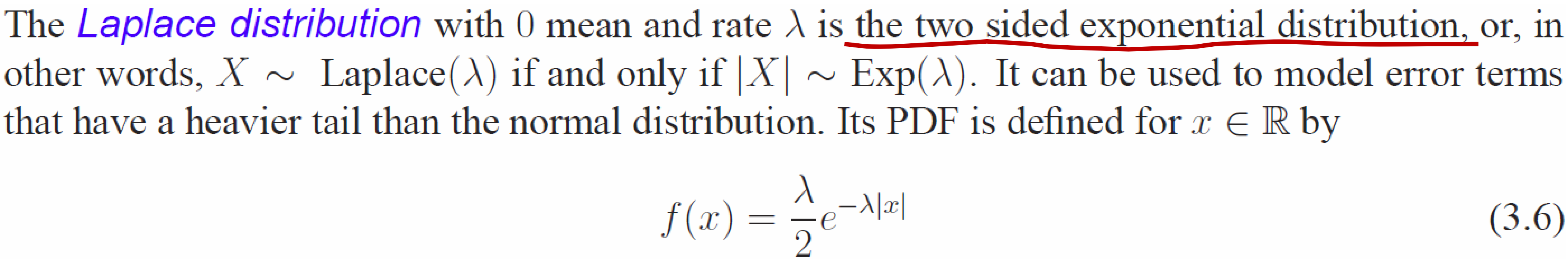
where $\hat{\sigma}^2(\vec{\beta}) = \frac{1}{I} \sum_i \left(y_i - f_i(\vec{\beta})\right)^2$ and ξ_1 is the γ quantile of the χ^2 distribution with 1 degree of freedom (for example, for $\gamma = 0.95$, $\xi_1 = 3.92$).

ℓ^1 Norm Minimization Corresponds to Laplace Noise

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

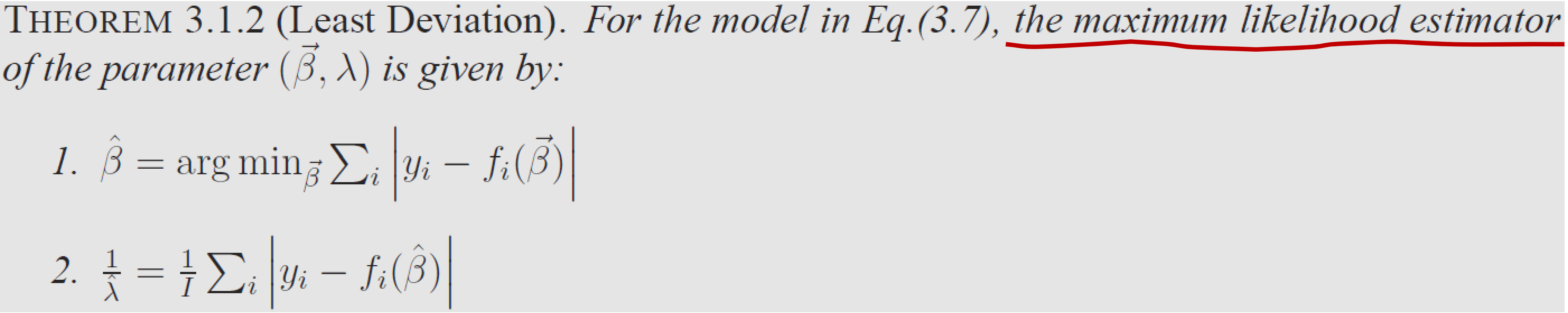


$$Y_i = f_i(\vec{\beta}) + \epsilon_i \text{ with } \epsilon_i \text{iid} \sim \text{ Laplace}(\lambda) \tag{3.7}$$



The *Laplace distribution* with 0 mean and rate λ is the two sided exponential distribution, or, in other words, $X \sim$ Laplace(λ) if and only if $|X| \sim$ Exp(λ). It can be used to model error terms that have a heavier tail than the normal distribution. Its PDF is defined for $x \in \mathbb{R}$ by

$$f(x) = \frac{\lambda}{2} e^{-\lambda |x|} \tag{3.6}$$

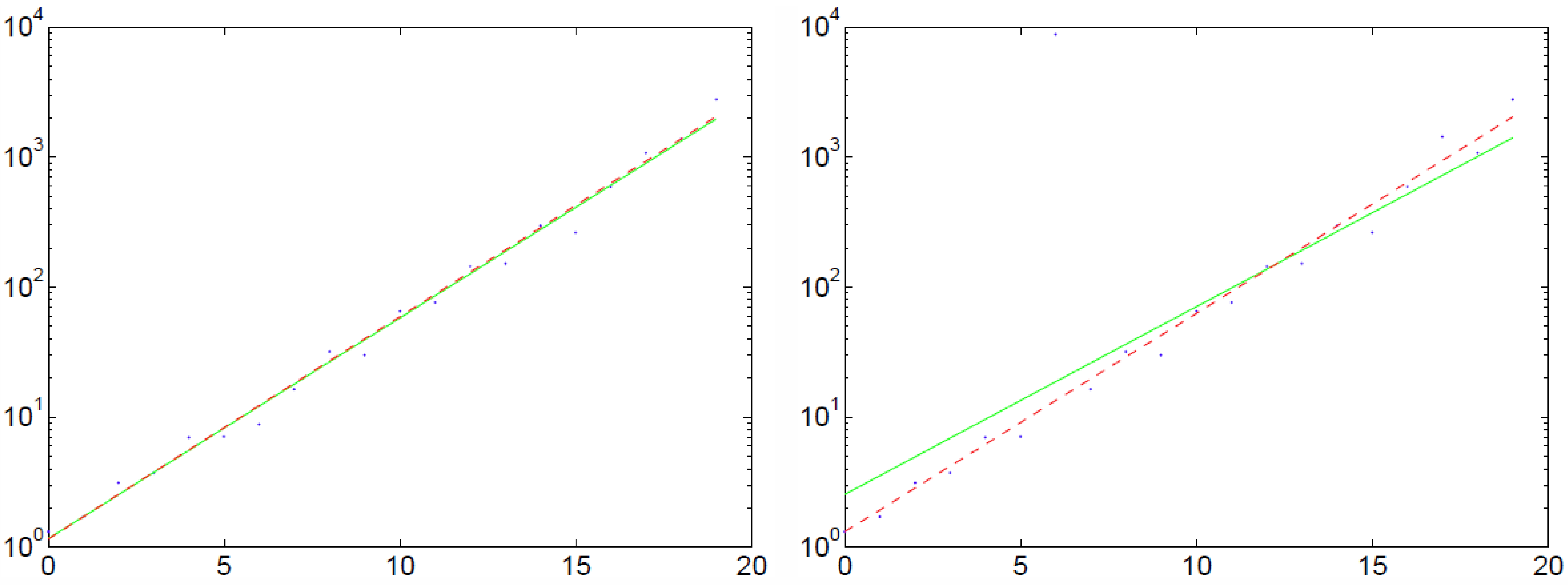


THEOREM 3.1.2 (Least Deviation). *For the model in Eq.(3.7), the maximum likelihood estimator of the parameter* $(\vec{\beta}, \lambda)$ *is given by:*

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

ℓ^1 norm minimization brushes away Laplace noise (dust).

Robustness to « Outliers »



with outlier

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

A Simple Example

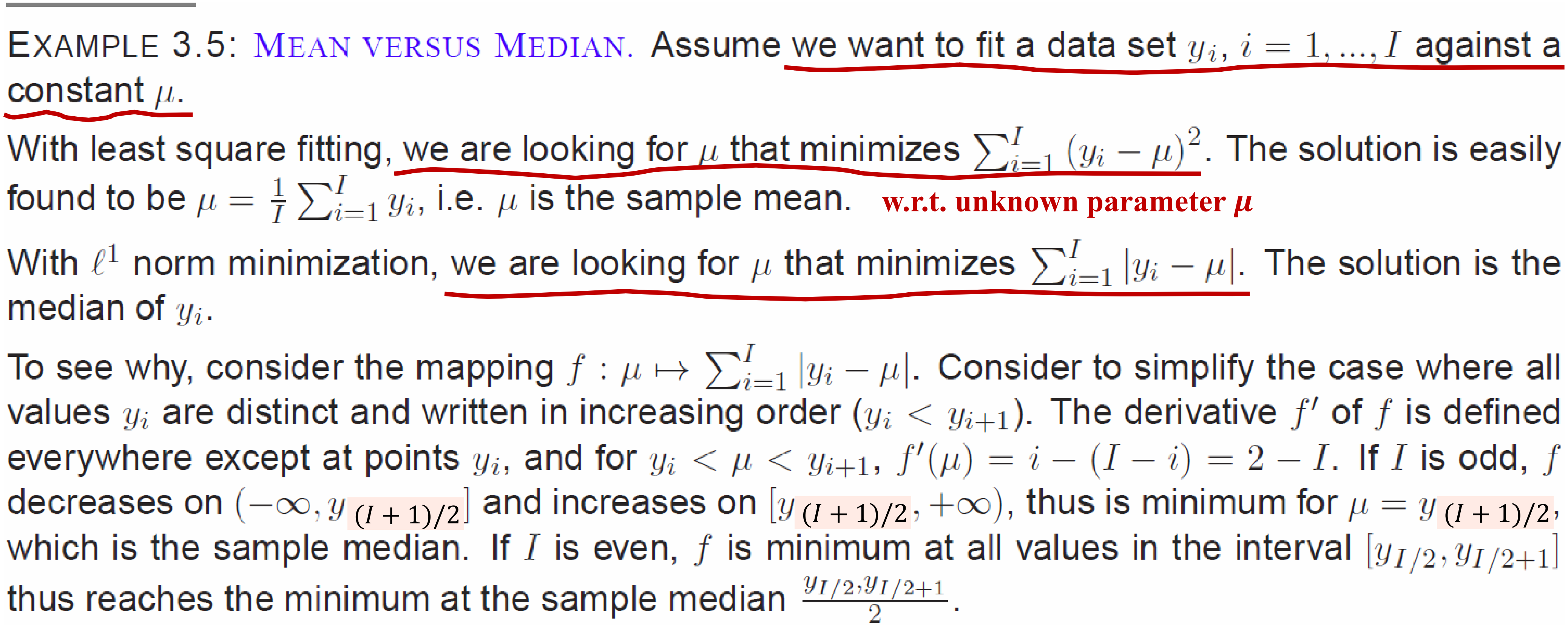
Least Square

- Model: $y_i \ = \ m \ + \text{noise}$
- What is m ?
  - Some true central value
- Confidence interval ?

L1 Norm Minimization

- Model : $y_i \ = \ m + \text{noise}$
- What is m ?
  - Some true central value
- Confidence interval ?

Mean vs. Median



EXAMPLE 3.5: MEAN VERSUS MEDIAN. Assume we want to fit a data set y_i, $i = 1, ..., I$ against a constant μ.

With least square fitting, we are looking for μ that minimizes $\sum_{i=1}^{I}(y_i - \mu)^2$. The solution is easily found to be $\mu = \frac{1}{I}\sum_{i=1}^{I} y_i$, i.e. μ is the sample mean. **w.r.t. unknown parameter *μ***

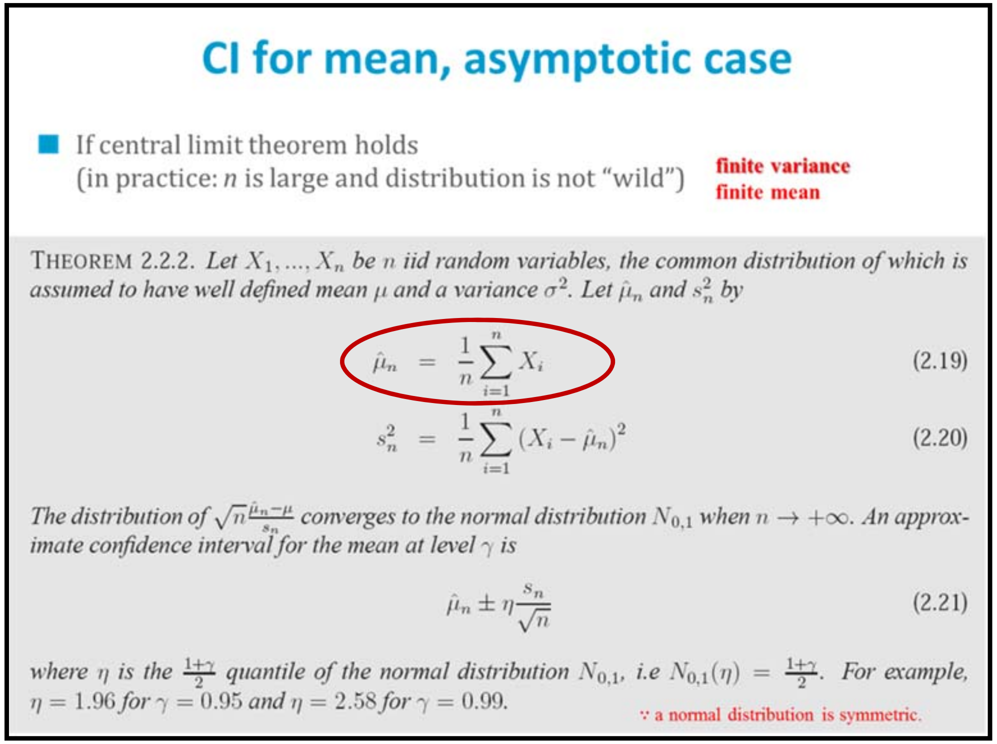
With ℓ^1 norm minimization, we are looking for μ that minimizes $\sum_{i=1}^{I}|y_i - \mu|$. The solution is the median of y_i.

To see why, consider the mapping $f : \mu \mapsto \sum_{i=1}^{I}|y_i - \mu|$. Consider to simplify the case where all values y_i are distinct and written in increasing order ($y_i < y_{i+1}$). The derivative f' of f is defined everywhere except at points y_i, and for $y_i < \mu < y_{i+1}$, $f'(\mu) = i - (I - i) = 2 - I$. If I is odd, f decreases on $(-\infty, y_{(I+1)/2}]$ and increases on $[y_{(I+1)/2}, +\infty)$, thus is minimum for $\mu = y_{(I+1)/2}$, which is the sample median. If I is even, f is minimum at all values in the interval $[y_{I/2}, y_{I/2+1}]$ thus reaches the minimum at the sample median $\frac{y_{I/2}, y_{I/2+1}}{2}$.

