Methods of Applied Statistics I

STA2101H F LEC9101

Week 2

September 21 2022

What's that got to do with statistics?

What's that got to do with statistics?

"BEARER-PARTY TURN LEFT!"

What's that got to do with statistics?

"BEARER-PARTY TURN LEFT!"

subtitle

What's that got to do with statistics?

"BEARER-PARTY TURN LEFT!"

subtitle

"Dr. Fauci turn left"

What's that got to do with statistics?

"BEARER-PARTY TURN LEFT!"

subtitle

"Dr. Fauci turn left"

data **quality** matters

Today

1. Grading scheme, Comments re texts 8/10 text website
2. Upcoming events
3. Steps in analysis; types of studies
4. Recap Week 1
5. Linear Regression Part 2: testing groups of variables, checking model assumptions, collinearity, $p > n$
6. In the news

- LM-1,2 – Linear Models with R by Faraway (1st and 2nd editions) LM (both)
- ELM-1,2 – Extending the Linear Model with R by Faraway (1st and 2nd editions)
- CD – Principles of Applied Statistics by Cox & Donnelly
- SM – Statistical Models by Davison highly rec'd for PhD

Upcoming Events

- Thursday Sep 22 3.30 UY 9014
  Full likelihood inference for abundance from capture-recapture data
  Pengfei Li, U Waterloo

- Thursday Nov 26 5 pm Toronto Data Workshop
  Mining Software Repositories
  Melina Vidoni Australian National U
  Online

Dear friends,

This week the Toronto Data Workshop meets on Zoom at **Thursday, 22 September, at 5pm**. Our guest is Dr Melina Vidoni, who is a lecturer at the Australian National University in the CECS School of Computing.

Abstract: Mining Software Repositories (MSR) is an increasingly common methodology based on extracting open, publicly available software-related data. Hence, it is considered Evidence-Based Research. Since their emergence in 2004, many investigations have analysed different aspects of MSR-based studies, such as validity of sources or data usage. This talk draws from Dr Vidoni's research experience using MSR approaches in several sources to investigate Technical Debt in different paradigms, with a special focus in scientific software. It will discuss common challenges, combining MSRs with developer surveys for mixed-methods approaches, and discuss when to consider Ethical Applications. Additionally, findings derived from MSR studies will be presented.

… upcoming events

- September 29: CANSSI Ontario Research Day Schedule and Registration

- Distinguished Lecture Series in Statistical Sciences
- Xihong Lin, Harvard U Details and Registration

- September 29 3.30 89 Chestnut Street, 3rd Floor
  Lessons learned from the COVID-19 Pandemic: a statistician's reflection

- September 30 3.30 UY9014
  Ensemble methods for testing a global null hypothesis

2022 DLSS: Xihong Lin

Professor, Department of BiostatisticsCoordinating Director, Program in Quantitative Genomics; Harvard T.H. Chan School of Public Health; Professor of Statistics, Department of Statistics, Harvard University

Today

1. Grading scheme, Comments re texts
2. Upcoming events
3. Steps in analysis; types of studies
4. Recap Week 1
5. Linear Regression Part 2: testing groups of variables, checking model assumptions, collinearity, $p > n$
6. In the news

Steps in Analysis

LM-2 §1.1

- understand the physical background
- understand the objective
- make sure you know what the client wants
- put the problem into statistical terms

Steps in Analysis

LM-2 §1.1

- understand the physical background
- understand the objective
- make sure you know what the client wants
- put the problem into statistical terms

- How were the data collected:
  - are the data observational or experimental? etc.
  - is there nonresponse
  - are there missing values
  - how are the data coded
  - what are the units of measurement
  - beware of data entry errors

Components of investigation

CD §1.2

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

Components of investigation

CD §1.2

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

Components of investigation

CD §1.2

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately

but not pessimistically

Components of investigation

CD §1.2

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately

but not pessimistically

- visualization strategies, conveyance of uncertainties

Experimental and observational studies

LM-2 Ch.5.3,5.4; CD Ch. 1

- experiment is a study in which all key elements are under the control of the investigator
- in an observational study features and responses of interest are measured, not assigned by the investigator
- in an experiment, there is typically one or more treatments, and treatment is usually assigned at random

using a randomization device

- LM-2 gives two reasons for randomizing treatment assignment

clinical trial

1. groups are balanced on other features

on average

2. can analyse using permutation test
3. elimination of personal judgement in assigning treatment to units in the experiment

