Chaos, Complexity, and Inference (36-462)

Lecture 17: Inference in General

Cosma Shalizi

16 March 2009

Some Basics About Inference in General

Further reading: Mayo (1996), the book which got me interested in statistics
Also: Abelson (1995); Kelly (1996)

What Is Statistics?

Reliable rule of inference: one which is unlikely to lead us into (big) errors

Statistics is:

a branch of applied mathematics studying models of inference from random data

a branch of methodology evaluating techniques for non-demonstrative inferences from ditto

a form of principled rhetoric a practice of persuasion, using *honest* arguments from ditto

Arguing

Deductive logic is a mode of principled argument:

> *"if you believe A and B, then you should also believe C"*

Bayesian statistics tries to extend this to fractions of a deduction:

> *"if you p-believe A and q-believe B, then you should r-believe C"*

Won't follow that path for several reasons — a big one is that it gives no reliability guarantees (Wasserman, 2006)
Instead, look at arguments about errors

Error Statistics

Simplest sort of error: parameter estimation
Want errors to be small, rare, detectable, etc.
Basic argument:

> *if you agree that the data were generated by P_θ for some $\theta \in \Theta$, then it's really unlikely that the population θ is very different from $\hat{\theta}$*

Important properties of estimators are error properties:

Unbiased No systematic error

Consistent Errors shrink, can be made arbitrarily small by taking enough data (in probability)

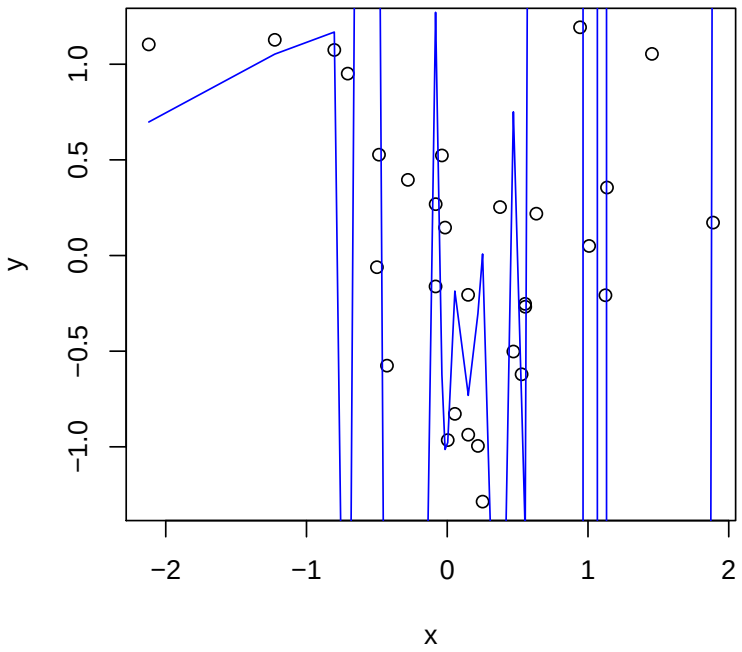
Unbiased minimum variance No systematic error, smallest achievable statistical error

etc.

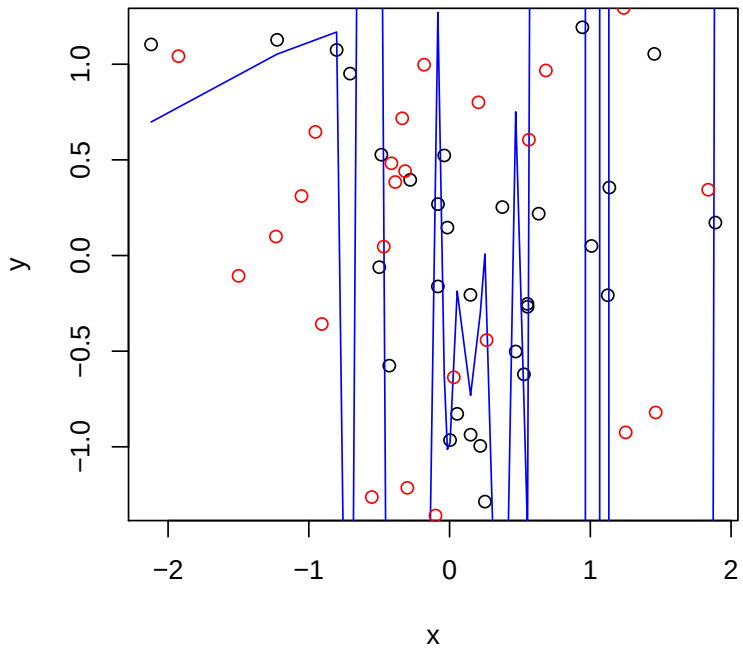
Maximum likelihood estimator is justified by its error properties (when it has them), not by the intrinsic indubitable rightness of the likelihood principle

