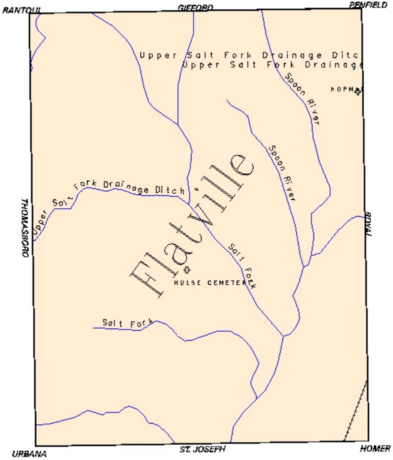
Under Appropriate Regularity Conditions: And Without Loss of Generality

Roger Koenker

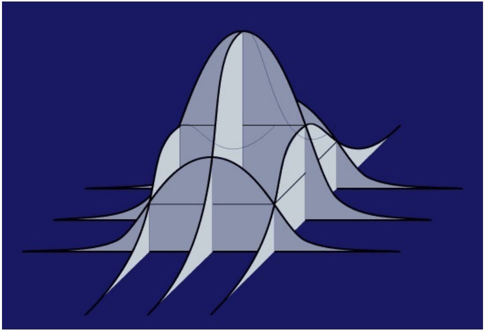
University of Illinois, Urbana-Champaign

MEG: 6 October 2011

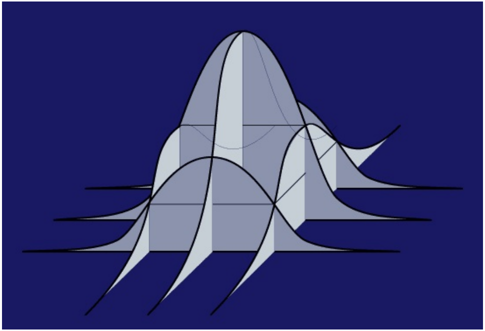
Where's the Hill?



The Imaginary Gaussian Hill



The Imaginary Gaussian Hill



Only a house of cards, if the truth were known.

The Devil and the Deep BLUE Theorem

Theorem (Gauss-Markov)

Given a random vector, $Y \in \mathbb{R}^n$ *with* $\mu = \mathbb{E}Y \in L$*, a linear subspace of* $\mathbb{R}^n$*, and* $\Omega = \mathbb{V}Y$*, the projection* $\hat{\mu} = P(Y)$ *onto* L *that maps the subspace* $K = \{u | u^\top \Omega v = 0,\ v \in L\}$ *conjugate to* L *into the origin, has a concentration ellipsoid contained in that of every other linear, unbiased estimator of* μ*.*

Appropriate regularity:

- μ and Ω must exist, heavy tails need not apply, "of particular interest in econometrics, since the distribution of the 'errors' is rarely known."
- The ghostly, $X :$ span$(X) = L$ can be singular, so too can Ω,
- "The usual estimators are linear."
- "However, we might be interested in allowing some bias . . . "

Why Are the Usual Estimators Linear?

Three possible explanations:

- Because $Y - \mu$ is Gaussian: uniquely $\frac{d}{dx} \log(\phi(x)) = -x$, any other distributional assumption yields a nonlinear estimator,

Why Are the Usual Estimators Linear?

Three possible explanations:

- Because $Y - \mu$ is Gaussian: uniquely $\frac{d}{dx} \log(\phi(x)) = -x$, any other distributional assumption yields a nonlinear estimator,
- Because Haavelmo (1944) told us that $Y - \mu$ should be approximately Gaussian by CLT considerations,

Why Are the Usual Estimators Linear?

Three possible explanations:

- Because $Y - \mu$ is Gaussian: uniquely $\frac{d}{dx} \log(\phi(x)) = -x$, any other distributional assumption yields a nonlinear estimator,
- Because Haavelmo (1944) told us that $Y - \mu$ should be approximately Gaussian by CLT considerations,
- Because we know how to solve linear equations.

Why Are the Usual Estimators Linear?

Three possible explanations:

- Because $Y - \mu$ is Gaussian: uniquely $\frac{d}{dx}\log(\phi(x)) = -x$, any other distributional assumption yields a nonlinear estimator,
- Because Haavelmo (1944) told us that $Y - \mu$ should be approximately Gaussian by CLT considerations,
- Because we know how to solve linear equations.

