Statistical Data Analysis 2021/22
Lecture Week 9

London Postgraduate Lectures on Particle Physics

University of London MSc/MSci course PH4515

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Course web page via RHUL moodle (PH4515) and also

www.pp.rhul.ac.uk/~cowan/stat_course.html

Statistical Data Analysis
Lecture 9-1

- Least squares with histogram data

LS with histogram data

The fit function in an LS fit is not a pdf, but it could be proportional to one, e.g., when we fit the “envelope” of a histogram.

Suppose for example, we have an i.i.d. data sample of n values $x_1,\dots, x_n$ sampled from a pdf $f(x;\boldsymbol{\theta})$. Goal is to estimate $\boldsymbol{\theta}$.

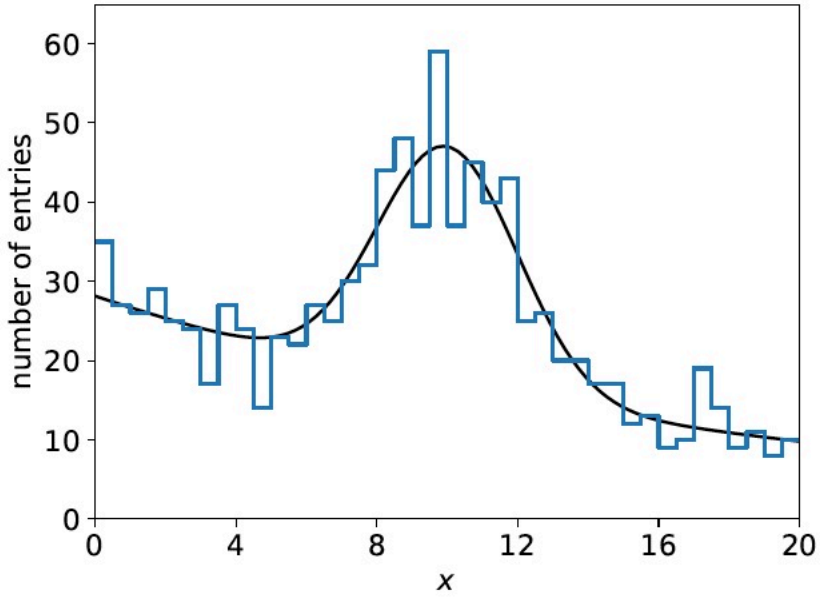
Instead of using all n values, put them in a histogram with N bins, i.e., y_i = number of entries in bin i: $\boldsymbol{y} = (y_1,\dots, y_N)$.

The model predicts mean values:

$$E[y_i] = \mu_i(\boldsymbol{\theta})$$

$$= n \int_{\mathrm{bin}\, i} f(x;\boldsymbol{\theta})\, dx$$

$$\approx n f(\underset{\text{bin centre}}{x_i};\boldsymbol{\theta})\, \underset{\text{bin width}}{\Delta x}$$



LS with histogram data (2)

The usual models:

for fixed sample size n, take $\boldsymbol{y} \sim$ multinomial,
if n not fixed, $y_i \sim$ Poisson(μ_i)

Suppose that the expected number of entries in each μ_i are all $\gg$ 1 and probability to be in any individual bin $p_i \ll 1$, one can show

$\rightarrow y_i$ indep. and $\sim$ Gauss with $\sigma_i \approx \sqrt{\mu_i}$. ($\rightarrow \sigma_i$ depends on $\boldsymbol{\theta}$).

The (log-) likelihood functions are then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i(\boldsymbol{\theta})} e^{-(y_i - \mu_i(\boldsymbol{\theta}))^2/2\sigma_i^2(\boldsymbol{\theta})}$$

$$\ln L(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\sigma_i(\boldsymbol{\theta})^2} - \sum_{i=1}^{N} \ln \sigma_i(\boldsymbol{\theta}) + C$$

LS with histogram data (3)

Still define the least-squares estimators to minimize

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\sigma_i(\boldsymbol{\theta})^2}$$

No longer equivalent to maximum likelihood (equal for $\mu_i \gg 1$).

Two possibilities for σ_i:

$\sigma_i = \sqrt{\mu_i(\boldsymbol{\theta})}$ (LS method)

$\sigma_i = \sqrt{y_i}$ (Modified LS method)

Modified LS can be easier computationally but not defined if any y_i = 0.

For either method, $\chi^2_{\text{min}} \sim$ chi-square pdf for $\mu_i \gg 1$, but this breaks down for when the μ_i are not large.

LS with histogram data — normalization

Do not “fit” the normalization, i.e., $n \longrightarrow$ free parameter ν:

$$\mu_i(\boldsymbol{\theta}, \nu) = \nu \int_{\mathrm{bin}\, i} f(x; \boldsymbol{\theta})\, dx$$

If you do this, one finds the LS estimator for ν is not n, but rather

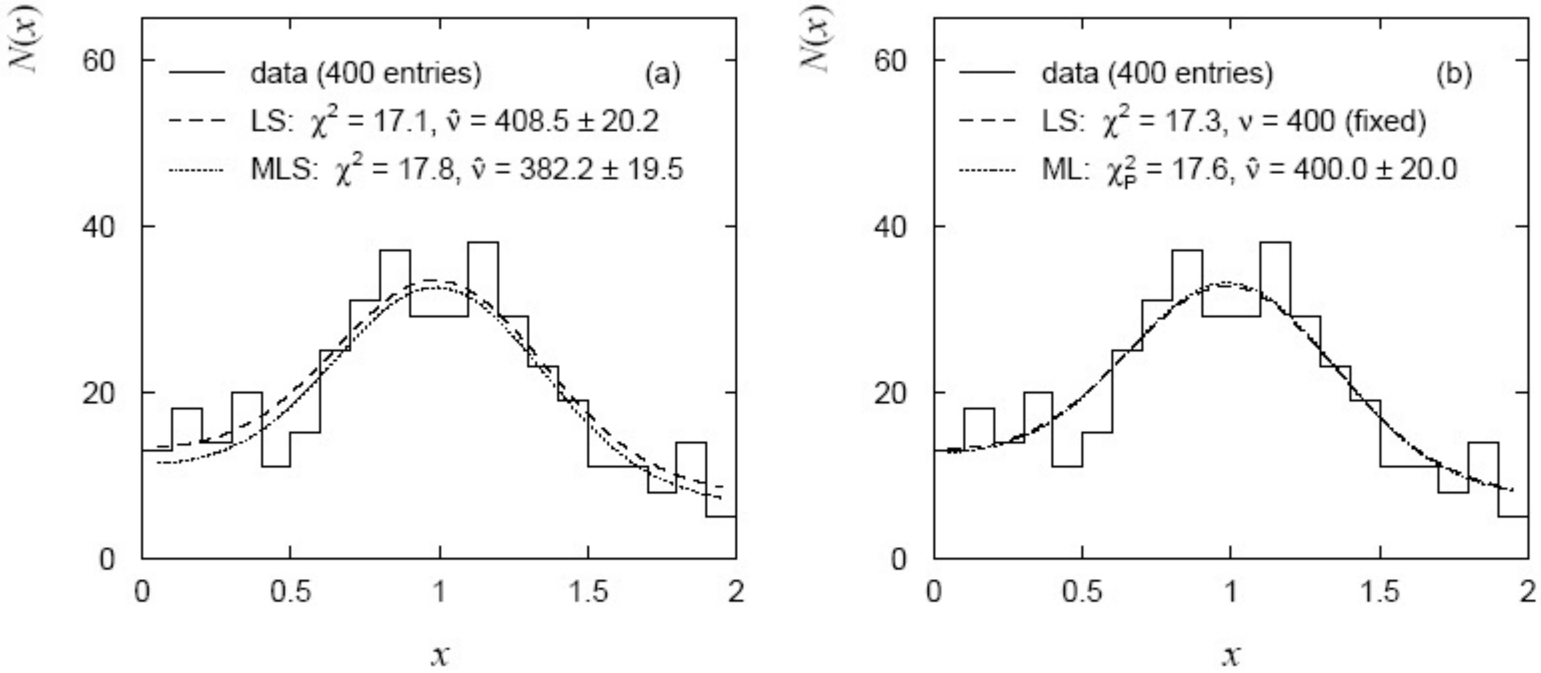
$$\hat{\nu}_{\mathrm{LS}} = n + \frac{\chi^2_{\mathrm{min}}}{2}$$

$$\hat{\nu}_{\mathrm{MLS}} = n - \chi^2_{\mathrm{min}}$$

Software may include adjustable normalization parameter as default; better to use known n.

LS normalization example

Example with $n = 400$ entries, $N = 20$ bins:



Expect $\chi^2_{\min}$ around $N - m$,

→ relative error in $\hat{\nu}$ large when N large, n small

Either get n directly from data for LS (or better, use ML).

Statistical Data Analysis
Lecture 9-2

- Goodness-of-fit from the likelihood ratio
- Wilks’ theorem
- MLE and goodness-of-fit all in one

Goodness of fit from the likelihood ratio

Suppose we model data using a likelihood $L(\boldsymbol{\mu})$ that depends on N parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$. Define the statistic

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu})}{L(\hat{\boldsymbol{\mu}})}$$

where $\hat{\boldsymbol{\mu}}$ is the ML estimator for $\boldsymbol{\mu}$. Value of $t_{\boldsymbol{\mu}}$ reflects agreement between hypothesized $\boldsymbol{\mu}$ and the data.

Good agreement means $\boldsymbol{\mu} \approx \hat{\boldsymbol{\mu}}$, so $t_{\boldsymbol{\mu}}$ is small;

Larger $t_{\boldsymbol{\mu}}$ means less compatibility between data and $\boldsymbol{\mu}$.

Quantify “goodness of fit” with p-value: $p_{\boldsymbol{\mu}} = \int_{t_{\boldsymbol{\mu},\mathrm{obs}}}^{\infty} \underset{\nwarrow \text{ need this pdf}}{f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu})} \, dt_{\boldsymbol{\mu}}$

Likelihood ratio (2)

Now suppose the parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)$ can be determined by another set of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_M)$, with $M < N$.

Want to test hypothesis that the true model is somewhere in the subspace $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ versus the alternative of the full parameter space $\boldsymbol{\mu}$. Generalize the LR test statistic to be

fit M parameters

$$t_{\boldsymbol{\mu}} = -2\ln\frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

fit N parameters

To get p-value, need pdf $f(t_{\boldsymbol{\mu}}|\boldsymbol{\mu}(\boldsymbol{\theta}))$.

Wilks' Theorem

Wilks' Theorem: if the hypothesized $\mu_i(\boldsymbol{\theta})$, $i = 1,\ldots,N$, are true for some choice of the parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$, then in the large sample limit (and provided regularity conditions are satisfied)

MLE of $(\theta_1,\ldots, \theta_M)$

$$t_{\boldsymbol{\mu}} = -2\ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

follows a chi-square distribution for $N - M$ degrees of freedom.

