Statistical Analysis of Experimental Data

Douglas W. Higinbotham (Jefferson Lab)

with many thanks to

Sanjoy Mahajan (M.I.T. & Olin College of Engineering)
Simon Širca (University of Slovenia)
& Dave Meekins (Jefferson Lab)

What's to know?

Name	Statistic
chi-squared distribution	$\sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$

Just fit until you get $\chi^2 / \nu = 1$ and your good? Right.... ?!

(where ν is the degrees of freedom in the fit N – j -1)

What could possibly go wrong?!

What if the weights (sigma's) are underestimated or overestimated?
What if I have the wrong model?
What if the data aren't normally distributed?
What if average redcued χ^2 is good, but one over-fits one area and under-fits another!!
(It is NOT as trivial and just getting a reduced χ^2 ~ 1 does NOT mean you have a good result.)

Highlighted Resources

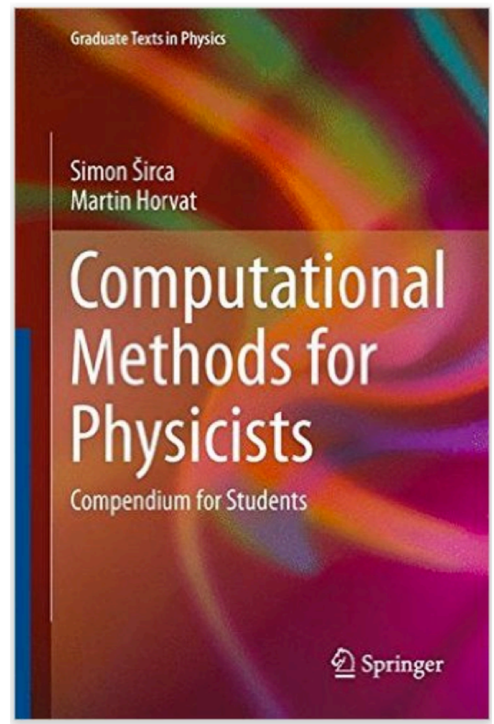
- Particle Data Handbook – Statistics Section
  - *http://pdg.lbl.gov/2015/reviews/rpp2015-rev-statistics.pdf*
- The Interpretation of Errors – Fredrick James
  - *http://seal.cern.ch/documents/**minuit**/mn**error**.pdf*
- Data Analysis Textbooks
  - Data Reduction and Error Analysis – Philip Bevington
  - Statistical Methods in Experimental Physics – Fredrick James
  - Computation Methods for the Physical Science – Simon Širca
  - Probability of Physics – Simon Širca
- R Programing Language
  - https://www.r-project.org/
- **Estimation**
  - Street-Fighting Mathematics – Sanjoy Mahajan
  - Guesstimation – Larry Weinstein

All Models Are Wrong

"The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful." - George Box (1919 – 2013)

"An ever increasing amount of computational work is being relegated to computers, and often we almost blindly assume that the obtained results are correct."

- Simon Širca & Martin Horvat



Some Wrong But Useful Models

- F = ma … but what about the friction
- pV = nRT … but what about Van der Waals
- F = kx … but what about the elongation
- $y = a_1 + a_2x$ … but what about a_2x^2, a_3x^3, etc.
  - sin(θ) for small θ ≅ θ
  - cos(θ) for small θ ≅ 1
  - tan(θ) for small angles goes to zero.
  - tan(θ) for large angle goes to infinity.
- And of course the spherical cows...

Charge Radius of the Proton

- Proton G_E has no measured diffractive minima and it is too light for the Fourier transformation to work in any kind of model independent way.
  - Jim Kelly, Phys.Rev. C66 (2002) 065203.
- Thus for the proton we make use of the theorem that as Q^2 goes to zero the charge radius is equal to the slope of G_E

$$G_E(Q^2) = 1 + \sum_{n \geq 1} \frac{(-1)^n}{(2n+1)!} \langle r^{2n} \rangle \; Q^{2n}$$

For small Q^2 (< 1 fm^{-2}), the higher order terms, ~ $Q^{2n}/(2n+1)!$, become less important.

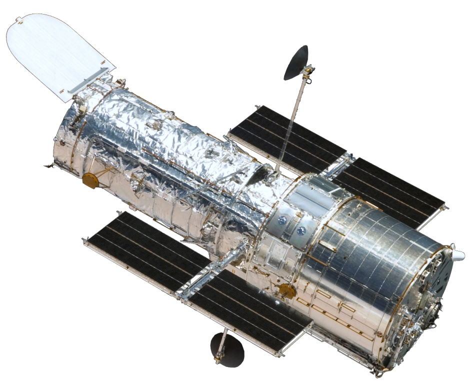
$$r_p \equiv \sqrt{\langle r^2 \rangle} = \left(-6 \, \frac{\mathrm{d}G_E(Q^2)}{\mathrm{d}Q^2} \bigg|_{Q^2=0} \right)^{1/2}$$

i.e. Experimentalists are trying to determine the slope of G_E as Q^2 goes to zero.

Measurement Is Often A Goldilocks Problem

From Deep Space

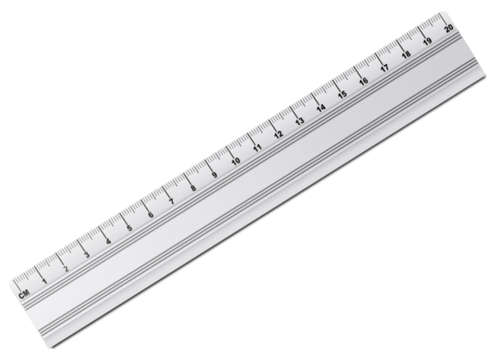
Too Far



A Modern Telescope

From Orbit

Just Right



Ruler & Some Geometry

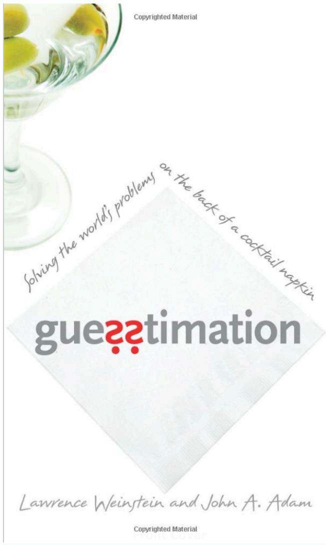
On The Planet

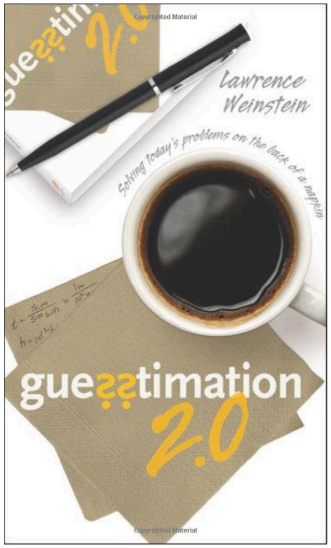
Too Close

Theodolite*

What is *just right* for the proton?!

