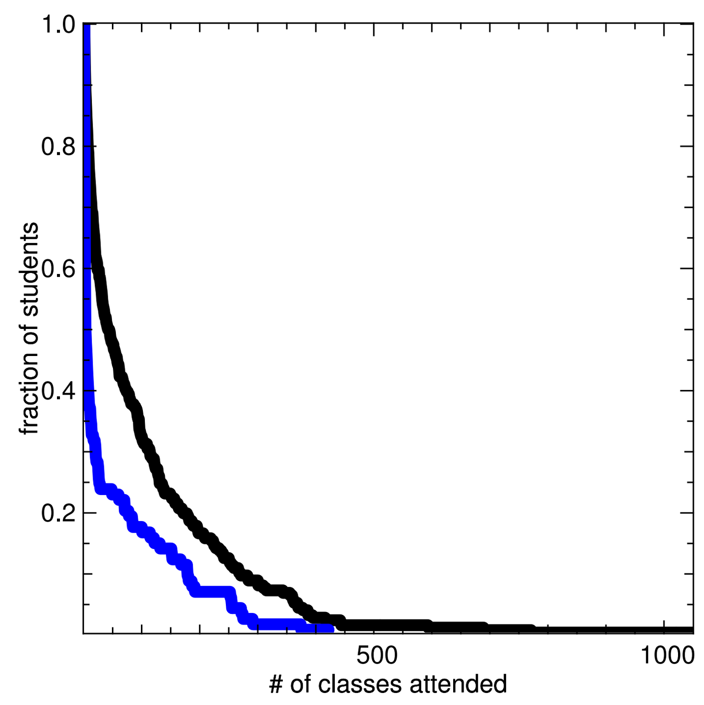
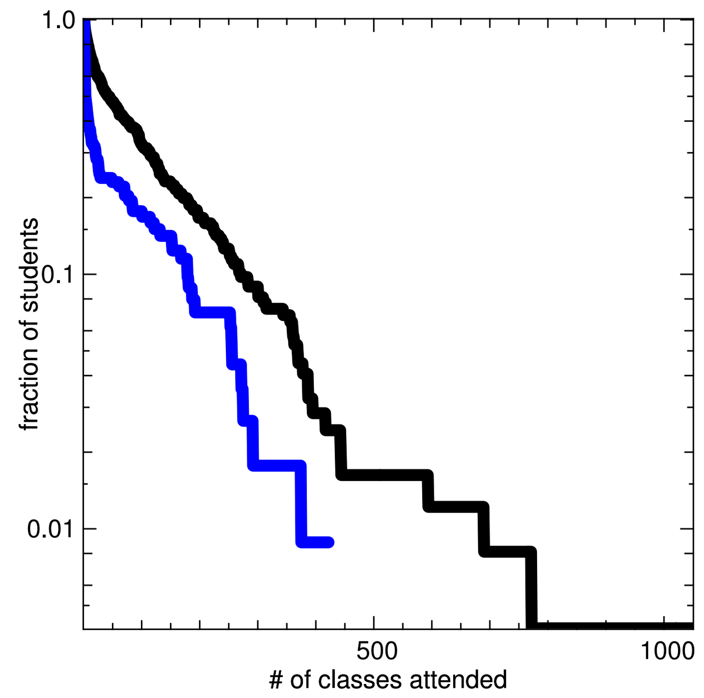
Attendance recorded by hand

	Tue Dec 2	Thu Dec 4	Tue Dec 9	Thu Dec 11	Tue Dec 16	Thu Dec 18	Tue Dec 23	Totals
Mary	✓				✓	✓		3
Noah		✓	✓					2
Helen	✓		✓	✓		✓	✓	5
Liam	✓		✓	✓		✓	✓	5
Dorothy		✓	✓		✓	✓		4
Mason	✓	✓	✓		✓	✓	✓	6
Margaret	✓	✓	✓	✓	✓	✓		6
Jacob						✓	✓	2

Raw data



Logarithmic y-axis



$$S(t) = \exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)$$

Weibull distribution

$$-\ln S(t) = \left(\frac{t}{\tau}\right)^{\gamma}$$

can be cast into a linear form

$$\ln(-\ln S(t)) = \gamma \ln\left(\frac{t}{\tau}\right) = \gamma \ln(t) - \gamma \ln(\tau)$$