Hold on.

Why can't we use *sample mean* for Laplace case?

Sample Mean vs. MLE of Mean



CI for mean, asymptotic case

- If central limit theorem holds
  (in practice: n is large and distribution is not "wild")

finite variance
finite mean

THEOREM 2.2.2. *Let $X_1, ..., X_n$ be n iid random variables, the common distribution of which is assumed to have well defined mean μ and a variance σ^2. Let $\hat{\mu}_n$ and s_n^2 by*

$$\boxed{\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i} \quad (2.19)$$

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2 \quad (2.20)$$

The distribution of $\sqrt{n}\frac{\hat{\mu}_n - \mu}{s_n}$ converges to the normal distribution $N_{0,1}$ when $n \to +\infty$. An approximate confidence interval for the mean at level γ is

$$\hat{\mu}_n \pm \eta \frac{s_n}{\sqrt{n}} \quad (2.21)$$

where η is the $\frac{1+\gamma}{2}$ quantile of the normal distribution $N_{0,1}$, i.e $N_{0,1}(\eta) = \frac{1+\gamma}{2}$. For example, $\eta = 1.96$ for $\gamma = 0.95$ and $\eta = 2.58$ for $\gamma = 0.99$.

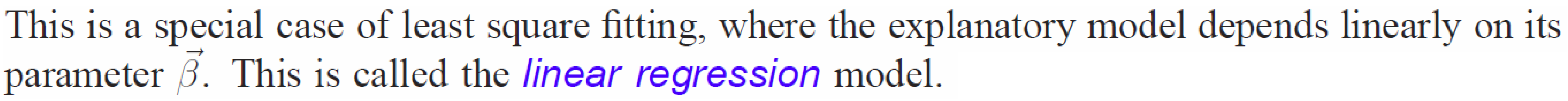
$\because$ a normal distribution is symmetric.

- Recall:
  - Sample mean is a **universal estimator** for all not "wild" distributions
  - Sample mean and MLE for mean **coincide** in case of **normal data**

- Which is superior between sample mean and MLE for mean?
  - If you know distribution type, **MLE is superior** because it's a **bespoke** method.
  - Roughly speaking, **MLE** tends to provide a more **sophisticated** estimate.
  - You can usually better estimate mean of measured data by MLE.

If the data follows a Laplace distribution, MLE for mean is the median of the data.

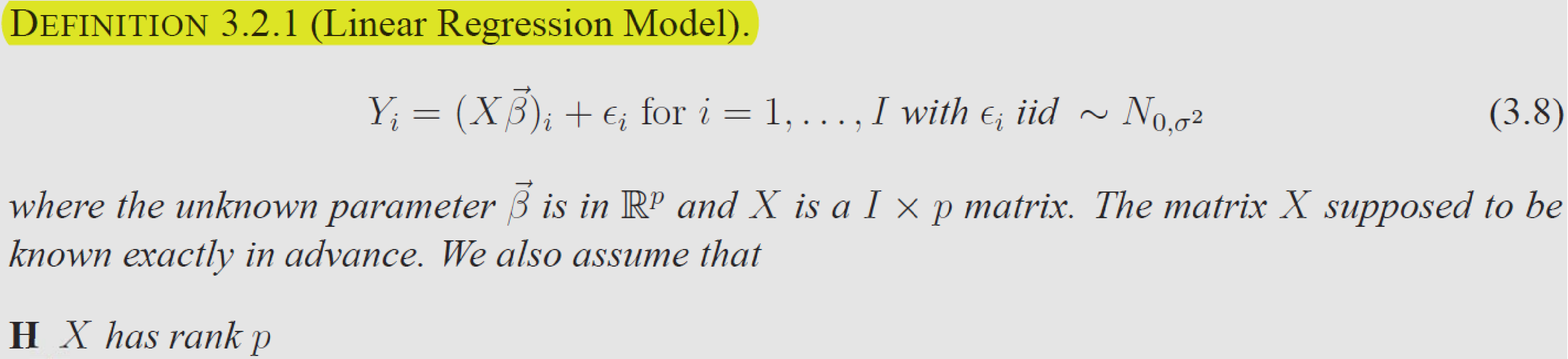
2. Linear Regression



This is a special case of least square fitting, where the explanatory model depends linearly on its parameter $\vec{\beta}$. This is called the *linear regression* model.

- Also called « **ANOVA** » (**AN**alysis **O**f **VA**riance)

- = least square + linear dependence on parameter



DEFINITION 3.2.1 (Linear Regression Model).

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

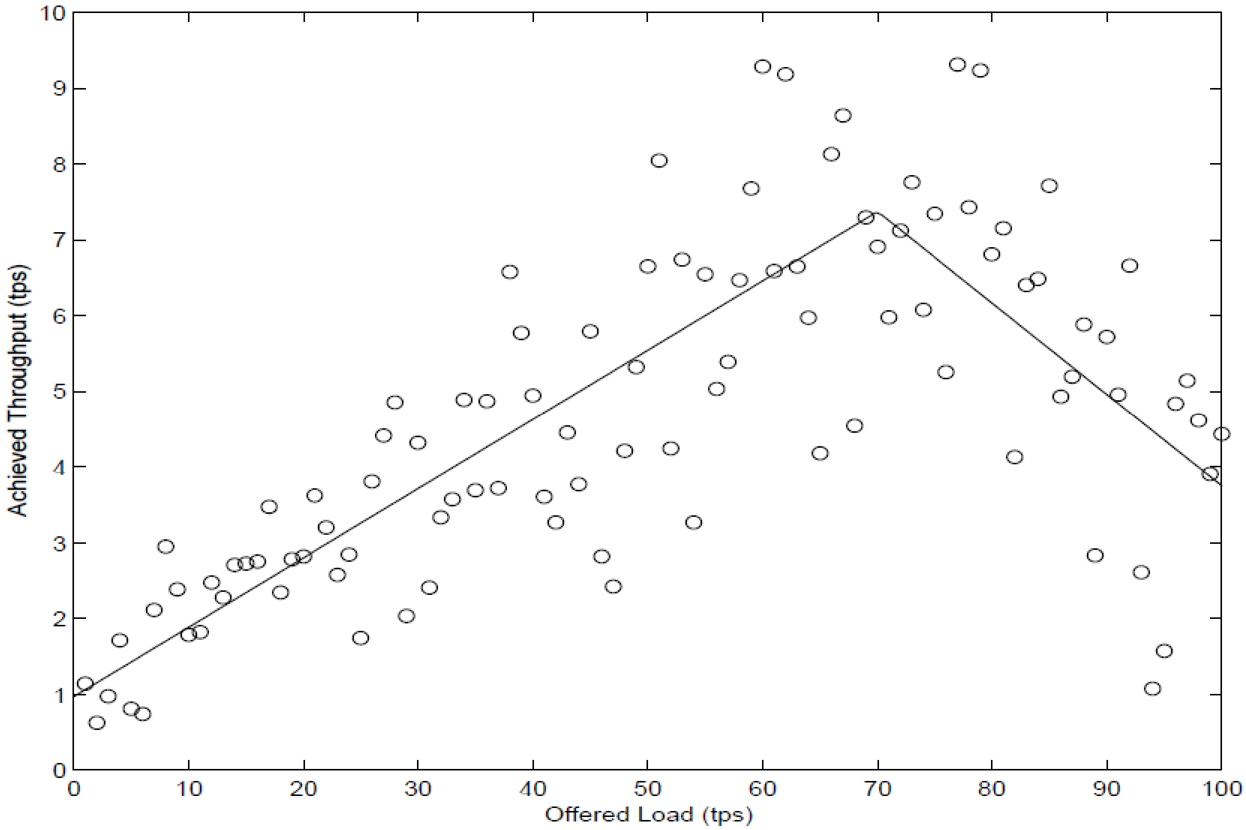
where the unknown parameter $\vec{\beta}$ is in $\mathbb{R}^p$ and X is a $I \times p$ matrix. The matrix X supposed to be known exactly in advance. We also assume that

H *X has rank p*

A necessary condition for which is $I \geq p$, i.e., the number of measurements ≥ the number of unknown parameter

- A special case where computations are easy

Example 3.6-7

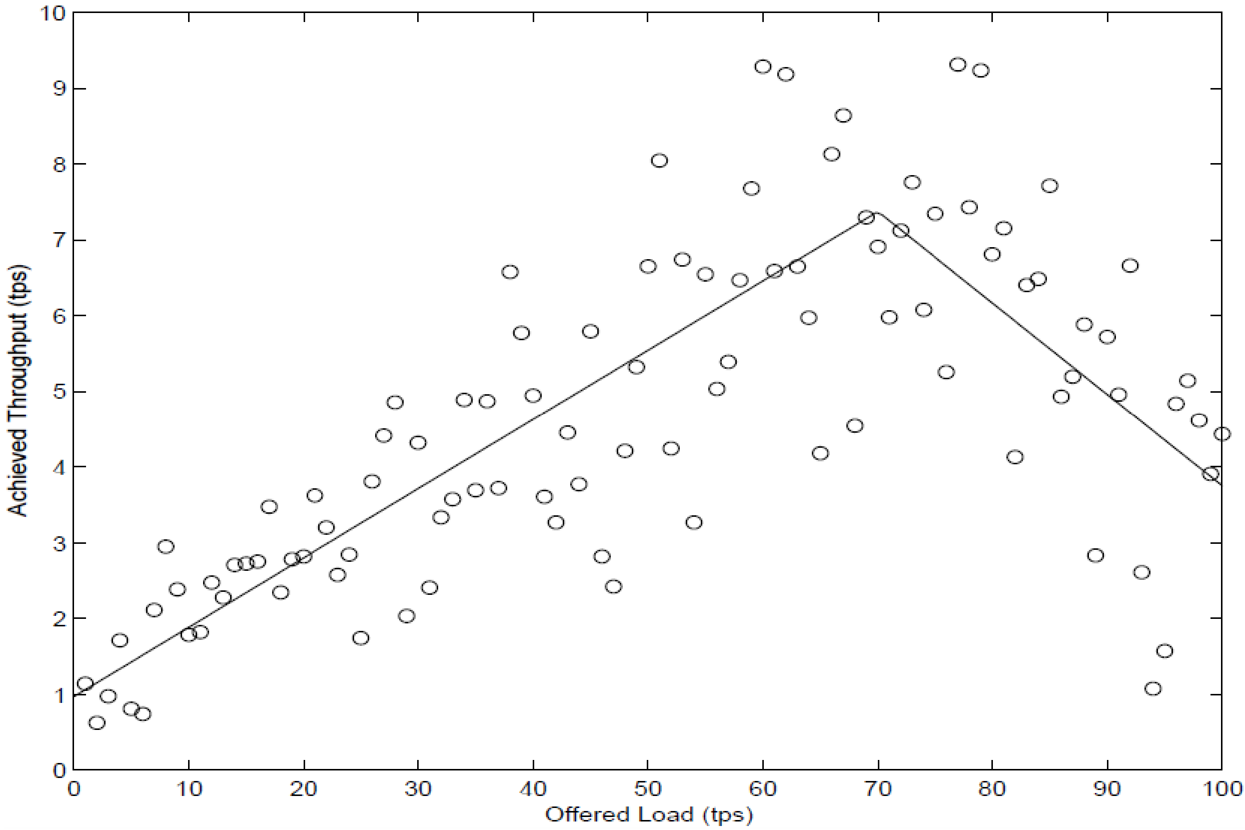


- What is the parameter β?
- Is it a ***linear*** model?
- How many degrees of freedom?
- What do we assume on ϵ_i?
- What is the matrix X?

$$Y_i = (a + bx_i)1_{x_i \le \xi} + (c + dx_i)1_{\{x_i > \xi\}} + \epsilon_i$$
$$a + b\xi = c + d\xi$$

$\vec{\beta} = (a, b, d)$ and ξ [zai] is taken to be 70.

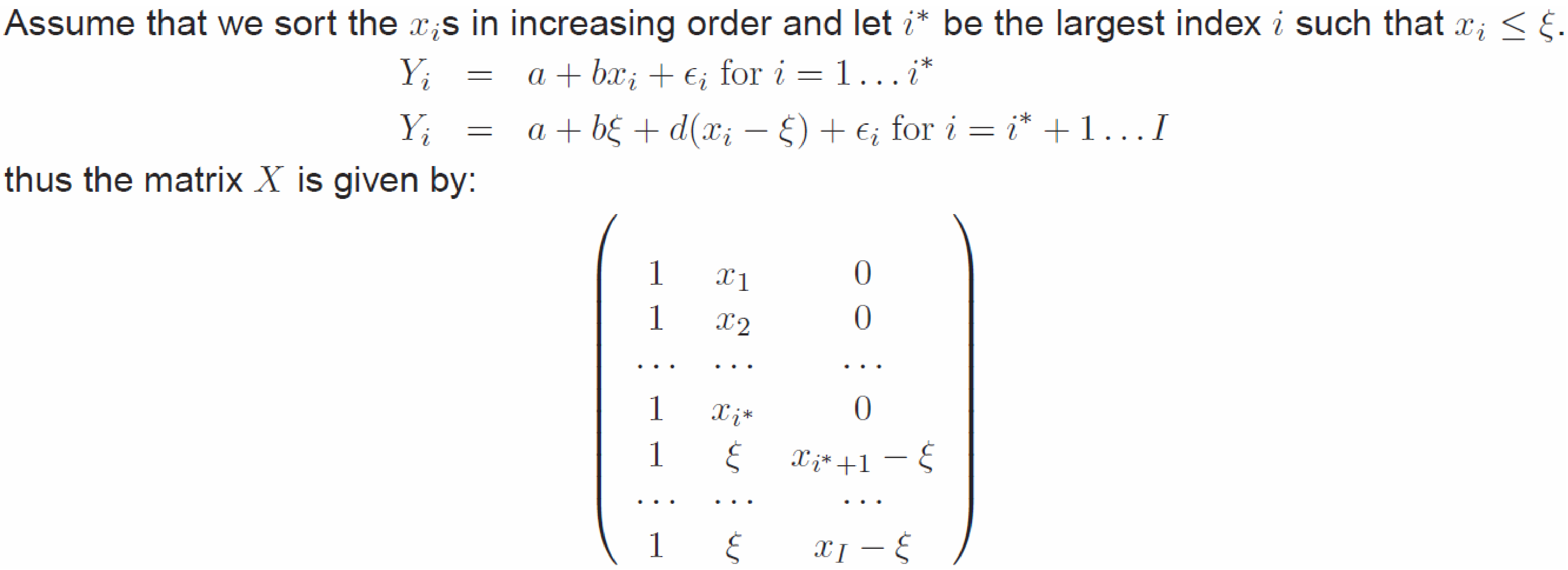
Some Terminology



- *X*'s elements are called **explanatory** variable
  - Assumed fixed and **known**
- *Y*'s elements are called **response** variables
  - They are « the data »
  - Assumed to be one sample output of the model
  - For they are corrupted, their **true values are unknown**.

$$Y_i = (a + bx_i)1_{x_i \leq \xi} + (c + dx_i)1_{\{x_i > \xi\}} + \epsilon_i$$
$$a + b\xi = c + d\xi$$

$\vec{\beta} = (a, b, d)$ (we can derive $c = a + (b - d)\xi$ from $a + b\xi = c + d\xi$



Assume that we sort the x_is in increasing order and let i^* be the largest index i such that $x_i \leq \xi$.