- We use **Plank's constant** one to relate energy to length in natural units:
  - **Q^2 of 1 GeV^2 = 25.7 fm^{-2}.**
- Radius of the proton is ~ 0.84 - 0.88 fm
- Thus one can immediately guesstimate that with electron scattering one needs:
  - $Q^2 < (1/0.88\ \text{fm})^2 < 1.2\ \text{fm}^{-2}$ to get the radius of the proton.
  - $Q^2 > 1.2\ \text{fm}^{-2}$ to understand the details of the edge of the proton (e.g. a pion cloud, CQCBM, etc.)
  - $Q^2 >> 1.2\ \text{fm}^{-2}$ to understand transition from hadronic to partonic (e.g. the bound light constitute quarks)





Guesstimation books by Larry Weinstein (ODU)

Test of Additional Term

F distribution table (alpha 0.05)

df_2	1
1	161.4
2	18.51
3	10.13
4	7.71
5	6.61
6	5.99
7	5.59
8	5.32
9	5.12
10	4.96
11	4.84
12	4.75
13	4.67
14	4.60
15	4.54
16	4.49
17	4.45
18	4.41
19	4.38
20	4.35
21	4.32
22	4.30
23	4.28
24	4.26
25	4.24
26	4.22
27	4.21
28	4.20
29	4.18
30	4.17
40	4.08
60	4.00
120	3.92
∞	3.84

A textbook statistics problem is to quantify when to stop adding terms to a fit of experimental data.

One way to do this is with an F-distribution test.

$$F = \frac{\chi^2(j-1) - \chi^2(j)}{\chi^2(j)}(N - j - 1)$$

where j is the order of the fit and N the number points being fit.

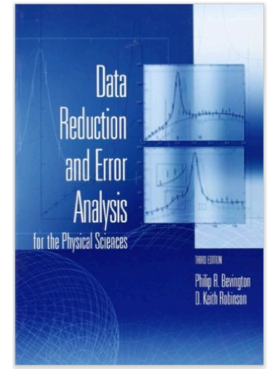
Table 10.2. Maximum degree needed in polynomial approximation.

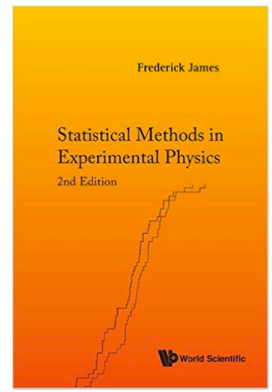
$N-j-1$	2	3	4	6	8	12	20	60	120
Reject j^{th} order to 95% confidence level if F is smaller than	18.5	10.1	7.7	6	5.3	4.7	4.3	4	3.9

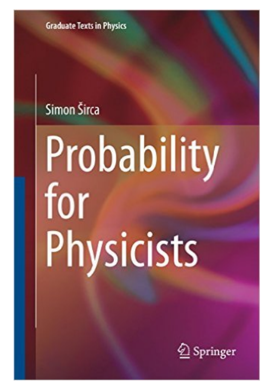
Quantifies a statement that adding a term doesn't significantly improve the fit.

One is free to pick a different alpha, alpha=0.05 is just typical to prevent over-fitting.

(see James 2nd edition page 282, Bevington 3rd edition page 207, or Širca page 95)





