3. Analysis of Big Data: Maximum Likelihood

Rod Little

Department of Biostatistics

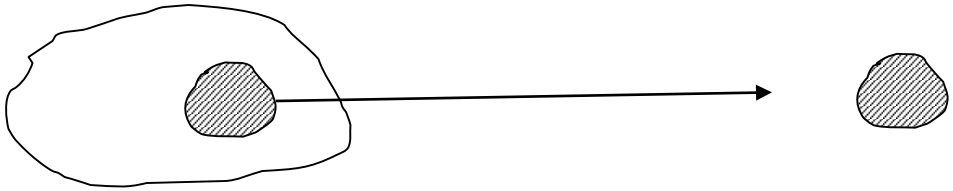
UNIVERSITY OF MICHIGAN

Parameter estimation/likelihood

- Seminal paper: Fisher, R.A. (1922). "On the Mathematical Foundations of Theoretical Statistics". Phil. Trans. Roy. Soc. A, pp. 309-368
- Statistical models -- examples
- Maximum Likelihood – key properties
- Bayesian methods
- I am covering material for several courses in statistics in 1.5 hours, so this will be very "broad brush"

Inference for a population based on a sample

- Statistical inference: the process of making inferences about parameters of a population based on sample data.



• Population		• Sample	
•Mean	μ	•Mean	$\overline{x}$
•SD	σ	•SD	s

- Crucial assumption: random sampling

Use "hat" for sample estimate: $\hat{\mu}$ is estimate of μ, etc.

Principles of Estimation

- Need a theory on how to estimate parameters from sample data, with associated measures of precision
- Until Fisher, two main approaches were least squares (Gauss) and the method of moments
- The method of maximum likelihood is a more general approach – applies to non-normal data

Statisticians Impacting Society #1

- Sir Ronald Fisher's experimental designs and analysis of variance have greatly increased the world food supply
- Fisher is generally viewed as the founder of modern statistics
- In particular, theory of maximum likelihood, a major tool of statistical inference – was laid out in Fisher (1924)

Sir Ronald Fisher

Fisher (1924) 1

“THE NEGLECT OF THEORETICAL STATISTICS.

SEVERAL reasons have contributed to the prolonged neglect into which the study of statistics, in its theoretical aspects, has fallen. In spite of the immense amount of fruitful labour which has been expended in its practical applications, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved.

This anomalous state of statistical science is strikingly exemplified by a recent paper (1) entitled " The Fundamental Problem of Practical Statistics," in which one of the most eminent of statisticians presents what purports to be a general proof of BAYES' postulate, which, in the opinion of a second statistician of equal eminence, " seems to rest on a very peculiar--not to say hardly supposable – relation.”

Fisher (1924) 2

"THE PURPOSE OF STATISTICAL METHODS

In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the **reduction of data.** A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall **contain as much as possible, ideally the whole, of the relevant information contained in the original data**."

Fisher (1924) 3

"THE PROBLEMS OF STATISTICS

The problems which arise in reduction of data may be conveniently divided into three types :

(1) Problems of Specification. These arise in the choice of the mathematical form of the population.

(2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.

(3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known."

Fisher (1924) 4

“CRITERIA OF ESTIMATION

The common-sense criterion employed in problems of estimation may be stated thus : That when applied to the whole population the derived statistic should be equal to the parameter. This may be called the **Criterion of Consistency.**

… Consideration of the above example will suggest a second criterion, namely :That in large samples, when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error. This may be called the **Criterion of Efficiency**.

… The complete criterion suggested by our work on the mean square error (7) is: That the statistic chosen should summarise the whole of the relevant information supplied by the sample. This may be called the **Criterion of Sufficiency**

Fisher (1924) 5

"FORMAL SOLUTION OF PROBLEMS OF ESTIMATION

…For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied.

Such a method is, I believe, provided by the **Method of Maximum Likelihood**, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect."

Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood inference for large samples
  - Bayesian inference for small samples:

    log(likelihood) + log(prior) = log(posterior)

Statistical Models

- Random variables: numerical characteristics measured in the sample
- A statistical model is a simplified representation of the measured random variables in population that captures main features
- Often relates an outcome variable or variables (Y) to a set of predictor variables (X) that are treated as fixed
- Y is assigned a probability distribution that captures variability – this is what makes the model *statistical*
- Developing a good statistical model is a key step in statistical analysis – what Fisher calls the "problem of specification."

Statistical Models

- More formally ....

$i =$ unit of observation (e.g. participant in a clinical trial)

$y_i =$ outcome measures for unit i

$x_i =$ predictor variables, covariates, treated as fixed.

Statistical model specifies

$f(y_i \mid x_i, \theta) =$ distribution of y_i given x_i

$\theta =$ unknown parameters, to be estimated

Ex.1: Normal model for a single sample

$$Y = (y_1, \ldots, y_n)$$

$y_i \sim_{\text{ind}} N(\mu, \sigma^2), \mu =$ mean, $\sigma^2 =$ variance

parameters are $\theta = (\mu, \sigma^2)$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2\right)$$

Ex. 2: Normal model for two independent samples

$Y = (y_1, ..., y_n)$

Define $x_i = 1$ if i belongs to group 1 (say $i = 1, ..., n_1$)

$y_i \mid x_i = 1 \sim N(\mu_1, \sigma_1^2)$

Define $x_i = 2$ if i belongs to group 2 ($i = n_1 + 1, ..., n_1 + n_2 = n$)

$y_i \mid x_i = 2 \sim N(\mu_2, \sigma_2^2)$

Sometimes we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$\mu_2 - \mu_1$ is the difference in means -- often of interest

$$f(Y \mid X, \mu_1, \mu_2, \sigma^2) =$$

$$\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_i - \mu_1\right)^2\right) \times \exp\left(-\frac{1}{2\sigma^2}\sum_{i=n_1+1}^{n_1+n_2}\left(y_i - \mu_2\right)^2\right)$$

Ex. 3: normal multiple linear regression model

y_i = outcome variable for unit i

$x_{i1}, x_{i2}, ..., x_{ip}$ = set of p predictor variables for unit i

$$(y_i \mid x_{i1}, ..., x_{ip}) \sim_{\text{iid}} N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2)$$