Three Contra-explanations:

- Linear estimators are qualitatively non-robust, and therefore can be highly inefficient in heavy tailed circumstances,
- Belief in a Lindeberg condition for unobservable contributions to model noise is just wishful thinking,
- More robust estimators are also easy to compute.

Normality: A Short Story

In the summer of 1872 Charles Saunders Peirce conducted a series of experiments designed to evaluate the applicability of the Gaussian law of errors, and thus of least squares methods, for observational data commonly used in astronomy.

- A young man, with no prior experience, was hired and asked to respond to “a signal consisting of a sharp sound” by depressing a telegraph key “nicely adjusted.”
- Response times were recorded in milliseconds with the aid of a Hipp chronoscope.
- For 24 days in July and early August, 1872, roughly 500 measurements were made for each day.

Peirce and the Hipp Chronoscope

(a) C.S. Peirce, (1839-1914) American scientist, philosopher, mathematician extra-ordinaire.

(b) Hipp Chronoscope (1848 –) Swiss instrument widely used in early experimental psychology experiments on reaction times.

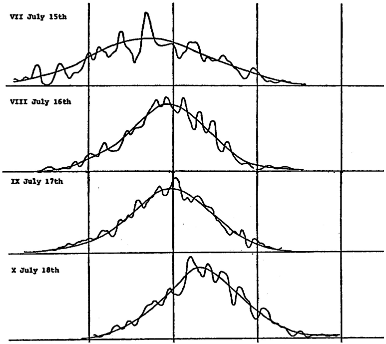
Day 6: The Experimental Data

SIXTH DAY, JULY 10, 1872

66	1	117	0	137	2	157	5	177	4	197	3	217	1	237	1	257	0
72	1	8	1	8	0	8	6	8	3	8	3	8	3	8	1	8	0
75	1	9	1	9	5	9	7	9	7	9	1	9	2	9	2	9	1
87	2	120	1	140	5	160	7	180	3	200	5	220	3	240	0	260	1
88	1	1	1	1	3	1	7	1	4	1	1	1	1	1	0	1	0
101	2	2	3	2	6	2	3	2	11	2	8	2	1	2	1	2	0
2	0	3	2	3	3	3	10	3	9	3	2	3	1	3	0	3	0
3	0	4	2	4	4	4	6	4	7	4	4	4	1	4	3	4	1
4	1	5	1	5	1	5	12	5	6	5	0	5	2	5	1	272	1
5	1	6	0	6	6	6	2	6	8	6	2	6	1	6	1	277	1
6	1	7	0	7	8	7	4	7	9	7	1	7	0	7	0	280	1
7	1	8	1	8	3	8	5	8	2	8	2	8	1	8	0	285	1
8	1	9	2	9	4	9	6	9	7	9	1	9	3	9	0	287	2
9	2	130	1	150	5	170	9	190	7	210	4	230	1	250	0	290	1
110	0	1	4	1	4	1	5	1	6	1	3	1	0	1	0	316	1
1	1	2	2	2	7	2	9	2	7	2	3	2	0	2	0	327	1
2	2	3	0	3	4	3	5	3	5	3	1	3	1	3	0	367	1
3	2	4	5	4	7	4	5	4	6	4	4	4	0	4	0	376	1
4	0	5	4	5	4	5	5	5	2	5	3	5	0	5	0	392	1
5	1	136	1	156	5	176	7	196	7	216	3	236	0	256	0	411	1
116	3																

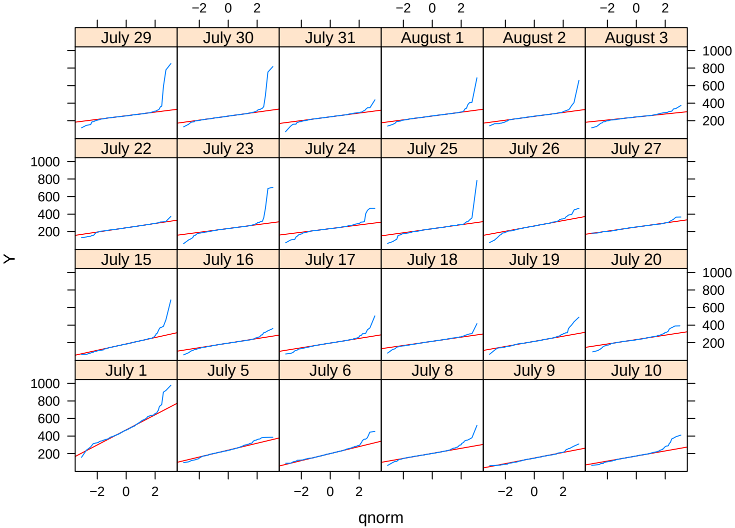
Times in milliseconds in odd columns, even columns report cell counts of the number of occurrences of the indicated timing. Source: Peirce(1873)

Peirce's Density Estimation



Not bad for 1873, Peirce concludes: "It was found that after the first two or three days the curves differed little from that derived from the theory of least squares."