Linear least-squares fitting

Model tests of

```
Will a student who's attended t classes
drop out and never attend a t+1th class?
```

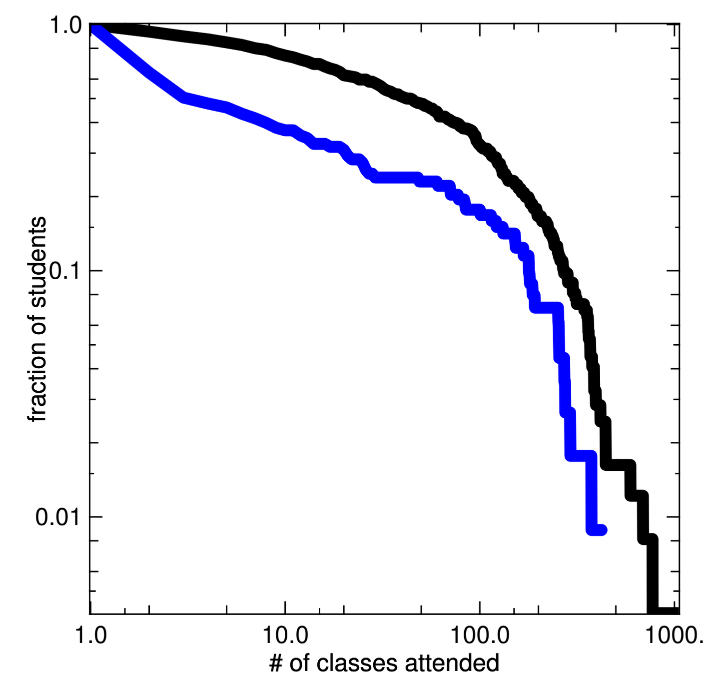
as independent Bernoulli trials

$$f(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$

$$f(y) = \begin{cases} p & y = 1 \\ 1-p & y = 0 \end{cases}$$

$$f(x,y) = \begin{cases} p^2 & x=1,\ y=1 \\ p(1-p) & x=0,\ y=1 \\ p(1-p) & x=1,\ y=0 \\ (1-p)^2 & x=0,\ y=0 \end{cases}$$