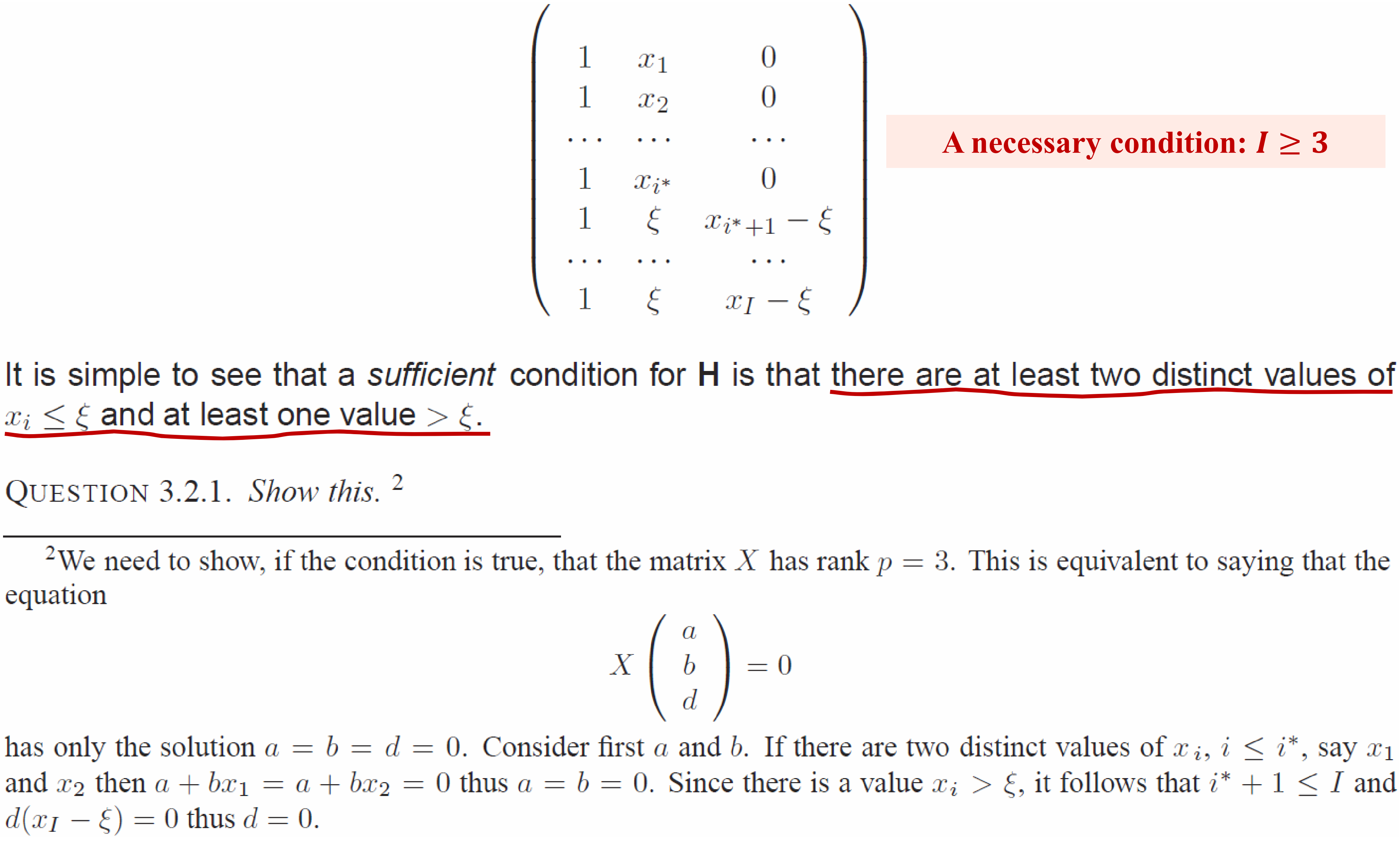
$$\begin{aligned} Y_i &= a + bx_i + \epsilon_i \text{ for } i = 1 \ldots i^* \\ Y_i &= a + b\xi + d(x_i - \xi) + \epsilon_i \text{ for } i = i^* + 1 \ldots I \end{aligned}$$

thus the matrix X is given by:

$$\begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \ldots & \ldots & \ldots \\ 1 & x_{i^*} & 0 \\ 1 & \xi & x_{i^*+1} - \xi \\ \ldots & \ldots & \ldots \\ 1 & \xi & x_I - \xi \end{pmatrix}$$

The model is linear because "response variables" Y is linear w.r.t. $\vec{\beta} = (a, b, d)$.

Does this model have full rank ?



$$\begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \cdots & \cdots & \cdots \\ 1 & x_{i^*} & 0 \\ 1 & \xi & x_{i^*+1} - \xi \\ \cdots & \cdots & \cdots \\ 1 & \xi & x_I - \xi \end{pmatrix}$$

A necessary condition: $I \geq 3$

It is simple to see that a *sufficient* condition for **H** is that there are at least two distinct values of $x_i \leq \xi$ and at least one value $> \xi$.

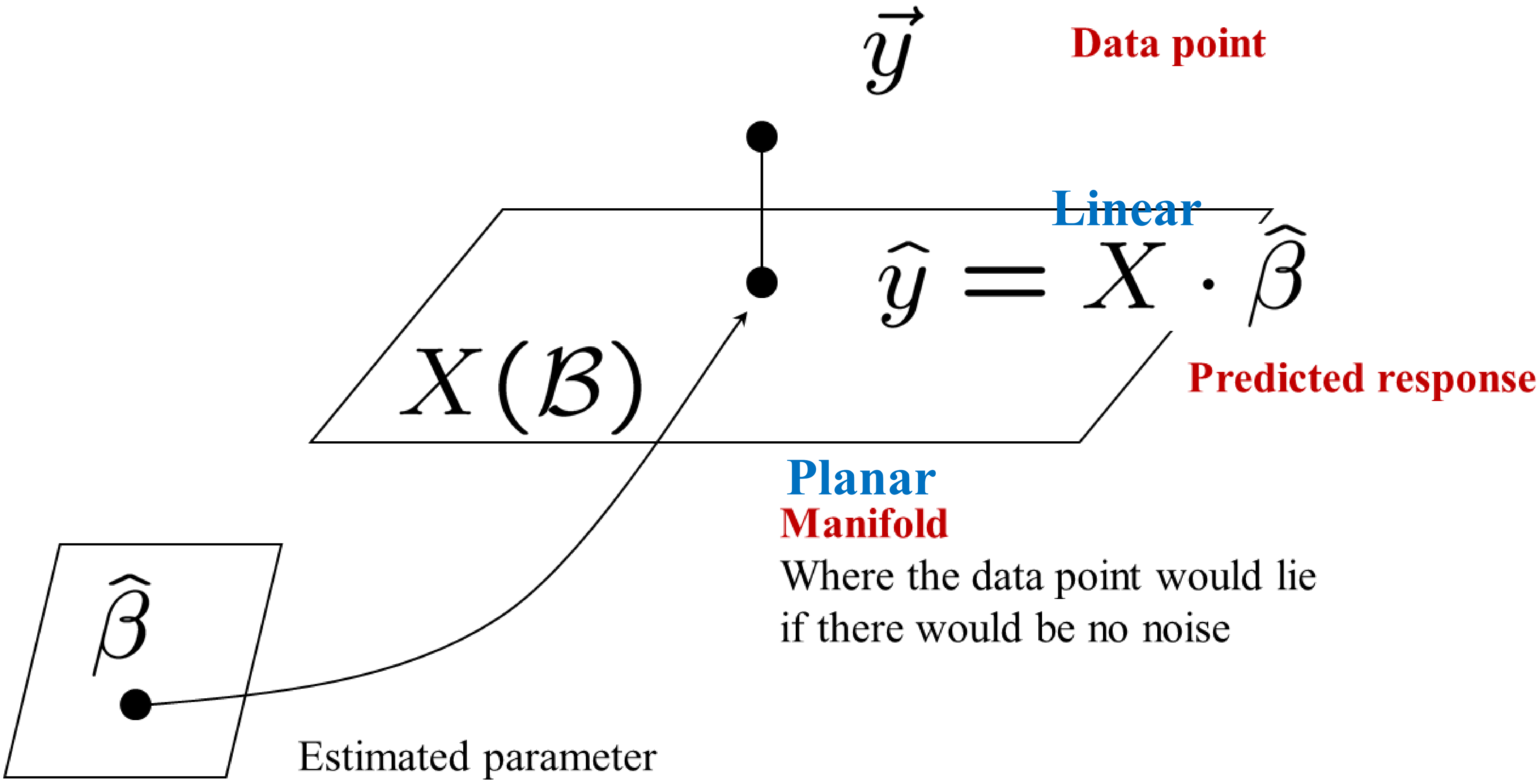
QUESTION 3.2.1. *Show this.* [2]

[2] We need to show, if the condition is true, that the matrix X has rank $p = 3$. This is equivalent to saying that the equation

$$X \begin{pmatrix} a \\ b \\ d \end{pmatrix} = 0$$

has only the solution $a = b = d = 0$. Consider first a and b. If there are two distinct values of x_i, $i \leq i^*$, say x_1 and x_2 then $a + bx_1 = a + bx_2 = 0$ thus $a = b = 0$. Since there is a value $x_i > \xi$, it follows that $i^* + 1 \leq I$ and $d(x_I - \xi) = 0$ thus $d = 0$.

Least Square and Projection



$\vec{y}$

Data point

Linear

$$\hat{y} = X \cdot \hat{\beta}$$

Predicted response

$X(\mathcal{B})$

Planar

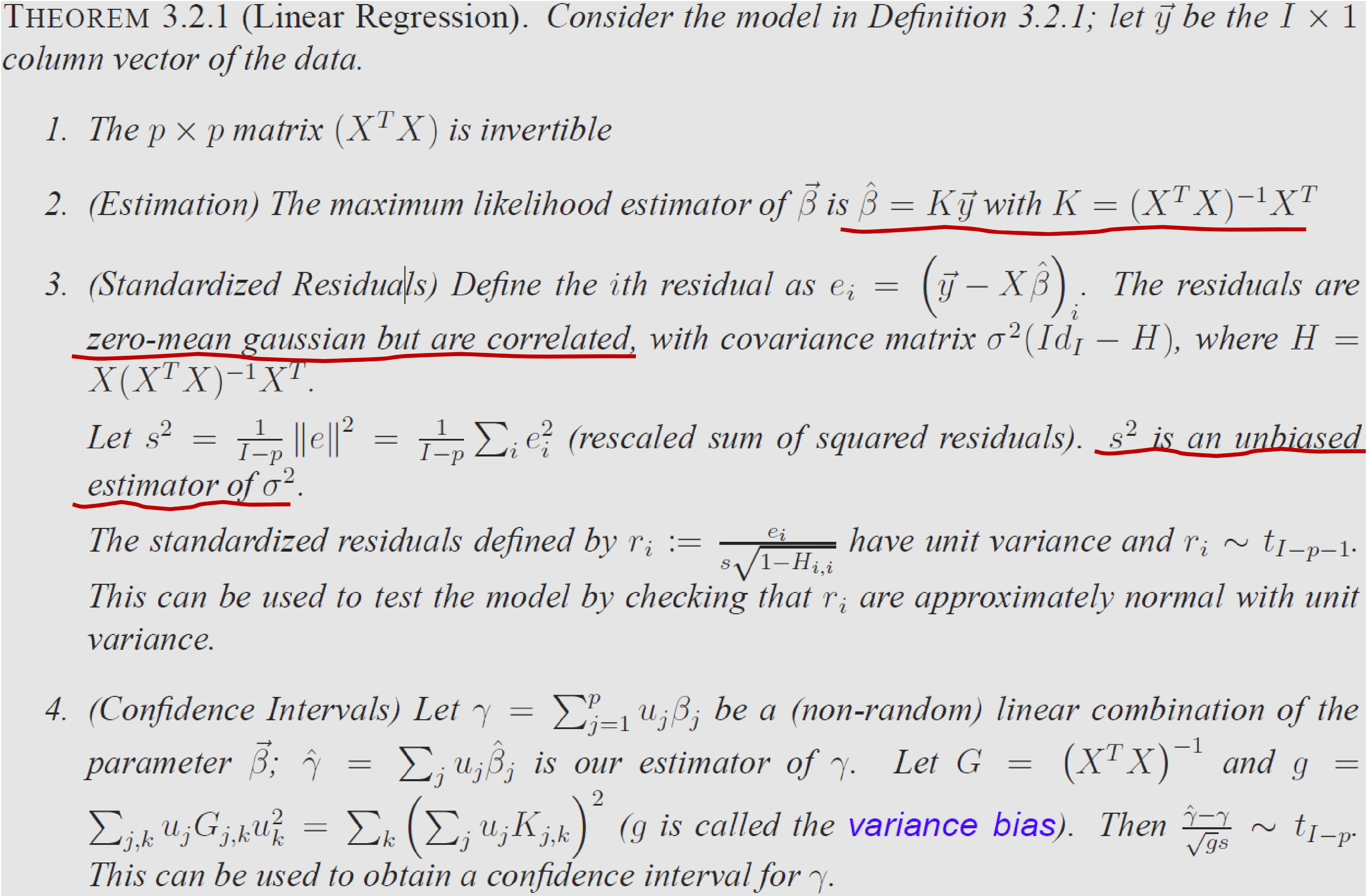
Manifold

Where the data point would lie
if there would be no noise

$\hat{\beta}$

Estimated parameter

Solution of the Linear Regression Model



THEOREM 3.2.1 (Linear Regression). *Consider the model in Definition 3.2.1; 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 $H = X(X^T X)^{-1} X^T$.*