MLE of $(\mu_1,\ldots, \mu_N)$

The regularity conditions include: the model in the numerator of the likelihood ratio is "nested" within the one in the denominator, i.e., $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a special case of $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$.

Proof boils down to having all estimators ~ Gaussian.

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

Goodness of fit with Gaussian data

Suppose the data are N independent Gaussian distributed values:

$$y_i \sim \text{Gauss}(\underset{\substack{\nearrow \\ \text{want to estimate}}}{\mu_i}, \underset{\substack{\nwarrow \\ \text{known}}}{\sigma_i}) \; , \qquad i = 1, \ldots, N$$

N measurements and N parameters (= "saturated model")

Likelihood: $$L(\boldsymbol{\mu}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2}$$

Log-likelihood: $$\ln L(\boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\sigma_i^2} + C$$

ML estimators: $$\hat{\mu}_i = y_i \qquad i = 1, \ldots, N$$

Likelihood ratio for Gaussian data

Now suppose $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$, e.g., in an LS fit with $\mu_i(\boldsymbol{\theta}) = \mu(x_i; \boldsymbol{\theta})$.

The goodness-of-fit statistic for the test of the hypothesis $\boldsymbol{\mu}(\boldsymbol{\theta})$ becomes

$$t_{\boldsymbol{\mu}} = -2 \ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})} = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\hat{\boldsymbol{\theta}}))^2}{\sigma_i^2} \sim \chi^2_{N-M}$$

chi-square pdf for N-M degrees of freedom

Here t_{μ} is the same as $\chi^2_{\min}$ from an LS fit.

So Wilks' theorem formally states the property that we claimed for the minimized chi-squared from an LS fit with N measurements and M fitted parameters.

Likelihood ratio for Poisson data

Suppose the data are a set of values $\boldsymbol{n} = (n_1,\ldots, n_N)$, e.g., the numbers of events in a histogram with N bins.

Assume $n_i \sim$ Poisson(ν_i), $i = 1,\ldots, N$, all independent.

First (for LR denominator) treat $\boldsymbol{\nu} = (\nu_1,\ldots, \nu_N)$ as all adjustable:

Likelihood:
$$L(\boldsymbol{\nu}) = \prod_{i=1}^{N} \frac{\nu_i^{n_i}}{n_i!} e^{-\nu_i}$$

Log-likelihood:
$$\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} [n_i \ln \nu_i - \nu_i] + C$$

ML estimators:
$$\hat{\nu}_i = n_i \ , \qquad i = 1, \ldots, N$$

Goodness of fit with Poisson data (2)

For LR numerator find $\boldsymbol{\nu}(\boldsymbol{\theta})$ with M fitted parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_M)$:

$$t_{\boldsymbol{\nu}} = -2\ln\frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2\sum_{i=1}^{N}\left[n_i \ln\frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i} - \nu_i(\hat{\boldsymbol{\theta}}) + n_i\right]$$

if $n_i = 0$, skip term

Wilks' theorem: in large-sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M}$

Exact in large sample limit; in practice good approximation for surprisingly small n_i (~several).

As before use t_ν to get p-value of $\boldsymbol{\nu}(\boldsymbol{\theta})$,

independent of $\boldsymbol{\theta}$

$$p_{\boldsymbol{\nu}} = \int_{t_{\boldsymbol{\nu}},\mathrm{obs}}^{\infty} f(t_{\boldsymbol{\nu}}|\boldsymbol{\nu}(\boldsymbol{\theta}))\,dt_{\boldsymbol{\nu}} = 1 - F_{\chi^2}(t_{\boldsymbol{\nu},\mathrm{obs}}; N-M)$$

Goodness of fit with multinomial data

Similar if data $\boldsymbol{n} = (n_1,\ldots, n_N)$ follow multinomial distribution:

$$P(\mathbf{n}|\mathbf{p}, n_{\rm tot}) = \frac{n_{\rm tot}!}{n_1! n_2! \ldots n_N!} p_1^{n_1} p_2^{n_2} \ldots p_N^{n_N}$$

E.g. histogram with N bins but fix: $n_{\rm tot} = \sum_{i=1}^{N} n_i$

Log-likelihood: $\ln L(\boldsymbol{\nu}) = \sum_{i=1}^{N} n_i \ln \frac{\nu_i}{n_{\rm tot}} + C$ $(\nu_i = p_i n_{\rm tot})$

ML estimators: $\hat{\nu}_i = n_i$ (Only N–1 independent; one is $n_{\rm tot}$ minus sum of rest.)

Goodness of fit with multinomial data (2)

The likelihood ratio statistics become:

$$t_{\boldsymbol{\nu}} = -2 \ln \frac{L(\boldsymbol{\nu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\nu}})} = -2 \sum_{i=1}^{N} n_i \ln \frac{\nu_i(\hat{\boldsymbol{\theta}})}{n_i}$$

if $n_i = 0$, skip term

Wilks: in large sample limit $t_{\boldsymbol{\nu}} \sim \chi^2_{N-M-1}$

One less degree of freedom than in Poisson case because effectively only $N\text{–}1$ parameters fitted in denominator of LR.

Estimators and g.o.f. all at once

Evaluate numerators with $\boldsymbol{\theta}$ (not its estimator); if any $n_i = 0$, omit the corresponding log terms:

$$\chi^2_{\rm P}(\boldsymbol{\theta}) \;=\; -2\sum_{i=1}^{N}\left[n_i \ln\frac{\nu_i(\boldsymbol{\theta})}{n_i} - \nu_i(\boldsymbol{\theta}) + n_i\right] \qquad \text{(Poisson)}$$

$$\chi^2_{\rm M}(\boldsymbol{\theta}) \;=\; -2\sum_{i=1}^{N} n_i \ln\frac{\nu_i(\boldsymbol{\theta})}{n_i} \qquad \text{(Multinomial)}$$

These are equal to the corresponding $-2 \ln L(\boldsymbol{\theta})$ plus terms not depending on $\boldsymbol{\theta}$, so minimizing them gives the usual ML estimators for $\boldsymbol{\theta}$.

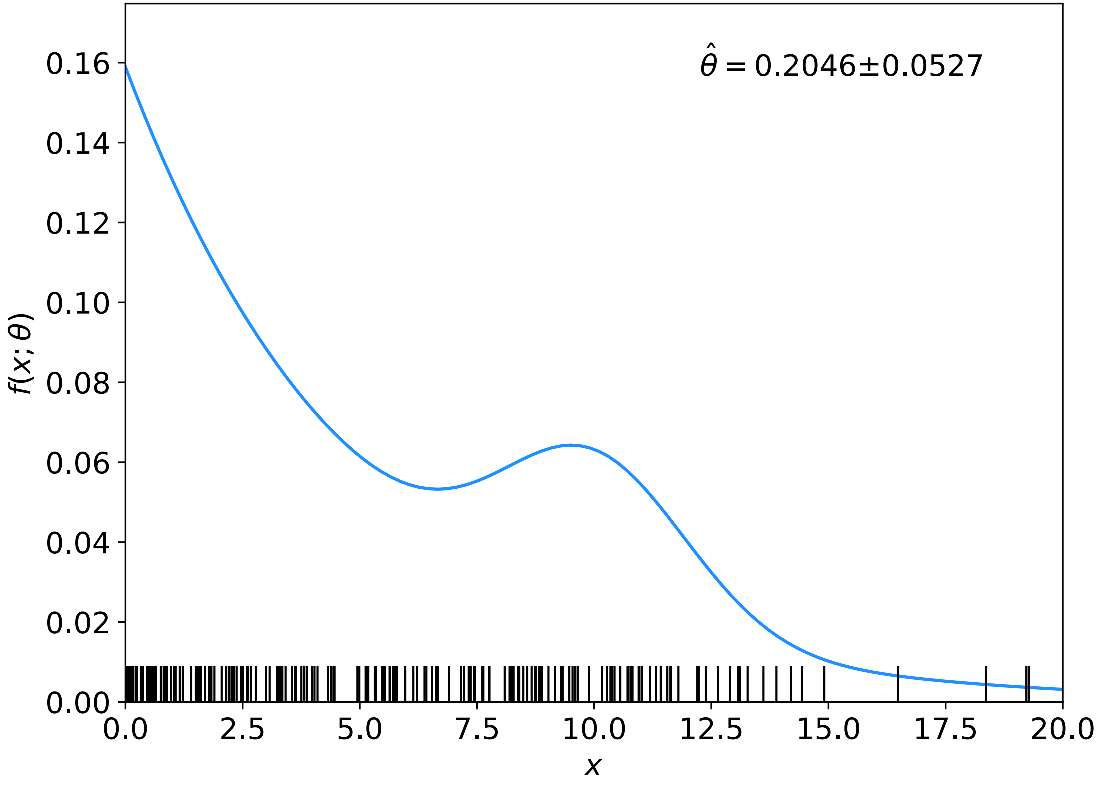
The minimized value gives the statistic t_ν, so we get goodness-of-fit for free.

Steve Baker and Robert D. Cousins, *Clarification of the use of the chi-square and likelihood functions in fits to histograms*, NIM **221** (1984) 437.