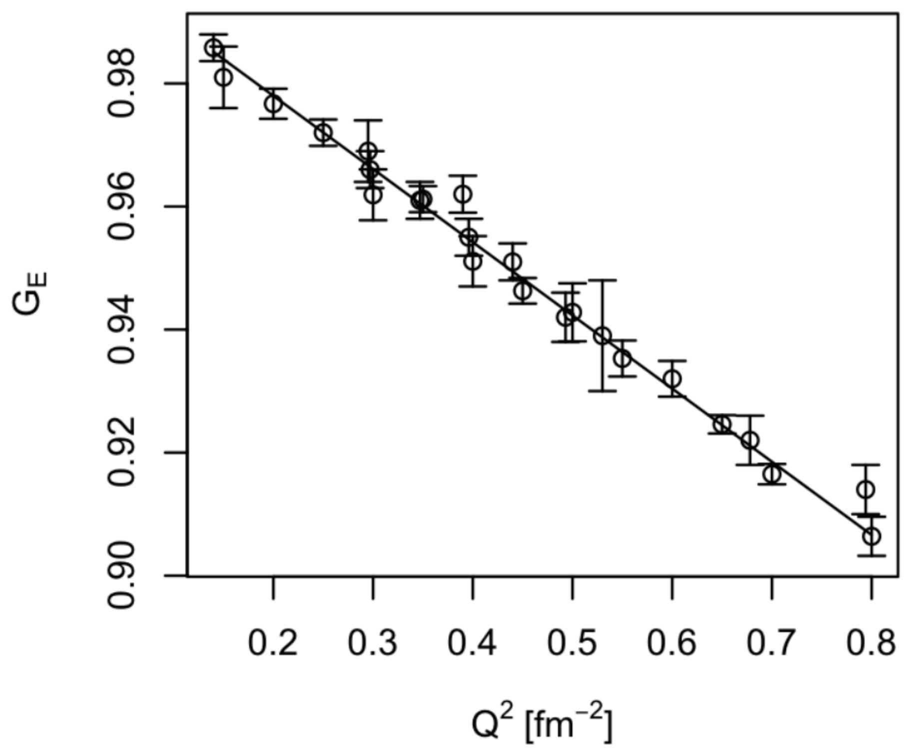


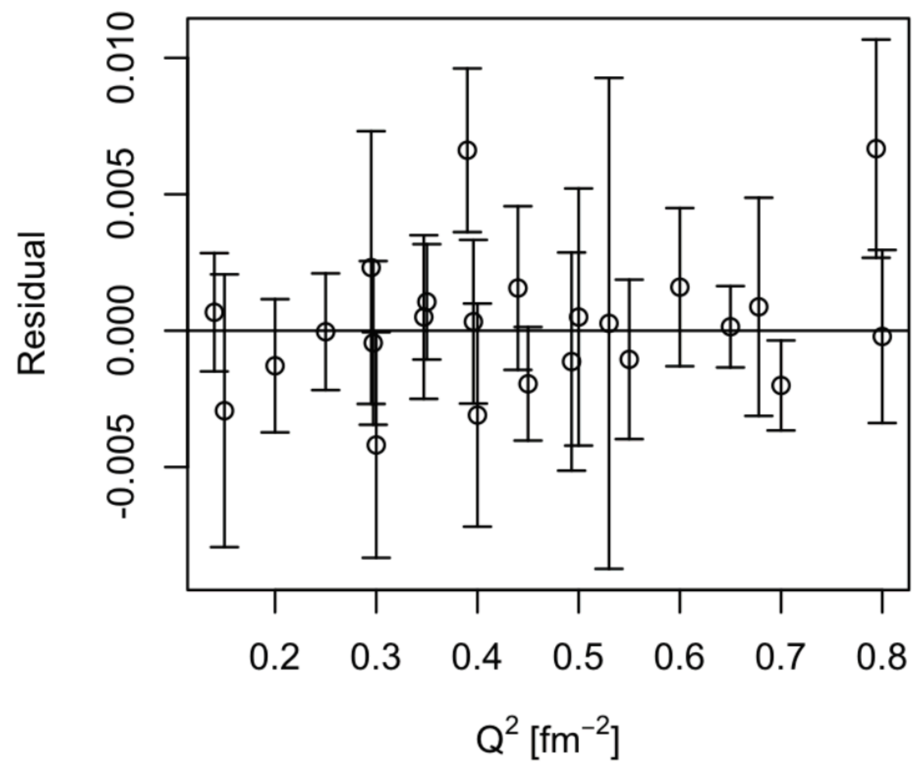
Simple Example

G. G. Simon, C. Schmitt, F. Borkowski, and V. H. Walther, Nucl. Phys. **A333** (1980) 381.
J. J. Murphy, Y. M. Shin, and D. M. Skopik, Phys. Rev. **C9** (1974) 2125.

$$f(Q^2) = n_0 G_E(Q^2) \approx n_0 \left(1 + \sum_{i=1}^{m} a_i Q^{2i}\right)$$

N	j	χ^2	χ^2/ν	n_0	a_1	a_2
24	2	13.71	0.623	1.002(2)	−0.119(4)	
24	3	13.71	0.652	1.002(5)	−0.120(20)	0.00(2)





F-test rejects **fitting** with the more complex j=3 function, that does NOT mean $a_2 = 0$.

F-Test Is Not An Acceptance Test

For a more complex example, F-Test will reject the j=7 fit, but you then need to examine the fits that weren't rejected. This is not an acceptance test!

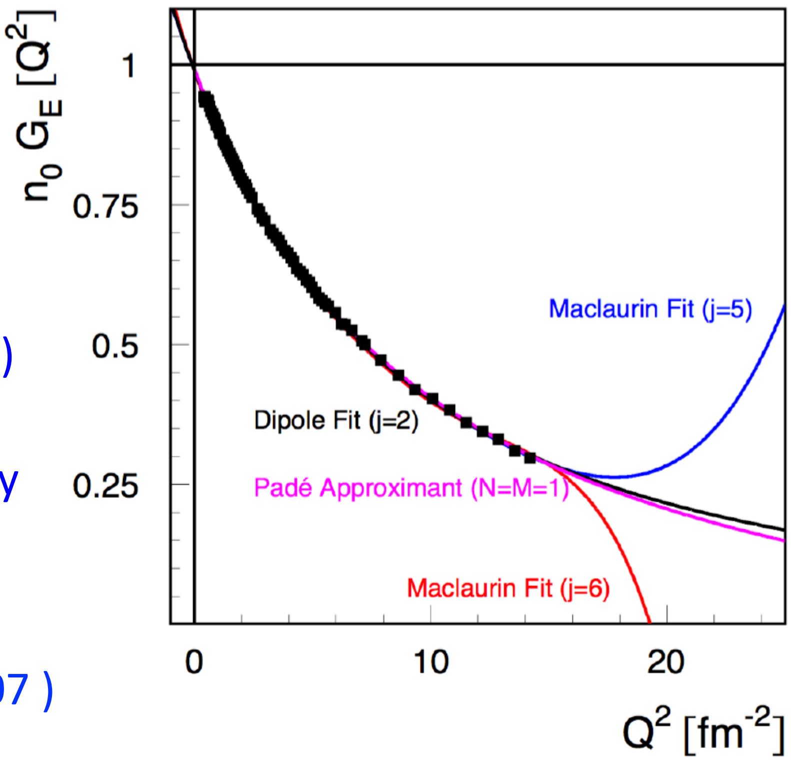
N	j	χ^2	χ^2/ν	n_0	a_1	a_2	a_3	a_4	a_5	a_6
77	5	49.57	0.688	0.991(2)	−0.113(1)	$0.88(1)\cdot10^{-2}$	$-0.44(2)\cdot10^{-3}$	$9.7(8)\cdot10^{-6}$		
77	6	41.34	0.582	0.996(2)	−0.121(1)	$1.25(1)\cdot10^{-2}$	$-1.14(2)\cdot10^{-3}$	$6.8(1)\cdot10^{-5}$	$-1.62(7)\cdot10^{-6}$	
77	7	41.32	0.590	0.995(3)	−0.119(1)	$1.18(1)\cdot10^{-2}$	$-0.93(2)\cdot10^{-3}$	$3.9(1)\cdot10^{-5}$	$0.12(6)\cdot10^{-6}$	$-4.2(5)\cdot10^{-8}$

$$f(Q^2) = n_0 G_E(Q^2) \approx n_0\left(1+\sum_{i=1}^{m} a_i Q^{2i}\right)$$

I find it interesting to note that the a_1 term between j=5 and j=6 bounds the Muonic Lamb shift result (i.e. 0.84fm -> a_1 of -0.1176)

Note you can get 0.88 from this same data by simply going higher order. (i.e. a battle of claims of under-fitting vs. over-fitting)

(for details see Phys. Rev. C **93** (2016) 055207)



In fact, it is clear from our knowledge of G_E than none of these power series fits extrapolate correctly.

Padé Approximant & Continued Fractions

Pade' Approximant

When it exists, the Pade' approximant (N,M) of a Tayler series is unique.

$$f(x) = \frac{a_0 + a_1 x^1 + a_2 x^2 \ldots + a^M * x^M}{1 + b_1 x^1 + b_2 x^2 \ldots + b^N * x^N}$$

In our case we want $f(x) = n_0 G_E(Q^2)$, so

$$f(x) = n_0 \frac{1 + a_1 Q_2 + a_2 Q^4 \ldots + a^{M*2} * Q^{M*2}}{1 + b_1 Q_2 + b_2 Q^4 \ldots + b^{N*2} * x^{N*2}}$$

(Henri Padé ~ 1860)