Constructing Reliable Hypotheses

Problem with making up a hypothesis after seeing the data: fooling yourself
Over-fitting is the most basic form of this
Tight fit to the data is *in general* a poor argument for the model



Fitting noise
with a high-order polynomial



. . . doesn't generalize

When *is* fitting the data a good argument?

Obvious situation: The model is pre-specified ("first plot your curve, then take your data")

Almost as obvious: The data are a new sample, not what you fit to

"Probably Approximately Correct"

Find best-fitting model $\hat{H}$ in class $\mathcal{H}$

Probably approximately correct (PAC) results say:

> *given enough data $N(\epsilon, \delta)$, then with probability at least $1 - \delta$ ("probably"), the generalization error of $\hat{H}$ is within ϵ of ("approximately") the best $H^* \in \mathcal{H}$ ("correct")*

The function $N(\epsilon, \delta)$ depends on the size/richness/complexity of $\mathcal{H}$, not of $\hat{H}$

Stronger than consistency because of the finite-sample bounds on errors

Resources: Kearns and Vazirani (1994); Vapnik (2000)

Process-Oriented Evaluation

PAC bounds are worst case, work by restricting the model class
Can get more optimistic (tighter) bounds by examining the *process* of picking out a model
Good fit much more impressive if you tried 10 models than if you tried 10^{10}
Details on **process-oriented evaluation**: Domingos (1999) and other papers

Confidence Sets

$1 - \alpha$ confidence set: all the hypotheses we cannot reject at level α

Either:

1. the true value is in the confidence set
2. OR we're very unlucky (something with prob. less than α happened)
3. OR the background assumptions are wrong

Automatic reliability (if the background assumptions hold)

Note: everyone agrees on confidence sets, no matter what priors

may not correspond to a $1 - \alpha$ *credible* set

Obviously coverage property isn't the *only* thing one wants!

Can be extended to different model structures as well as parameters

e.g., TETRAD project here at CMU, returns (samples from) the set of causal structures compatible with data;

`http://www.phil.cmu.edu/projects/tetrad/`

Severity

Reliability (again): the probability of our getting reaching this conclusion, if it's wrong, is low

Hypothesis testing: size and power are bounds on error probabilities that hold regardless of the data

Why power matters: good fit to the null only gives you *evidence* if you could have *noticed* if the null was wrong

Probability of noticing $\equiv$ power

Severity extends size and power in a data-dependent way

Definition (Severity of Rejection)

1 − Pr(Probability of getting results which fit the model at least as badly as the data did, if the model were right)

Basically, the $(1 - p)$-value!

Definition (Severity of Acceptance)

1 − Pr(Probability of getting results which fit the model at least as well as the data did, if the model were wrong)

a.k.a. **achieved power**
Severity of acceptance depends on the alternative, just like the power

What Kind of Null Hypothesis?

"Nothing to see here, move along" null: any apparent pattern is due to noise or some *boring* process

Outstanding example: neutral models in evolution and ecology; apparent adaptation is caused by non-adaptive evolutionary processes

Requires careful specification of the *boring* mechanism

"My linear regression coefficient is zero" is usually *not* a plausible neutral model

Null hypotheses should be carefully crafted to probe for specific errors

Mis-Specification and Residuals

Mis-specification: Functional form or distributional assumptions are wrong

Specifications are often in terms of the residuals ("take Gaussian noise, add this, square that...")

Small residuals are NOT a good sign of model fit

at least not if you believe your statistical model!