randomized, double-blind

- Example: hydroxychloroquine as a treatment for COVID

Hydroxychloroquine

New England J Medicine, June 2020

ORIGINAL ARTICLE

Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid-19

Joshua Geleris, M.D., Yifei Sun, Ph.D., Jonathan Platt, Ph.D., Jason Zucker, M.D., Matthew Baldwin, M.D., George Hripcsak, M.D., Angelena Labella, M.D., Daniel K. Manson, M.D., Christine Kubin, Pharm.D., R. Graham Barr, M.D., Dr.P.H., Magdalena E. Sobieszczyk, M.D., M.P.H., and Neil W. Schluger, M.D.

Article Figures/Media Metrics

June 18, 2020
N Engl J Med 2020; 382:2411-2418
DOI: 10.1056/NEJMoa2012410

14 References 300 Citing Articles

Chinese Translation 中文翻译

Abstract

Hydroxychloroquine

New England J Medicine, June 2020

ORIGINAL ARTICLE

Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid-19

Joshua Geleris, M.D., Yifei Sun, Ph.D., Jonathan Platt, Ph.D., Jason Zucker, M.D., Matthew Baldwin, M.D., George Hripcsak, M.D., Angelena Labella, M.D., Daniel K. Manson, M.D., Christine Kubin, Pharm.D., R. Graham Barr, M.D., Dr.P.H., Magdalena E. Sobieszczyk, M.D., M.P.H., and Neil W. Schluger, M.D.

Article Figures/Media Metrics

14 References 300 Citing Articles

June 18, 2020
N Engl J Med 2020; 382:2411-2418
DOI: 10.1056/NEJMoa2012410
Chinese Translation 中文翻译

Abstract

response

"In this observational study involving patients with Covid-19 who had been admitted to the hospital, hydroxychloroquine administration was not associated with either a greatly lowered or an increased risk of the composite end point of intubation or death. Randomized, controlled trials of hydroxychloroquine in patients with Covid-19 are needed."

… hydroxychloroquine

New England J Medicine, August 2020

ORIGINAL ARTICLE

A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19

David R. Boulware, M.D., M.P.H., Matthew F. Pullen, M.D., Ananta S. Bangdiwala, M.S., Katelyn A. Pastick, B.Sc., Sarah M. Lofgren, M.D., Elizabeth C. Okafor, B.Sc., Caleb P. Skipper, M.D., Alanna A. Nascene, B.A., Melanie R. Nicol, Pharm.D., Ph.D., Mahsa Abassi, D.O., M.P.H., Nicole W. Engen, M.S., Matthew P. Cheng, M.D., et al.

Article Figures/Media Metrics

18 References 128 Citing Articles Letters 11 Comments

August 6, 2020
N Engl J Med 2020; 383:517-525
DOI: 10.1056/NEJMoa2016638

"We conducted a randomized, double-blind, placebo-controlled trial across the United States and parts of Canada testing hydroxychloroquine as postexposure prophylaxis."

"This randomized trial did not demonstrate a significant benefit of hydroxychloroquine as postexposure prophylaxis for Covid-19. "