Continued Fraction

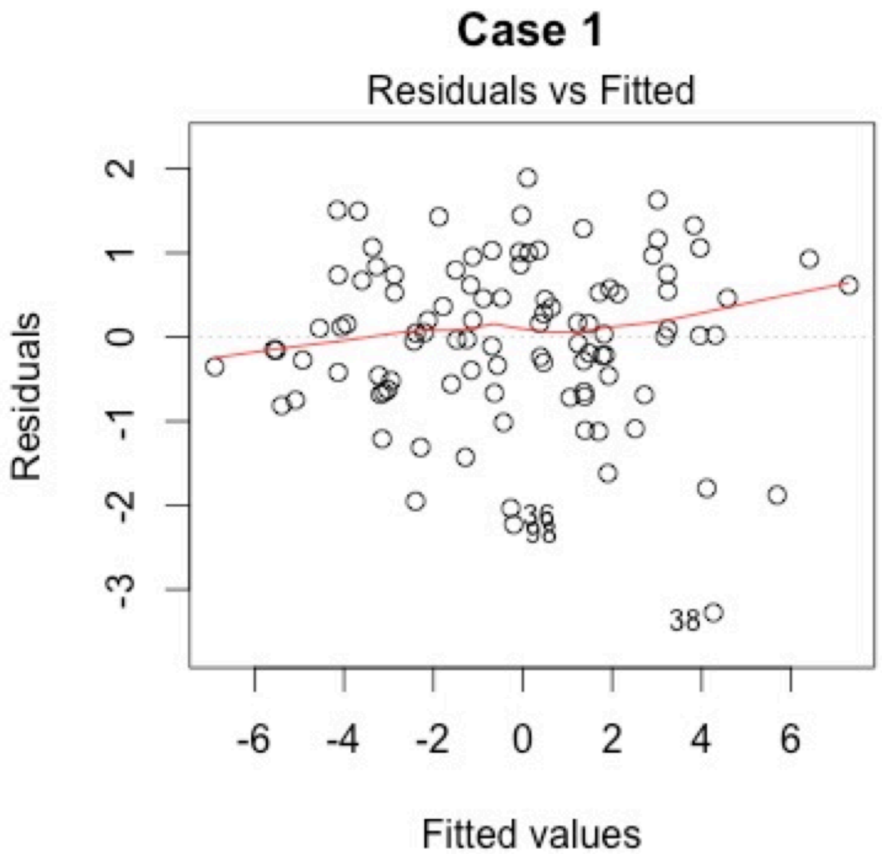
$$f(Q^2) = \cfrac{c_1}{1 + \cfrac{c_2 Q^2}{1 + \cfrac{c_3 Q^2}{1 + \cfrac{c_4 Q^2}{1 + \ldots}}}}$$

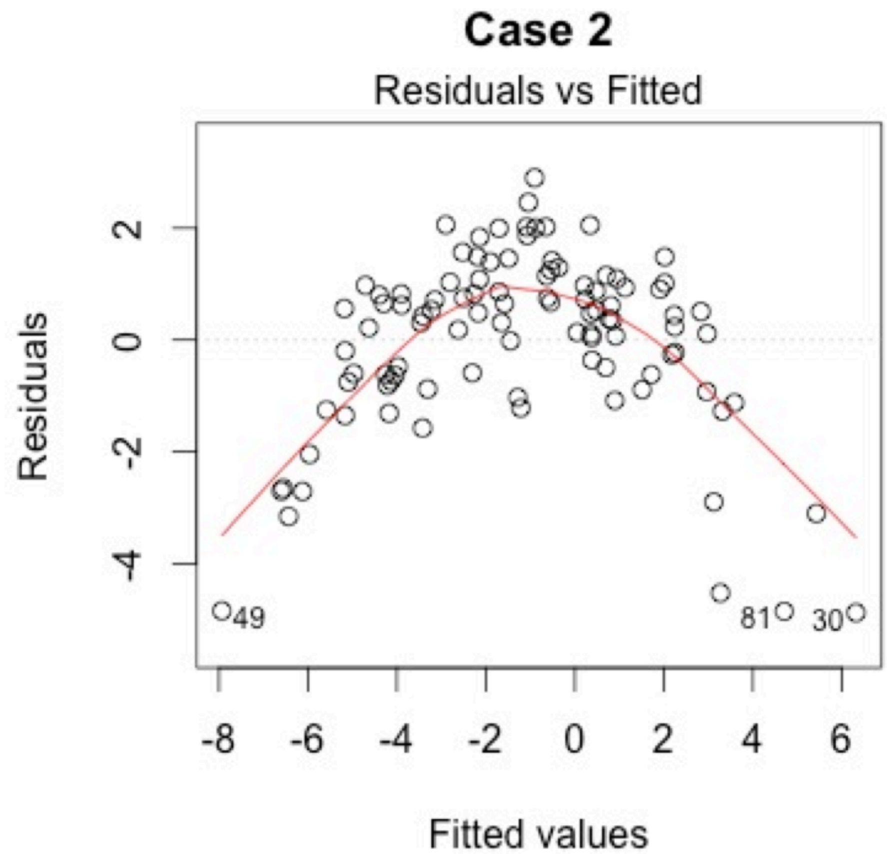
(Ancient Greeks)

Further reading: **Extrapolation algorithms and Padé approximations: a historical survey** C. Brezinski, Applied Numerical Mathematics 20 (1996) 299.