Examples of ML/LS fits

Unbinned maximum likelihood (mlFit.py, minimize negLogL)

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i; \boldsymbol{\theta})$$

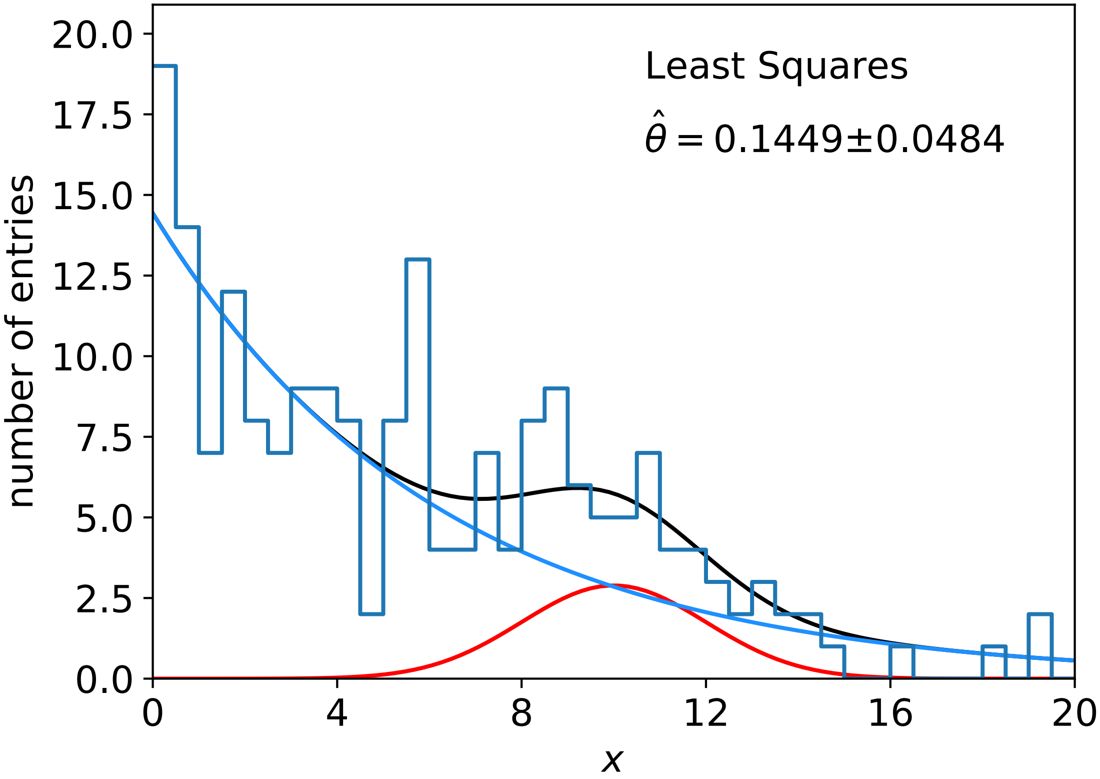


No useful measure of goodness-of-fit from unbinned ML.

Examples of ML/LS fits

Least Squares fit (histFit.py, minimize chi2LS)

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{(y_i - \mu_i(\boldsymbol{\theta}))^2}{\mu_i(\boldsymbol{\theta})}$$



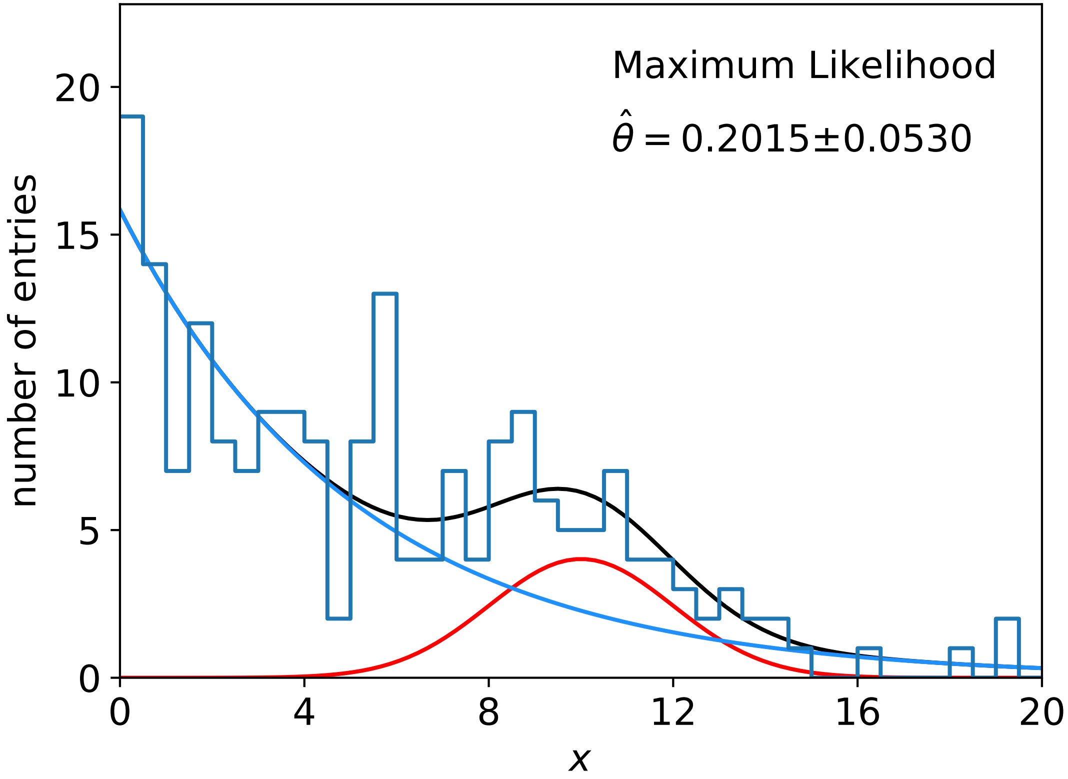
$\chi^2_{\text{min}} = 32.7$
$n_{\text{dof}} = 38$
$p = 0.71$

Many bins with few entries, LS not expected to be reliable.

Examples of ML/LS fits

Multinomial maximum likelihood fit (histFit.py, minimize chi2M)

$$\chi^2_{\rm M}(\boldsymbol{\theta}) = -2\sum_{i=1}^{N} n_i \ln \frac{\nu_i(\boldsymbol{\theta})}{n_i}$$



$\chi^2_{\rm min} = 35.3$

$n_{\rm dof} = 37$

$p = 0.55$

Essentially same result as unbinned ML.

Statistical Data Analysis
Lecture 9-3

- Interval estimation
- Confidence interval from inverting a test
- Example: limits on mean of Gaussian

Confidence intervals by inverting a test

In addition to a ‘point estimate’ of a parameter we should report an interval reflecting its statistical uncertainty.

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are ‘disfavoured’ by θ (critical region) such that $P(\text{data in critical region} \,|\, \theta) \leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

set of θ values that are not rejected in a test of size α (confidence level CL is $1-\alpha$).

Relation between confidence interval and p-value

Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, p_θ.

If $p_\theta \leq \alpha$, then we reject θ.

The confidence interval at CL $= 1 - \alpha$ consists of those values of θ that are not rejected.

E.g. an upper limit on θ is the greatest value for which $p_\theta > \alpha$.

In practice find by setting $p_\theta = \alpha$ and solve for θ.

For a multidimensional parameter space $\boldsymbol{\theta} = (\theta_1, \ldots\ \theta_M)$ *use* same idea – result is a confidence “region” with boundary determined by $p_{\boldsymbol{\theta}} = \alpha$.

Coverage probability of confidence interval

If the true value of θ is rejected, then it's not in the confidence interval. The probability for this is by construction (equality for continuous data):

$$P(\text{reject } \theta | \theta) \leq \alpha = \text{type-I error rate}$$

Therefore, the probability for the interval to contain or "cover" θ is

$$P(\text{conf. interval “covers” } \theta | \theta) \geq 1 - \alpha$$

This assumes that the set of θ values considered includes the true value, i.e., it assumes the composite hypothesis $P(\boldsymbol{x}|H,\theta)$.

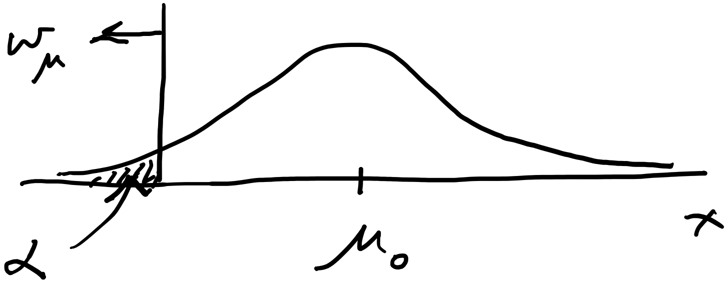
Example: upper limit on mean of Gaussian

When we test the parameter, we should take the critical region to maximize the power with respect to the relevant alternative(s).

Example: $x \sim$ Gauss(μ, σ) (take σ known)

Test $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu < \mu_0$

→ Put w_μ at region of x-space characteristic of low μ (i.e. at low x)

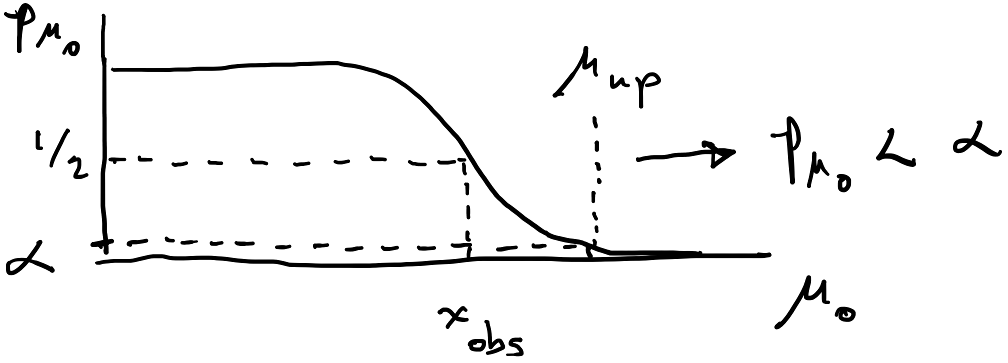


Equivalently, take the p-value to be

$$p_{\mu_0} = P(x \leq x_{\rm obs} | \mu_0) = \int_{-\infty}^{x_{\rm obs}} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu_0)^2/2\sigma^2} \, dx = \Phi\left(\frac{x_{\rm obs} - \mu_0}{\sigma} \right)$$

Upper limit on Gaussian mean (2)

To find confidence interval, repeat for all μ_0, i.e., set $p_{\mu 0} = \alpha$ and solve for μ_0 to find the interval's boundary



