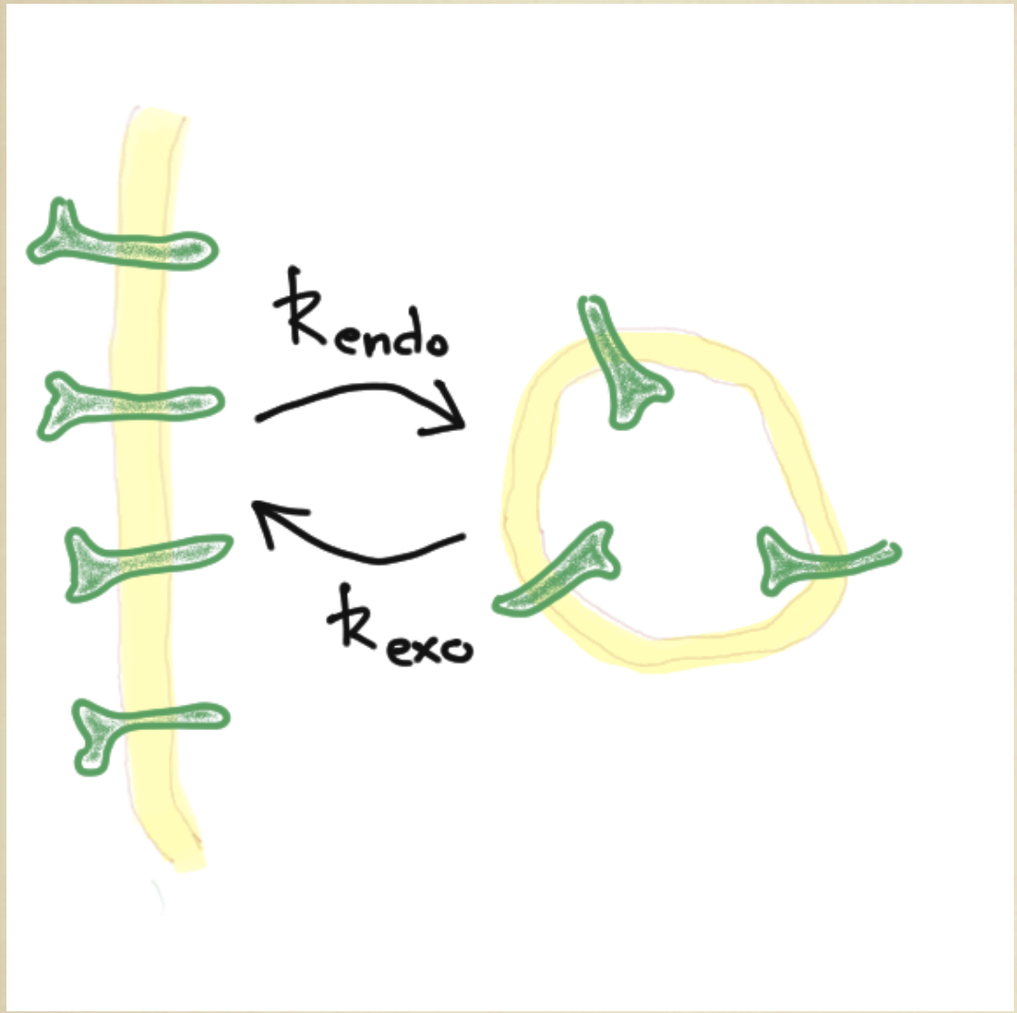
Experimental data analysis
Lecture 1

Dodo Das

Motivation

- Model for integrin turnover based on FRAP data (Week 1)