Logarithmic x- and y-axis



Survival Function

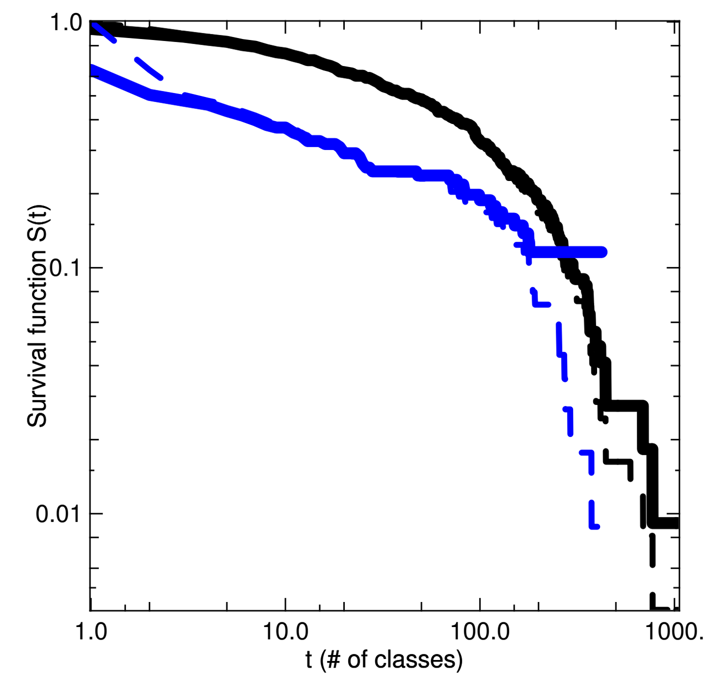
Kaplan-Meier Estimator

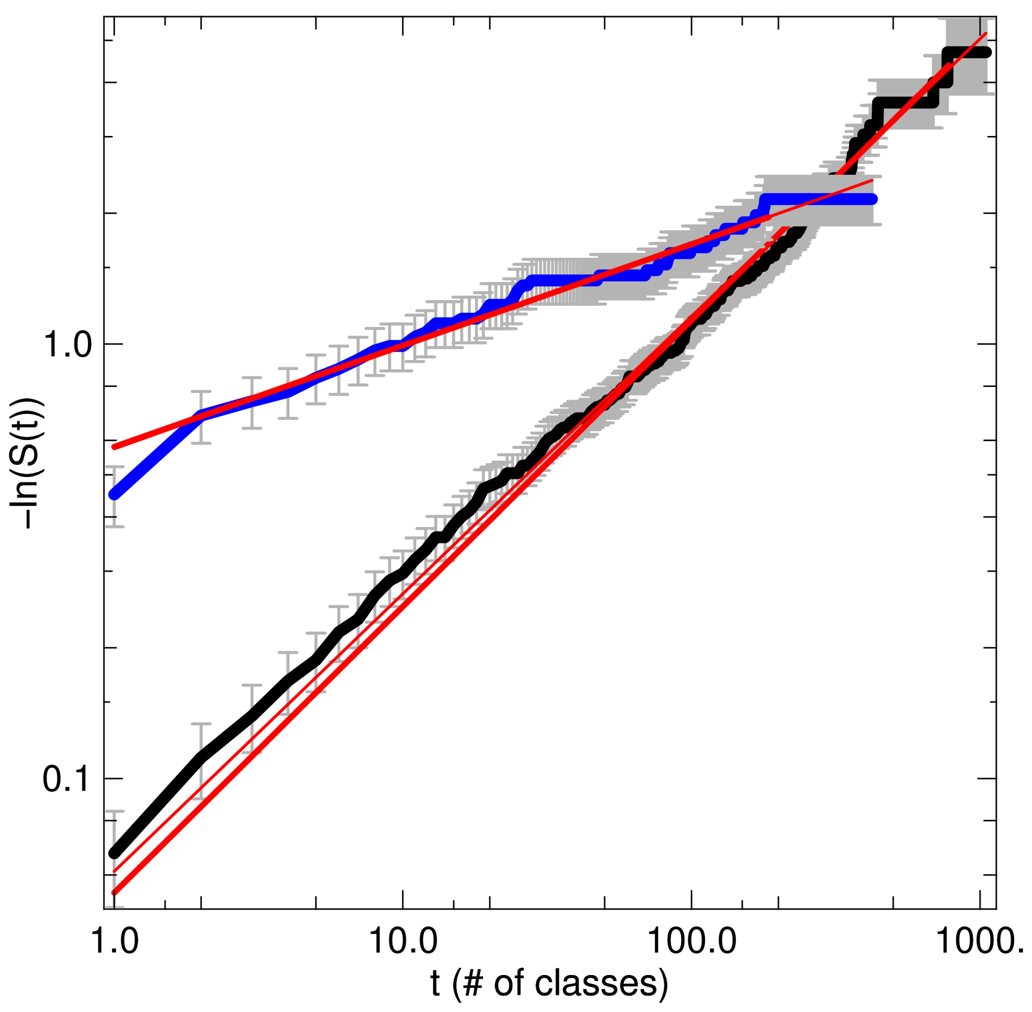
also called product-limit estimator

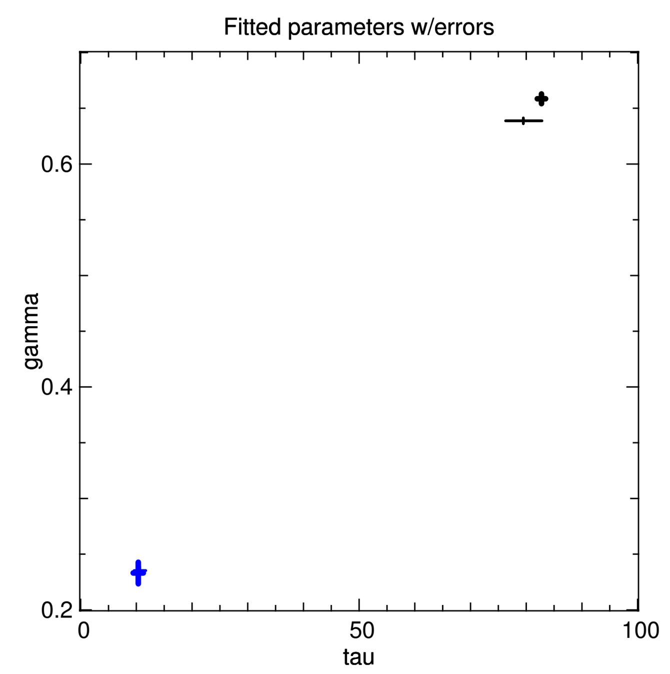
accounts for censoring of data

$$\hat{S}(t) = \prod_{i=1}^{t} \left(\frac{n_i - d_i}{n_i}\right)$$

$$\ln\left(\hat{S}(t)\right) = \sum_{i=1}^{t} \ln\left(\frac{n_i - d_i}{n_i}\right)$$