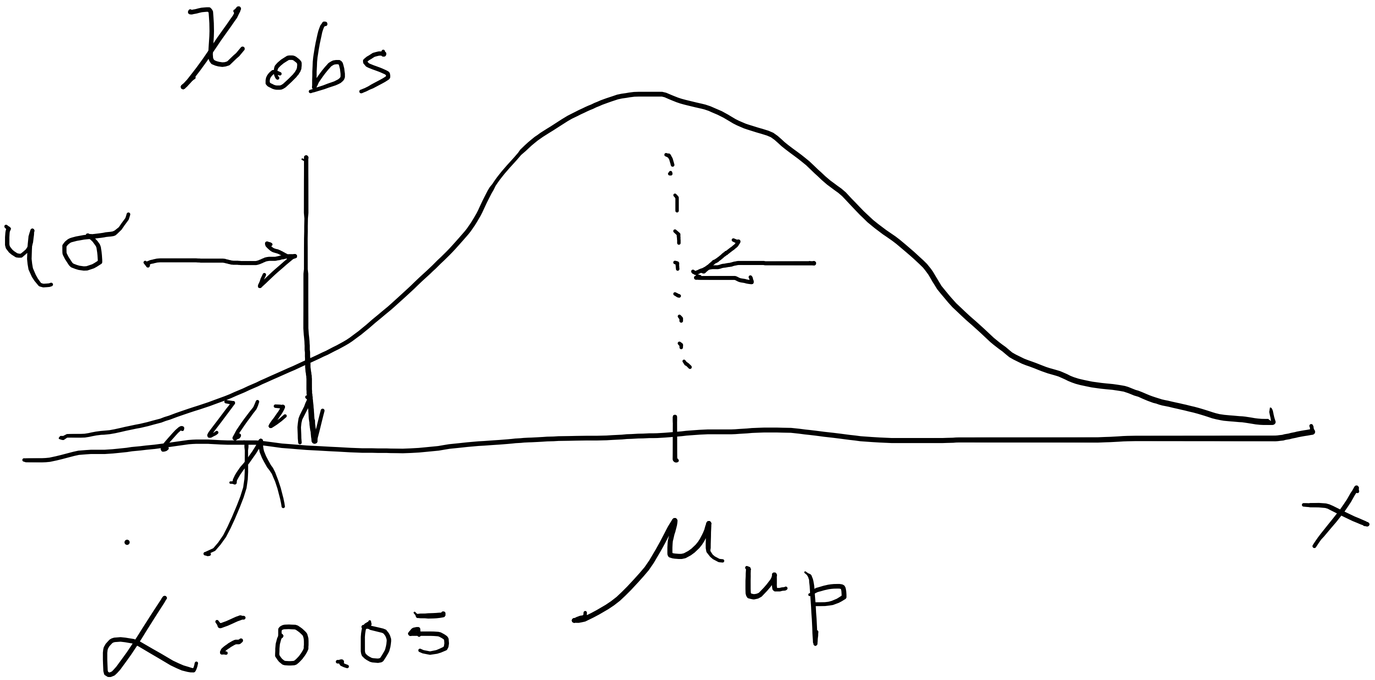
$$\mu_0 \to \mu_{\rm up} = x_{\rm obs} - \sigma\Phi^{-1}(\alpha) = x_{\rm obs} + \sigma\Phi^{-1}(1-\alpha)$$

This is an upper limit on μ, i.e., higher μ have even lower p-value and are in even worse agreement with the data.

Usually use $\Phi^{-1}(\alpha) = –\Phi^{-1}(1–\alpha)$ so as to express the upper limit as $x_{\rm obs}$ plus a positive quantity. E.g. for α = 0.05, $\Phi^{-1}(1–0.05)$ = 1.64.

Upper limit on Gaussian mean (3)

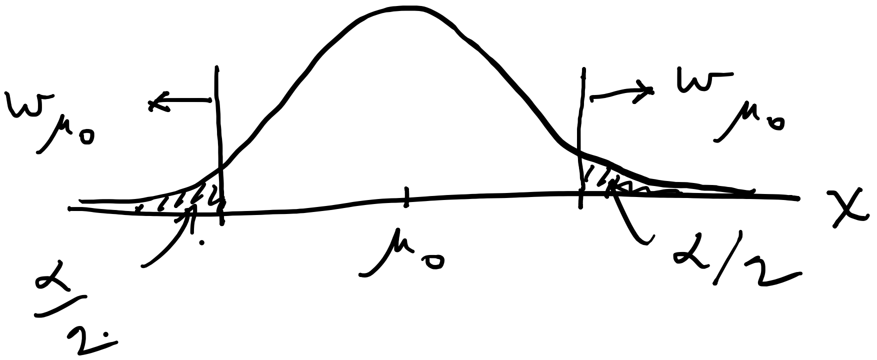
μ_{up} = the hypothetical value of μ such that there is only a probability α to find $x < x_{\text{obs}}$.



1- vs. 2-sided intervals

Now test: $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu \neq \mu_0$

I.e. we consider the alternative to μ_0 to include higher and lower values, so take critical region on both sides:



Result is a "central" confidence interval $[\mu_{\rm lo}, \mu_{\rm up}]$:

$$\mu_{\rm lo} = x_{\rm obs} - \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

$$\mu_{\rm up} = x_{\rm obs} + \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

E.g. for $\alpha = 0.05$

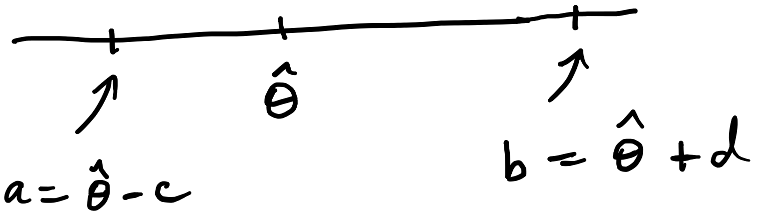
$$\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 1.96 \approx 2$$

Note upper edge of two-sided interval is higher (i.e. not as tight of a limit) than obtained from the one-sided test.

On the meaning of a confidence interval

Often we report the confidence interval $[a,b]$ together with the point estimate as an “asymmetric error bar”, e.g.,

$$\hat{\theta}^{+d}_{-c}$$



E.g. (at CL = 1 – α = 68.3%): $\hat{\theta} = 80.25^{+0.31}_{-0.25}$

Does this mean P(80.00 < θ < 80.56) = 68.3%? No, not for a frequentist confidence interval. The parameter θ does not fluctuate upon repetition of the measurement; the endpoints of the interval do, i.e., the endpoints of the interval fluctuate (they are functions of data):

$$P(a(x) < \theta < b(x)) = 1 - \alpha$$

Statistical Data Analysis
Lecture 9-4

- Confidence intervals from the likelihood function

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad\qquad 0 \leq \lambda(\boldsymbol{\theta}) \leq 1$$

Lower $\lambda(\boldsymbol{\theta})$ means worse agreement between data and hypothesized $\boldsymbol{\theta}$. Equivalently, usually define

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher t_θ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore

$$p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and provided certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

chi-square dist. with # d.o.f. = # of components in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

Assuming this holds, the p-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_n}(t_{\boldsymbol{\theta}})$$

← set equal to α

To find boundary of confidence region set $p_\theta = \alpha$ and solve for t_θ:

$$t_{\boldsymbol{\theta}} = F^{-1}_{\chi^2_n}(1 - \alpha)$$

Recall also

$$t_{\boldsymbol{\theta}} = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

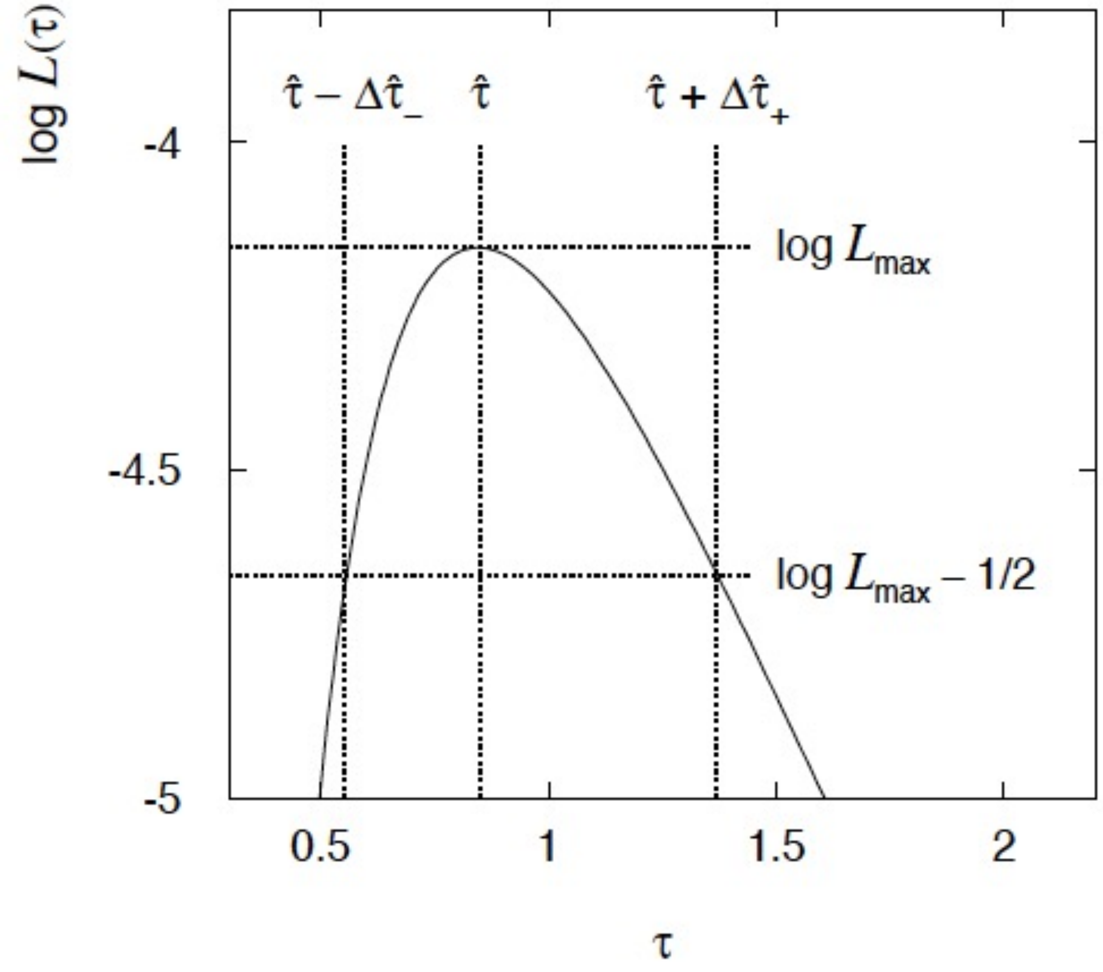
$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from $\ln L(\theta)$

For n=1 parameter, CL = 0.683, $Q_\alpha = 1$.



Our exponential example, now with only $n = 5$ events.