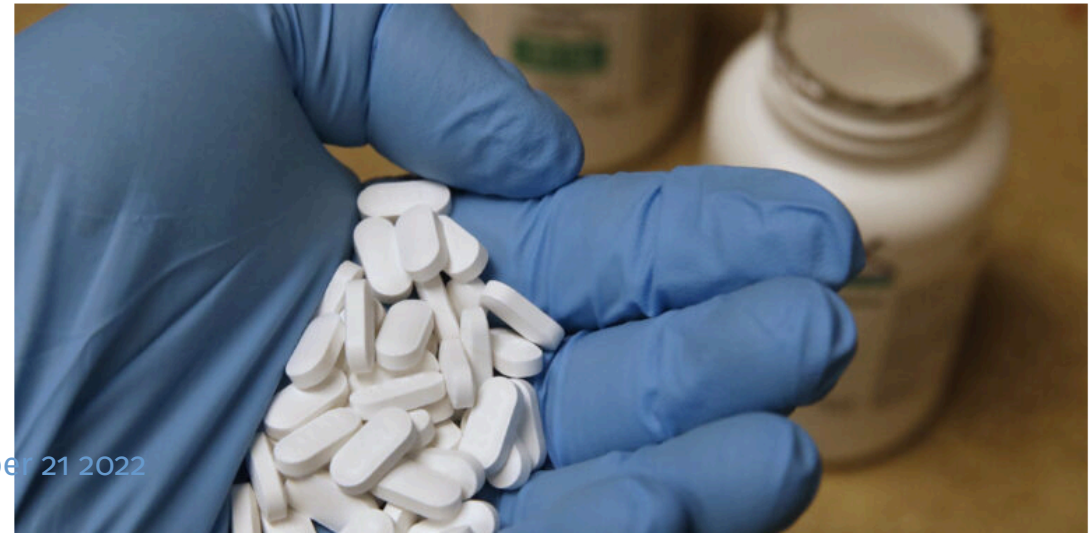
... hydroxychloroquine

also June 2020

HEALTH

Lancet, New England Journal retract Covid-19 studies, including one that raised safety concerns about malaria drugs

By Andrew Joseph June 4, 2020 Reprints



data quality

… hydroxychloroquine

March 2021

Cochrane

Trusted evidence.
Informed decisions.
Better health.

Search…

Our evidence | About us | Join Cochrane | News and jobs | Cochrane Library

Coronavirus (COVID-19) resources

Is chloroquine or hydroxychloroquine useful in treating people with COVID-19, or in preventing infection in people who have been exposed to the virus?

Cochrane Reviews

… hydroxychloroquine

March 2021

Cochrane Reviews ▾ Trials ▾ Clinical Answers ▾ About ▾ Help ▾

Cochrane Database of Systematic Reviews | Review - Intervention

Chloroquine or hydroxychloroquine for prevention and treatment of COVID-19

Bhagteshwar Singh, Hannah Ryan, Tamara Kredo, Marty Chaplin, Tom Fletcher Authors' declarations of interest

Version published: 12 February 2021 Version history

https://doi.org/10.1002/14651858.CD013587.pub2

Collapse all Expand all

Abstract

Hydroxychloroquine does not reduce deaths from COVID-19, and probably does not reduce the number of people needing mechanical ventilation.

Hydroxychloroquine caused more unwanted effects than a placebo treatment, though it did not appear to increase the number of serious unwanted effects.

We do not think new studies of hydroxychloroquine should be started for treatment of COVID

… hydroxychloroquine

March 2021

Cochrane Library

Trusted evidence.
Informed decisions.
Better health.

Title Abstra

Cochrane Reviews ▾ Trials ▾ Clinical Answers ▾ About ▾ Help ▾

Cochrane Database of Systematic Reviews | Editorial

Contested effects and chaotic policies: the 2020 story of (hydroxy) chloroquine for treating COVID-19

Susan Gould, Susan L Norris Authors' declarations of interest

Version published: 25 March 2021

https://doi.org/10.1002/14651858.ED000151

English Deutsch Español فارسی Français 日本語 Bahasa Malaysia Português Русский ▼ Media | Contact us | Community | My Account

Cochrane

Trusted evidence.
Informed decisions.
Better health.

Search...

Our evidence | About us | Join Cochrane | News and jobs | Cochrane Library ▶

Ivermectin for preventing and treating COVID-19

Published:
21 June 2022

Authors:
Popp M, Reis S, Schießer S, Hausinger RIlona, Stegemann M, Metzendorf M-I, Kranke P, Meybohm P, Skoetz N, Weibel S

Primary Review Group:
Infectious Diseases Group,

Is ivermectin effective for COVID-19?

Key messages

We found no evidence to support the use of ivermectin for treating COVID-19 or preventing SARS-CoV-2 infection. The evidence base improved slightly in this update, but is still limited.

Evaluation of ivermectin is continuing in 31 ongoing trials, and we will update this review again when their results become available.

Am score 835

Who is talking about this article?

Video: Systematic reviews explained

How our health evidence can help you

$$\frac{\partial \log L(\beta, \sigma^2)}{\partial \sigma^2} = 0$$

$\tilde{\sigma}^2_{\text{not mle}} = \Sigma(\ )^2/(n-p) \leftarrow$ unb.