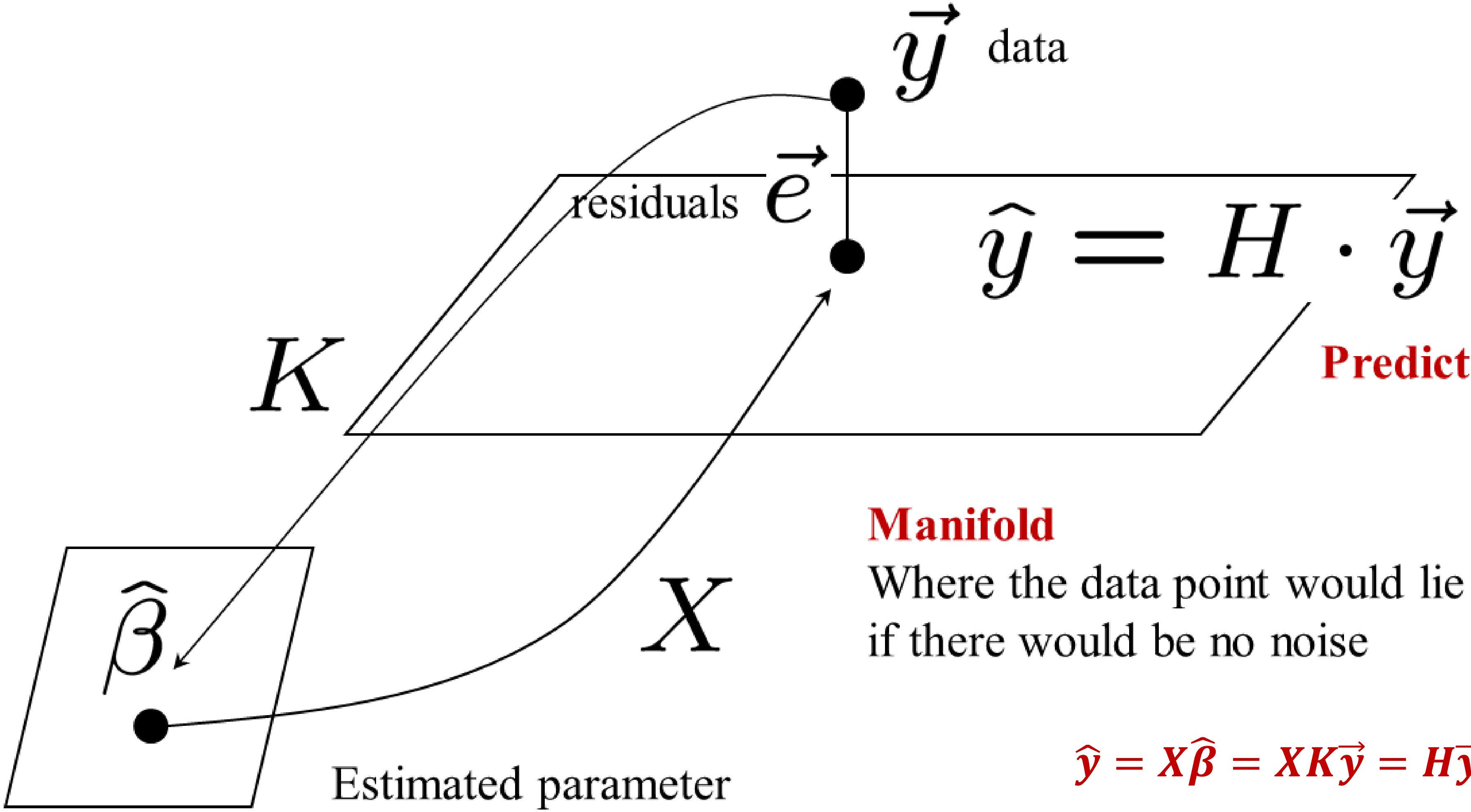
   Let $s^2 = \frac{1}{I-p}\|e\|^2 = \frac{1}{I-p}\sum_i e_i^2$ (rescaled sum of squared residuals). s^2 is an unbiased estimator of σ^2.

   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 $\gamma = \sum_{j=1}^{p} u_j \beta_j$ be a (non-random) linear combination of the parameter $\vec{\beta}$; $\hat{\gamma} = \sum_j u_j \hat{\beta}_j$ is our estimator of γ. Let $G = \left(X^T X\right)^{-1}$ and $g = \sum_{j,k} u_j G_{j,k} u_k^2 = \sum_k \left(\sum_j u_j K_{j,k}\right)^2$ (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.*

Least Square and Projection

- The theorem gives $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$ and $\boldsymbol{K} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$



Predicted response

Manifold

Where the data point would lie
if there would be no noise

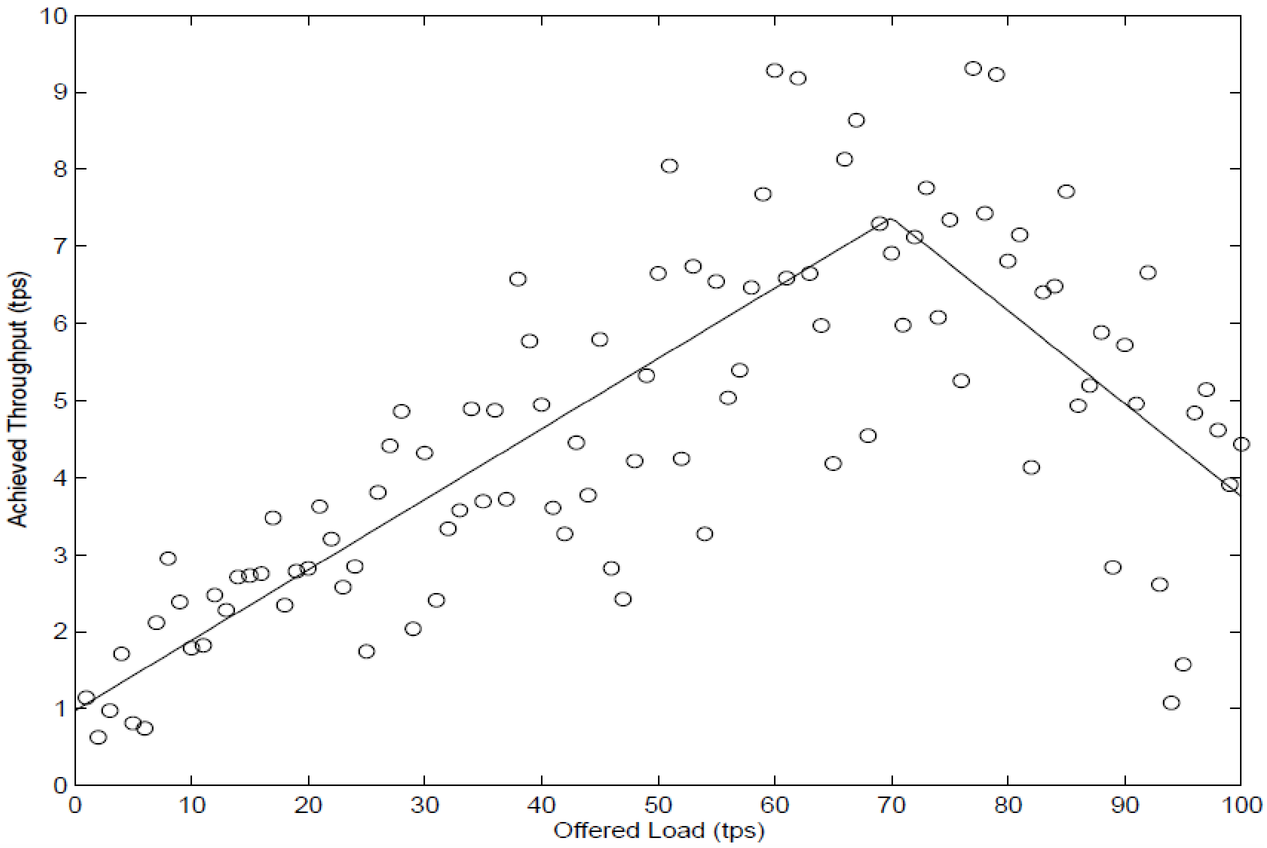
$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{K}\vec{\boldsymbol{y}} = \boldsymbol{H}\vec{\boldsymbol{y}}$

What we estimate (predict) is $\boldsymbol{\beta}$, and its estimate $\hat{\boldsymbol{\beta}}$ in turn re-adjusts $\boldsymbol{y}$ to $\hat{\boldsymbol{y}} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}}$.

The Theorem Gives β with Confidence Interval

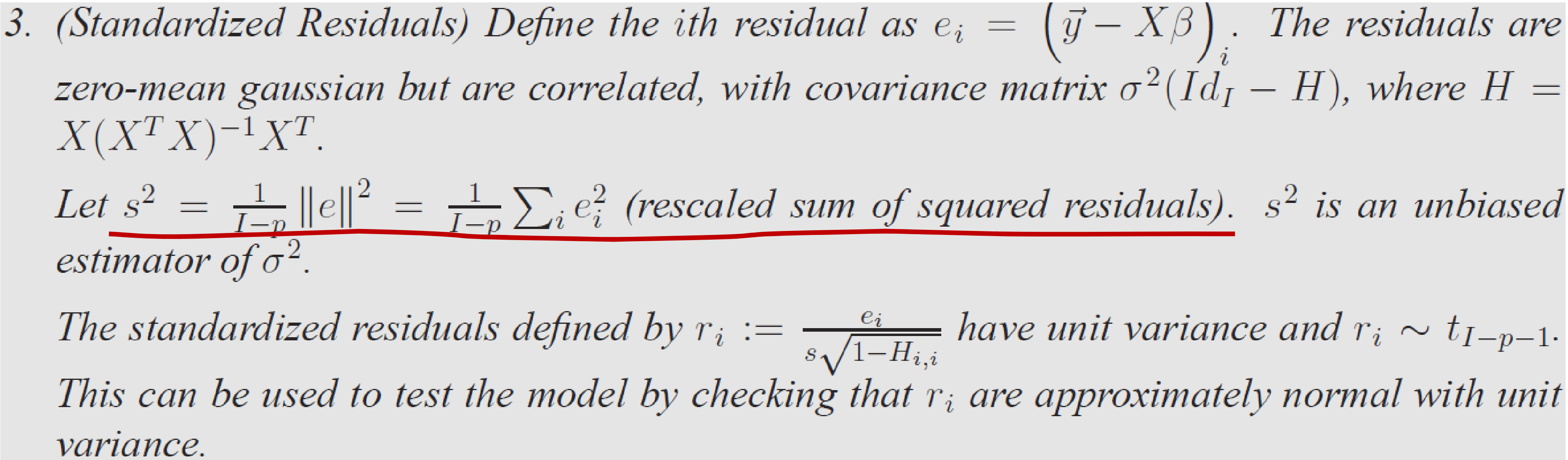
4. *(Confidence Intervals) Let* $\gamma = \sum_{j=1}^{p} u_j \beta_j$ *be a (non-random) linear combination of the parameter* $\vec{\beta}$; $\hat{\gamma} = \sum_j u_j \hat{\beta}_j$ *is our estimator of* γ. *Let* $G = \left(X^T X\right)^{-1}$ *and* $g = \sum_{j,k} u_j G_{j,k} u_k^2 = \sum_k \left(\sum_j u_j K_{j,k}\right)^2$ *(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.

a	0.978 ± 0.609
b	0.0915 ± 0.0137
c	15.8 ± 2.99
d	-0.121 ± 0.037



SSR

- Confidence Intervals use the quantity s



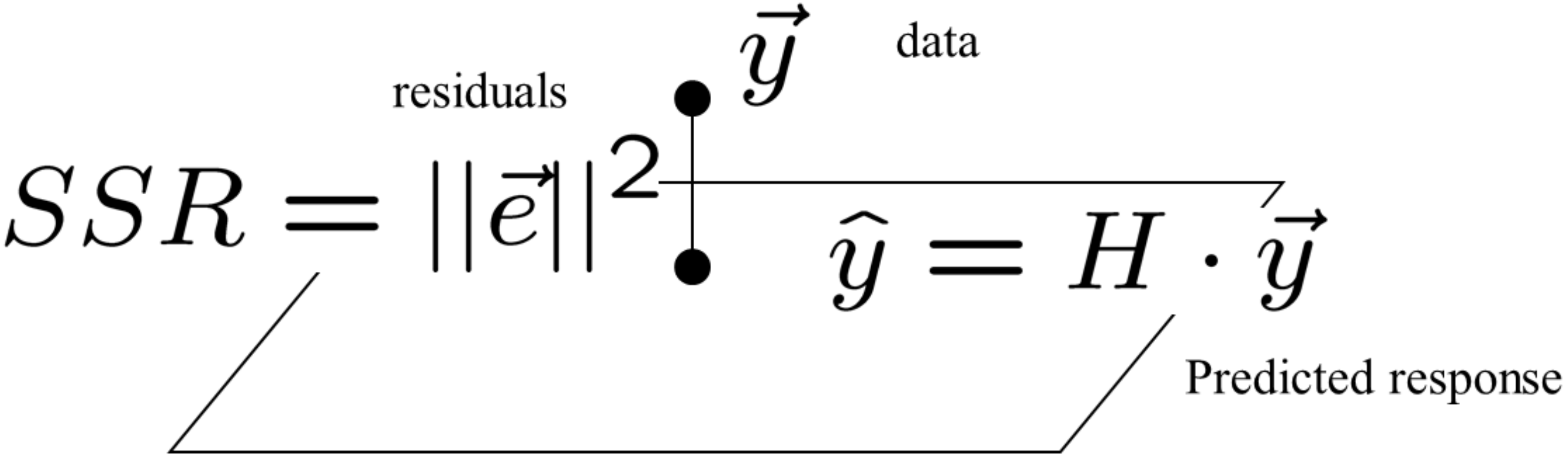
3. *(Standardized Residuals) Define the ith residual as* $e_i = \left(\vec{y} - X\beta\right)_i$. *The residuals are zero-mean gaussian but are correlated, with covariance matrix* $\sigma^2(Id_I - H)$, *where* $H = X(X^TX)^{-1}X^T$.