Partial derivatives with respect to the fit parameters

$$S(t,\tau,\gamma) = \exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)$$

$$\frac{\partial S}{\partial \tau} = \exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)\frac{\partial}{\partial \tau}\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right) = -\exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right) t^{\gamma} \frac{\partial}{\partial \tau}\left(\tau^{-\gamma}\right) = \left(\frac{t}{\tau}\right)^{\gamma} S(t,\tau,\gamma)\frac{\gamma}{\tau}$$

$$S(t,\tau,\gamma) = \exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)$$

$$\frac{\partial S}{\partial \gamma} = \exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)\frac{\partial}{\partial \gamma}\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right) = -\exp\left(-\left(\frac{t}{\tau}\right)^{\gamma}\right)\left(\frac{t}{\tau}\right)^{\gamma}\ln\left(\frac{t}{\tau}\right) = -\left(\frac{t}{\tau}\right)^{\gamma} S(t,\tau,\gamma)\ln\left(\frac{t}{\tau}\right)$$

Now change variables to combine multiple trials

$$u = x + y$$
$$v = x - y$$
$$x = h_1(u,v) = \frac{u+v}{2}$$
$$y = h_2(u,v) = \frac{u-v}{2}$$

$$f(u,v) = \begin{cases} p^2 & u=2,\ v=0 \\ p(1-p) & u=1,\ v=-1 \\ p(1-p) & u=1,\ v=1 \\ (1-p)^2 & u=0,\ v=0 \end{cases}$$

$$f(u) = \begin{cases} p^2 & u=x+y=2 \\ 2p(1-p) & u=1 \\ (1-p)^2 & u=0 \end{cases}$$

$$f(v) = \begin{cases} p(1-p) & v=x-y=1 \\ \left(p^2+(1-p)^2\right) & v=0 \\ p(1-p) & v=-1 \end{cases}$$

Call d_i the number of students who attended exactly i classes and then dropped out

$$\text{Mean } d_i = n_i p_i$$
$$\text{Var } d_i = n_i p_i (1-p_i)$$
$$\text{Var } \frac{d_i}{n_i(1-p_i)} = \frac{p_i}{n_i(1-p_i)}$$

When we can't derive a closed form for the probability of some function of our variables, we can approximate with a Taylor expansion around the mean

$$f(d_i) \approx f(\langle d_i \rangle) + (d_i - \langle d_i \rangle)\left.\frac{\mathrm{d}f}{\mathrm{d}d_i}\right|_{\langle d_i \rangle}$$

How many karate classes will a new student end up attending?

Matt Borthwick

April 12, 2017

This generalizes to the binomial distribution,

$$f\left(u = \sum_{i=1}^{n} x_i \,|n, p\right) = \frac{n!}{u!\,(n-u)!}p^u (1-p)^{n-u}$$

Moment-generating function

$$M_x(t) = \left\langle e^{tx} \right\rangle = \sum_x e^{tx} f(x)$$

$$\left.\frac{\mathrm{d}^n M_x(t)}{\mathrm{d}t^n}\right|_{t=0} = \left.\sum_x x^n e^{tx} f(x)\right|_{t=0} = \sum_x x^n f(x) = \langle x^n \rangle$$