$$f(Y \mid X, \beta_0, \beta_1, ...\beta_p, \sigma^2) =$$

$$\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip} \right)^2 \right)$$

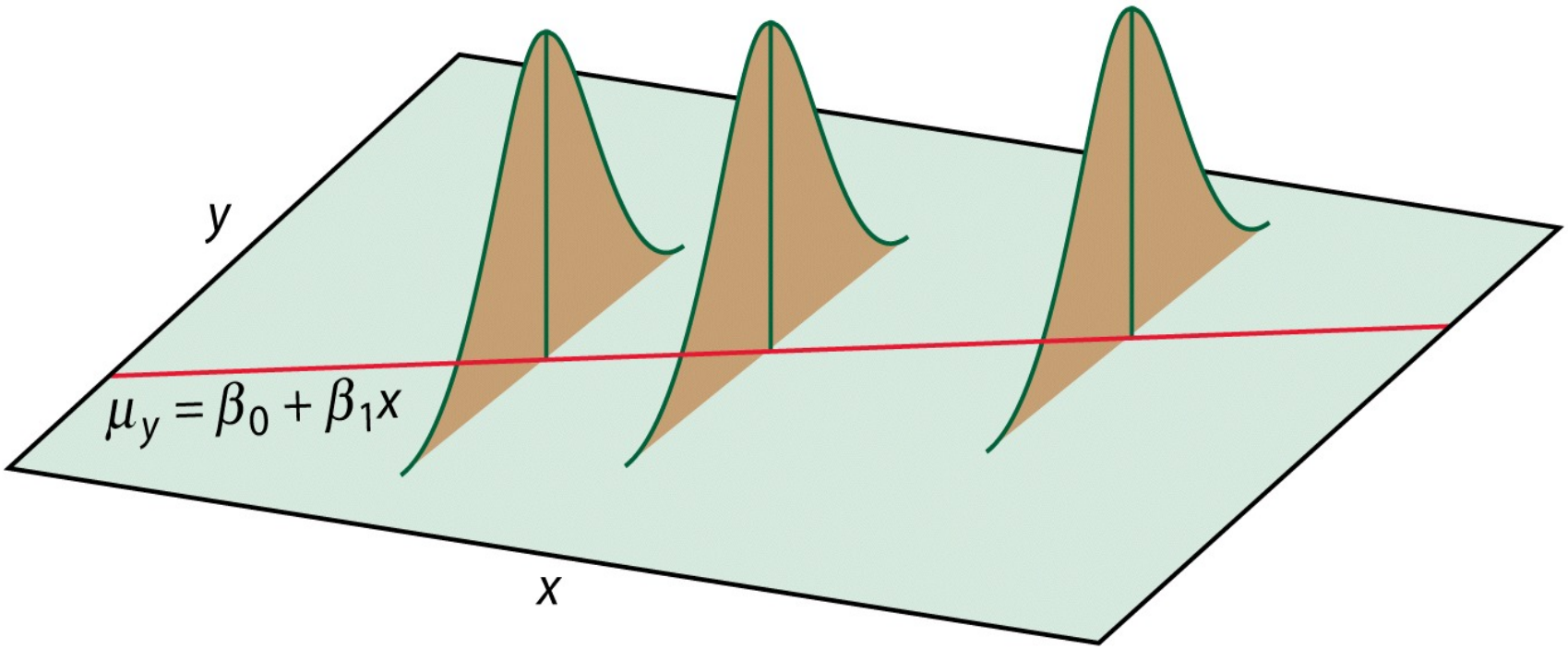
Special case: simple linear regression

$p = 1, x_{i1} = x_i$

$$(y_i \mid x_i) \sim_{\text{iid}} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

When x_i is binary, we get Example 2

Linear regression: mean response depending linearly on quantitative x



Ex. 4: Binomial sample

$$Y = (y_1, ..., y_n), y_i = 0 \text{ or } 1$$

$$\Pr(y_i = 1) = \pi$$

$$\sum_{i=1}^{n} y_i = n_1 = \text{number of successes,}$$

$$n_0 = n - n_1 = \text{number of failures}$$

$$f(n_1 \mid \pi) = \left(\frac{n!}{n_1! n_2!} \right) \pi^{n_1} (1-\pi)^{n-n_1} \text{ (Binomial distribution)}$$

Ex. 5: logistic regression model

y_i = binary outcome variable for unit i (0 or 1)

$x_{i1}, x_{i2}, ..., x_{ip}$ = set of p predictor variables for unit i

$$\Pr(y_i = 1 | x_{i1}, ..., x_{ip}) \sim_{\text{iid}} \pi_i(\beta)$$

$$\log\left(\pi_i(\beta) / (1 - \pi_i(\beta))\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$f(Y \mid X, \beta_0, \beta_1, ...\beta_p) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{(1-y_i)}$$

Ex. 6: generalized linear models

y_i = outcome variable for unit *i*

$x_{i1}, x_{i2}, ..., x_{ip}$ = set of p predictor variables for unit *i*

$$f_Y(y_i \mid x_i, \beta, \phi) = \exp\left[\left(y_i \delta(x_i, \beta) - b\left(\delta(x_i, \beta)\right)\right) / \phi + c(y_i, \phi)\right]$$

$$E(y_i \mid x_i, \beta, \phi) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})$$

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, g = \text{ link function}$$

A natural choice is the *canonical* link g_c, for which

$$g_c(\mu_i) = \delta(x_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Ex. 6: generalized linear models

- Important specific models with their canonical links include:

Normal linear regression: $g_c = \text{identity}, b(\delta) = \delta^2 / 2, \phi = \sigma^2$

Logistic regression: $g_c = \text{logit}, b(\delta) = \log(1 + \exp(\delta)), \phi = 1$

Poisson regression: $g_c = \log, b(\delta) = \log(\delta), \phi = 1$

Other problems

- Statistical models have also been developed for
  - Clustered data
  - Repeated measures data
  - Survival analysis
  - Latent variables
  - Time series
  - Etc. etc.

Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood inference for large samples
  - Bayesian inference for small samples:

    log(likelihood) + log(prior) = log(posterior)

Likelihood function

- Data Y
- Statistical model yields probability density $f(Y \mid \theta)$

  for Y with unknown parameters θ
- Likelihood function is $f(Y \mid \theta)$ treated as a function of θ

$$L(\theta \mid Y) = \text{const} \times f(Y \mid \theta)$$

- Loglikelihood is often easier to work with:

$$\ell(\theta \mid Y) = \log L(\theta \mid Y) = \text{const} + \log\{ f(Y \mid \theta)\}$$

Constants can depend on data but not on parameter θ

Ex. 1: Normal model for a single sample

$$Y = (y_1, \ldots, y_n)$$

$y_i \sim N(\mu, \sigma^2), \mu =$ mean, $\sigma^2 =$ variance

parameters are $\theta = (\mu, \sigma^2)$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2\right)$$

$$\ell(\mu, \sigma^2 \mid Y) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2$$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of θ maximizes the likelihood

$$L(\hat{\theta} \mid Y) \geq L(\theta \mid Y) \text{ for all } \theta$$

- The ML estimate is the

  "value of the parameter that makes the data most likely"

- The ML estimate is not always unique, but is for many regular problems given enough data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the score equation(s)

$$S(\theta \mid Y) \equiv \frac{\partial \log L(\theta \mid Y)}{\partial \theta} = 0$$

where *S* is the score function.

Explicit solutions for some models (normal regression, multinomial, …)

Iterative methods – e.g. Newton-Raphson, Scoring, EM algorithm -- required for other problems (logistic regression, repeated measures models, non-monotone missing data)

Normal Examples

- Univariate Normal sample $Y = (y_1, ..., y_n)$ $\theta = (\mu, \sigma^2)$

$$\hat{\mu} = \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

(Note the lack of a correction for degrees of freedom)

- Multivariate Normal sample

$$\hat{\mu} = \bar{y}, \ \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T$$

- Normal Linear Regression

$$(y_i \mid x_{i1}, ..., x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2)$$

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \text{ least squares estimates}$$

$$\hat{\sigma}^2 = (\text{residual sum of squares})/n$$

Ex. 4. Binomial Sample

$Y = (y_1, ..., y_n), y_i = 0$ or 1

$\Pr(y_i = 1) = \pi$

$\sum_{i=1}^{n} y_i = n_1 =$ number of successes,

$n_0 = n - n_1 =$ number of failures

$$f(n_1 \mid \pi) = \left(\frac{n!}{n_1! n_2!} \right) \pi^{n_1} (1-\pi)^{n-n_1}$$

$$L(\pi \mid n_1) = \pi^{n_1}(1-\pi)^{n-n_1}, \ell(\pi \mid n_1) = n_1 \log \pi + (n - n_1) \log(1-\pi)$$

Score equation $\frac{\partial \ell}{\partial \pi} = \frac{n_1}{\pi} - \frac{n - n_1}{1-\pi} = 0$

yields ML estimate $\hat{\pi} = n_1 / n$ (sample proportion)

Ex. 5: logistic regression model

y_i = binary outcome variable for unit i (0 or 1)

$x_{i1}, x_{i2}, ..., x_{ip}$ = set of p predictor variables for unit i

$$\Pr(y_i = 1 | x_{i1}, ..., x_{ip}) \sim_{\text{iid}} \pi_i(\beta)$$

$$\log\left(\pi_i(\beta) / (1 - \pi_i(\beta))\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$f(Y | X, \beta_0, \beta_1, ...\beta_p) = L(\beta_0, \beta_1, ..., \beta_p | Y, X)$$

$$= \prod_{i=1}^{n} \left(\pi_i(\theta)\right)^{y_i} \left(1 - \pi_i(\theta)\right)^{(1-y_i)}$$

Score equation does not have explicit solution --

Need iterative maximization algorithms --

Newton Raphson, scoring)

Properties of ML estimates

- Under assumed model, ML estimate is:
  - Consistent (not necessarily unbiased)
  - Efficient for large samples
  - not necessarily the best for small samples