   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* $r_i \sim t_{I-p-1}$.

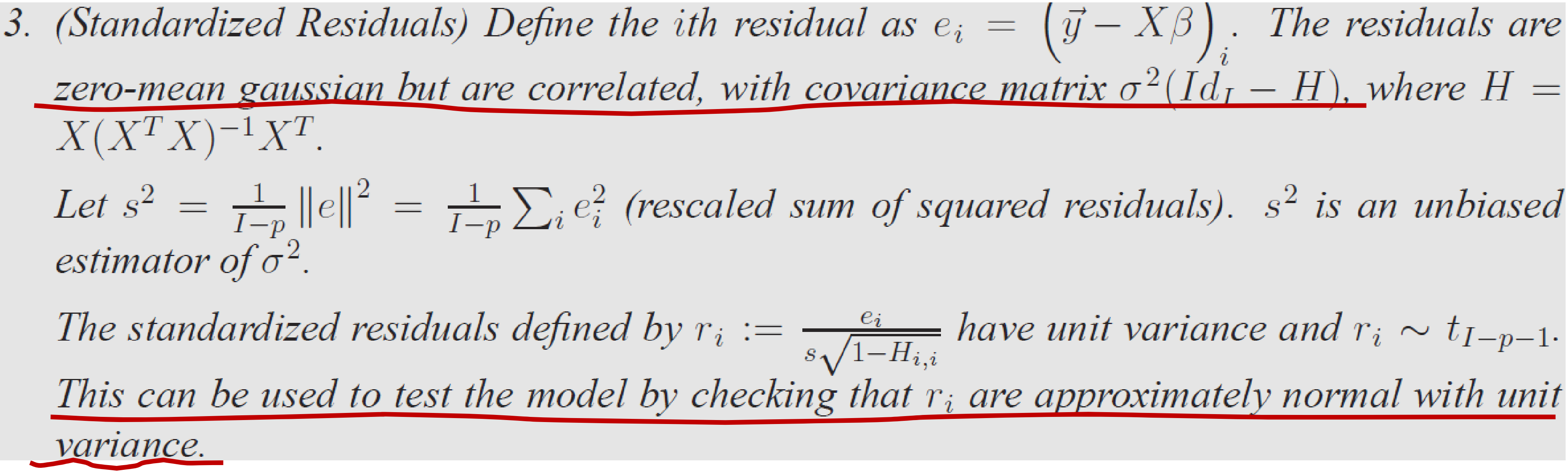
   This can be used to test the model by checking that r_i *are approximately normal with unit variance.*

- s^2 is called « (Rescaled) Sum of Squared Residuals »



Residuals

- Residuals are given by the theorem

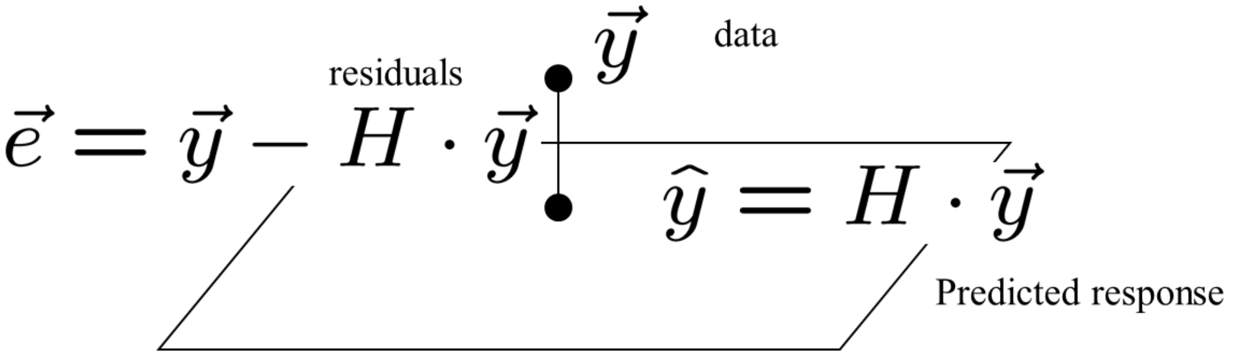


3. (Standardized Residuals) Define the i*th residual as* $e_i = \left(\vec{y} - X\beta\right)_i$. *The residuals are zero-mean gaussian but are correlated, with covariance matrix* $\sigma^2(Id_I - H)$, *where* $H = X(X^TX)^{-1}X^T$.

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.*



$\vec{y}$ data

residuals

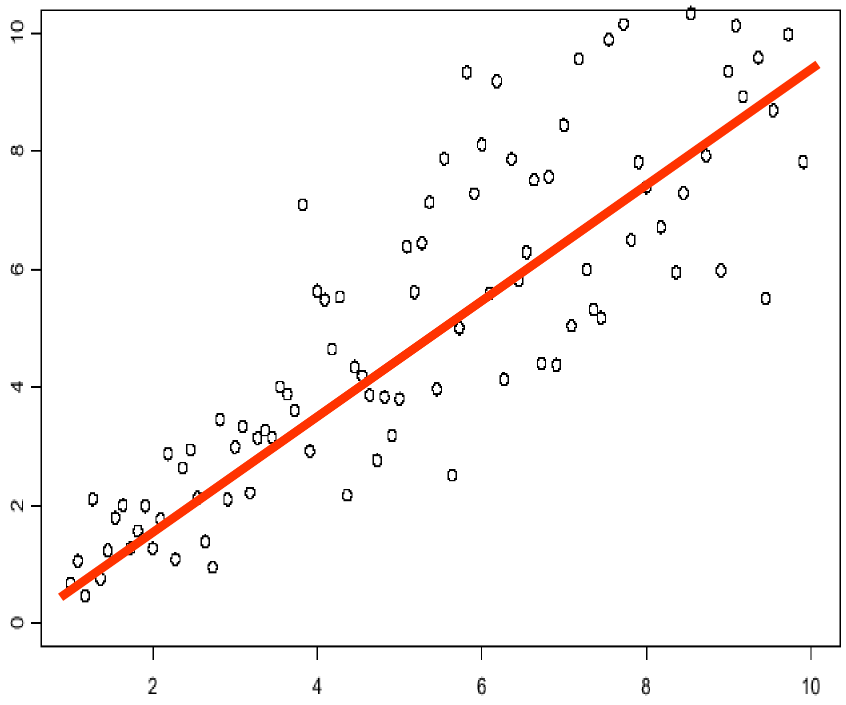
$$\vec{e} = \vec{y} - H \cdot \vec{y}$$

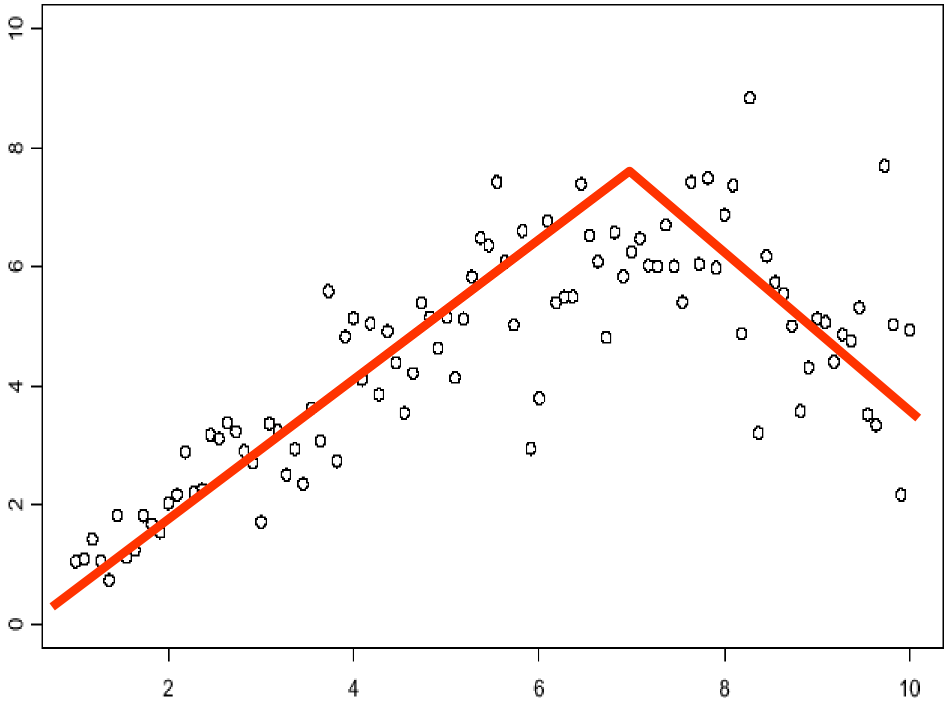
$$\hat{y} = H \cdot \vec{y}$$

Predicted response

- Even **standardized** residuals are not (exactly) normal iid.
  → **violation of homoscedasticity can be checked**

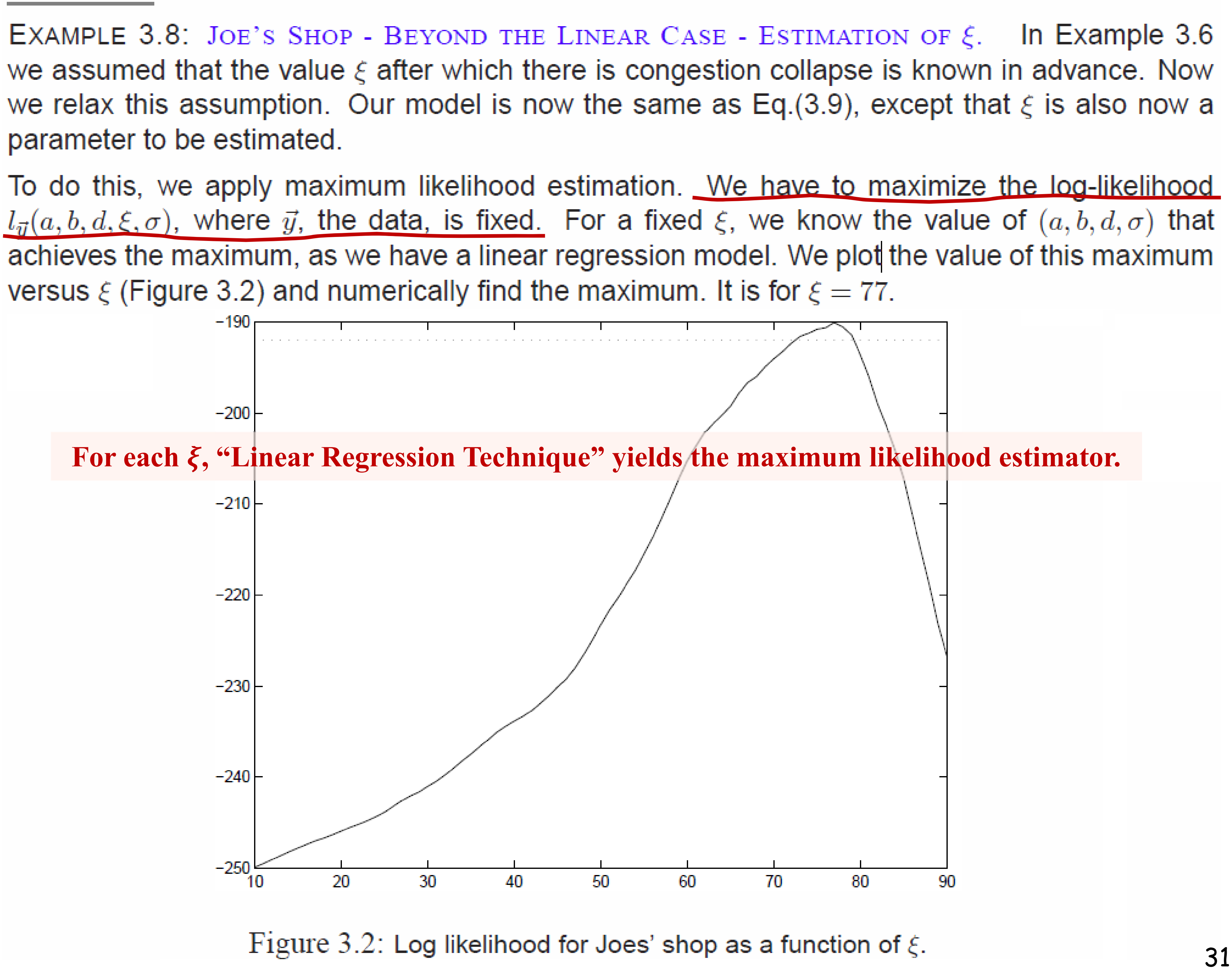
Which of these two models could be a linear regression model ?





- A: both

- Linear regression does *not* mean that y_i is a linear function of x_i
- Caution: There is a hidden assumption
  - Noise is iid Gaussian -> homoscedasticity



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.