Residuals vs. Fitted Values

Examples taken from http://data.library.virginia.edu/diagnostic-plots/



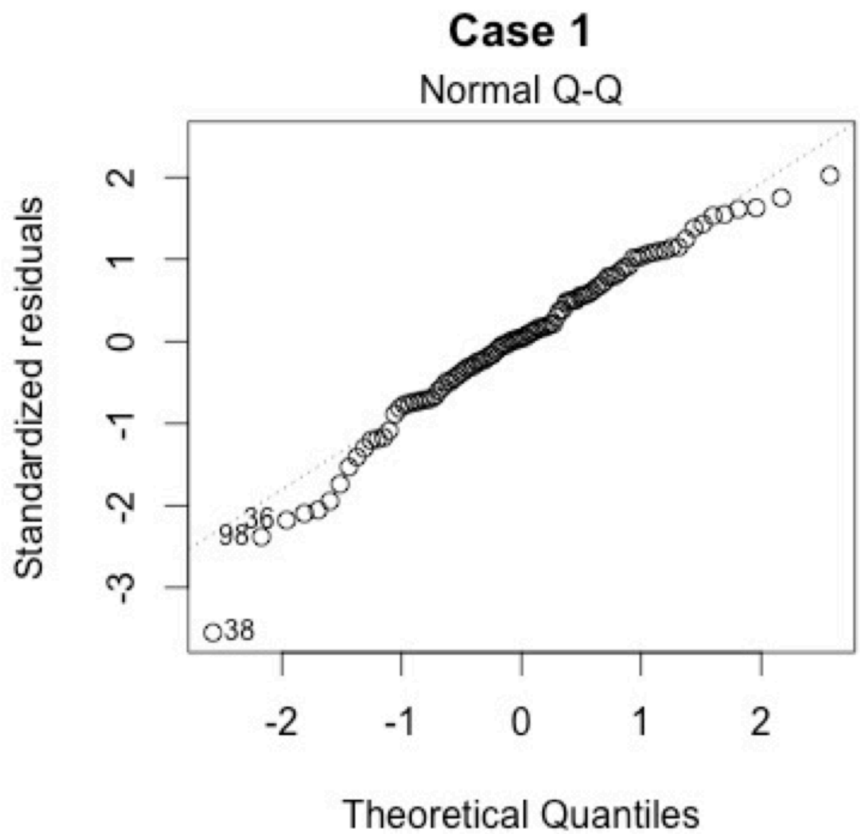


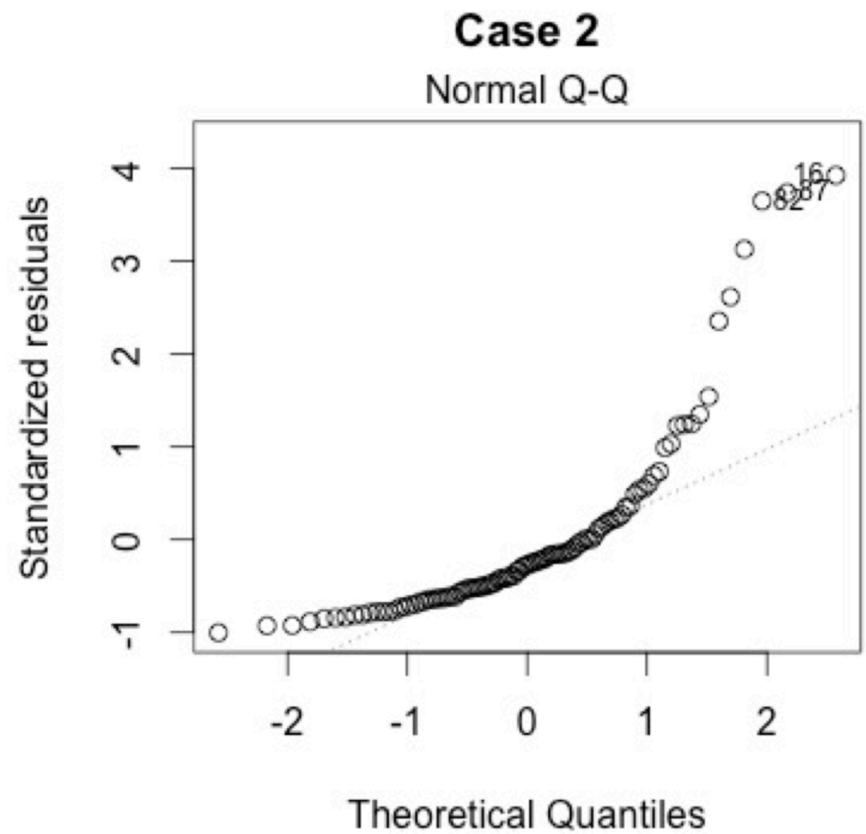
Am I fitting with a reasonable model to describe the data?

Normal Q-Q Plots

Examples taken from http://data.library.virginia.edu/diagnostic-plots/

(also see http://data.library.virginia.edu/understanding-q-q-plots/)



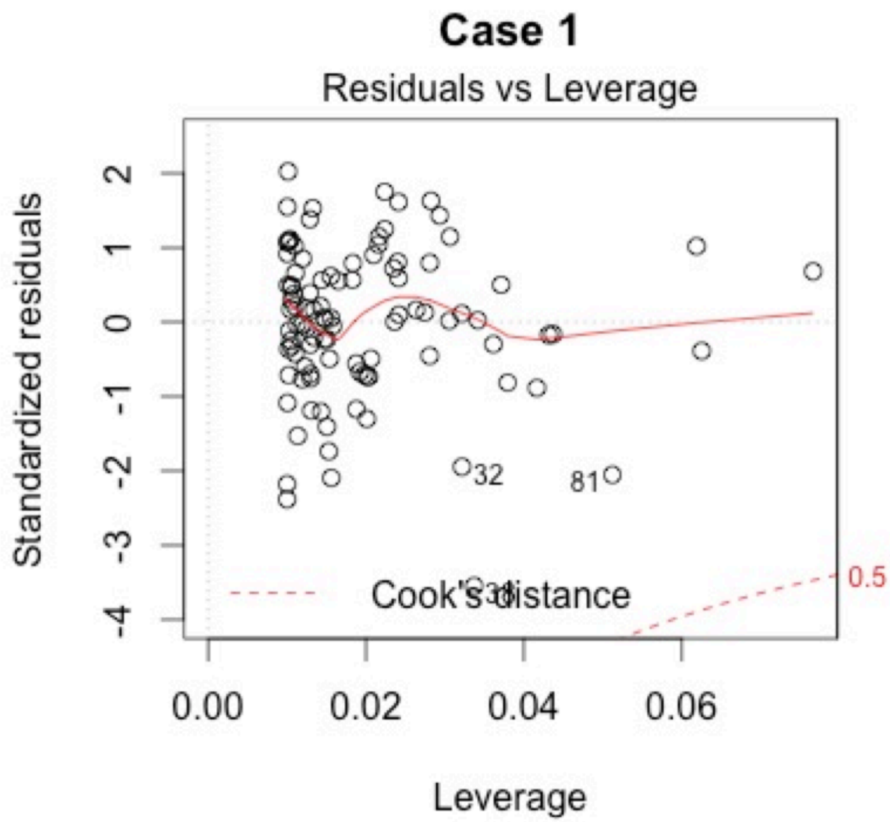


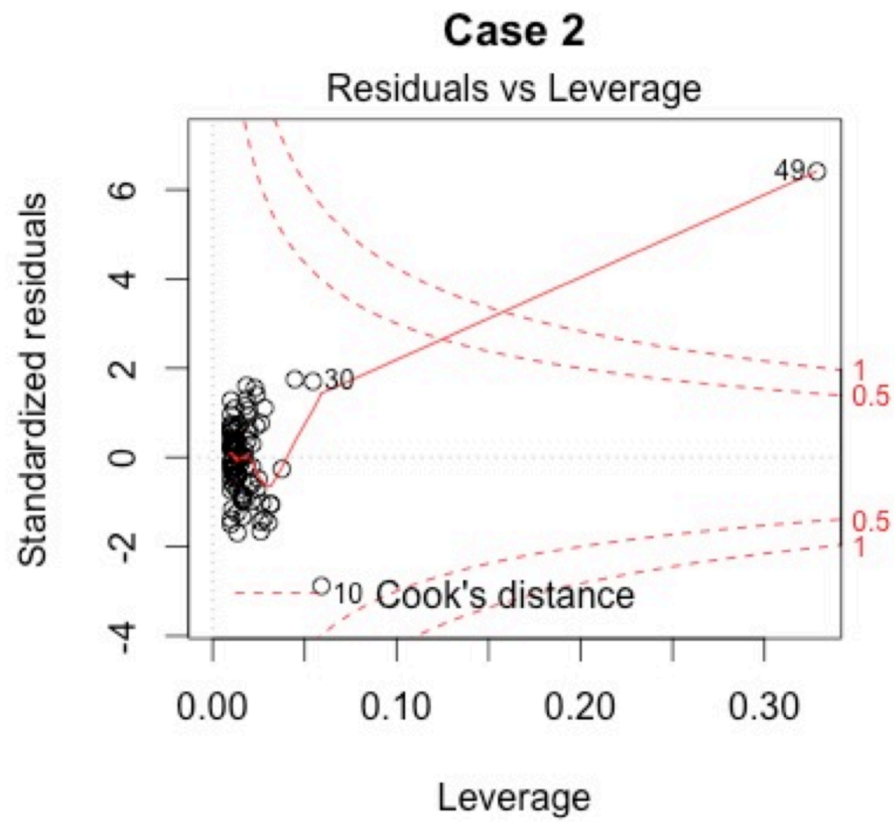
Are the data normally distributed?