PATTERNLESS residuals are a good sign of model fit

Remember: "any signal distinguishable from noise is insufficiently compressed"

residual patterns mean statistical inadequacy

Demonstration that Small R^2 Does Not Indicate a Bad Model

True model: $Y = X + \eta$, $\eta \sim \mathcal{N}(0, 1)$

```
> x1 <- runif(1000,min=0,max=100)
> y1 <- x1 + rnorm(1000)
> summary(lm(y1 ~ x1))
> x2 <- runif(1000,min=50,max=50.1)
> y2 <- x2 + rnorm(1000)
> summary(lm(y2 ~ x2))
```

With the *same model*, R^2 can be either 0.9988 or 0.0006513, depending on range of input variable

Mis-Specification Testing

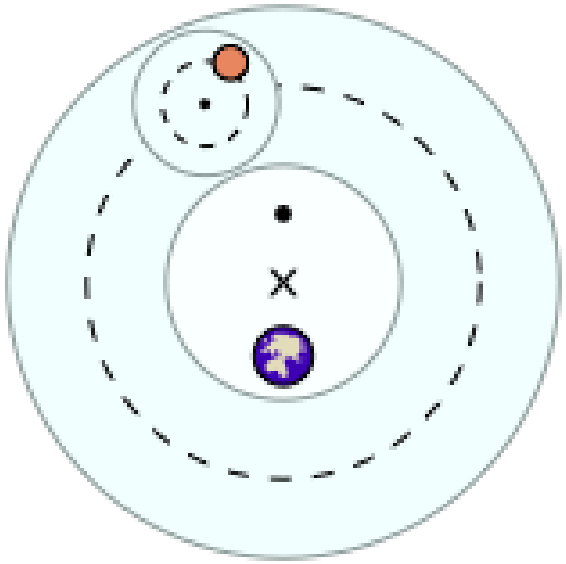
Mis-specification testing in general: if the form of the model is right, then such-and-such should follow (no matter what the actual parameter value); check whether it does

General mis-specification tests exist, but are fairly advanced (White, 1994)

Particular assumptions like Gaussian, IID disturbances can be tested more easily, more powerfully, and tell you more

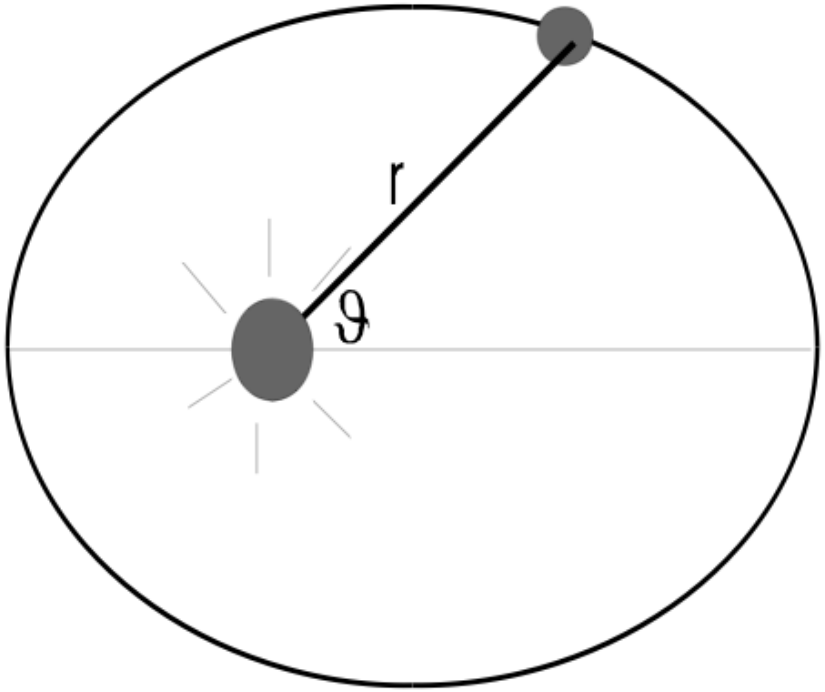
Case Study: The Two Chief World Systems

PTOLEMY: planets move around a point slightly off ("eccentric") the Earth; basic motion is a uniform circle; additional smaller circular motions ("epicycles") imposed on top of that



Wikipedia, s.v. "Deferent and Epicycle"