$\hat{\beta}$ max. lik. est. (LS)

Recap Week 1

Linear regression

$\hat{\sigma}^2$ " " " of σ^2 $E(\hat{\sigma}^2) \neq \sigma^2$

$y = \beta_0 + \beta_1 x + \varepsilon$

- model
- estimation
- inference

$E(\varepsilon) = 0$

$var(\varepsilon) = \sigma^2 I$

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \dots$$

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y \quad p\times 1$$

$$\min_{\beta} (y - X\beta)^T (y - X\beta)$$

$$E(\hat{\beta}) = \beta; \quad var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

X full rank $\iff X^T X$ is invertible

$f(x, y) \neq f(x) f(y)$ even tho $r = 0$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$E(y) = X\beta$

$var(y) = \sigma^2 I$

$e^{-\frac{1}{2}x^2}$ $e^{-\frac{1}{2}y^2}$

$$\frac{\partial}{\partial \beta}\{SS(\beta)\}\Big|_{\hat{\beta}} = 0$$

Recap Week 1

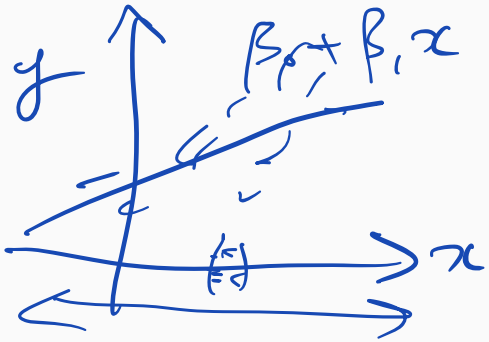
$\min_{\beta} (y - X\beta)^T (y - X\beta)$

- model

$$y = X\beta + \epsilon$$

- estimation

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y$$



- inference

$$\mathrm{E}(\hat{\beta}) = \beta; \quad \mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

- if $\epsilon \sim N(0, \sigma^2 I)$:

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}),$$

$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$

- estimate of σ^2:

$$\boxed{\tilde{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - p}}$$

minimum

ow f

- leads to t-tests for individual components β_j and confidence intervals

... Recap Week 1

- X is an $n \times p$ matrix of explanatory variables, which may be:
- measured in the sample (obs'l data) LM-2 §2.6; LM-1 §2.8; SM Ex 8.3
- fixed by design (exp't) HW1; LM-1 §3.6; LM-2 §2.11; SM Ex 8.4,
- introduced to make the model more flexible LM-2 Ch.9; LM-2 Ch.7; SM Ex 8.2

... Recap Week 1

- X is an $n \times p$ matrix of explanatory variables, which may be:
- measured in the sample — LM-2 §2.6; LM-1 §2.8; SM Ex 8.3
- fixed by design — HW1; LM-1 §3.6; LM-2 §2.11; SM Ex 8.4,
- introduced to make the model more flexible — LM-2 Ch.9; LM-2 Ch.7; SM Ex 8.2
- X often called the design matrix — in `R`, `model.matrix`

... Recap Week 1

$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1}$

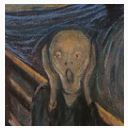
intercept

- X is an $n \times p$ matrix of explanatory variables, which ~~may~~ be:
- measured in the sample — LM-2 §2.6; LM-1 §2.8; SM Ex 8.3
- fixed by design — HW1; LM-1 §3.6; LM-2 §2.11; SM Ex 8.4,
- introduced to make the model more flexible — LM-2 Ch.9; LM-2 Ch.7; SM Ex 8.2
- X often called the design matrix — in `R`, `model.matrix`
- p or $p+1$?

$$\underset{n \times p}{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & \cdot & & \\ 1 & \cdot & & \\ \vdots & \cdot & & \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$$

$\boxed{\text{assume } x_{i1} = 1}$

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$



$\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})$

Example

$\in \mathbb{R}^p$

```
summary(model1)
Call:
lm(formula = lpsa ~ ., data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7331 -0.3713 -0.0170  0.4141  1.6381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

not of interest

t-test

$$\frac{\hat{\beta}_j}{\tilde{\sigma}\,(X^TX)^{-1}_{jj}} \sim t \text{ if } \beta_j = 0$$

$$F = \frac{Reg\ SS/p-1}{RSS/n-p} = 20.86$$

library(faraway)