Normal QQ Plots for the Peirce Experiment



Wilson and Hilferty's (1929) Reanalysis of Peirce Data

E.B. Wilson and Margaret Hilferty published an extensive reanalysis of the Peirce data in the PNAS. They found:

- Most day's data is skewed to the right, and all days have excess kurtosis.
- Comparing the precision of the median and the mean, they remark that: Although for normal data, the median is known to be about 25% worse than the mean, for the Peirce data, "the median and the mean are on the whole about equally well determined."
- Maurice Fréchet, at about the same time, had a diploma student who reached very similar conclusions.

Wilson and Hilferty's (1929) Reanalysis of Peirce Data

E.B. Wilson and Margaret Hilferty published an extensive reanalysis of the Peirce data in the PNAS. They found:

- Most day's data is skewed to the right, and all days have excess kurtosis.
- Comparing the precision of the median and the mean, they remark that: Although for normal data, the median is known to be about 25% worse than the mean, for the Peirce data, "the median and the mean are on the whole about equally well determined."
- Maurice Fréchet, at about the same time, had a diploma student who reached very similar conclusions.

A Mystery: How did Wilson and Hilferty estimate the precision of the median? In 1929 there was no agreed "standard deviation" for the median.

The Median is the Message?

Day	n	median	mean	Day	n	median	mean
1	495	468 ± 2.5	475.6 ± 4.1	13	489	244 ± 1.3	244.5 ± 1.2
2	490	237 ± 2.1	241.5 ± 2.1	14	500	236 ± 1.3	236.7 ± 1.9
3	489	200 ± 1.7	203.2 ± 2.1	15	498	235 ± 1.1	236.0 ± 1.5
4	499	201 ± 1.2	205.6 ± 1.8	16	498	233 ± 1.6	233.2 ± 1.7
5	490	147 ± 2.0	148.5 ± 1.6	17	507	264 ± 1.8	265.5 ± 1.7
6	489	172 ± 1.9	175.6 ± 1.8	18	495	254 ± 1.3	253.0 ± 1.1
7	496	184 ± 1.7	186.9 ± 2.2	19	500	255 ± 0.9	258.7 ± 2.0
8	490	194 ± 1.3	194.1 ± 1.4	20	494	253 ± 1.4	255.4 ± 2.0
9	495	195 ± 1.5	195.8 ± 1.6	21	502	245 ± 1.7	245.0 ± 1.2
10	498	215 ± 1.6	215.5 ± 1.3	22	499	255 ± 1.6	255.6 ± 1.4
11	499	213 ± 2.1	216.6 ± 1.7	23	498	252 ± 1.2	251.4 ± 1.4
12	396	233 ± 1.8	235.6 ± 1.7	24	497	244 ± 0.9	243.4 ± 1.1

Summary Statistics for the Peirce (1872) Experiments: An attempt to reproduce a portion of the Wilson and Hilferty (1929) analysis of the Peirce experiments.

The Standard Deviation of the Median?

Day	WH	Laplace	Yule	Siddiqui	Exact I	Exact II	Jeffreys	Boot
Mean	1.538	1.155	1.567	1.549	1.573	1.531	1.594	1.584
MAE	0.000	0.393	0.129	0.135	0.180	0.166	0.191	0.103
MSE	0.000	0.219	0.027	0.029	0.064	0.056	0.079	0.025
MXE	0.000	0.896	0.457	0.306	0.827	0.777	0.827	0.553

Standard Deviations for the Medians: Wilson and Hilferty's daily estimates of the standard deviation and seven attempts to reproduce their estimates. Column means and three measures of discrepancy between the original estimates and the new ones are given: mean absolute error, mean squared error, and maximal absolute error.

Koenker, R. (2009) The Median is the Message, Am.Statistician, contains some further details, and all the data and code is available from my R package for quantile regression. This is a homework exercise in forensic statistics, or reverse engineering.

Why Should We Be Interested in Allowing Some Bias?