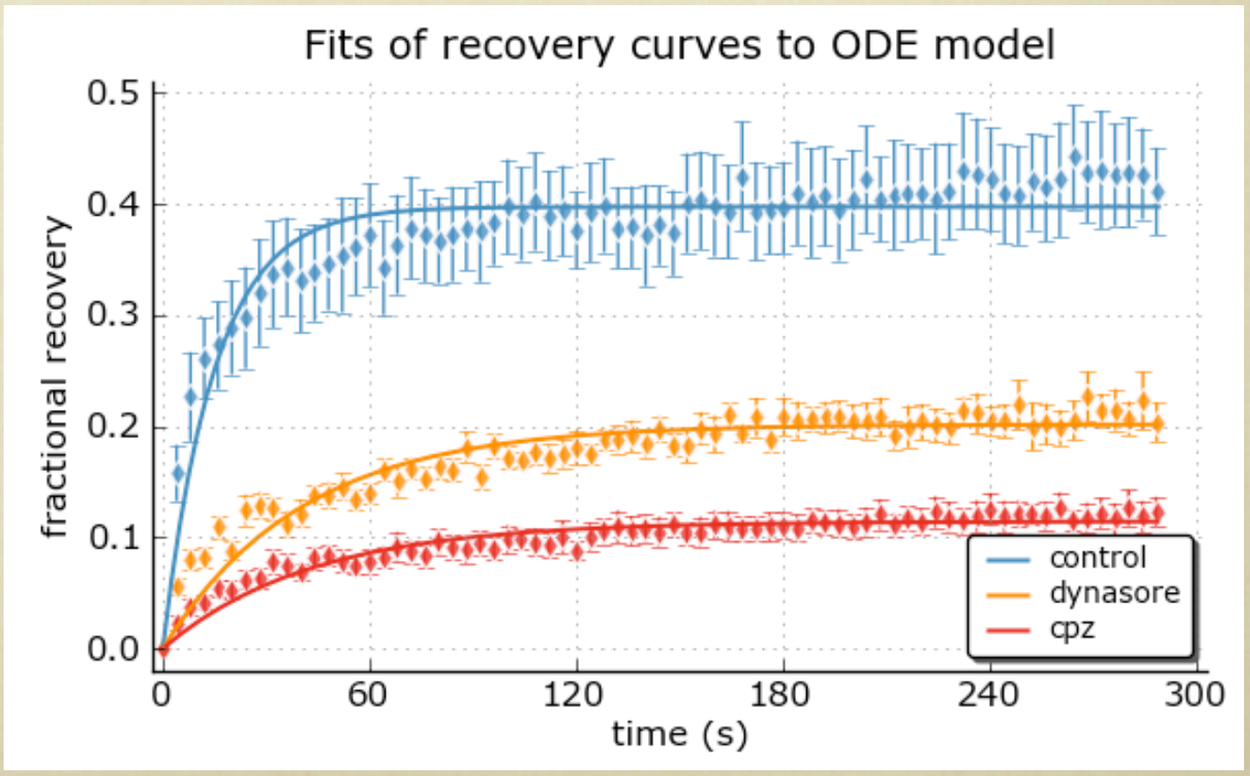


$$f(t) = f_{max}\left(1 - e^{-t/\tau}\right)$$

$$k_{endo} = f_{max}/\tau$$

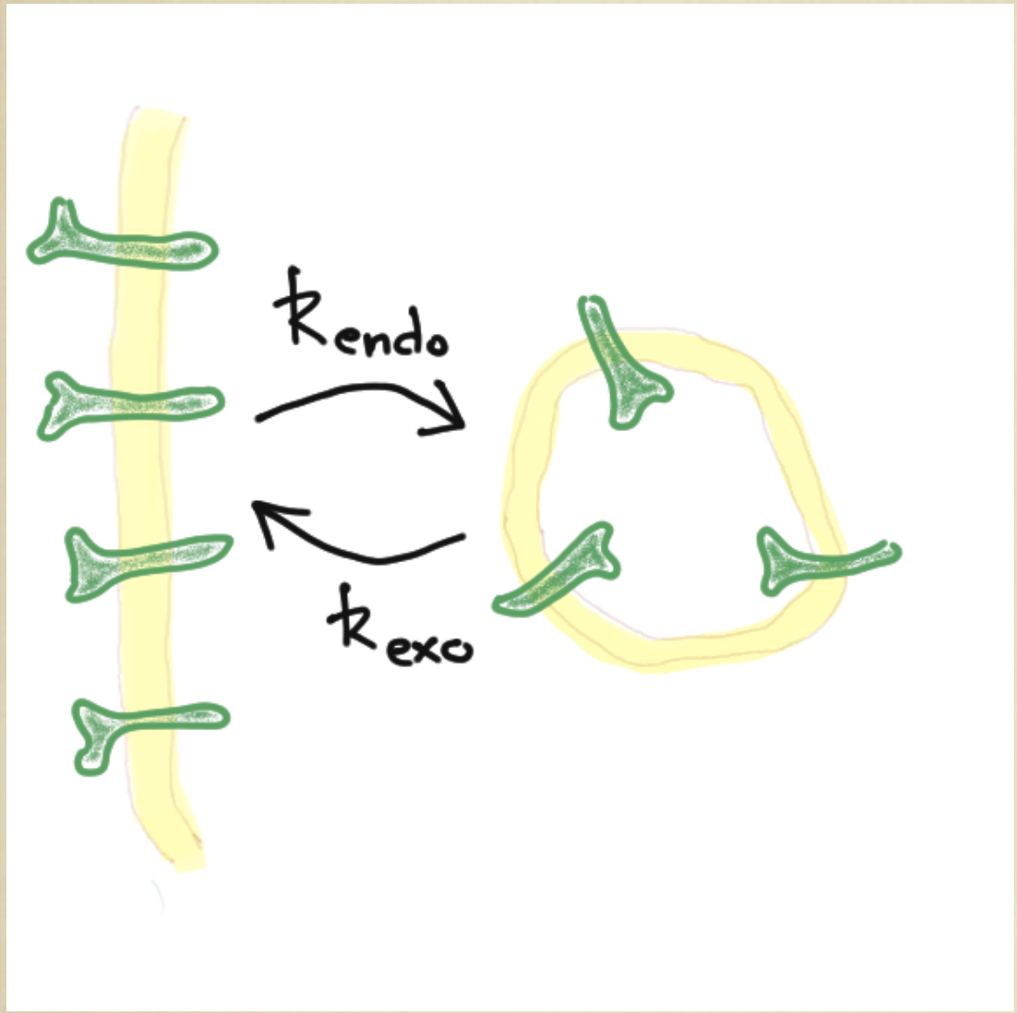
$$k_{exo} = (1 - f_{max})/\tau$$

$$\frac{dM}{dt} = -k_{endo}M + k_{exo}V$$



Motivation

- Model for integrin turnover based on FRAP data (Week 1)