```
> summary(model1)
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.6693367  1.2963875  0.5163  0.606934
lcavol       0.5870218  0.0879203  6.6767 2.111e-09
lweight      0.4544674  0.1700124  2.6731  0.008955
age         -0.0196372  0.0111727 -1.7576  0.082293
lbph         0.1070540  0.0584492  1.8316  0.070398
svi          0.7661573  0.2443091  3.1360  0.002329
lcp         -0.1054743  0.0910135 -1.1589  0.249638
gleason      0.0451416  0.1574645  0.2867  0.775033
pgg45        0.0045252  0.0044212  1.0235  0.308860

n = 97, p = 9, Residual SE = 0.70842, R-Squared = 0.65
```

$\tilde{\sigma}$

$$n - p = 97 - 9 = 88$$

$$\tilde{\sigma}^2 = \frac{SS(\hat{\beta})}{88}$$

Estimation

$H_0: \beta_1 = \cdots = \beta_8 = 0$

fitted

- hat matrix H

$$\hat{y}_{n\times 1} = Hy = X\hat{\beta} = \underbrace{\boxed{\overset{n\times p}{X}\overset{p\times p}{(X^TX)^{-1}}\overset{p\times n}{X^T}}}_{\text{Hat matrix}} y_{n\times 1}$$

- residual sum of squares *RSS* $(y - X\hat{\beta})^T(y - X\hat{\beta})$ $SS(\hat{\beta})$

minimized value of obj. fn

$$= (y - Hy)^T(y - Hy)$$

$$(y - Hy)^T$$

$$y^T - (Hy)^T$$

$$y^T - y^TH^T$$

$$= y^T(I - H)^T(I - H)y$$

$E(y) = X\beta$

$$\underline{H^T = H}$$ "idempotent"

$$(I - H)^T = (I - H)$$

Estimation

- hat matrix $H = X(X^T X)^{-1} X^T$ $\hat{y} = Hy$ extract elements of H in $\mathbb{R}$
- residual sum of squares *RSS* $y^T(I-H)y$
- coefficient of determination R^2 ← a measure of how ~~good~~ well the model fits

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = R^2 \qquad (y-\hat{y})^T(y-\hat{y})$$

residual SS in model $y_i = \beta_0 + \varepsilon_i$

Estimation

- hat matrix
- residual sum of squares *RSS*
- coefficient of determination R^2
- identifiability
- orthogonality

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\operatorname{var}\hat{\beta} = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} (X_1^T X)^{-1} & 0 \\ 0 & (X_2^T X)^{-1} \end{bmatrix}$$

$$\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = 0$$

← can we estimate β well (all components)

← columns of X matrix have 0 corr

$$[x_{col\,3} \cdot x_{col\,6}] = 0$$

$$X_1 \perp X_2$$

$$X^T X = \begin{bmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{bmatrix}$$

$$[\quad \cdot \quad] \simeq 1$$