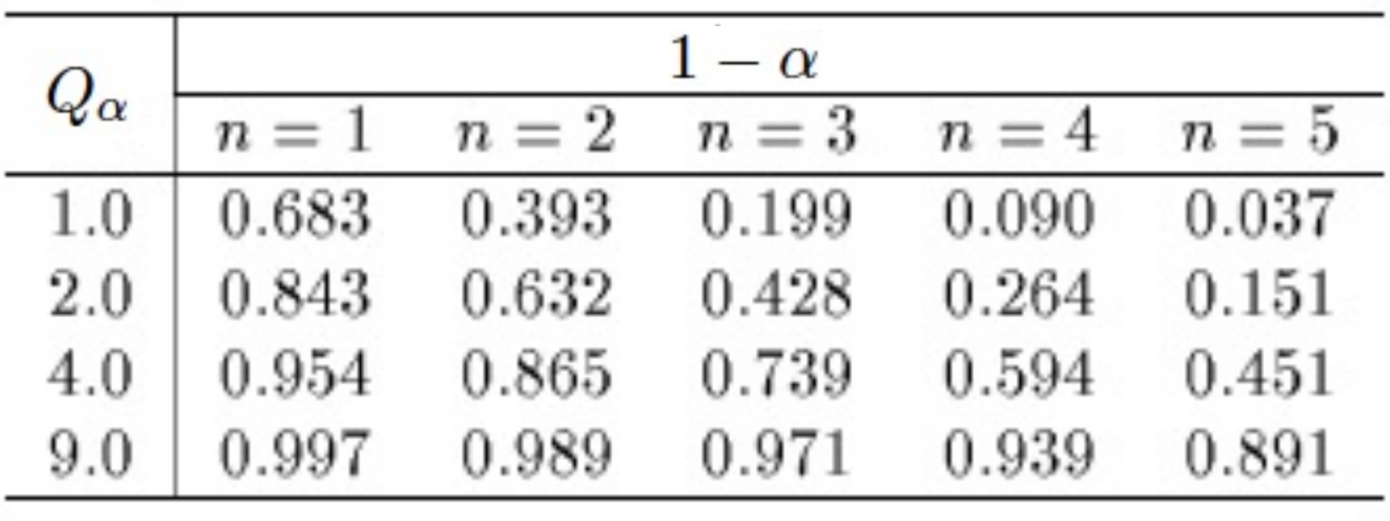
Can report ML estimate with approx. confidence interval from $\ln L_{\max} - 1/2$ as "asymmetric error bar":

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

For increasing number of parameters, CL $= 1 - \alpha$ decreases for confidence region determined by a given

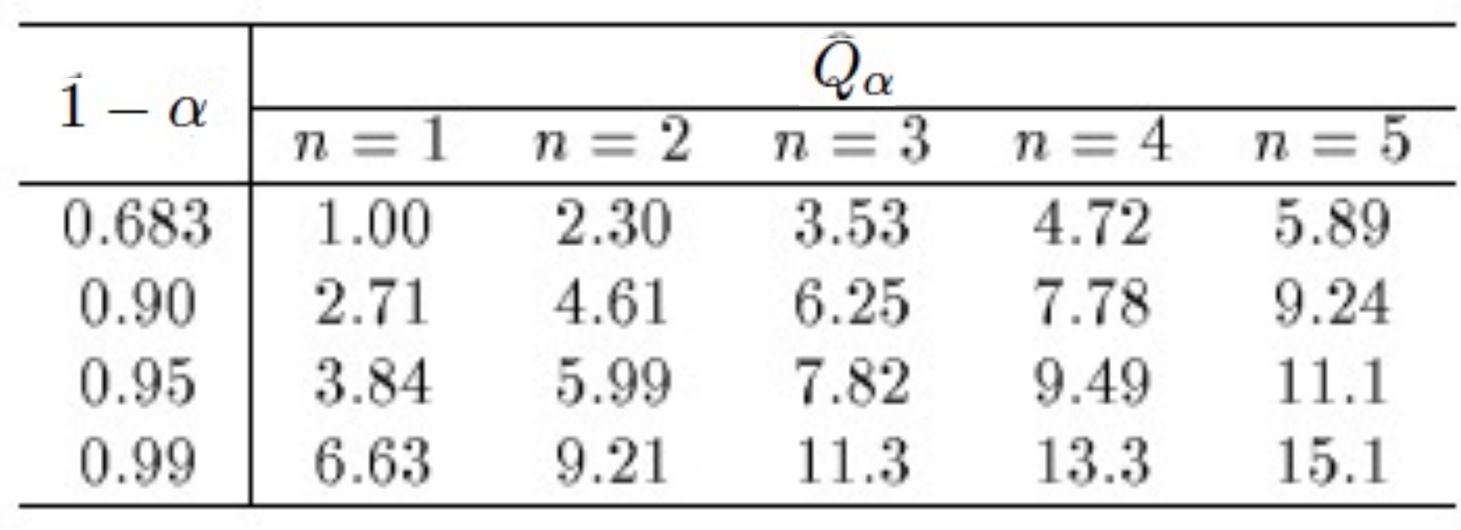
$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$



Q_α	$1-\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given CL $= 1 - \alpha$.

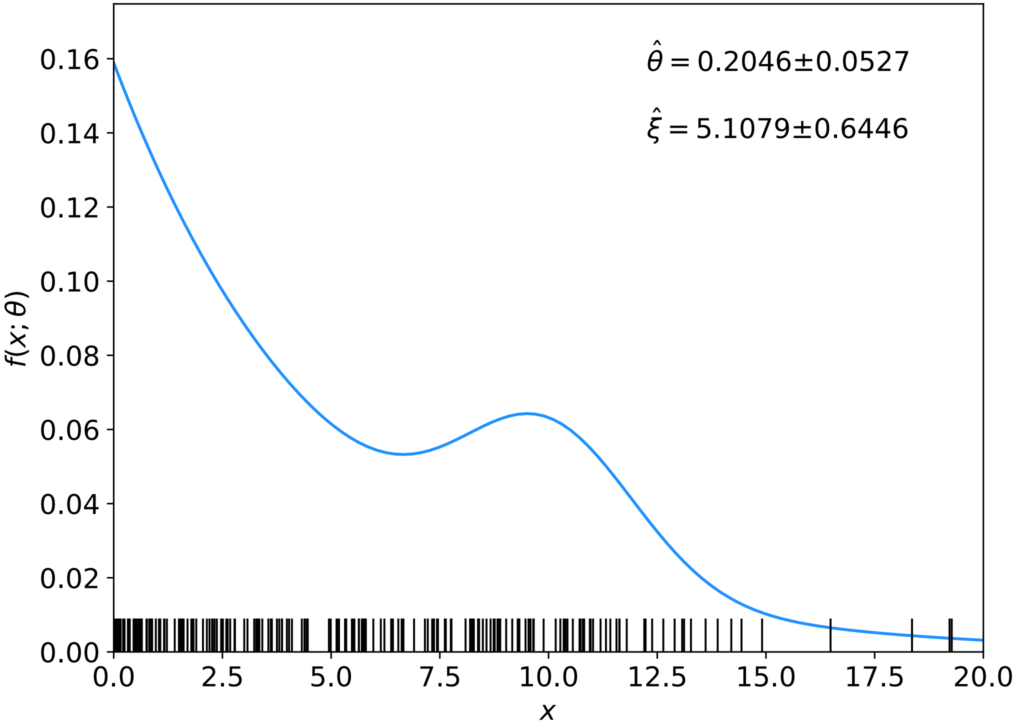


$1 - \alpha$	Q_α				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Example: 2 parameter fit:

Example from problem sheet 8, i.i.d. sample of size 200

$$x \sim f(x;\theta,\xi) = \theta \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} + (1-\theta)\frac{1}{\xi} e^{-x/\xi}$$

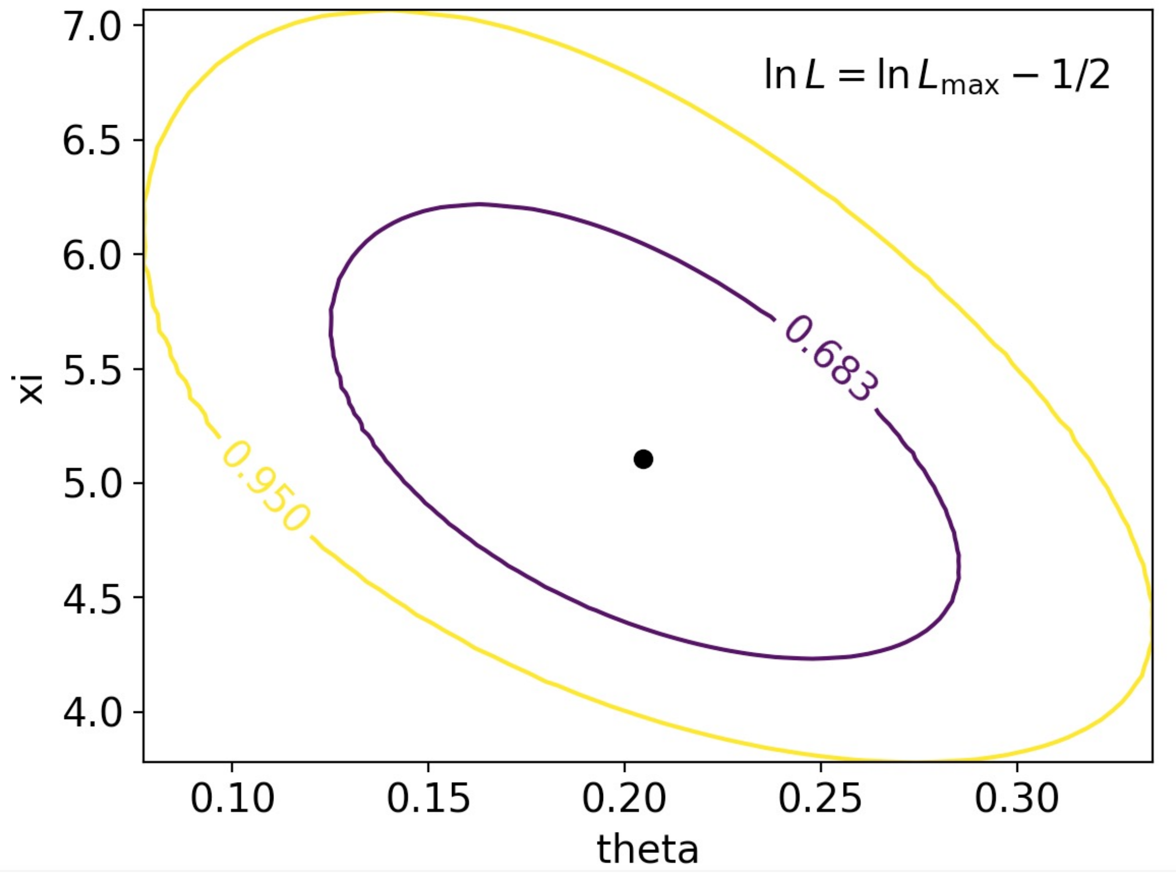


Here fit two parameters: θ and ξ.