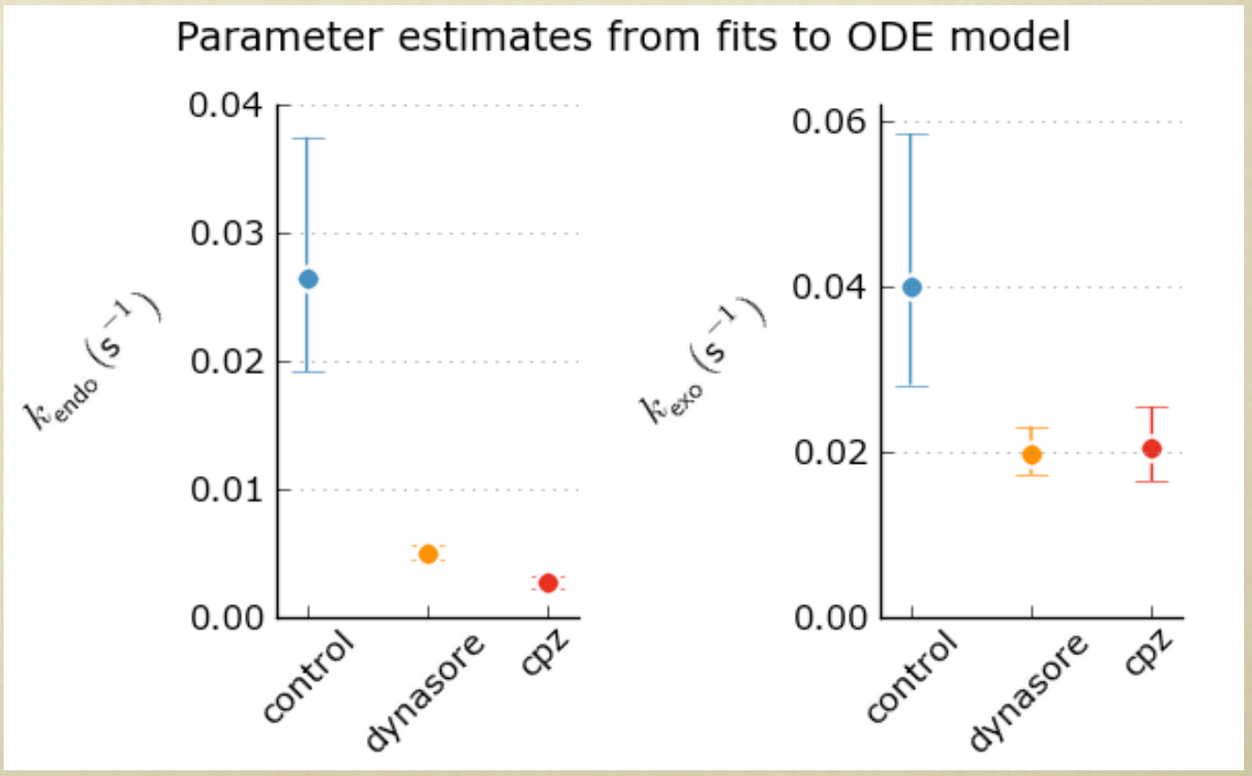


$$f(t) = f_{max}\left(1 - e^{-t/\tau}\right)$$

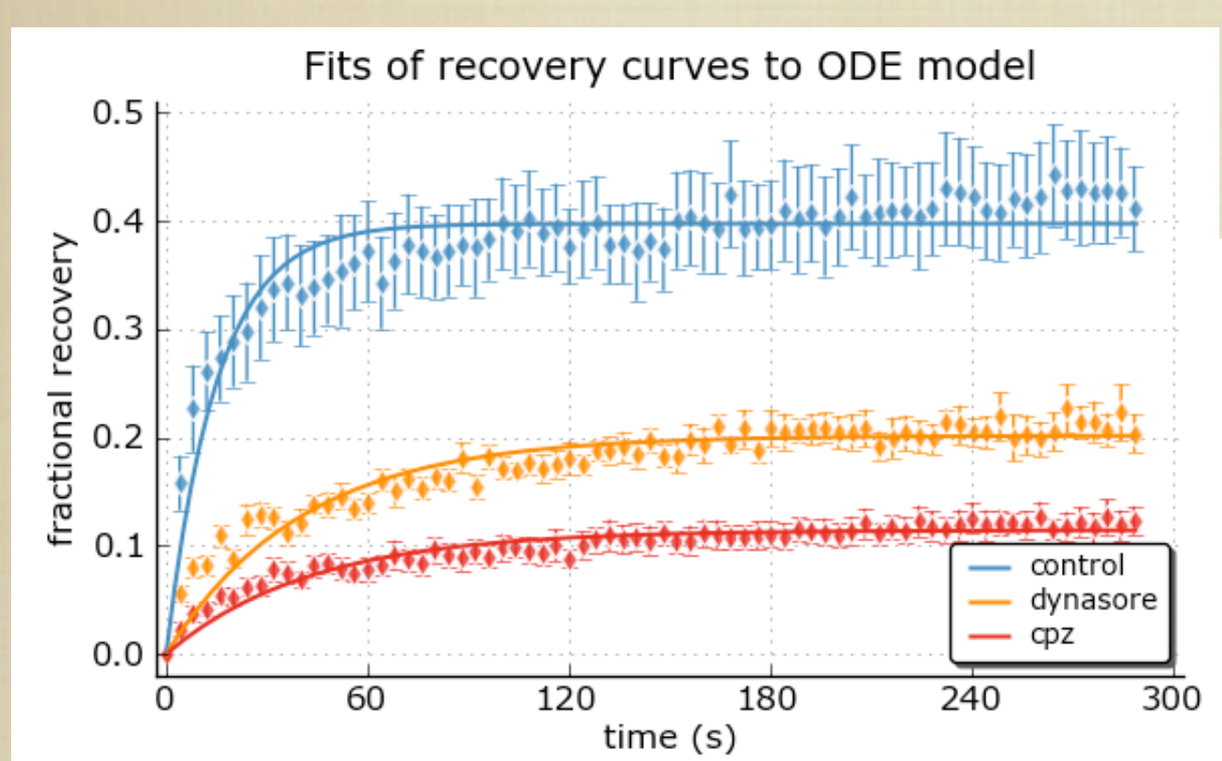
$$k_{endo} = f_{max}/\tau$$

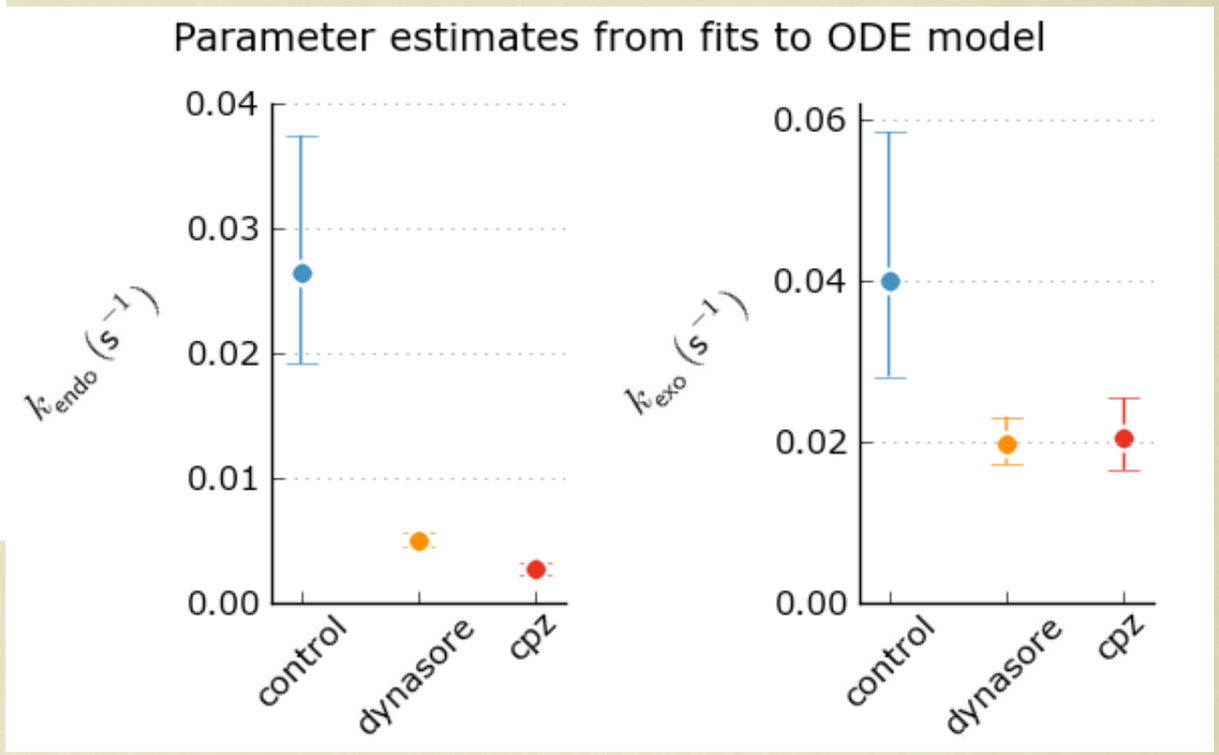
$$k_{exo} = (1 - f_{max})/\tau$$

$$\frac{dM}{dt} = -k_{endo}M + k_{exo}V$$



Questions





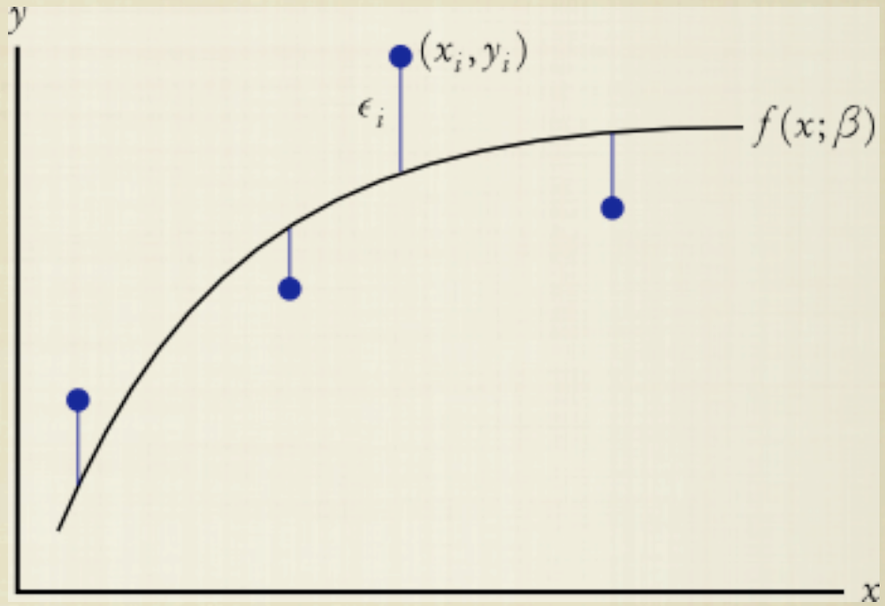
- How did we generate the best fits to experimental data?
- How did we compute the errors in our parameter estimates?

Practical goals

- "Fit" noisy experimental data to a parametric mathematical model.
- Estimate parameter values and errors in those estimates.
- Compare quality of fits with different models and rank the models.