KEPLER: Planets move in ellipses, with the Sun at one focus, and variable speeds (equal areas swept out in equal times)



from Spanos (2007)

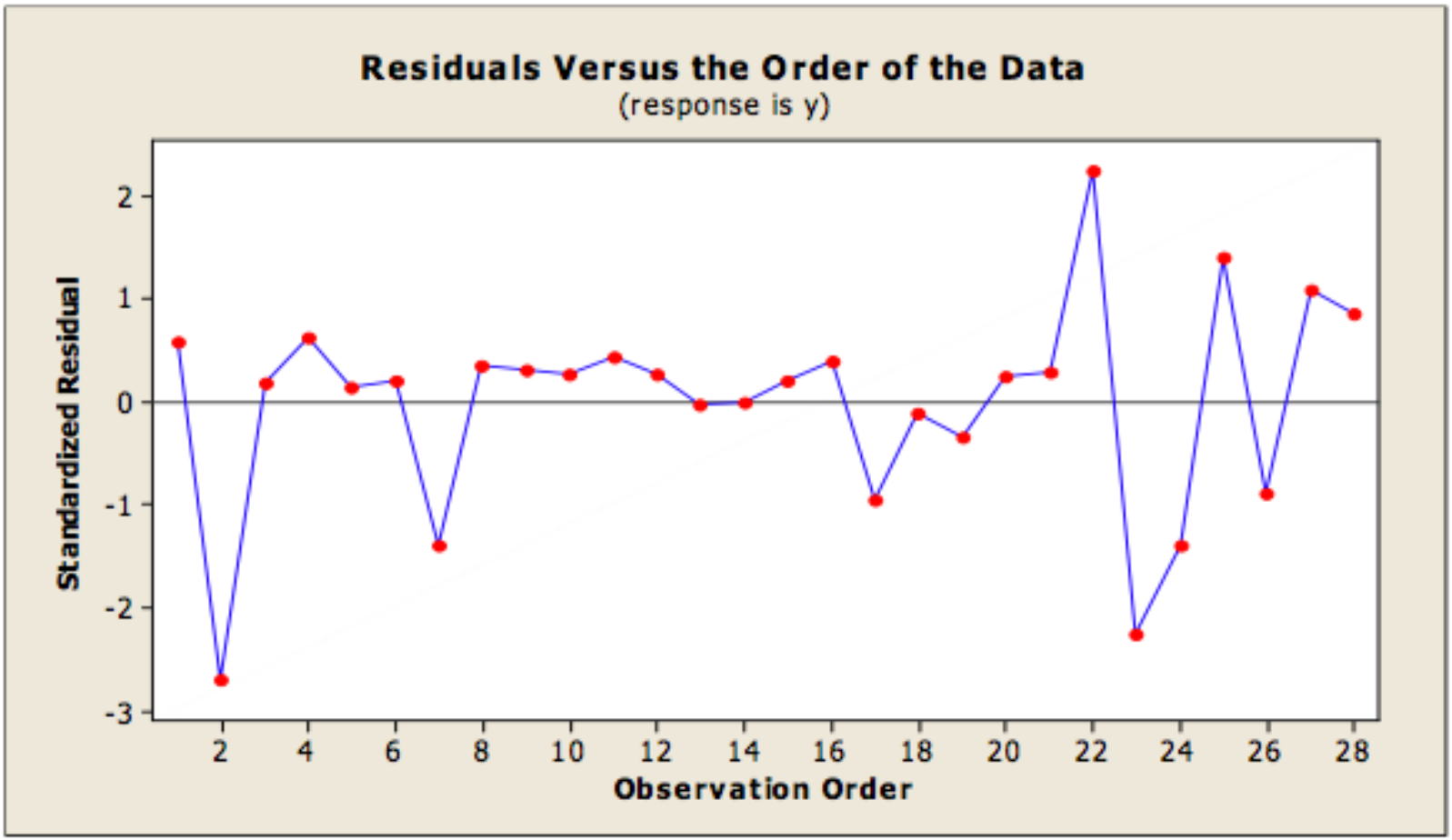
Kepler is right (to an *excellent* approximation), and Ptolemy is wrong

First Proof: Galilei (1632/1953)