(a requirement for many of the other stat. tests to be valid!)

Residuals vs. Leverage

Examples taken from http://data.library.virginia.edu/diagnostic-plots/





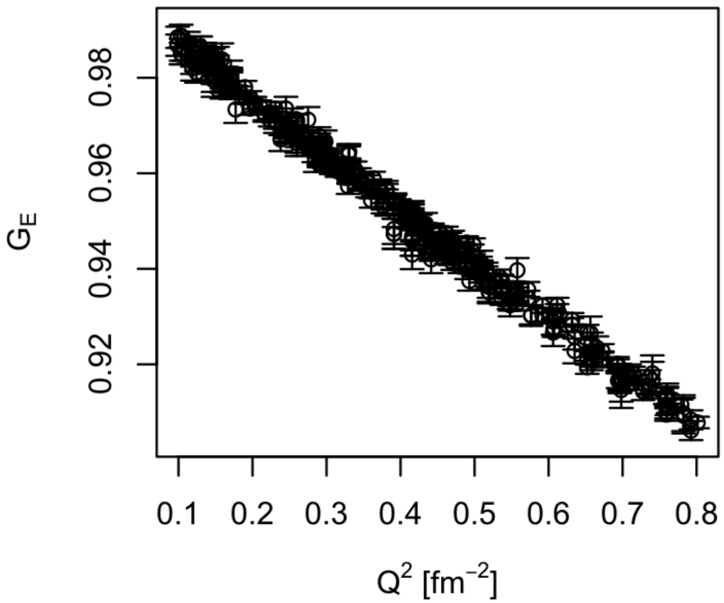
Is a single data point dramatically influencing the fit?

R Programming Language

Language Rank	2015 Spectrum Ranking	2014 Spectrum Ranking
1. Java	100.0	100.0
2. C	99.9	99.3
3. C++	99.4	95.5
4. Python	96.5	93.5
5. C#	91.3	92.4
6. R	84.8	84.8
7. PHP	84.5	84.5
8. JavaScript	83.0	78.9
9. Ruby	76.2	74.3
10. Matlab	72.4	72.8

IEEE Rankings are based mostly on CPU usage (i.e. big data)

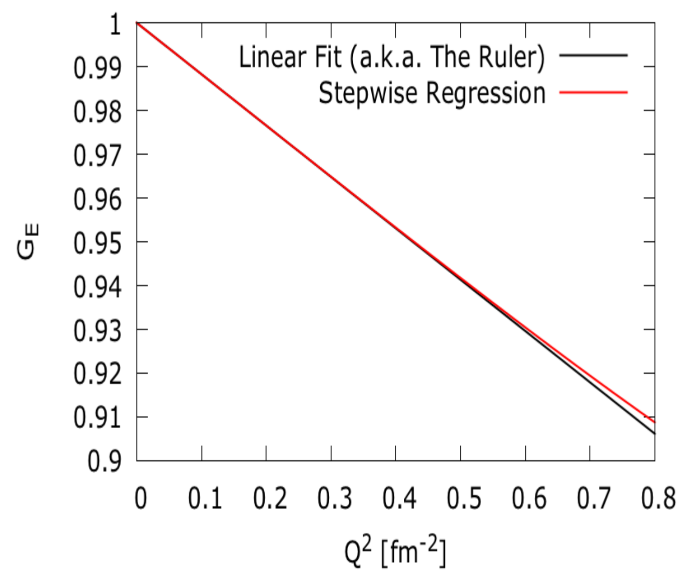
Stepwise Regression of G_E from Carl & Keith

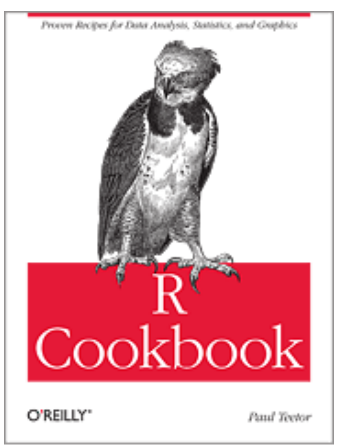



```
Start:  AIC=36.77
data$y ~ data$x

               Df Sum of Sq    RSS    AIC
+ I(data$x^4)   1   10.3725 358.06 29.236
+ I(data$x^3)   1   10.2911 358.14 29.312
+ I(data$x^5)   1   10.2718 358.16 29.330
+ I(data$x^6)   1   10.0519 358.38 29.535
+ I(data$x^2)   1    9.9568 358.48 29.624
+ I(data$x^7)   1    9.7627 358.67 29.804
+ I(data$x^8)   1    9.4401 359.00 30.105
+ I(data$x^9)   1    9.1075 359.33 30.414
+ I(data$x^10)  1    8.7790 359.66 30.719
+ I(data$x^11)  1    8.4620 359.97 31.013
<none>                      368.44 36.774

Step:  AIC=29.24
data$y ~ data$x + I(data$x^4)

               Df Sum of Sq    RSS    AIC
<none>                      358.06 29.236
+ I(data$x^2)   1 0.0088531 358.05 31.228
+ I(data$x^3)   1 0.0028516 358.06 31.233
+ I(data$x^11)  1 0.0007801 358.06 31.235
+ I(data$x^5)   1 0.0006383 358.06 31.236
+ I(data$x^6)   1 0.0004668 358.06 31.236
+ I(data$x^7)   1 0.0003015 358.06 31.236
+ I(data$x^10)  1 0.0001705 358.06 31.236
+ I(data$x^8)   1 0.0001061 358.06 31.236
+ I(data$x^9)   1 0.0000000 358.06 31.236
```