Use the moment-generating function to find the mean and variance of the binomial distribution,

$$\begin{aligned}
M_u(t) &= \sum_u e^{tu} f(u) \\
&= \sum_u e^{tu} \frac{n!}{u!\,(n-u)!}p^u (1-p)^{n-u} \\
&= \sum_u \frac{n!}{u!\,(n-u)!}\left(pe^t\right)^u (1-p)^{n-u} \\
&= \left(pe^t + 1 - p\right)^n \\
\text{Mean } u = \langle u \rangle &= \left.\frac{\mathrm{d}}{\mathrm{d}t}\left(pe^t+1-p\right)^n\right|_{t=0} \\
&= \left. npe^t \left(pe^t+1-p\right)^{n-1}\right|_{t=0} \\
&= np \\
\langle u^2 \rangle &= \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(pe^t+1-p\right)^n\right|_{t=0} \\
& \left.\frac{\mathrm{d}}{\mathrm{d}t} npe^t \left(pe^t+1-p\right)^{n-1}\right|_{t=0} \\
&= \left. npe^t \left(pe^t+1-p\right)^{n-1} + np^2(n-1)\left(pe^t+1-p\right)^{n-2}\right|_{t=0} \\
&= \left. np\left(e^t\left(pe^t+1-p\right)+(n-1)p\right)\left(pe^t+1-p\right)^{n-2}\right|_{t=0} \\
&= np\left(1+(n-1)p\right) \\
\text{Var } u = \langle u^2 \rangle - \langle u \rangle^2 &= np\left(1+(n-1)p\right) - n^2p^2 \\
&= np + n^2p^2 - np^2 - n^2p^2 \\
&= np - np^2 \\
&= np(1-p)
\end{aligned}$$

$$\begin{aligned}
f(d_i) &= \ln\left(\frac{n_i-d_i}{n_i}\right) \\
\frac{\mathrm{d}f}{\mathrm{d}d_i} &= \frac{n_i}{n_i-d_i}\left(-\frac{1}{n_i}\right) \\
&= -\frac{1}{n_i-d_i} \\
\left.\frac{\mathrm{d}f}{\mathrm{d}d_i}\right|_{\langle d_i\rangle} &= \left.\frac{\mathrm{d}f}{\mathrm{d}d_i}\right|_{n_ip_i} \\
&= -\frac{1}{n_i-n_ip_i} \\
&= -\frac{1}{n_i(1-p_i)}
\end{aligned}$$

$$\begin{aligned}
f(d_i) = \ln\left(\frac{n_i-d_i}{n_i}\right) &\approx \ln\left(\frac{n_i-n_ip_i}{n_i}\right) - (d_i-n_ip_i)\frac{1}{n_i(1-p_i)} \\
&= -\frac{1}{n_i(1-p_i)}d_i + \left(\ln(1-p_i) + \frac{n_ip_i}{n_i(1-p_i)}\right) \\
\text{Var } \ln\left(\frac{n_i-d_i}{n_i}\right) &= \text{Var } \frac{d_i}{n_i(1-p_i)} \\
&= \frac{p_i}{n_i(1-p_i)} \\
&\approx \frac{\widehat{p}_i}{n_i(1-\widehat{p}_i)} \\
&= \frac{\frac{d_i}{n_i}}{n_i\left(1-\frac{d_i}{n_i}\right)} \\
&= \frac{d_i}{n_i(n_i-d_i)} \\
\text{Var } \ln\left(\hat{S}(t)\right) &= \text{Var } \sum_{i=1}^{t}\ln\left(\frac{n_i-d_i}{n_i}\right) \\
&\approx \sum_{i=1}^{t}\frac{d_i}{n_i(n_i-d_i)}
\end{aligned}$$

This is known as Greenwood's Formula

Nonlinear least-squares fitting

Find parameters that minimize

$$\chi^2(\tau,\gamma) = \frac{1}{2}\sum_t \frac{\left(S(t,\tau,\gamma) - \hat{S}(t,\tau,\gamma)\right)^2}{\text{Var } \hat{S}(t,\tau,\gamma)}$$

by approximating

$$\hat{S}(t,\tau+h_\tau,\gamma+h_\gamma) \approx \hat{S}(t,\tau,\gamma) + \frac{\partial \hat{S}(t,\tau,\gamma)}{\partial \tau}h_\tau + \frac{\partial \hat{S}(t,\tau,\gamma)}{\partial \gamma}h_\gamma = \hat{S}(t,\tau,\gamma) + \sum_{\alpha=\tau,\gamma} J_\alpha(t,\tau,\gamma)\,h_\alpha$$

from *Methods For Non-Linear Least Squares Problems* by K. Madsen, H.B. Nielsen, and O. Tingleff