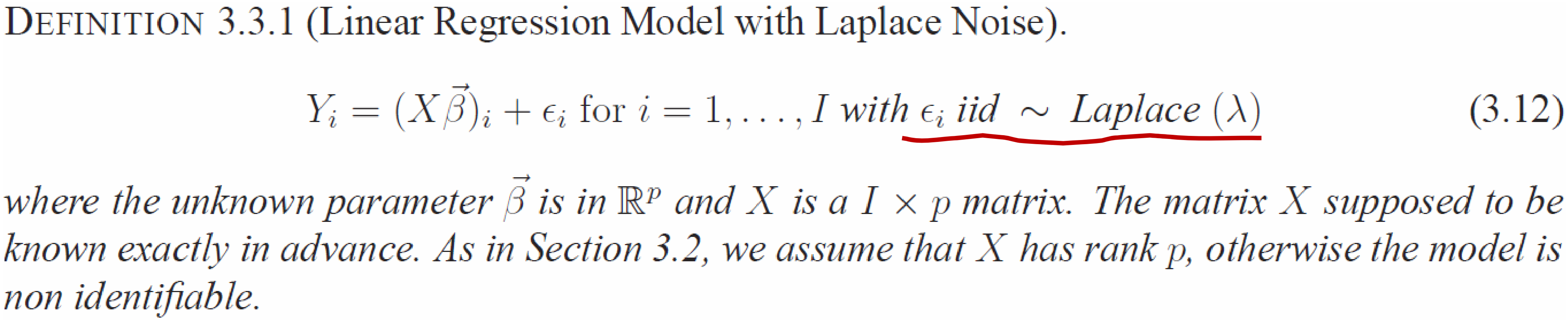
To do this, we apply maximum likelihood estimation. We have to maximize the log-likelihood $l_{\vec{y}}(a, b, d, \xi, \sigma)$, where $\vec{y}$, the data, is fixed. For a fixed ξ, we know the value of (a, b, d, σ) that achieves the maximum, as we have a linear regression model. We plot the value of this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

For each ξ, "Linear Regression Technique" yields the maximum likelihood estimator.

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

3. Linear Regression with ℓ^1 norm minimization

- **ℓ^1 norm minimization + linear dependency on parameter**
- More robust
- Less traditional

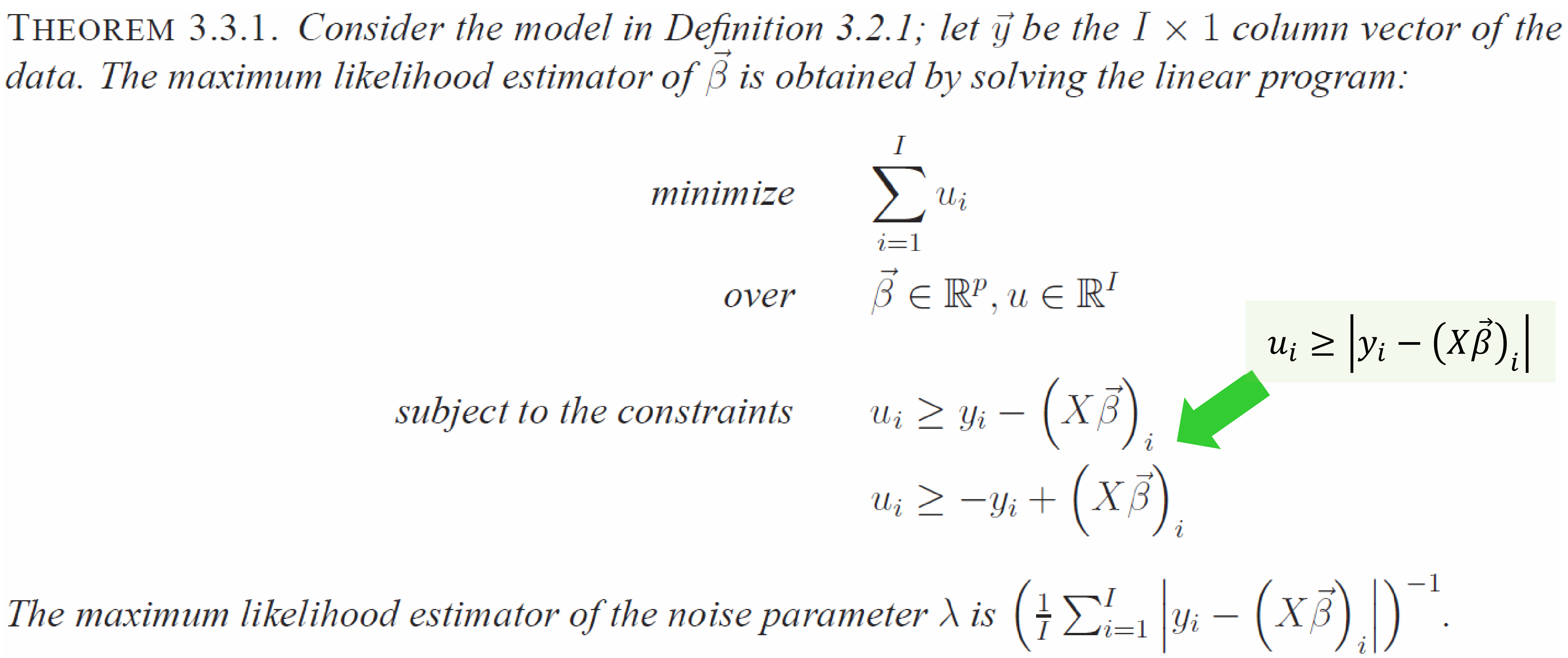


DEFINITION 3.3.1 (Linear Regression Model with Laplace Noise).

$$Y_i = (X\vec{\beta})_i + \epsilon_i \text{ for } i = 1, \dots, I \textit{ with } \underline{\epsilon_i \textit{ iid } \sim \textit{ Laplace } (\lambda)} \tag{3.12}$$

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. As in Section 3.2, we assume that X has rank p, otherwise the model is non identifiable.

This is convex programming



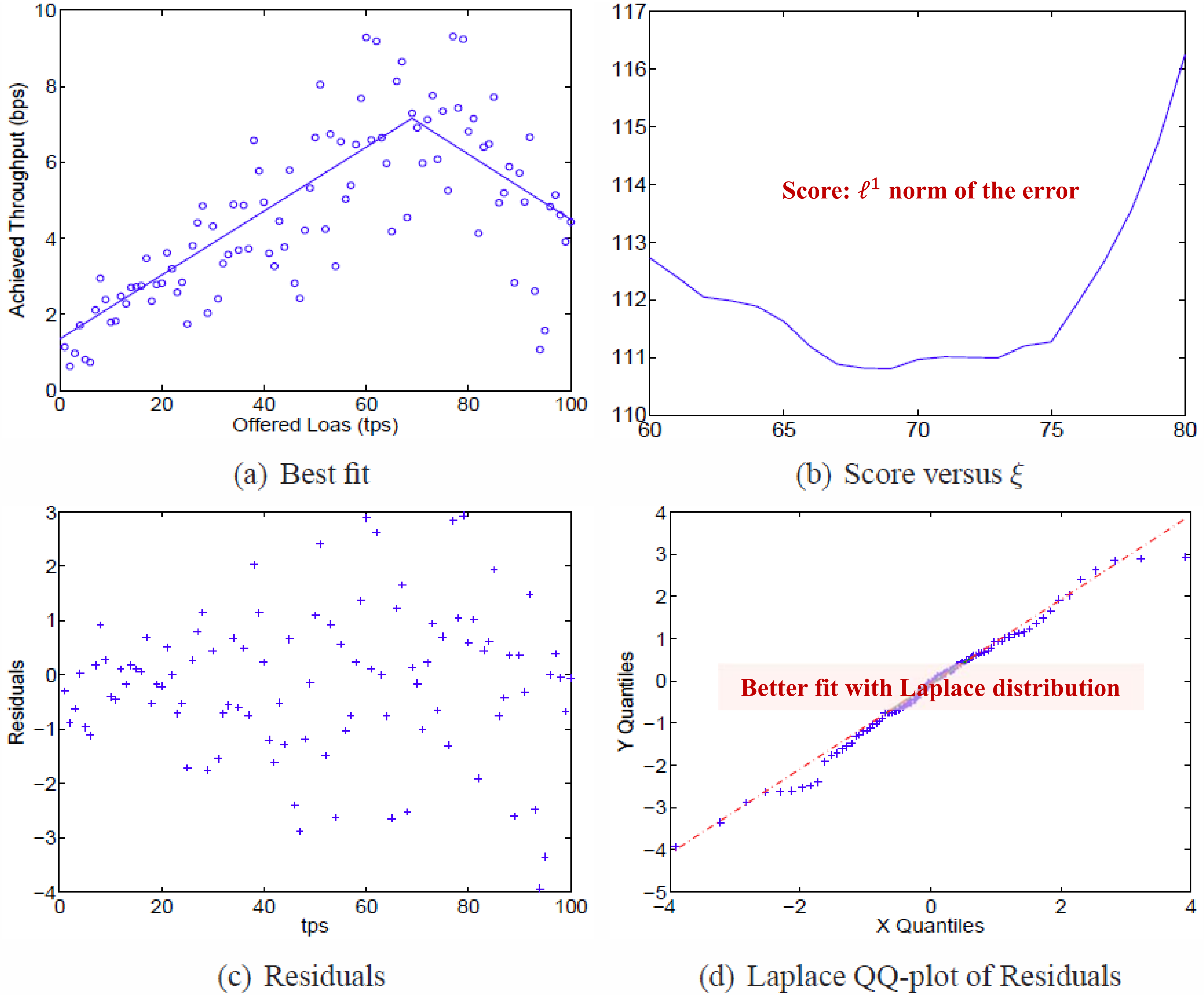
THEOREM 3.3.1. *Consider the model in Definition 3.2.1; let $\vec{y}$ be the $I \times 1$ column vector of the data. The maximum likelihood estimator of $\vec{\beta}$ is obtained by solving the linear program:*

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

$$u_i \geq \left|y_i - \left(X\vec{\beta}\right)_i\right|$$

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

Maximum likelihood estimator must be obtained by solving this convex optimization.



Score: ℓ^1 norm of the error

Better fit with Laplace distribution

(a) Best fit

(b) Score versus ξ

(c) Residuals

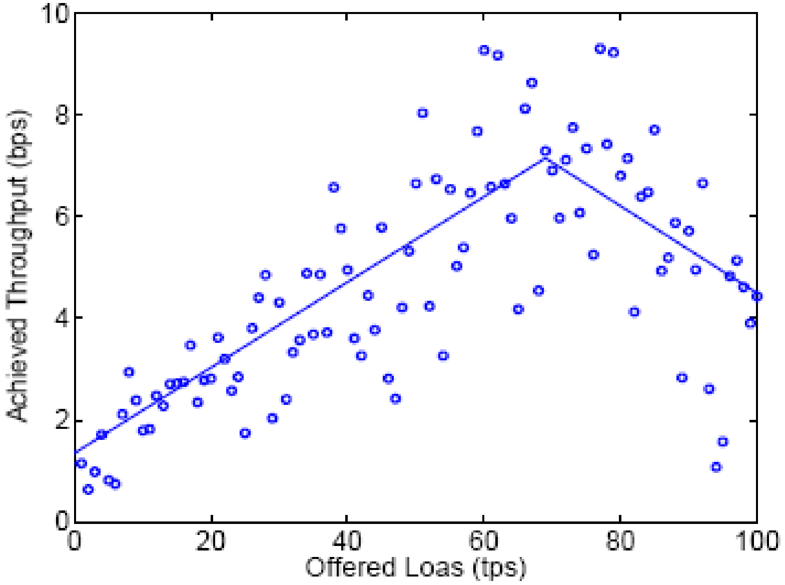
(d) Laplace QQ-plot of Residuals

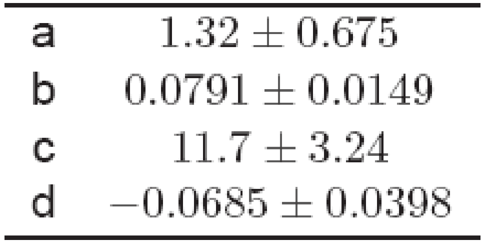
Figure 3.3: Modelling congestion collapse in Joe's shop with a piecewise linear function and ℓ^1 norm minimization of the errors.

Confidence Intervals

- No closed form
  - Compare to median !

- Bootstrap method!

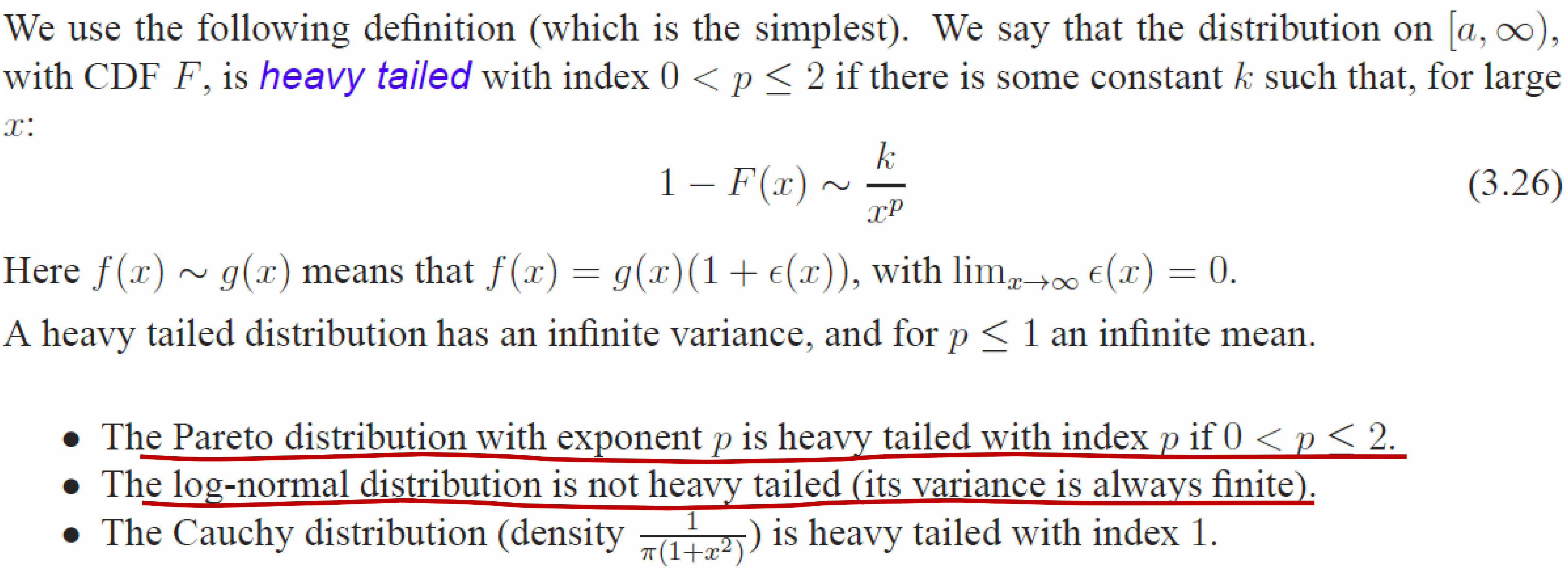




a	1.32 ± 0.675
b	0.0791 ± 0.0149
c	11.7 ± 3.24
d	-0.0685 ± 0.0398

4. Heavy Tails

- Probably helpful when reading papers on measurement studies



We use the following definition (which is the simplest). We say that the distribution on $[a, \infty)$, with CDF F, is *heavy tailed* with index $0 < p \leq 2$ if there is some constant k such that, for large x:

$$1 - F(x) \sim \frac{k}{x^p} \tag{3.26}$$

Here $f(x) \sim g(x)$ means that $f(x) = g(x)(1 + \epsilon(x))$, with $\lim_{x \to \infty} \epsilon(x) = 0$.