Akaike Information Criterion Selected Model

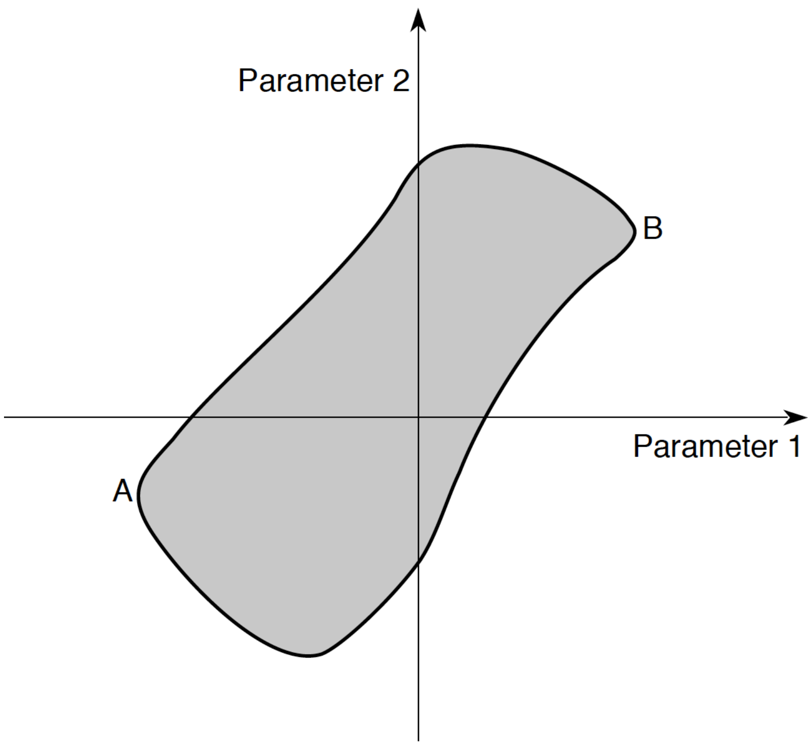
```
Call:
lm(formula = data$y ~ data$x + I(data$x^4), weights = 1/data$dy^2)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-3.02110 -0.73469 -0.08639  0.66588  3.08298

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.9988419  0.0003534 2826.253  < 2e-16 ***
data$x      -0.1172672  0.0010936 -107.229  < 2e-16 ***
I(data$x^4)  0.0063583  0.0020534    3.097  0.00213 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 331 degrees of freedom
Multiple R-squared:  0.9932,    Adjusted R-squared:  0.9932
F-statistic: 2.434e+04 on 2 and 331 DF,  p-value: < 2.2e-16
```





Pohl *et.al's* 0.84 fm radius would predict a slope of - 0.1176

Multivariate Errors



The Interpretation of Errors in Minuit (2004 by James)

seal.cern.ch/documents/minuit/mnerror.pdf

As per the particle data handbook, one should be using a co-variance matrix and calculating the probably content of the hyper-contour of the fit. Default setting of Minuit of "up"(often call $\Delta\chi^2$ is one.

Also note standard Errors often underestimate true uncertainties. (manual of gnuplot fitting has an explicate warning about this)

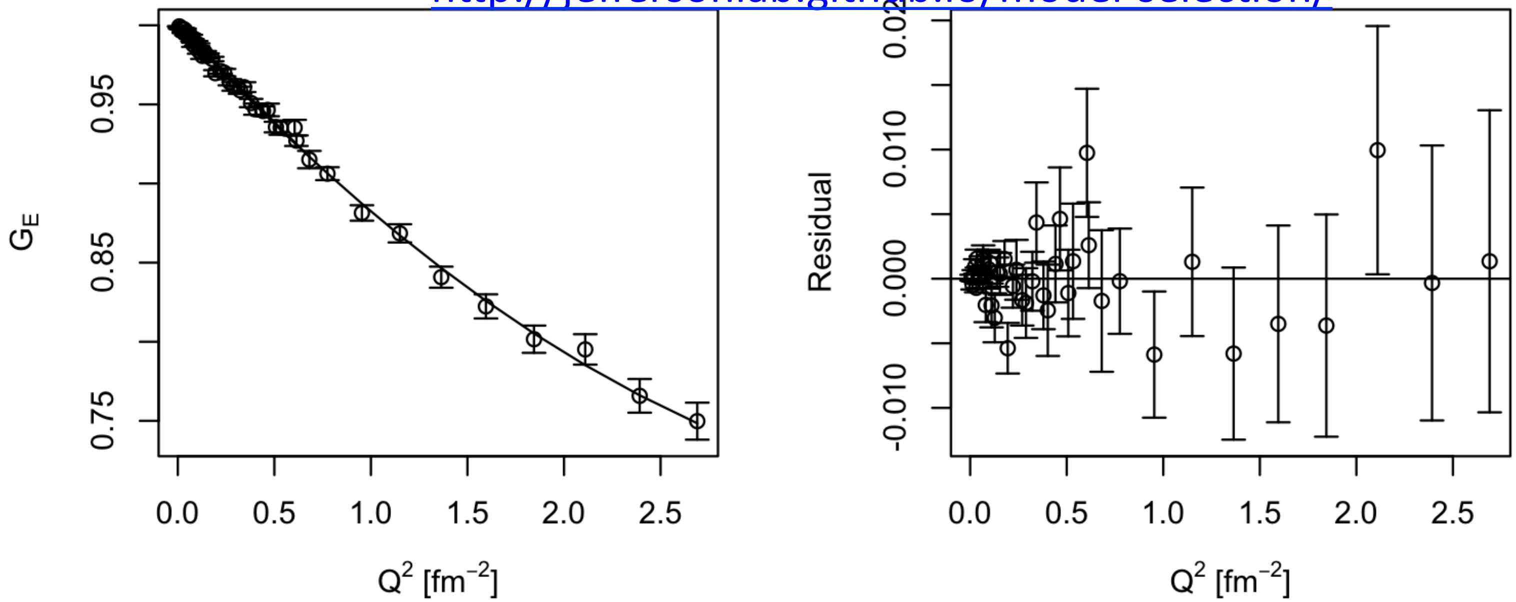
Number of Parameters	Confidence level (probability contents desired inside hypercontour of $\chi^2 = \chi^2_{min} + up$)				
	50%	70%	90%	95%	99%
1	0.46	1.07	2.70	3.84	6.63
2	1.39	2.41	4.61	5.99	9.21
3	2.37	3.67	6.25	7.82	11.36
4	3.36	4.88	7.78	9.49	13.28
5	4.35	6.06	9.24	11.07	15.09
6	5.35	7.23	10.65	12.59	16.81
7	6.35	8.38	12.02	14.07	18.49
8	7.34	9.52	13.36	15.51	20.09
9	8.34	10.66	14.68	16.92	21.67
10	9.34	11.78	15.99	18.31	23.21
11	10.34	12.88	17.29	19.68	24.71
	If FCN is $-\log(\text{likelihood})$ instead of χ^2, all values of up should be divided by 2.				

In ROOT: **SetDefaultErrorDef(X.X)**

Default is 1 and doesn't change unless you change it!