Large-sample ML Inference

- Basic large-sample approximation:

  for regular problems,

$$\hat{\theta} \sim N(\theta, C)$$

  where C is a covariance matrix estimated from the sample

Forms of precision matrix

- The precision of the ML estimate is measured by C^{-1} Some forms for this are:
  - Observed information (recommended)

$$C^{-1} = I(\hat{\theta} \mid Y) = -\left. \frac{\partial^2 \log L(\theta \mid Y)}{\partial\theta \, \partial\theta} \right|_{\theta=\hat{\theta}}$$

  - Expected information (not as good, may be simpler)

$$C^{-1} = J(\hat{\theta}) = E\left[I(\theta \mid Y, \theta) \right]_{\theta=\hat{\theta}}$$

  - Some other approximation to curvature of loglikelihood in the neighborhood of the ML estimate

Interval estimation

- 95% (confidence, probability) interval for scalar θ is:

$\hat{\theta} \pm 1.96\, C^{1/2}$, where 1.96 is 97.5 pctile of normal distribution

- Example: univariate normal sample

$$I = J = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{bmatrix} \Rightarrow C = \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix}$$

Hence some 95% intervals are:

$$\bar{y} \pm 1.96\, s/\sqrt{n} \text{ for } \mu$$

$$s^2 \pm 1.96\, s^2/\sqrt{n/2} \text{ for } \sigma^2$$

$$\ln(s) \pm 1.96\, \sqrt{2/n} \text{ for } \ln(\sigma)$$

Significance Tests

Tests based on likelihood ratio (LR) or Wald (W) statistics:

$\theta = (\theta_{(1)}, \theta_{(2)})$; $\theta_{(1)0}$ = null value of $\theta_{(1)}$; θ_2 = other parameters

$\hat{\theta}$ = unrestricted ML estimate

$\tilde{\theta} = (\theta_{(1)0}, \tilde{\theta}_{(2)})$; $\tilde{\theta}_{(2)}$ = ML estimate of $\theta_{(2)}$ given $\theta_{(1)} = \theta_{(1)0}$

LR statistic: $\mathrm{LR}(\hat{\theta}, \tilde{\theta}) = 2\left[\ell(\hat{\theta} \mid Y) - \ell(\tilde{\theta} \mid Y)\right]$

Wald statistic: $W(\hat{\theta}, \tilde{\theta}) = (\theta_{(1)0} - \hat{\theta}_{(1)})^T C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)})$

$C_{(11)}$ = covariance matrix of $(\theta_{(1)} - \hat{\theta}_{(1)})$

yield P-values $P = pr\left(\chi_q^2 > D(\hat{\theta}, \tilde{\theta})\right)$

D = LR or Wald statistic; q = dimension of θ_0

χ_q^2 = Chi-squared distribution with q degrees of freedom

Application

- Regression models like logistic regression are an important tool for controlling confounding variables in observational studies

  y_i = binary outcome variable for unit i (0 or 1)

  $x_{i1}, x_{i2}, ..., x_{ip}$ = set of p predictor variables for unit i

  $$\Pr(y_i = 1 | x_{i1}, ..., x_{ip}) \sim_{\text{iid}} \pi_i(\beta)$$

  $$\log\left(\pi_i(\beta) / (1 - \pi_i(\beta))\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

  β_j is effect of X_j, holding other X's fixed

- Environmental Health: Lead exposure and hypertension

The Relationship of Bone and Blood Lead to Hypertension; The Normative Aging Study

Howard Hu, MD, ScD; Antonio Aro, PhD; Marinelle Payton, MD, PhD; Susan Korrick, MD, MPH; David Sparrow, DSc; Scott T. Weiss, MD, MS; Andrea Rotnitzky. PhD

- Objective.-To test the hypothesis that long-term lead accumulation, as reflected by levels of lead in bone (as opposed to blood, which reflects recent lead exposure), is associated with an increased odds of developing hypertension.
- Design.-Case-control study of participants in the Veterans Administration (now Department of Veterans Affairs) Normative Aging Study, a 30-year longitudinal study of men.
- Participants.--Of 1171 active subjects who were seen between August 1991 and December 1994,590 (50%) participated in this investigation and had data on all variables of interest.