A heavy tailed distribution has an infinite variance, and for $p \leq 1$ an infinite mean.

- The Pareto distribution with exponent p is heavy tailed with index p if $0 < p \leq 2$.
- The log-normal distribution is not heavy tailed (its variance is always finite).
- The Cauchy distribution (density $\frac{1}{\pi(1+x^2)}$) is heavy tailed with index 1.

Heavy Tail means Central Limit does not hold

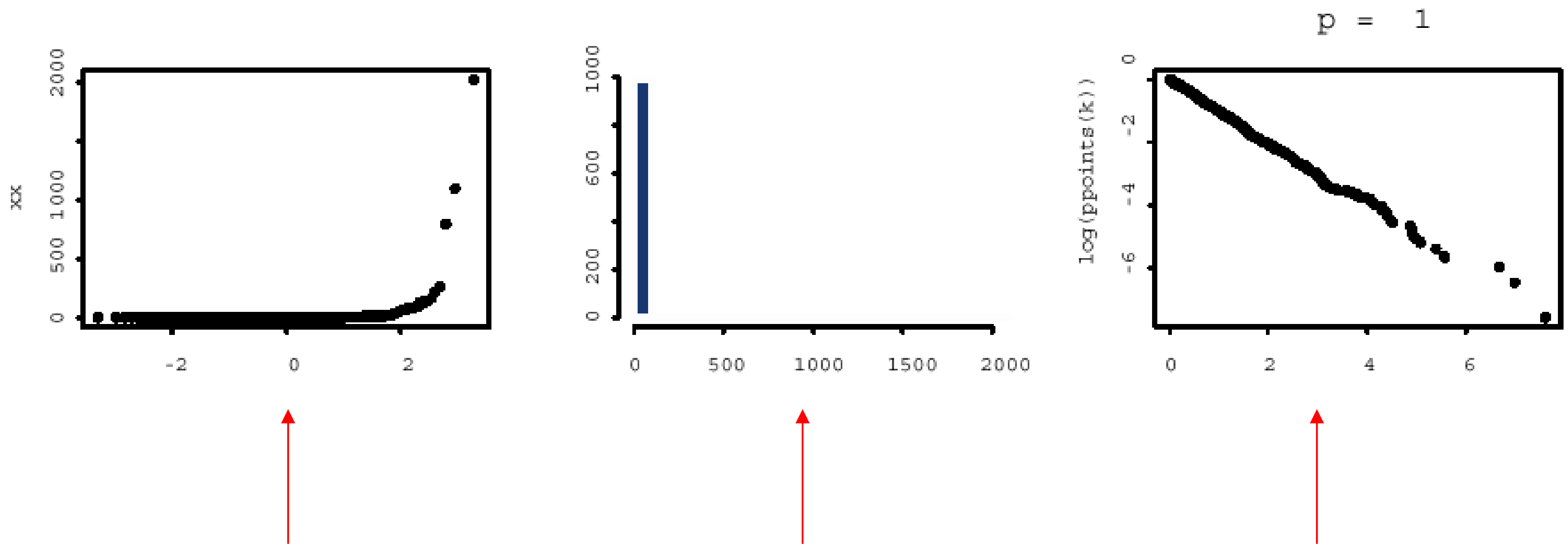
- Central limit theorem:

  a sum of n independent random variables with finite second moment tends to have a normal distribution, when n is large

  explains **why we can often use normal assumption**

- But it does not always hold. It does not hold if random variables have **infinite second moment**.

Central Limit Theorem for Heavy Tails

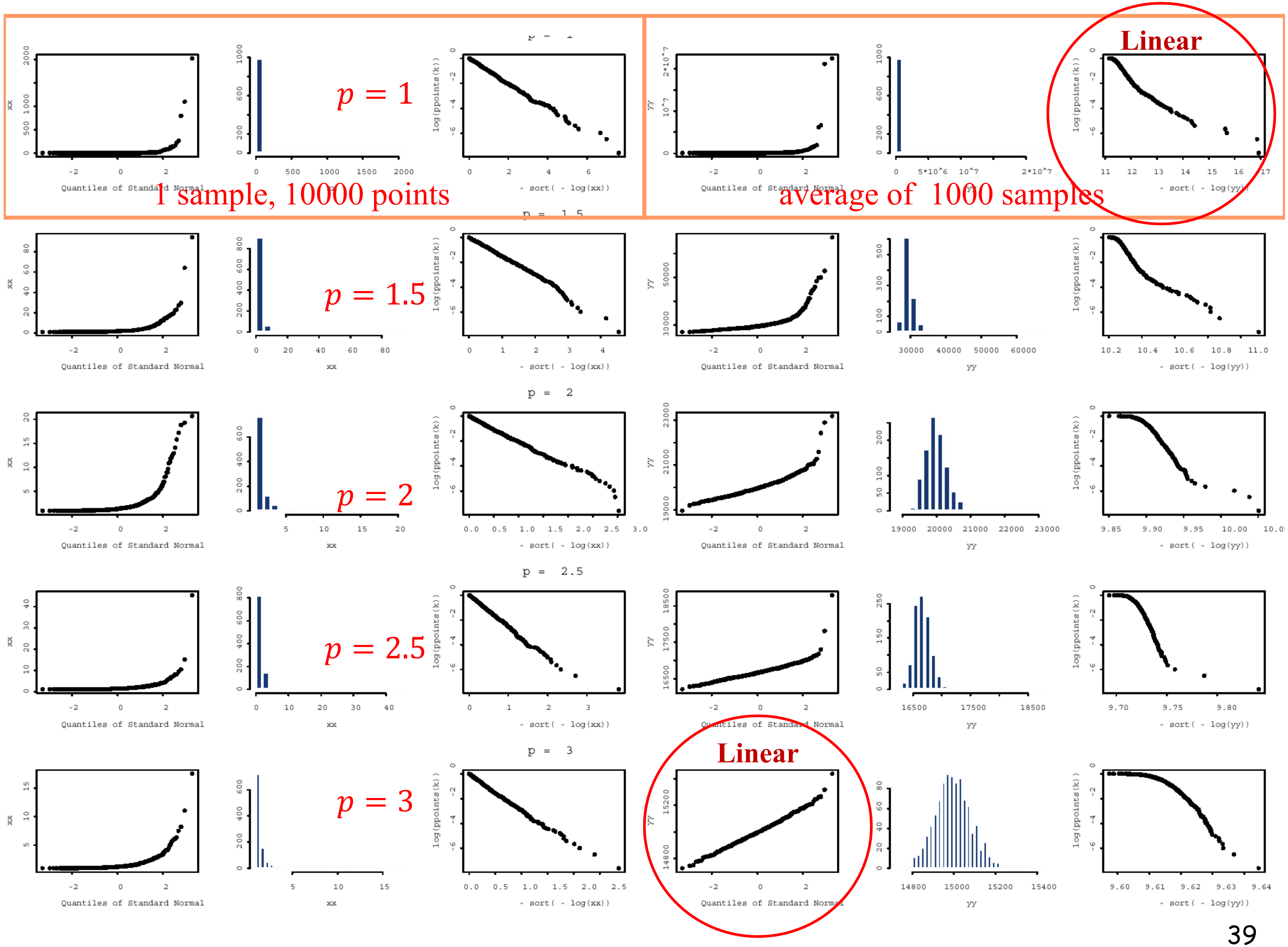


normal qqplot

histogram

complementary cdf
log-log

One Sample of 10000 points
Pareto $p = 1$



p = 1

p = 1.5

p = 2

p = 2.5

p = 3

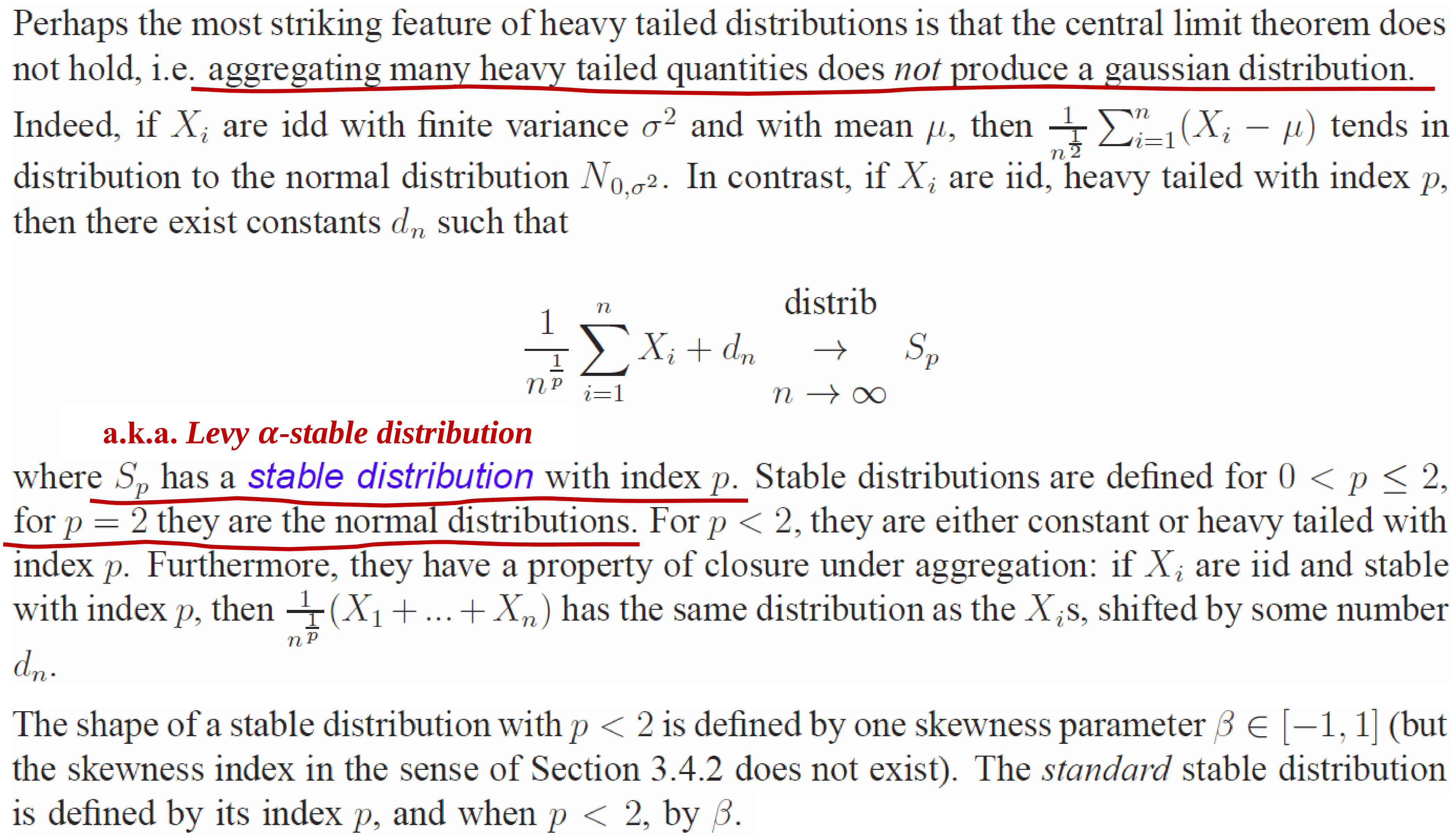
1 sample, 10000 points

average of 1000 samples

Linear

Linear

Convergence for heavy tailed distributions



Perhaps the most striking feature of heavy tailed distributions is that the central limit theorem does not hold, i.e. aggregating many heavy tailed quantities does *not* produce a gaussian distribution.