The case for bias:

- Stein: Even under strictly Gaussian regression conditions some bias is desirable when $p \geqslant 3$, and p is almost always greater than three.
- Vapnik: In non-parametric settings bias is essential, without regularization of some form we're in the Dirac swamp.
- Leamer: Model selection (pre-testing) is the poor man's shrinkage.
- And the lasso and the lariat have made coef roping a growth industry.

Why Should We Be Interested in Allowing Some Bias?

The case for bias:

- Stein: Even under strictly Gaussian regression conditions some bias is desirable when $p \geqslant 3$, and p is almost always greater than three.
- Vapnik: In non-parametric settings bias is essential, without regularization of some form we're in the Dirac swamp.
- Leamer: Model selection (pre-testing) is the poor man's shrinkage.
- And the lasso and the lariat have made coef roping a growth industry.

Insisting on unbiasedness is a little like insisting on Type I error of 0.05 regardless of the sample size.

Illicit Priors

Ever since Kant, people have been wondering “Where does the synthetic a priori come from?”

Illicit Priors

Ever since Kant, people have been wondering "Where does the synthetic a priori come from?"

- "I agree with Professor Bernardo that prior elicitation is nearly impossible in complex models." [Malay Ghosh, Stat. Sci. 2011]

Illicit Priors

Ever since Kant, people have been wondering "Where does the synthetic a priori come from?"

- "I agree with Professor Bernardo that prior elicitation is nearly impossible in complex models." [Malay Ghosh, Stat. Sci. 2011]
- Jeffrey's $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ is fine, unless there are nuisance parameters, but there are almost always nuisance parameters.

Illicit Priors

Ever since Kant, people have been wondering "Where does the synthetic a priori come from?"

- "I agree with Professor Bernardo that prior elicitation is nearly impossible in complex models." [Malay Ghosh, Stat. Sci. 2011]
- Jeffrey's $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ is fine, unless there are nuisance parameters, but there are almost always nuisance parameters.
- Sometimes the data can provide workable priors, Stein rules, Tweedie's formula, hierarchical models, Kiefer-Wolfowitz (Heckman-Singer).

Illicit Priors

Ever since Kant, people have been wondering "Where does the synthetic a priori come from?"

- "I agree with Professor Bernardo that prior elicitation is nearly impossible in complex models." [Malay Ghosh, Stat. Sci. 2011]
- Jeffrey's $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ is fine, unless there are nuisance parameters, but there are almost always nuisance parameters.
- Sometimes the data can provide workable priors, Stein rules, Tweedie's formula, hierarchical models, Kiefer-Wolfowitz (Heckman-Singer).
- Empirical Bayes is the wave of the future – waving while drowning in a sea of data.

Illicit Priors

Ever since Kant, people have been wondering "Where does the synthetic a priori come from?"

- "I agree with Professor Bernardo that prior elicitation is nearly impossible in complex models." [Malay Ghosh, Stat. Sci. 2011]
- Jeffrey's $\pi(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ is fine, unless there are nuisance parameters, but there are almost always nuisance parameters.
- Sometimes the data can provide workable priors, Stein rules, Tweedie's formula, hierarchical models, Kiefer-Wolfowitz (Heckman-Singer).
- Empirical Bayes is the wave of the future – waving while drowning in a sea of data.
- Lindley: "No one is less Bayesian than an empirical Bayesian."

The Last Slide – All Downhill from Here

- Beware of linear estimators, they are fragile like the house of cards they are built upon.

The Last Slide – All Downhill from Here

- Beware of linear estimators, they are fragile like the house of cards they are built upon.
- A little bias is usually a good thing, at least when $p \geqslant 3$.

The Last Slide – All Downhill from Here

- Beware of linear estimators, they are fragile like the house of cards they are built upon.
- A little bias is usually a good thing, at least when $p \geqslant 3$.
- Computation is more important than it appears.

The Last Slide – All Downhill from Here

- Beware of linear estimators, they are fragile like the house of cards they are built upon.
- A little bias is usually a good thing, at least when $p \geqslant 3$.
- Computation is more important than it appears.
- Anything worth doing is worth being able to do again.

The Last Slide – All Downhill from Here

- Beware of linear estimators, they are fragile like the house of cards they are built upon.
- A little bias is usually a good thing, at least when $p \geqslant 3$.
- Computation is more important than it appears.
- Anything worth doing is worth being able to do again.
- Nunc est Bibendum!