Example: 2 parameter fit:

In iminuit v2, user can set $\mathrm{CL} = 1 - \alpha$

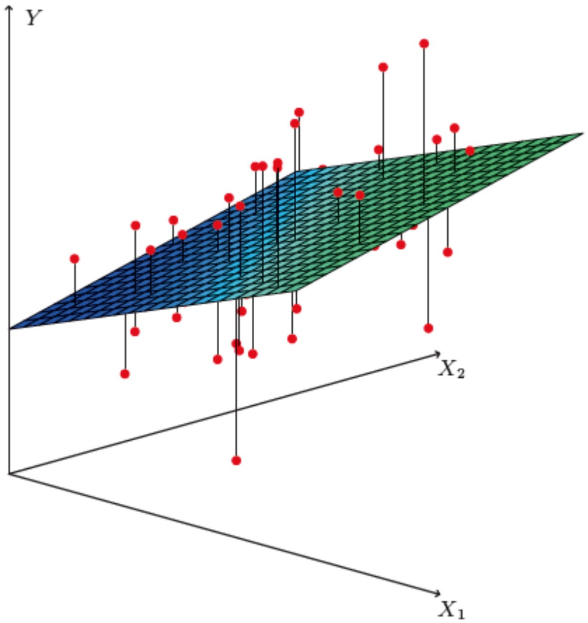
```
m.draw_mncontour('theta', 'xi', cl=[0.683, 0.95], size=200)
```



Extra slides

Brief intro to multiple regression

Multiple regression* can be seen as an extension of curve fitting to the case where the variable x is replaced by a multi-dimensional $\boldsymbol{x} = (x_1,\ldots,x_n)$, e.g., fitting a surface. Here suppose the data are points $(\boldsymbol{x}_i, y_i)$, $i = 1,\ldots,N$ (no error bars) and $\boldsymbol{x}$ is usually a random variable, often called the explanatory or predictor variable.



Equivalently, we can view it as an extension to classification with the discrete class label $y = 0, 1$ replaced by a continuous target y (and in this context $\boldsymbol{x}$ can also be called the feature vector).

*Note the term ”multivariate” regression refers to a vector target variable y; here we treat only scalar y.

Target (fit) function and loss function

As in the case of curve fitting, we assume some parametric function of $\boldsymbol{x}$ that represents the mean of the target variable

$$E[y] = f(\mathbf{x}; \mathbf{w})$$

where $\boldsymbol{w}$ is a vector of adjustable parameters (“weights”).

Suppose we have training data consisting of $(\boldsymbol{x}_i, y_i)$, $i = 1,\dots,N$.

Use these to determine the weights by minimizing a loss function (analogous to the χ^2), e.g.,

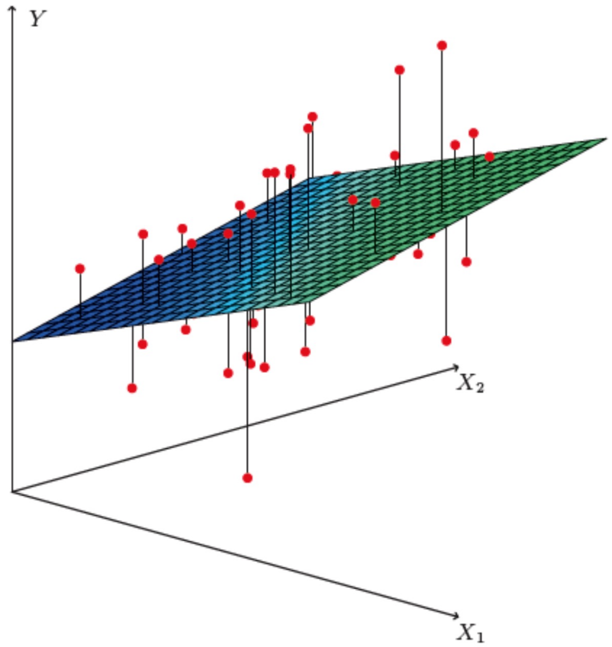
$$L(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{n} |y_i - f(\mathbf{x}_i; \mathbf{w})|^2$$

http://www-bcf.usc.edu/~gareth/ISL/

Linear regression

In linear regression, the fit function is of the form

$$f(\mathbf{x};\mathbf{w}) = w_0 + \sum_{i=1}^{n} w_i x_i$$



i.e. the problem is equivalent to an unweighted least-squares fit of a (hyper-)plane:

Can be generalized to a nonlinear surface with higher order terms,

$$f(\mathbf{x};\mathbf{w}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i,j=1}^{n} w_{ij} x_i x_j + \sum_{i,j,k=1}^{n} w_{ijk} x_i x_j x_k + \dots$$

Nonlinear regression

Other examples of nonlinear regression include:

MLP (multilayer perceptron) regression

Boosted decision tree regression

Support vector regression

For MLP regression, as with classification, regard the feature vector as the layer $k = 0$; i.e., $\varphi_i^{(0)} = x_i$.

The ith node of hidden layer k is

$$\varphi_i^{(k)} = h\left(w_{i0}^{(k)} + \sum_{j=1}^{n} w_{ij}^{(k)} \varphi_j^{(k-1)}\right)$$

where h is the activation function (tanh, relu, sigmoid,...).

MLP Regression (cont.)

For the final layer (k=K), in MLP regression (in contrast to classification), one omits the activation function, i.e.,

$$f(\mathbf{x};\mathbf{w}) = w_0^{(K)} + \sum_{j=1}^{n} w_j^{(K)} \varphi_j^{(K-1)}$$

where $\varphi_j^{(K-1)} =$ are the nodes of the last hidden layer ($k = K-1$).

For info on other types of multiple regression see, e.g.,

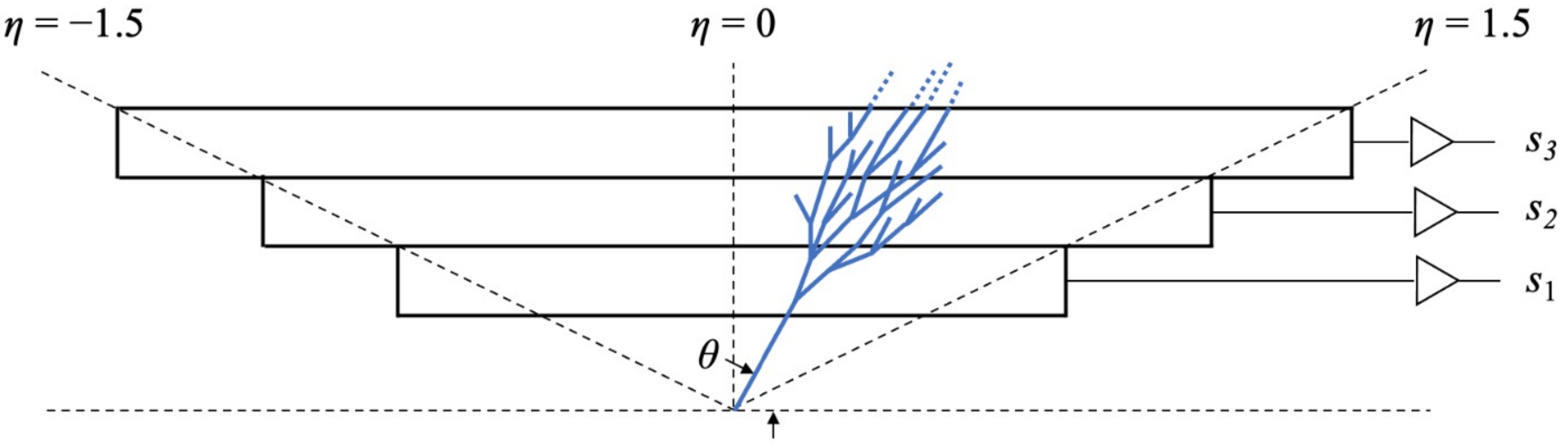
Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013; http://www-bcf.usc.edu/~gareth/ISL/

and the scikit-learn documentation.

Multiple regression example

Suppose particles with different energies E and angles θ (or equivalently $\eta = -\ln\tan(\theta/2)$) enter a calorimeter and create a particle showers that gives signals in three layers, s_1, s_2 and s_3, as well as an estimate of η.

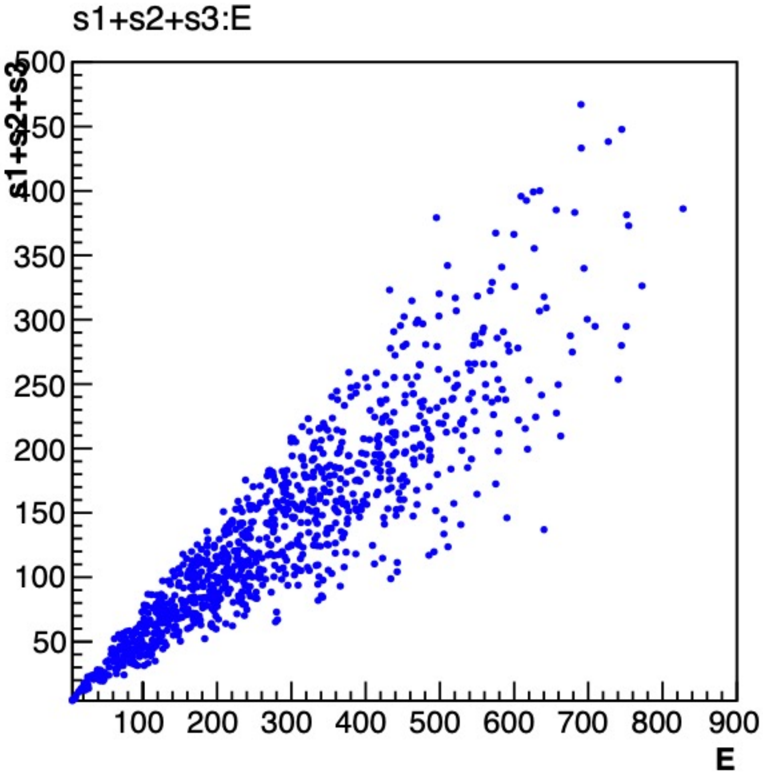
Some of the energy leaks through, with increased leakage for higher energy and more oblique angles (higher η).



The goal is to estimate the target $y_i = E_i$ given feature vectors $\boldsymbol{x}_i = (\eta,\, s_1,\, s_2, s_3)_i$ for $i = 1,\ldots,N$ training events.