Indeed, if X_i are idd with finite variance σ^2 and with mean μ, then $\frac{1}{n^{\frac{1}{2}}}\sum_{i=1}^{n}(X_i - \mu)$ tends in distribution to the normal distribution N_{0,σ^2}. In contrast, if X_i are iid, heavy tailed with index p, then there exist constants d_n such that

$$\frac{1}{n^{\frac{1}{p}}}\sum_{i=1}^{n} X_i + d_n \underset{n \to \infty}{\overset{\text{distrib}}{\to}} S_p$$

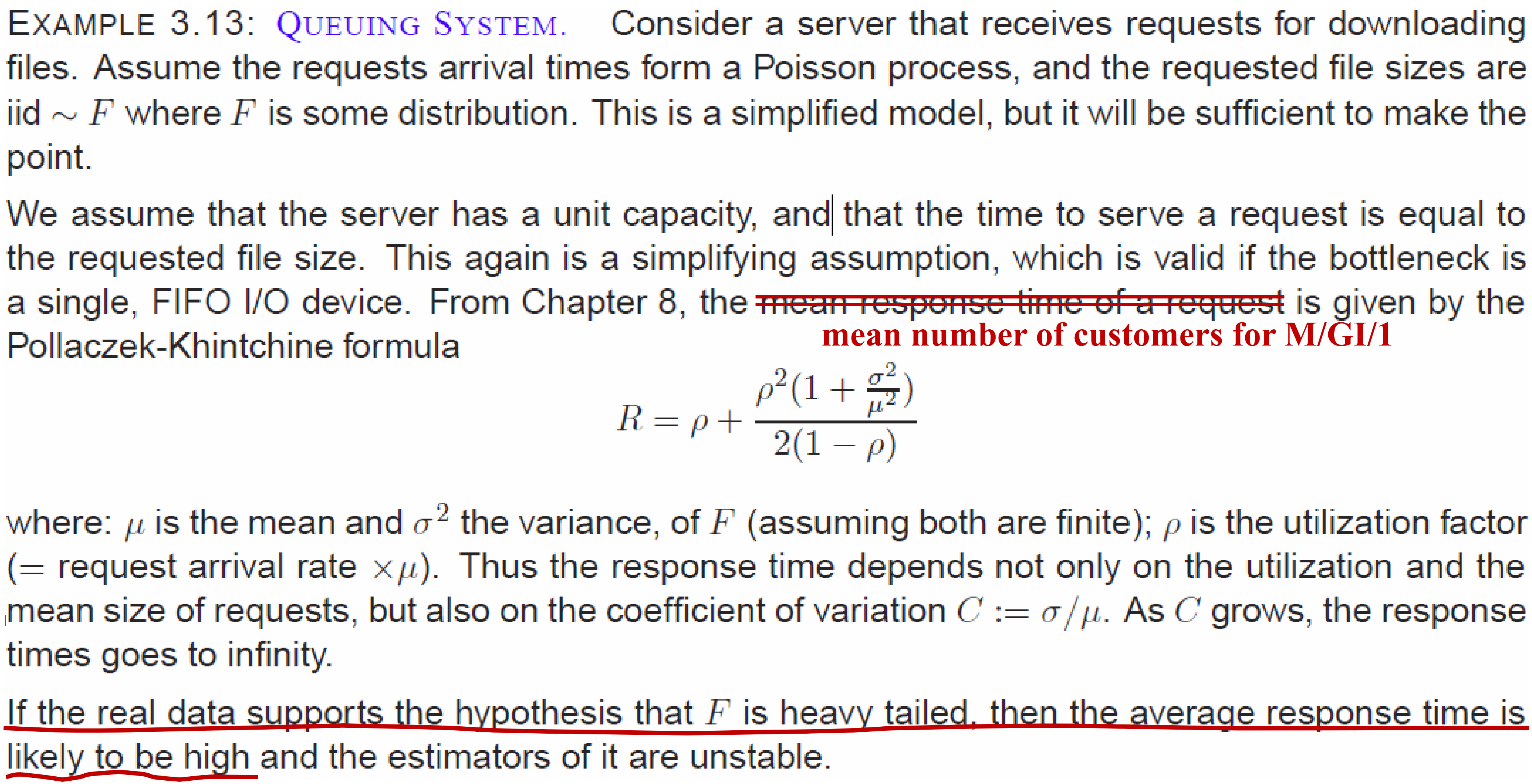
a.k.a. *Levy α-stable distribution*

where S_p has a *stable distribution* with index p. Stable distributions are defined for $0 < p \leq 2$, for $p = 2$ they are the normal distributions. For $p < 2$, they are either constant or heavy tailed with index p. Furthermore, they have a property of closure under aggregation: if X_i are iid and stable with index p, then $\frac{1}{n^{\frac{1}{p}}}(X_1 + ... + X_n)$ has the same distribution as the X_is, shifted by some number d_n.

The shape of a stable distribution with $p < 2$ is defined by one skewness parameter $\beta \in [-1, 1]$ (but the skewness index in the sense of Section 3.4.2 does not exist). The *standard* stable distribution is defined by its index p, and when $p < 2$, by β.

Generalized Central Limit Theorem (No closed-form expression for S_p)

Importance of Second Moment



EXAMPLE 3.13: QUEUING SYSTEM. Consider a server that receives requests for downloading files. Assume the requests arrival times form a Poisson process, and the requested file sizes are iid $\sim F$ where F is some distribution. This is a simplified model, but it will be sufficient to make the point.

We assume that the server has a unit capacity, and that the time to serve a request is equal to the requested file size. This again is a simplifying assumption, which is valid if the bottleneck is a single, FIFO I/O device. From Chapter 8, the ~~mean response time of a request~~ **mean number of customers for M/GI/1** is given by the Pollaczek-Khintchine formula

$$R = \rho + \frac{\rho^2(1 + \frac{\sigma^2}{\mu^2})}{2(1-\rho)}$$

where: μ is the mean and σ^2 the variance, of F (assuming both are finite); ρ is the utilization factor ($=$ request arrival rate $\times \mu$). Thus the response time depends not only on the utilization and the mean size of requests, but also on the coefficient of variation $C := \sigma/\mu$. As C grows, the response times goes to infinity.

If the real data supports the hypothesis that F is heavy tailed, then the average response time is likely to be high and the estimators of it are unstable.

Distribution Fitting Example : Censored Data

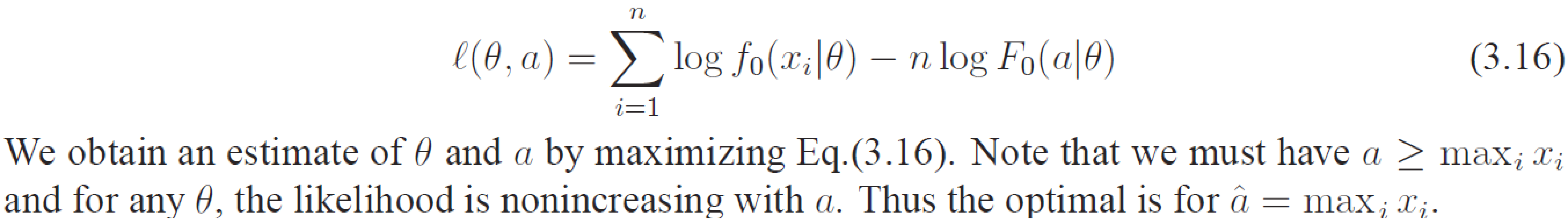
- We want to fit a log normal distrib, but we have only data samples with values less than some max.
- Lognormal is fat-tailed so we cannot ignore the tail
  - Not heavy-tailed but “heavier” than exponential & normal

- Idea: use the model

$$f_X(x) = \frac{1}{F_0(a)} f_0(x) \mathbf{1}_{\{x \le a\}}$$

and estimate

1. Normazlied CDF $F_0(a)$
2. truncation threshold a



$$\ell(\theta, a) = \sum_{i=1}^{n} \log f_0(x_i|\theta) - n \log F_0(a|\theta) \tag{3.16}$$

We obtain an estimate of θ and a by maximizing Eq.(3.16). Note that we must have $a \geq \max_i x_i$ and for any θ, the likelihood is nonincreasing with a. Thus the optimal is for $\hat{a} = \max_i x_i$.

Here, F_0 is the log-normal distribution with parameters μ and σ. Instead of brute force optimization, we can have more insight as follows. We have to maximize $\ell(\mu, \sigma)$ over $\mu \in \mathbb{R}$, $\sigma > 0$, with

$$\begin{aligned} \ell(\mu, \sigma) &= -n \ln(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln x_i - \mu)^2 - n \ln N_{0,1}\left(\mu + \sigma \ln a\right) \\ & \cancel{-\frac{n}{2} \ln(2\pi) - \sum_{i=1}^{n} \ln x_i} \end{aligned} \tag{3.17}$$

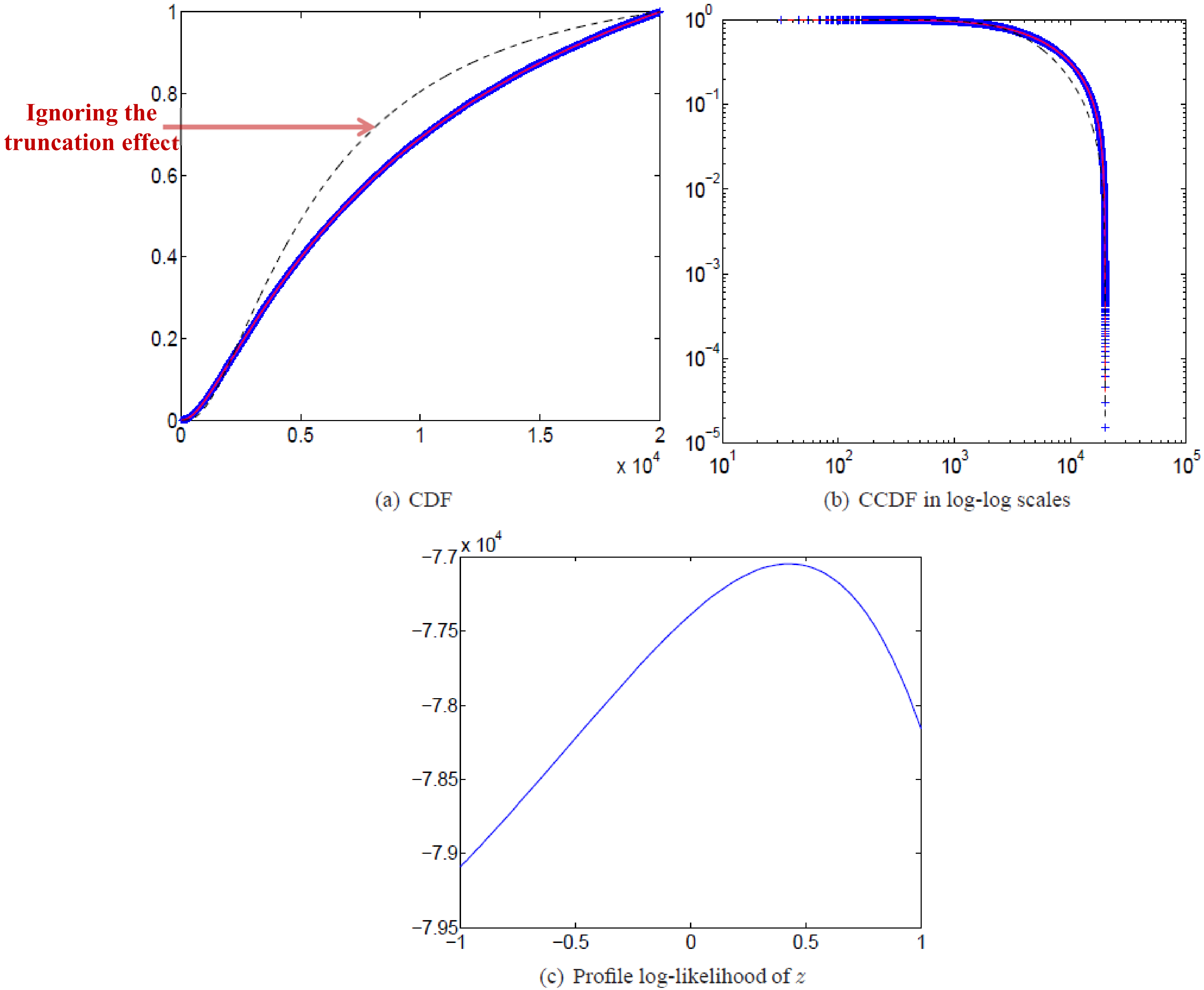
We can ignore the last two terms, which do not depend on (μ, σ). We can also do a change of variables by taking as parameters σ, z instead of σ, μ, with

$$z = \frac{\ln a - \mu}{\sigma} \tag{3.18}$$

For a fixed z, the optimization problem has a closed form solution (obtained by computing the derivative with respect to σ); the maximum likelihood is obtained for $\sigma = \hat{\sigma}(z)$ with

$$\underline{\hat{\sigma}(z) = \frac{-\beta z + \sqrt{4s^2 + \beta^2(4 + z^2)}}{2}} \tag{3.19}$$

$$\text{with } \beta = \ln a - y_1, \quad y_1 = \frac{1}{n} \sum_{i=1}^{n} \ln x_i, \quad s^2 = \frac{1}{n} \sum_{i=1}^{n} (\ln x_i - y_1)^2$$



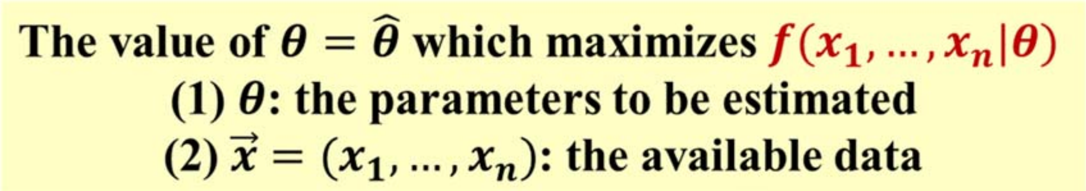
(a) CDF

(b) CCDF in log-log scales

(c) Profile log-likelihood of z

Figure 3.7: Fitting Censored Data in Example 3.10. The data set is an iid sample of a truncated log-normal distribution. Thick lines: data set; plain lines: fit obtained with a technique for censored data; dashed lines: fit obtained when ignoring the censored data.

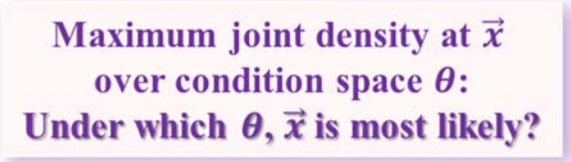
Conclusion



The value of $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ which maximizes $\boldsymbol{f(x_1, \dots, x_n|\theta)}$

(1) $\boldsymbol{\theta}$: the parameters to be estimated

(2) $\vec{\boldsymbol{x}} = (\boldsymbol{x_1, \dots, x_n})$: the available data



Maximum joint density at $\vec{x}$
over condition space θ:
Under which θ, $\vec{x}$ is most likely?

<u>"Maximum Likelihood Estimator"</u>

Simplistic yet very versatile model fitting technique.

<u>Another example of MLE on distribution fitting in Chapter 3.4</u>

Combinations of Distribution

e.g., Sometimes it is impossible to find a distribution that fits both the tail and body of the data.