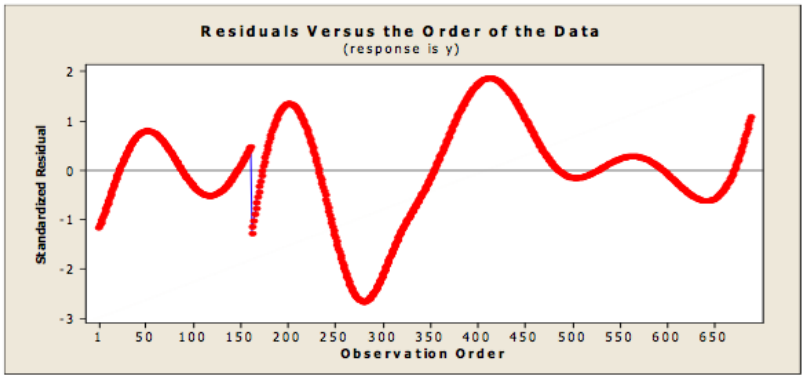
Second Proof: We can send our robots there!

Can our statistics get this right?

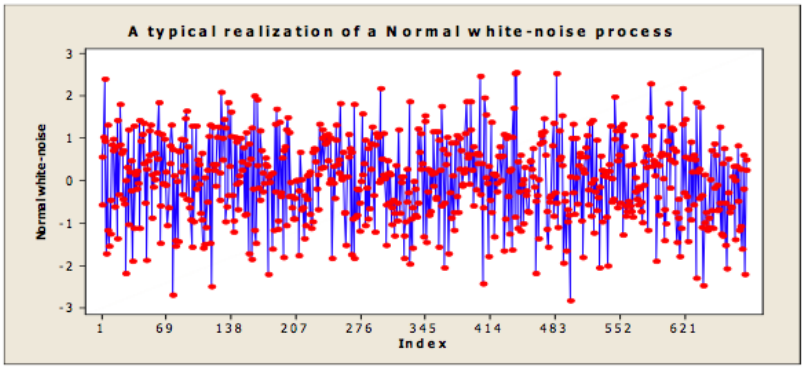
Kepler: excellent prediction, but more importantly *white noise residuals*



Ptolemy: at least as accurate a prediction, but *very* patterned residuals



Ptolemaic residuals



Comparable white noise

Fails formal mis-specification tests with p values $< 10^{-5}$

Abelson, Robert P. (1995). *Statistics as Principled Argument*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Domingos, Pedro (1999). "The Role of Occam's Razor in Knowledge Discovery." *Data Mining and Knowledge Discovery*, **3**: 409–425. URL `http://www.cs.washington.edu/homes/pedrod/papers/dmkd99.pdf`.

Galilei, Galileo (1632/1953). *Dialogue Concerning the Two Chief World Systems, Ptolemaic and Copernican*. Berkeley: University of California Press. Translated by Stillman Drake from *Dialogo di Galileo Galilei Linceo, Matematico Sopraordinario dello Studio di Pisa: E Filosofo, e Matematico primario del Serenissimo Gr. Duca di Toscana: Douce ne i congressi di quattro giornate si discorre sopra i due Massimi Sistemi del Mondo, Tolemaico, e Copernicano; Proponendo*

indeterminatamente le ragioni Filosofiche, e Naturali tanto per l'una, quanto per l'altra parte (Fiorenza: Gio. Batista Landini); with a forword by Albert Einstein.

Kearns, Michael J. and Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory*. Cambridge, Massachusetts: MIT Press.

Kelly, Kevin T. (1996). *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Spanos, Aris (2007). "Curve Fitting, the Reliability of Inductive Inference, and the Error-Statistical Approach." *Philosophy of Science*, **74**: 1046–1066. doi:10.1086/525643.

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory*. Berlin: Springer-Verlag, 2nd edn.

Wasserman, Larry (2006). "Frequentist Bayes is Objective." *Bayesian Analysis*, **1**: 451–456. URL `http://ba.stat.cmu.edu/journal/2006/vol01/issue03/wasserman.pdf`.

White, Halbert (1994). *Estimation, Inference and Specification Analysis*. Cambridge, England: Cambridge University Press.