Abstract continued

- Results.--Blood lead levels were low, ranging from less than 0.05 to 1.35 µmol/L (<I to 28 µg/dL), with a mean (SD) of 0.30 (0.20) µmol/L (6.3 [4.1] µg/dL). Bone lead levels were similar to those described in other general populations. In comparison to non-hypertensives, mean levels of lead in blood and both tibia and patella bone lead levels were significantly higher in hypertensive subjects.

Abstract continued

- In a logistic regression model of hypertensive status that began with age, race, body mass index, family history of hypertension, history of ethanol ingestion, pack-years of smoking, dietary sodium intake, dietary calcium intake, blood lead, tibia lead, and patella lead, the variables that remained after backward elimination were body mass index, family history of hypertension, and level of lead in the tibia. An increase from the midpoint of the lowest quintile to the midpoint of the highest quintile of tibia lead from 8 to 37 μg per gram of bone mineral was associated with an increased odds ratio of hypertension of 1.5.

Abstract continued

- Conclusion.--Our findings suggest that long-term lead accumulation, as reflected by levels of lead in bone, may be an independent risk factor for developing hypertension in men in the general population.

Descriptive analysis of bone lead levels

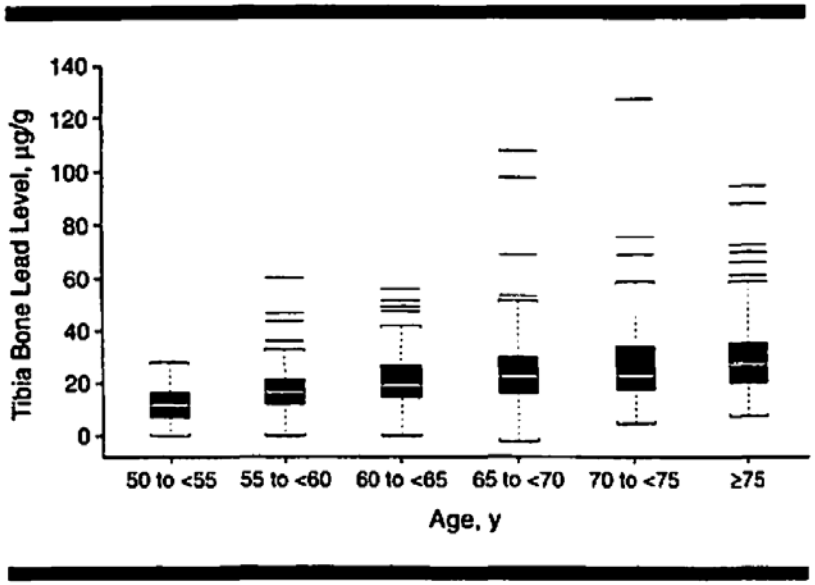


Figure 1.—Box plot of tibia lead levels measured by K x-ray fluorescence vs age among participants in the Normative Aging Study. The horizontal line in the interior of the box is located at the median of the data. The box describes the interquartile distance (IQD) between the third quartile of the data and the first quartile. The dotted lines extend a distance of 1.5 × IQD from the center to the "whiskers." The data bars that fall outside the whiskers may be considered outliers.