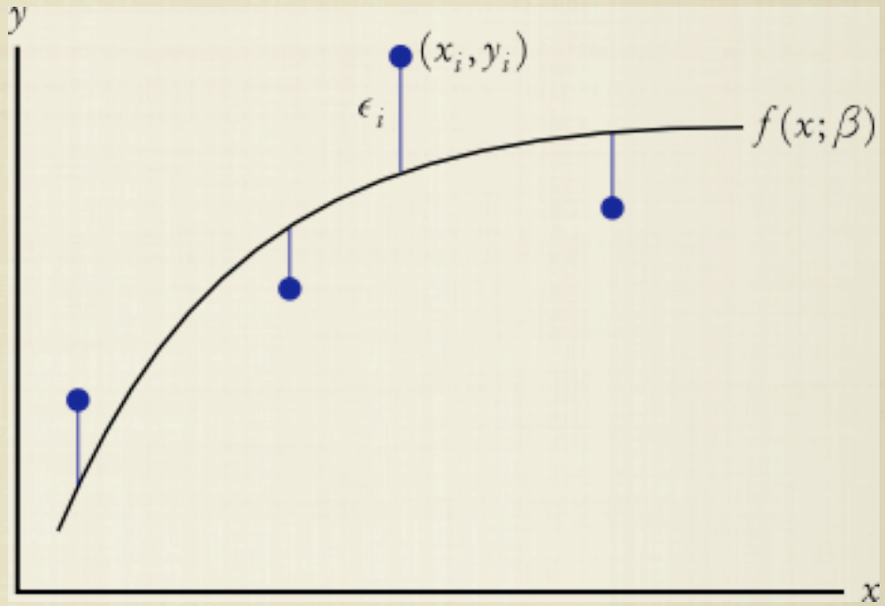
The likelihood of a model

- Notation:
  - Data/observations: $(\mathbf{X}_i, y_i)$, where $\mathbf{X}_i$ may be vectors.
  - Parametric model: $y_i = f(\mathbf{X}_i; \boldsymbol{\beta}) + \varepsilon_i$
  - $\boldsymbol{\beta}$ is a vector of (undetermined) parameters for the model.
  - $\varepsilon_i = y_i - f(\mathbf{X}i; \boldsymbol{\beta})$ are 'residuals', assumed to follow some error distribution, eg: a normal distribution.



The likelihood of a model

- Notation:
  - Data/observations: $(\mathbf{X}_i, y_i)$, where $\mathbf{X}_i$ may be vectors.
  - Parametric model: $y_i = f(\mathbf{X}_i; \beta) + \varepsilon_i$
  - β is a vector of (undetermined) parameters for the model.
  - $\varepsilon_i = y_i - f(\mathbf{X}i; \beta)$ are 'residuals', assumed to follow some error distribution, eg: a normal distribution.



How can we choose β such that the predictions $f(\mathbf{X}i; \beta)$ most closely resemble the data y_i?

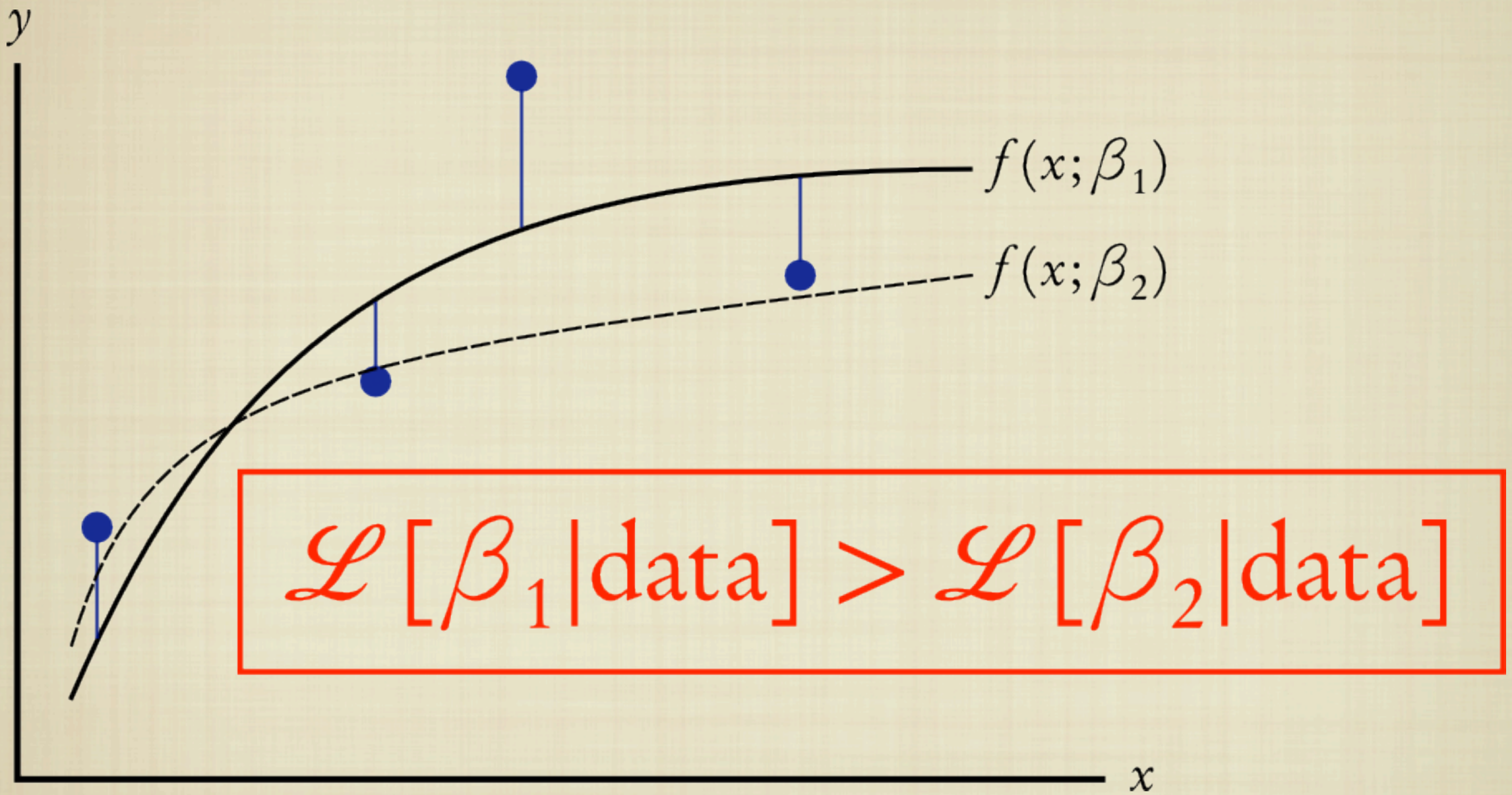
Data fitting = Likelihood maximization

- Define the likelihood of a model, given some data:

$$\begin{aligned}\mathcal{L}\left[\text{model} \mid \text{data}\right] &= \mathcal{L}\left[\beta \mid \{(\mathbf{X_i}, y_i)\}\right] \\ &\propto P\left[\{\epsilon_i\}\right] \\ &= \prod_i P(\epsilon_i)\end{aligned}$$

‘Fitting’ the data to the model is equivalent to maximizing this likelihood function with respect to the model parameters.

Data fitting = Likelihood maximization



'Fitting' the data to the model is equivalent to maximizing this likelihood function with respect to the model parameters.

Likelihood maximization for normal errors

If we assume that the errors are independent, and identically distributed normal random variables, then:

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

$$P(\epsilon_i) \sim e^{-1/2\left(\epsilon_i/\sigma_i\right)^2}$$

$$\mathcal{L}\left[\beta|\text{data}\right] \propto \prod_i e^{-1/2\left(\epsilon_i/\sigma_i\right)^2}$$

$$\log\left(\mathcal{L}\left[\beta|\text{data}\right]\right) = -\frac{1}{2}\sum_i \left(\frac{\epsilon_i}{\sigma_i}\right)^2 + \text{const. terms}$$

Therefore, maximizing the likelihood function is equivalent to minimizing the sum of squared residuals:

Likelihood maximization for normal errors

Therefore, maximizing the likelihood function is equivalent to minimizing the sum of squared residuals:

$$\mathrm{SSR} = \sum_i \left(\frac{\epsilon_i}{\sigma_i} \right)^2 = \sum_i \left[\frac{y_i - f(x_i; \beta)}{\sigma_i} \right]^2$$

Least squares regression ⇔ Likelihood maximization assuming independent, normally distributed measurement errors.

Likelihood maximization for normal errors

Therefore, maximizing the likelihood function is equivalent to minimizing the sum of squared residuals:

$$\mathrm{SSR} = \sum_i \left(\frac{\epsilon_i}{\sigma_i} \right)^2 = \sum_i \left[\frac{y_i - f(x_i; \beta)}{\sigma_i} \right]^2$$

Least squares regression ⇔ Likelihood maximization assuming independent, normally distributed measurement errors.

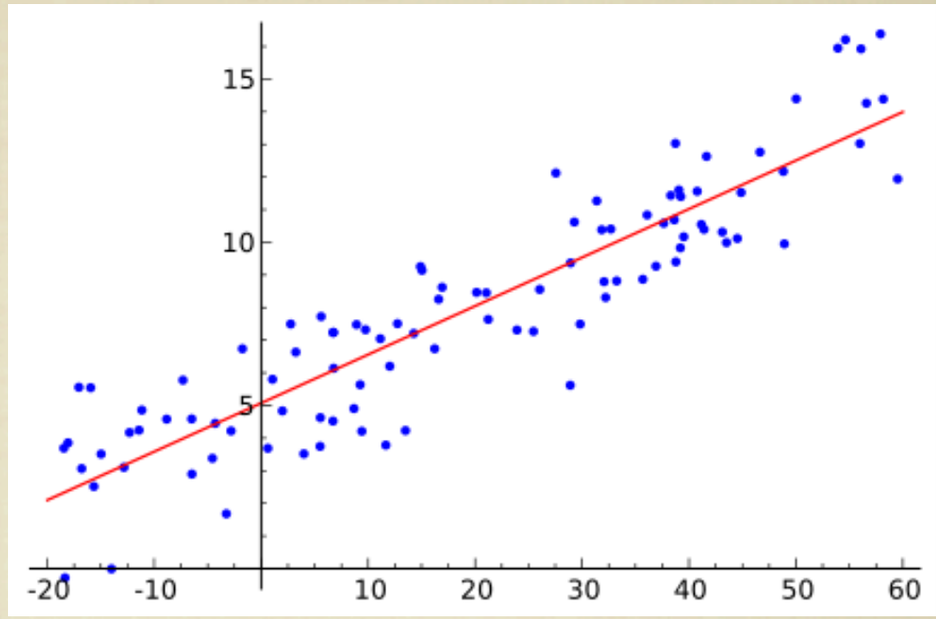
The maximum-likelihood parameter estimate is defined to be:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i \left[\frac{y_i - f(x_i; \beta)}{\sigma_i} \right]^2$$