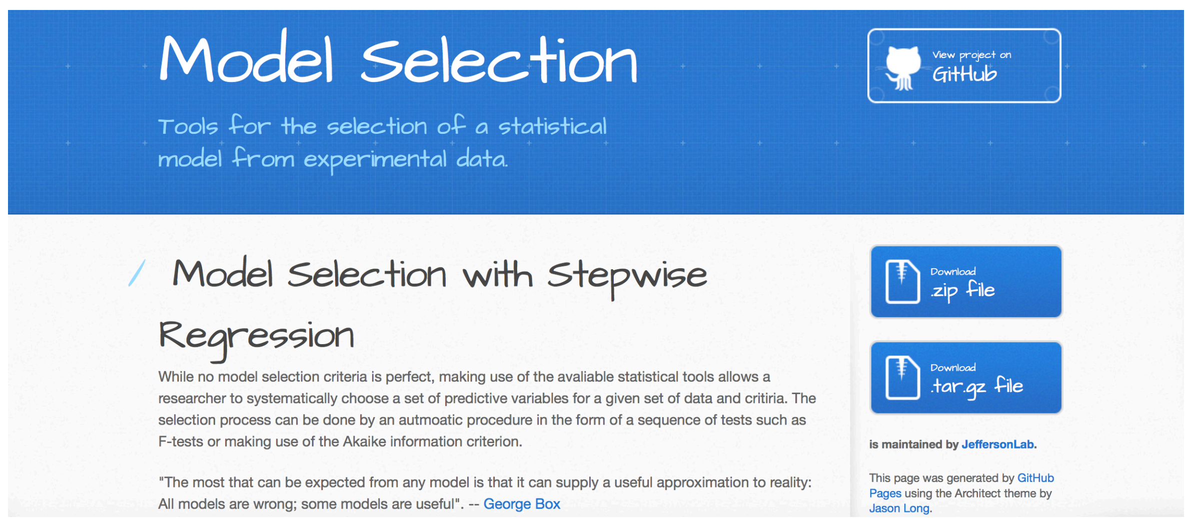
Expected PRad Results (for 0.88 fm radius)

http://jeffersonlab.github.io/model-selection/



Show is a stepwise regression using Monte Carlo of the expected PRad data for a 0.88 fm radius.

This is a range of data very similar to the HAND *et al.* 1963 review article.



Bayesian Priors (The Star Wars Example)

https://www.countbayesie.com/blog/2015/2/18/hans-solo-and-bayesian-priors

- C3PO can calculate the odds of a pilot navigating an asteroid field (20,000:1)

$$P(\text{RateOfSuccess}|\text{Successes}) = Beta(\alpha, \beta)$$

- But Han Solo is one of the best pilots in the galaxy. (i.e. C3P0 ignored a Bayesian Prior)

$$Beta(\alpha_{\text{posterior}}, \beta_{\text{posterior}}) = Beta(\alpha_{\text{likelihood}} + \alpha_{\text{prior}}, \beta_{\text{likelihood}} + \beta_{\text{prior}})$$

- So C3PO actually correctly predicts that average pilots will not successfully navigate the field while incorrectly predicting Han's chances. (estimated as 75% in the article)
- Ignoring A Bayesian Prior Can Lead To Wrong Conclusions

Warning: Danger of Confirmation Bias

In psychology and cognitive science, confirmation bias is a tendency to search for or interpret information in a way that confirms one's preconceptions, leading to statistical errors.

Believe Your Data !!

- **Electric and Magnetic Form Factors of the Nucleon**
  - L.N. Hand, D.G. Miller, Richard Wilson, Rev. Mod. Phys. **35** (1963) 335
  - Easy data to play with and see if you can get Hand's results.
- Particle Data Handbook – Statistics Section
  - *http://pdg.lbl.gov/2015/reviews/rpp2015-rev-statistics.pdf*
- The Interpretation of Errors – Fredrick James
  - *http://seal.cern.ch/documents/**minuit**/mn**error**.pdf*
- Data Analysis Textbooks
  - Data Reduction and Error Analysis – Philip Bevington
  - Statistical Methods in Experimental Physics – Fredrick James
  - Computation Methods for the Physical Science – Simon Širca
  - Probability of Physics – Simon Širca
- R Programing Language
  - https://www.r-project.org/
- Estimation
  - Street-Fighting Mathematics (open source) – Sanjoy Mahajan
  - Guesstimation – Larry Weinstein

"Proton Radius Puzzle" in 1975 !?

F. Borkowski, G.G. Simon, V. H. Walther, and R. D. Wendling, Nucl. Phys. **B93** (1975) 461.

$$G_{E,M}(q^2) = 1 - \tfrac{1}{6}\langle r^2_{E,M}\rangle |q|^2 + \tfrac{1}{120}\langle r^4_{E,M}\rangle |q|^4 - + \ldots, \qquad (6)$$

For $q^2 < 0.9\ \mathrm{fm}^{-2}$ the contributions of the higher terms in the expansion (6) are negligable and the series can be truncated to give $G_E(q^2) = \delta + \beta q^2$. From fitting this expression to the form factors of fig. 5, the solid line of fig. 5 has been obtained. The best fit parameters were $\delta = 0.994 \pm 0.002$ and $\beta = -0.118 \pm 0.004\ \mathrm{fm}^2$. The reduced χ^2 was 0.5. The result of the fit did not depend significantly on the fitted q^2 range. This was checked by fitting additionally the G_E values of table 2 up to $1.2\ \mathrm{fm}^{-2}$. The addition of a q^4 term to the fit formula did not improve the fit, moreoever the error of the additional parameter turned out to be larger than its value. The best fit value of the parameter δ is well within the normalization error of the G_E values. The best fit value of the parameter β gives a proton r.m.s. radius of $\langle r_E^2\rangle^{\frac{1}{2}} = 0.84 \pm 0.02$ fm. This value is higher than the dipole value of 0.81 fm, but within the error limits it is compatible with the result (0.81 ± 0.04 fm) of a similar experiment carried out at Saskatoon [7].

This is the same conclusions one gets with stepwise regression using the new data Mainz though with much smaller uncertainties.

Particle Data Handbook

By setting “ErrorDef” to 2.71 ROOT would report an m=1 90% coverage probalitiy instead of 68%.

Table 38.2: Values of $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

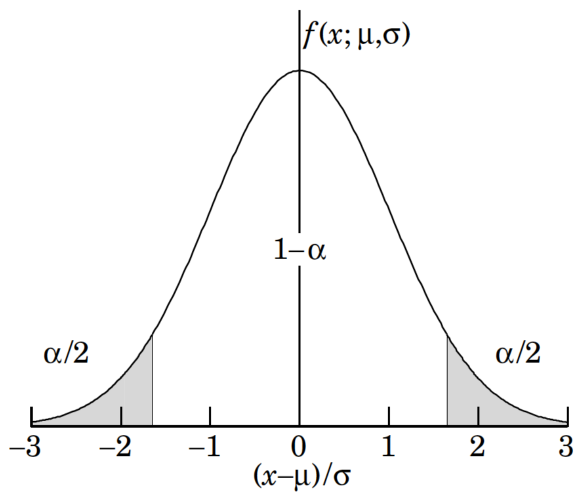


Figure 38.4: Illustration of a symmetric 90% confidence interval (unshaded) for a measurement of a single quantity with Gaussian errors. Integrated probabilities, defined by $\alpha = 0.1$, are as shown.

Confidence interval + alpha = 1