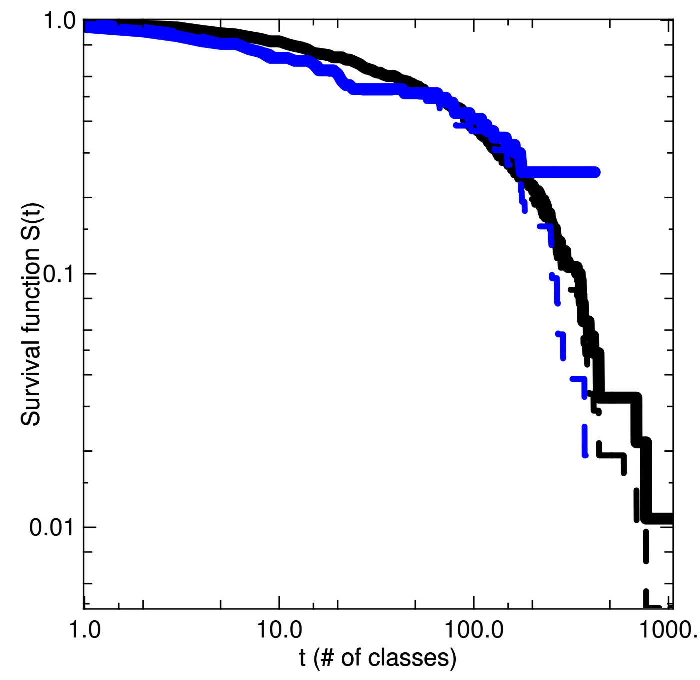
Algorithm 3.16. Levenberg–Marquardt method

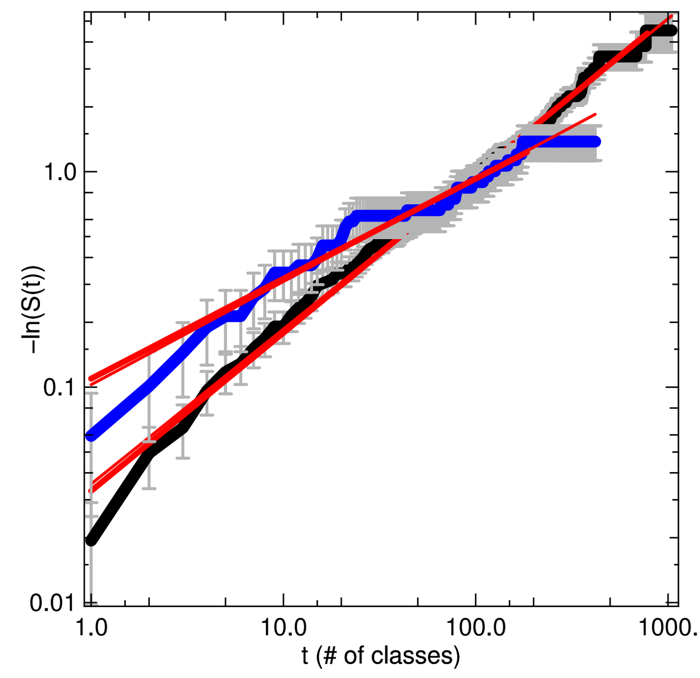
begin
  $k := 0$; $\nu := 2$; $\mathbf{x} := \mathbf{x}_0$
  $\mathbf{A} := \mathbf{J}(\mathbf{x})^\top \mathbf{J}(\mathbf{x})$; $\mathbf{g} := \mathbf{J}(\mathbf{x})^\top \mathbf{f}(\mathbf{x})$
  found $:= (\|\mathbf{g}\|_\infty \le \varepsilon_1)$; $\mu := \tau * \max\{a_{ii}\}$
  while (**not** *found*) **and** $(k < k_{\max})$
    $k := k+1$; Solve $(\mathbf{A} + \mu\mathbf{I})\mathbf{h}_{\text{lm}} = -\mathbf{g}$
    if $\|\mathbf{h}_{\text{lm}}\| \le \varepsilon_2(\|\mathbf{x}\| + \varepsilon_2)$
      found := **true**
    else
      $\mathbf{x}_{\text{new}} := \mathbf{x} + \mathbf{h}_{\text{lm}}$
      $\varrho := (F(\mathbf{x}) - F(\mathbf{x}_{\text{new}}))/(L(\mathbf{0}) - L(\mathbf{h}_{\text{lm}}))$
      if $\varrho > 0$ {step acceptable}
        $\mathbf{x} := \mathbf{x}_{\text{new}}$
        $\mathbf{A} := \mathbf{J}(\mathbf{x})^\top \mathbf{J}(\mathbf{x})$; $\mathbf{g} := \mathbf{J}(\mathbf{x})^\top \mathbf{f}(\mathbf{x})$
        found $:= (\|\mathbf{g}\|_\infty \le \varepsilon_1)$
        $\mu := \mu * \max\{\frac{1}{3}, 1-(2\varrho-1)^3\}$; $\nu := 2$
      else
        $\mu := \mu * \nu$; $\nu := 2 * \nu$
end