Analytical solution: Simple linear regression

- Fit (x_i, y_i) pairs to a straight line

$$y_i = f(x_i; a, b)$$
$$= a + b x_i + \epsilon_i$$



$$\mathrm{SSR} = \chi^2(a,b) = \sum_{i=1}^{n} \left(\frac{y_i - a - b x_i}{\sigma_i} \right)^2$$

$$\frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{n} \frac{y_i - a - b x_i}{\sigma_i^2} = 0$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{n} \frac{x_i (y_i - a - b x_i)}{\sigma_i^2} = 0$$

Analytical solution: Simple linear regression

Can rewrite equations in the following form (HW?)

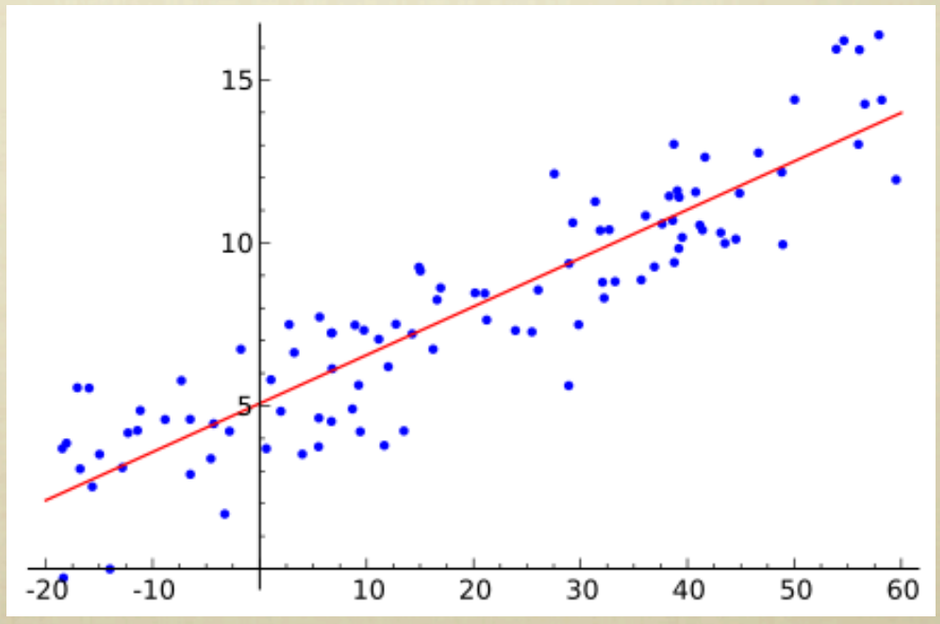
$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

where

$$S = \sum 1/\sigma_i^2 \qquad S_x = \sum x_i/\sigma_i^2 \qquad S_y = y_i/\sigma_i^2$$

$$S_{xx} = \sum x_i^2/\sigma_i^2 \qquad S_{xy} = \sum x_i y_i/\sigma_i^2$$

and solve for the maximum likelihood parameter estimates.



Multiple linear regression: Normal equations

- In general, we can have more than one independent variable.

$$y_i = f(\mathbf{x}_i; \beta) = \sum_k \beta_k (x_i)_k + \epsilon_i$$

- To estimate the best-fit values of β_k create a 'Design Matrix'

$$A = \begin{pmatrix} x_{11}/\sigma_1 & x_{12}/\sigma_1 & \dots & x_{1n}/\sigma_1 \\ x_{21}/\sigma_2 & x_{22}/\sigma_2 & \dots & x_{2n}/\sigma_2 \\ \dots & \dots & \dots & \dots \\ x_{m1}/\sigma_m & x_{m2}/\sigma_m & \dots & x_{mn}/\sigma_m \end{pmatrix}$$

- Solve the 'Normal equations'

$$\boxed{\left(A^T A\right) \hat{\beta} = A^T \mathbf{y}}$$

Extension: General linear regression

- The 'linear' in linear regression refers to linearity with respect to the parameters.

- The parameters themselves can be coefficients to nonlinear basis functions, eg: Polynomial regression:

$$y(\mathbf{X}|\beta) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots \beta_n x_i^n + \epsilon_i$$

- Can construct design matrix as before, and solve the normal equations to estimate maximum likelihood parameter.

Tomorrow

- Sample MATLAB code.
- Nonlinear least squares regression.
- Levenberg-Marquardt algorithm.
- Other likelihood maximization methods.
- Asymptotic error estimates.