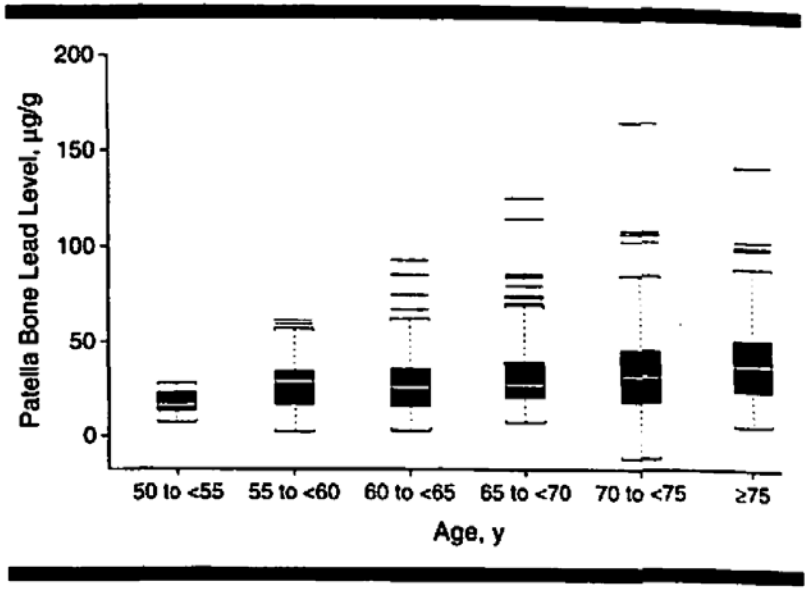


Figure 2.—Box plot of patella lead levels measured by K x-ray fluorescence vs age among participants in the Normative Aging Study. See the legend to Figure 1 for explanation of boxes, lines, "whiskers," and data bars.

Comparison of hypertensives and nonhypertensives

Table 2.—Characteristics of Hypertensives and Nonhypertensives in the Bone Lead Project of the Normative Aging Study

Characteristic	Hypertensives (n=146) Mean (SD)	Hypertensives (n=146) Median	Nonhypertensives (n=444) Mean (SD)	Nonhypertensives (n=444) Median
Age, y	67.2 (6.5)	67	66.4 (7.3)	66
Body mass index, kg/m²	29.0 (4.2)	31.4	27.4 (3.7)*	26.8
Dietary sodium, mg/d	3603 (1519)	3632	3799 (1776)	3573
Dietary calcium, mg/d	838 (437)	770	834 (385)	771
Blood lead levels, μmol/L [μg/dL]	0.33 (0.21) [6.9 (4.3)]	0.29 [6]	0.29 (0.19)* [6.1 (4.0)]*	0.24 [5]
Tibia bone lead level, μg/g	23.7 (14.0)	22	20.9 (11.4)*	19
Patella bone lead level, μg/g	35.1 (19.5)	31.5	31.1 (18.3)*	27
	Hypertensives, No. (%)		Nonhypertensives, No. (%)	
Average ethanol consumption†				
0-10	88 (60)		280 (63)	
11-50	33 (23)		91 (21)	
>50	25 (17)		73 (16)	
Family history of hypertension†				
Yes	60 (41)		96 (22)	
No	86 (59)		348 (78)*	
Pack-years of smoking				
0	38 (26)		144 (32)	
1-20	46 (32)		121 (27)	
21-40	28 (19)		79 (18)	
>40	34 (23)		100 (23)	
Race				
White	141 (97)		438 (99)	
African American	5 (3)		6 (1)	

P<.05

Adjusting for confounders

Logistic regression is used to assess relationship between binary outcome (hypertensive of not) on bone lead, adjusting for confounders

Table 3.—Results of Logistic Regression Models of Hypertensive Status in Relation to Lead Biomarkers, Age, Race, Body Mass Index, Pack-Years of Cumulative Smoking, Cumulative Ethanol Ingestion, Dietary Sodium, and Dietary Calcium*

	Model											
	A			B			C			D		
	β	SE	*P*	β	SE	*P*	β	SE	*P*	β	SE	*P*
Age	.366	0.0149	.01	.0358	0.0149	.02	.0279	0.0159	.08	.0285	0.0158	.07
Race	.4531	0.4032	.26	.3678	0.4081	.37	.3134	0.4167	.45	.1050	0.0266	<.001
Body mass index	.1080	0.0265	<.001	.1086	0.0265	<.001	.1046	0.0266	<.001	.1050	0.0266	<.001
Family history of hypertension	.8889	0.2136	<.001	.8638	0.2146	<.001	.8978	0.2141	<.001	.9048	0.2143	<.001
Pack-years of smoking	.0294	0.0618	.63	.0307	0.0616	.62	.0203	0.0620	.74	.0171	0.0623	.78
Ethanol ingestion	.2137	0.3167	.50	.1825	0.3175	.57	.1664	0.3186	.60	.1620	0.3189	.26
Dietary sodium	−.0002	0.0001	.04	−.0002	0.0001	.045	−.0002	0.0001	.04	−.0002	0.0001	.04
Dietary calcium	0	0.0003	.95	0	0.0003	.89	0	0.0003	.89	.0001	0.0002	.84
Blood lead level, μmol/L	...	...	...	.0017	0.0011	.15	...	...	...	...	...	...
(μg/dL)				(.0344	0.0237	.15)	...	...	...	...	...	...
Tibia bone lead level	...	...	...	...	...	...	.0136	0.0085	.11	...	...	...
Patella bone lead level	...	...	...	...	...	...	...	...	...	.0087	0.0055	.11