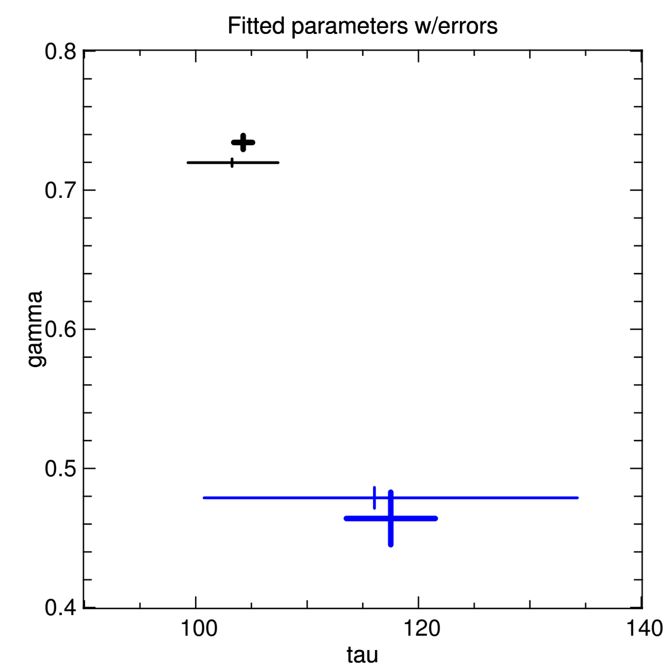
Variance of the fit parameters is given by

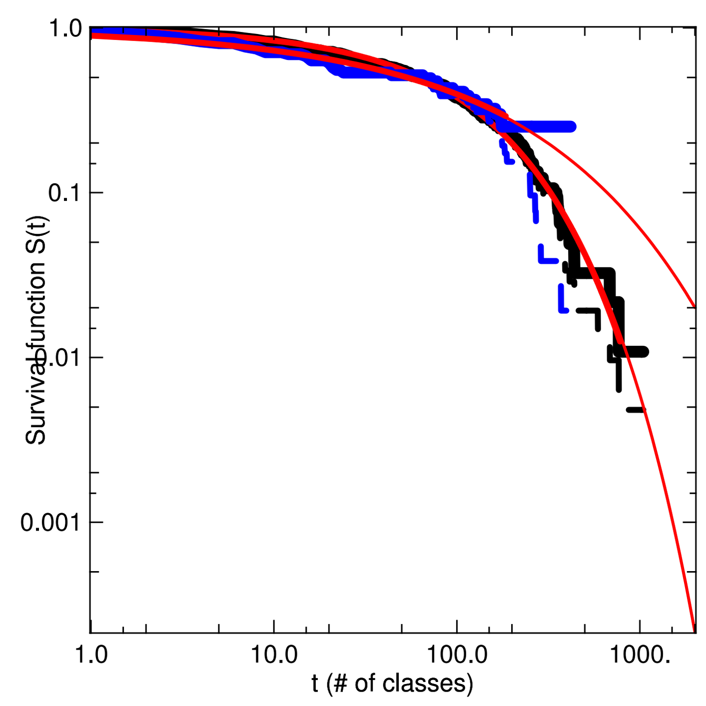
$$\text{Var } \alpha = \text{diag}\left(\left[\mathbf{J}^\mathrm{T}\mathbf{W}\mathbf{J}\right]^{-1}\right)$$

How many more classes will a student end up attending after attending his or her first four classes?









Predict future membership as a function of recruiting and survival

$$M(t) = \int_0^t R(t')\,S(t-t')\,\mathrm{d}t'$$