Energy estimate from sum of signals

Naively, one could try just summing the signals: $\hat{E} = s_1 + s_2 + s_3$

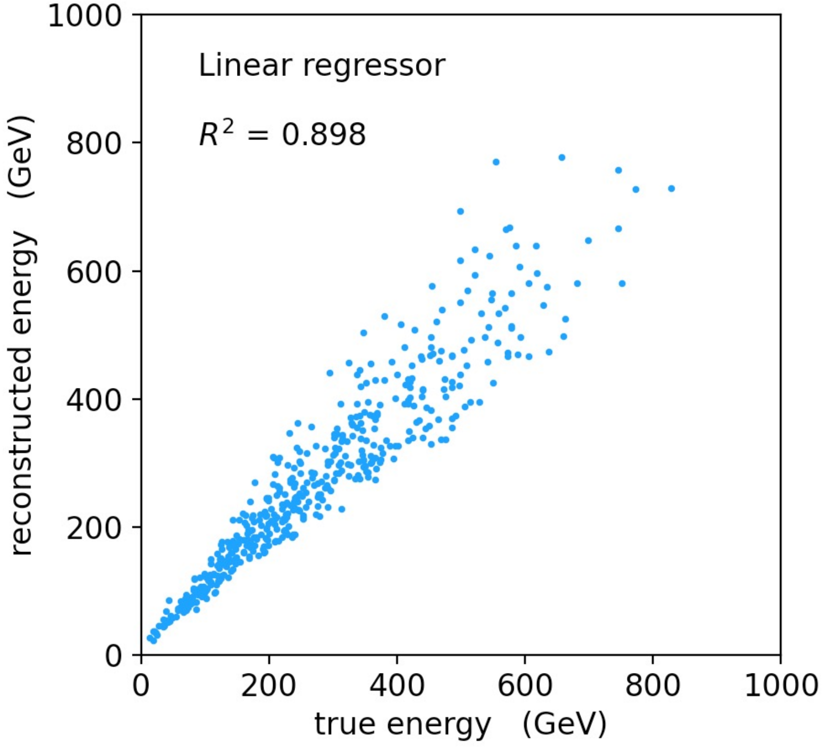


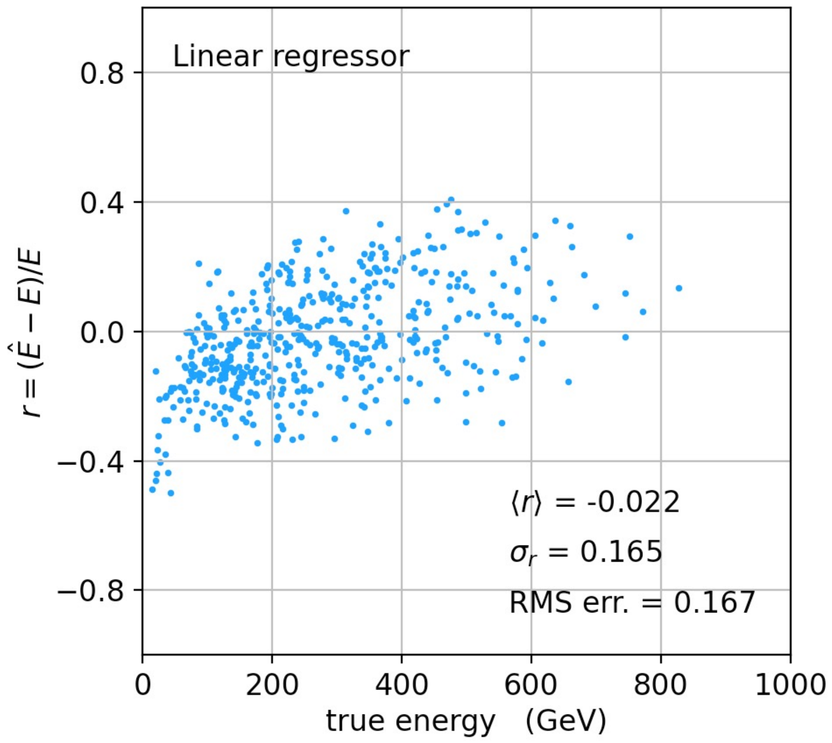
Gives very poor resolution because the particles have a distribution of energies and angles and hence differing amounts of the energy leak through undetected.

Linear regression

See MVRegressor.py, here using

```
regr = linear_model.LinearRegression()
regr.fit(X_train, y_train)
```

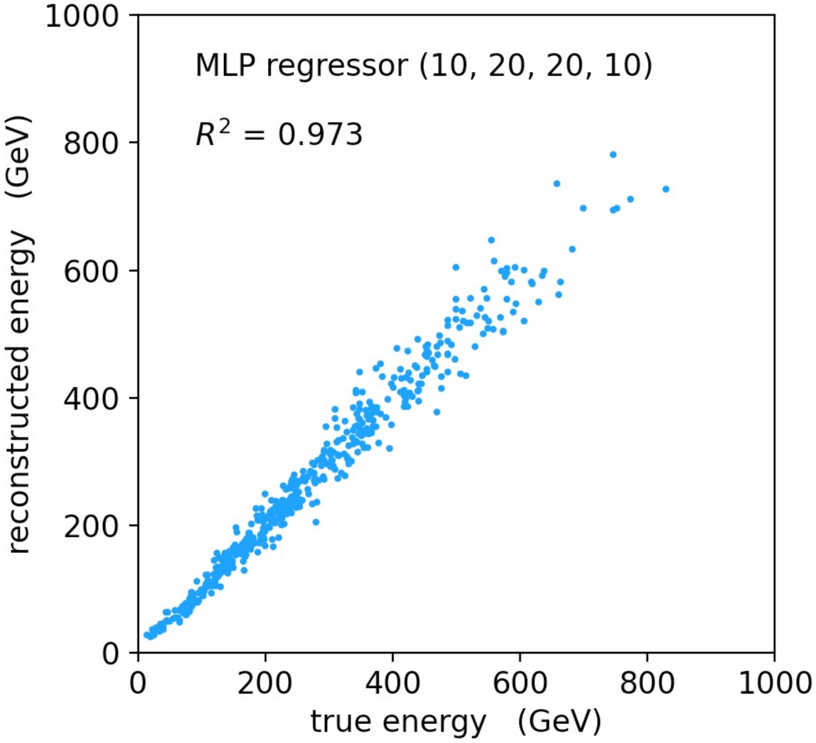


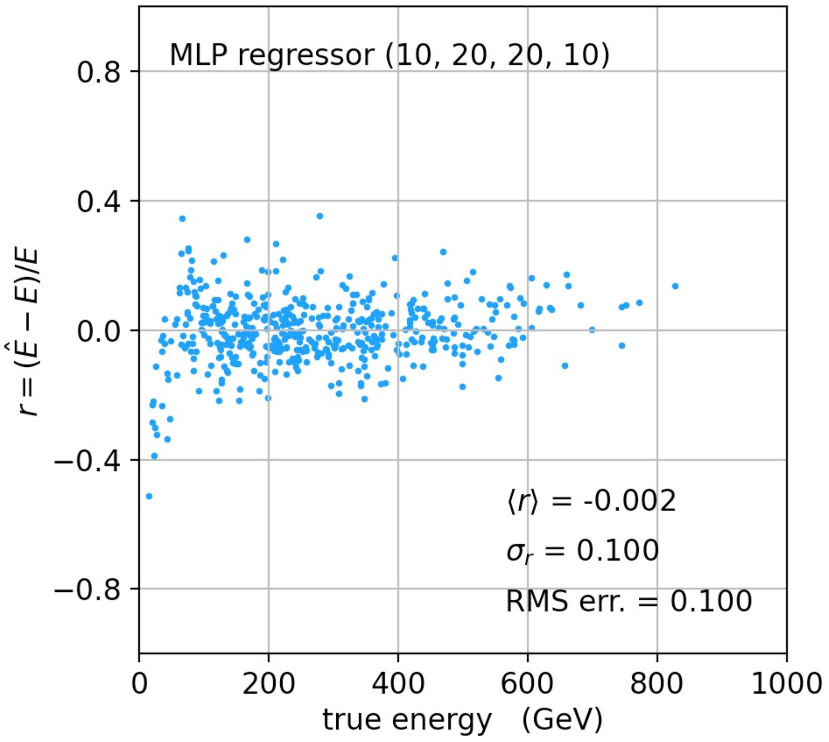


Average relative resolution 16.7%.

MLP Regression

```
regr = MLPRegressor(hidden_layer_sizes=(10,20,20,10), activation='relu'
regr.fit(X_train, y_train)
```





Better resolution (10%), here significant bias at low energies.

Refinements for multiple regression

One can try many improvements:

Scaling of predictor and target variables, e.g., standardize to zero mean and unit variance.

Use cross-validation to assess accuracy (and hence use entire sample of events for training.

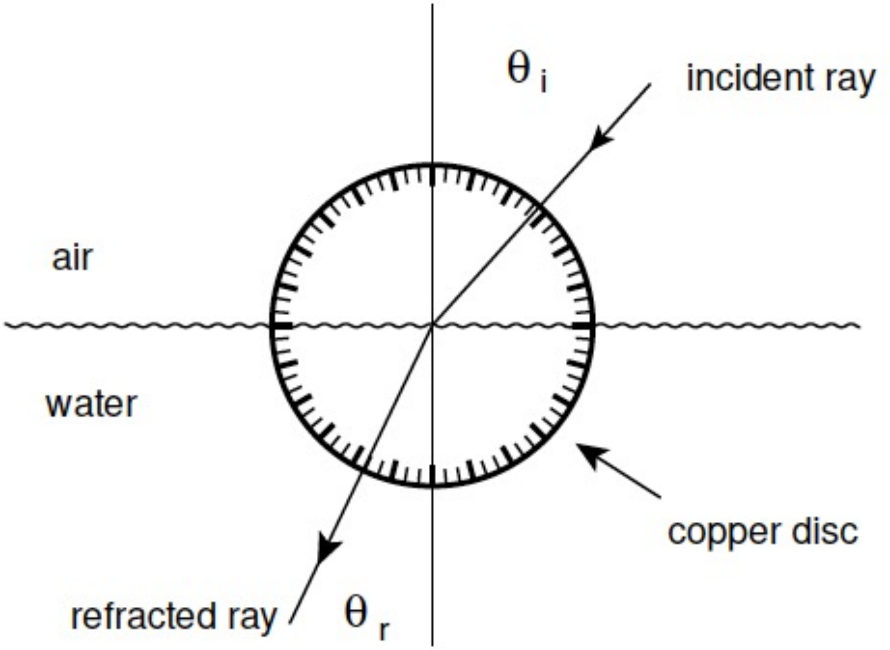
Try different loss functions.

Try different regression algorithms (ridge regression, lasso, decision tree, support vector regression,...).

Some simple code using scikit-learn and a short write-up (from a year-3 project) is on the course webpage.

LS example: refraction data from Ptolemy

Astronomer Claudius Ptolemy obtained data on refraction of light by water in around 140 A.D.:



Angles of incidence and refraction (degrees)

θ_i	θ_r
10	8
20	$15\frac{1}{2}$
30	$22\frac{1}{2}$
40	29
50	35
60	$40\frac{1}{2}$
70	$45\frac{1}{2}$
80	50

Suppose the angle of incidence is set with negligible error, and the measured angle of refraction has a standard deviation of ½° .