*Ellipses indicate variable was not included in this model.

Adjusting for confounders– final model

Table 4.—Final Logistic Regression Model of Hypertensive Status After Backward Elimination*

Variable	Parameter Estimate	SE	P	Odds Ratio Estimate (95% Confidence Interval)
Intercept	−4.379	0.7482	<.001	0.013 (0.003-0.054)
Body mass index, kg/m²	0.092	0.0254	<.001	1.096 (1.043-1.152)
Family history of hypertension	0.846	0.2099	<.001	2.329 (1.544-3.514)
Tibia bone lead level, μg/g	0.019	0.0078	.01	1.019 (1.004-1.035)

*The χ^2 for covariates of the final model was 39.9 with 3 *df* ($P<.001$).

Bayes inference

- Given a prior distribution $\pi(\theta)$ for the parameters, inference can be based on the posterior distribution using Bayes' theorem:

$$p(\theta \mid Y) = \text{const.} \times \pi(\theta) \times L(\theta \mid Y)$$

- For small samples, we prefer Bayes' inference based on the posterior to the large sample ML approximation.
  - In important standard problems with non-informative priors, Bayes yields inference comparable to small-sample frequentist inference
  - In many non-standard problems, Bayes yields answers where no exact frequentist answer exists

Example: linear regression

The normal linear regression model:

$$(y_i \mid x_{i1},\ldots,x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2)$$

with non-informative "Jeffreys'" prior:

$$\pi(\beta_0,\ldots,\beta_p,\sigma^2) \propto 1/\sigma^2$$

yields the posterior distribution of $(\beta_0,\ldots,\beta_p)$ as multivariate *T* with mean given by the least squares estimates $(\hat{\beta}_0,\ldots,\hat{\beta}_p)$, covariance matrix $(X^T X)^{-1} s^2$, where *X* is the design matrix, and degrees of freedom *n* - *p* - 1.

Resulting posterior credibility intervals are equivalent to standard t confidence intervals.

Simulating Draws from Posterior Distribution

- With problems with high-dimensional θ , it is often easier to draw values from the posterior distribution, and base inferences on these draws
- For example, if

$$(\theta_1^{(d)} : d = 1, \ldots, D)$$

is a set of draws from the posterior distribution for a scalar parameter θ_1, then

$\bar{\theta}_1 = D^{-1} \sum_{d=1}^{D} \theta_1^{(d)}$ approximates posterior mean

$s_\theta^2 = (D-1)^{-1} \sum_{d=1}^{D} (\theta_1^{(d)} - \bar{\theta}_1)^2$ approximates posterior variance

$(\bar{\theta}_1 \pm 1.96 s_\theta)$ or 2.5th to 97.5th percentiles of draws approximates 95% posterior credibility interval for θ

Example: Posterior Draws for Normal Linear Regression

$(\hat{\beta}, s^2) =$ ls estimates of slopes and resid variance

$\sigma^{(d)2} = (n - p - 1)s^2 / \chi^2_{n-p-1}$

$\beta^{(d)} = \hat{\beta} + A^T z \sigma^{(d)}$

χ^2_{n-p-1} = chi-squared deviate with $n - p - 1$ df

$z = (z_1, ..., z_{p+1})^T$, $z_i \sim N(0,1)$

$A =$ upper triangular Cholesky factor of $(X^T X)^{-1}$:

$A^T A = (X^T X)^{-1}$

- Easily extends to weighted regression: see Example 6.19

Summary

- We have reviewed basic ML results in the context of standard complete-data problems
- Bayes particularly useful for small sample problems