Laws of refraction

A commonly used law of refraction was

$$\theta_{\rm r} = \alpha \theta_{\rm i} \ ,$$

although it is reported that Ptolemy preferred

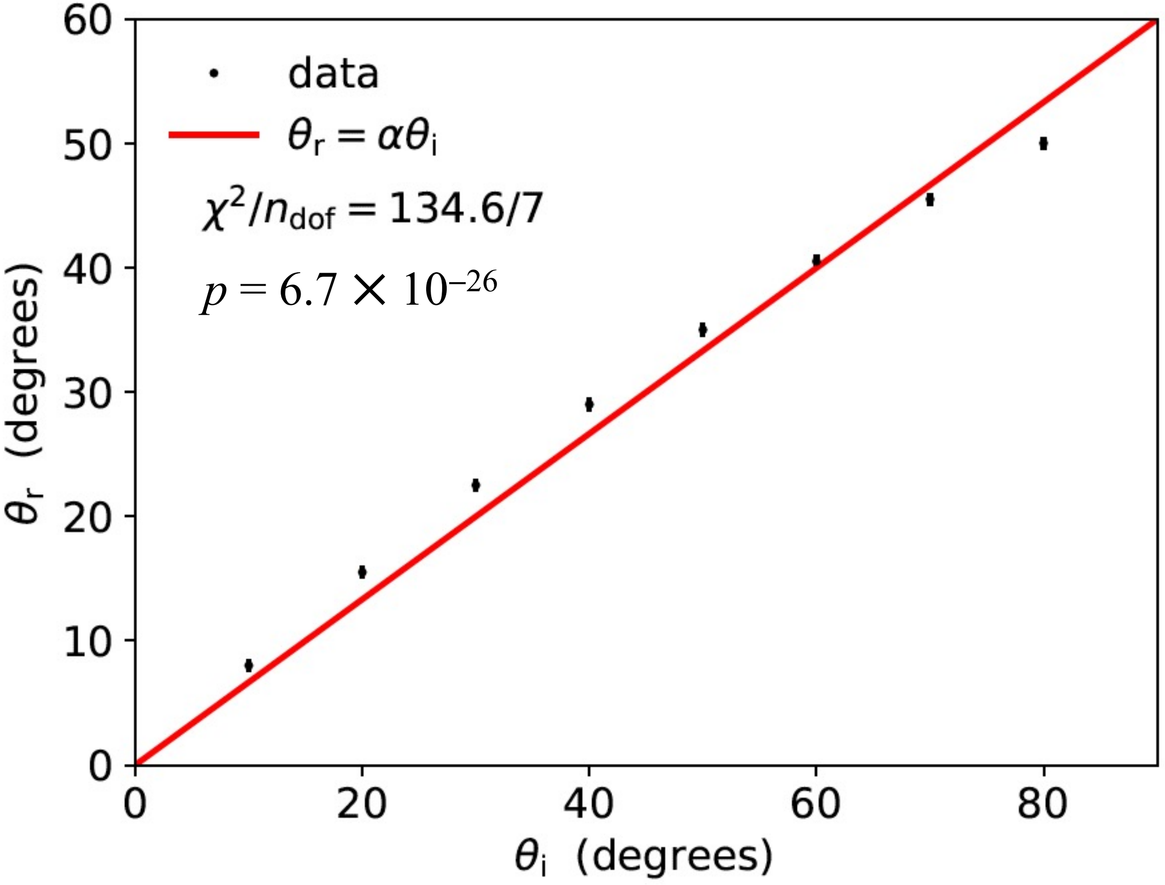
$$\theta_{\rm r} = \alpha \theta_{\rm i} \ - \ \beta \theta_{\rm i}^2 \ .$$

The law of refraction discovered by Ibn Sahl in 984 (and rediscovered by Snell in 1621) is

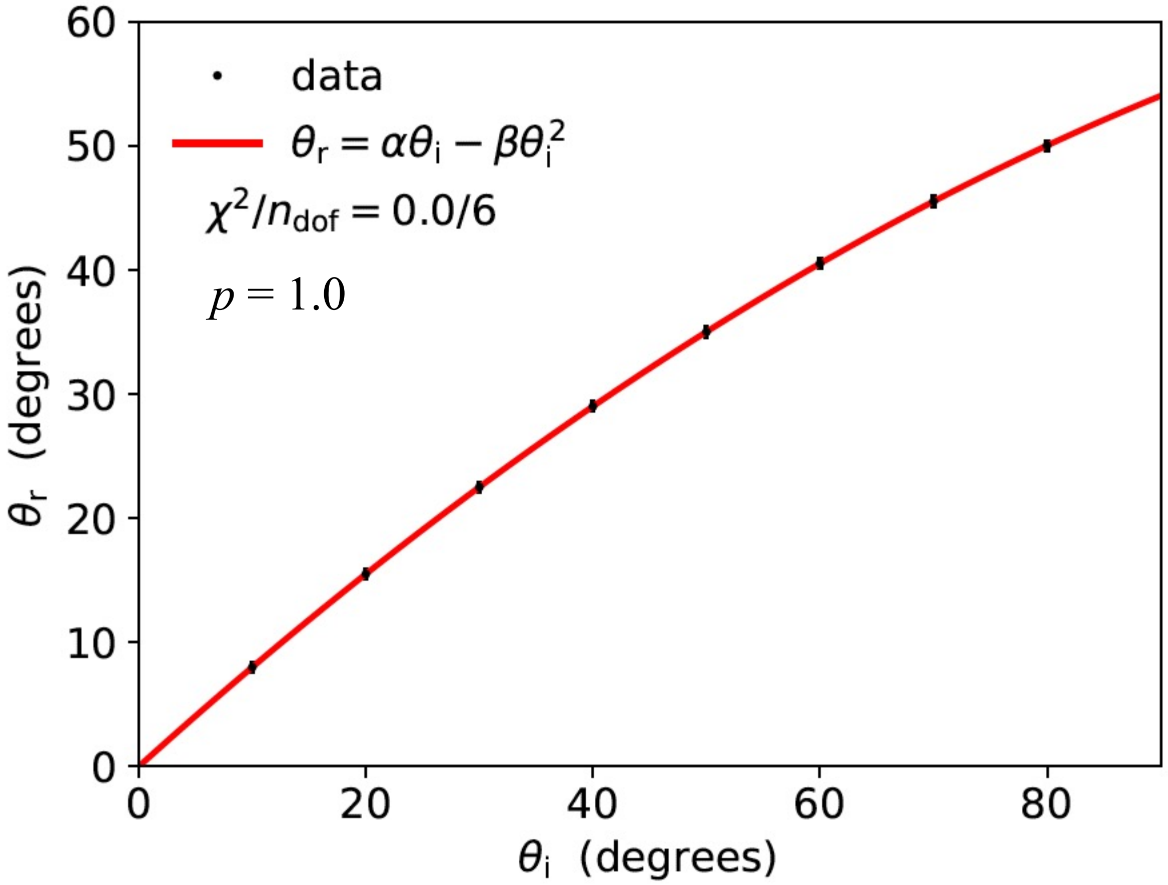
$$\theta_{\rm r} = \sin^{-1} \left(\frac{\sin \theta_{\rm i}}{r} \right) .$$

where $r = n_{\rm r}/n_{\rm i}$ is the ratio of indices of refraction of the two media.

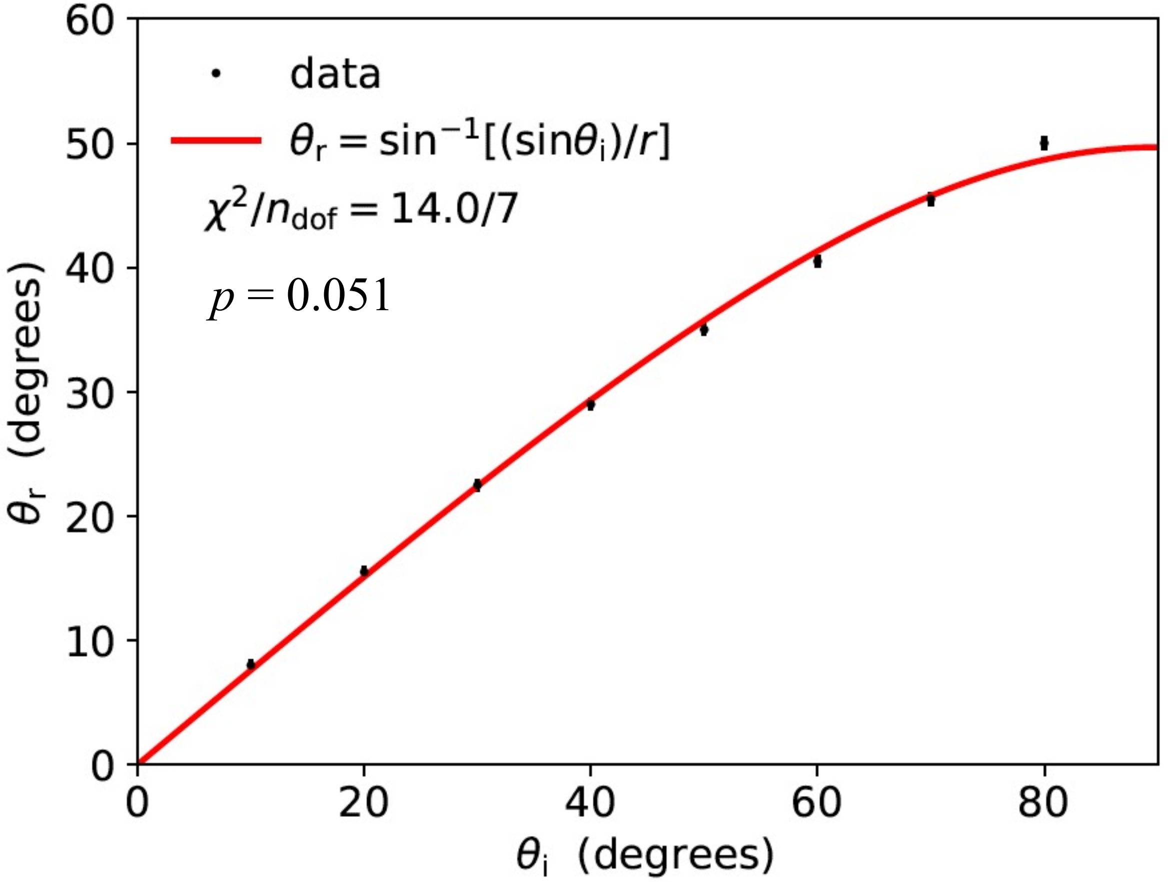
LS fit: $\theta_r = \alpha\theta_i$



LS fit: $\theta_\mathrm{r} = \alpha\theta_\mathrm{i} - \beta\theta_\mathrm{i}^2$



LS fit: Snell’s Law



Fitted index of refraction of water r = 1.3116 ± 0.0056 found not quite compatible with currently known value 1.330.