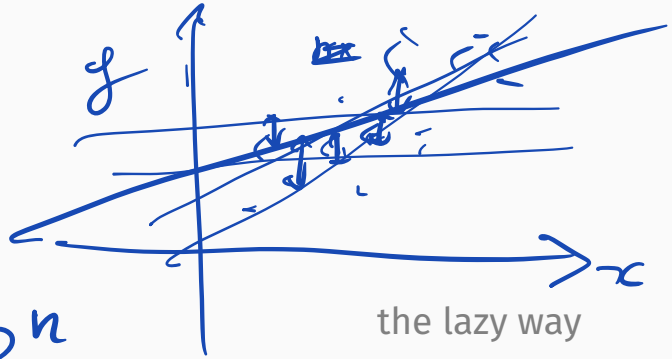
0	1	2	3	4
1	0	0	⋮	
1	1	0		
1	0	1		
⋮	1	1		
	⋮			
1				

Aside: Lazy Notation

- $y_i = x_i^{\mathrm{T}} \beta + \epsilon_i, \quad i = 1, \ldots, n$
- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y}, \boldsymbol{\epsilon}$ vectors of length n

Aside: Lazy Notation

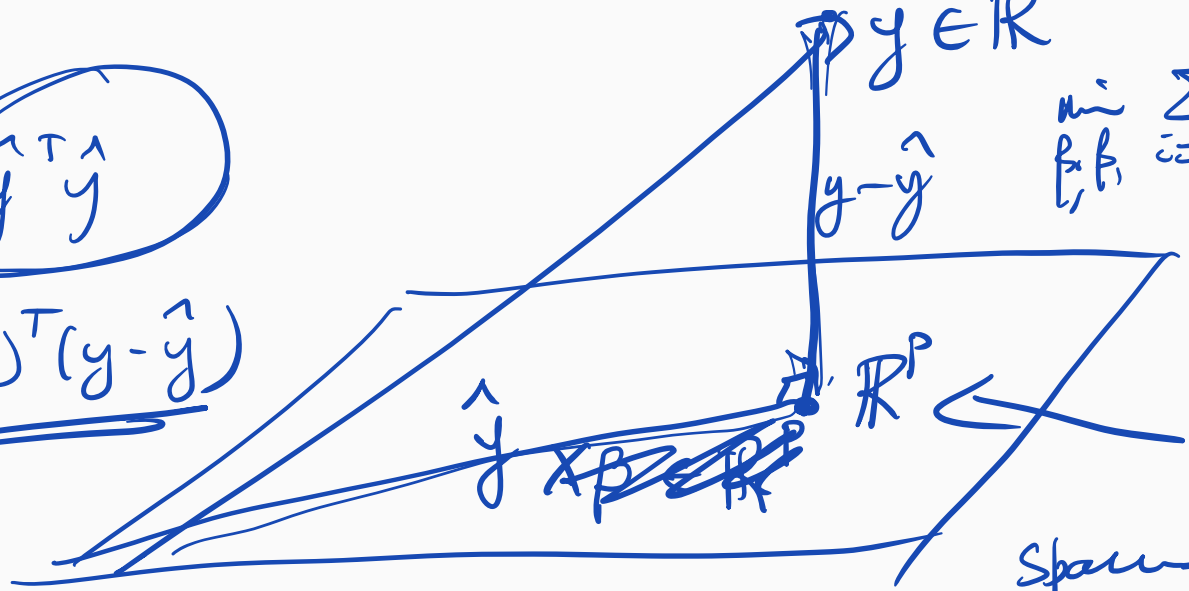
- $y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$
- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y}, \boldsymbol{\epsilon}$ vectors of length n
- $y = X\beta + \epsilon, \qquad$ also vectors of length n



the lazy way

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$y^T y = \hat{y}^T \hat{y} + (y - \hat{y})^T (y - \hat{y})$$



linear space spanned by col's of X

Aside: Lazy Notation

- $y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$
- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y}, \boldsymbol{\epsilon}$ vectors of length n
- $y = X\beta + \epsilon, \qquad$ also vectors of length n

the lazy way

- a generic observation $y \in \mathbb{R}$ for a generic vector of covariates $x \in \mathbb{R}^1$ often written

$$y = x^{\mathrm{T}}\beta + \epsilon$$

or even $x\beta + \epsilon$

- "where we hope there is no confusion"

Comparing Models

LM 3.1,2; SM 8.5

- Residual sum of squares $(y-\hat{y})^T(y-\hat{y})$ RSS

$$\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$$

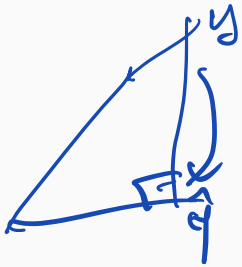
- Decomposition of variance TSS

Total SS

$$\sum y_i^2 = y^T y = (y-\hat{y}+\hat{y})^T(y-\hat{y}+\hat{y})$$

$$= \sum(y_i - x_i^T\hat{\beta})^2 + \sum_{i=1}^{n}(x_i^T\hat{\beta})^2$$

$$= (y-\hat{y})^T(y-\hat{y}) + \hat{y}^T\hat{y}$$



$$H^T H = H$$

Comparing Models

LM 3.1,2; SM 8.5

- Residual sum of squares $(y-\hat{y})^T(y-\hat{y})$ RSS
- Decomposition of variance $y^Ty - n\bar{y}^2 = (y-\hat{y})^T(y-\hat{y}) + (\hat{y}^T\hat{y}) - n\bar{y}^2$ TSS
- Corrected TSS $\sum(y_i - \bar{y})^2$

$$(y-\bar{y}\mathbf{1})^{\mathrm{T}}(y-\bar{y}\mathbf{1}) = \overbrace{(y-X\hat{\beta})^{\mathrm{T}}(y-X\hat{\beta})}^{RSS} + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{y}^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\hat{\beta})^2 + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}_2$$

var in y about $\bar{y}$

var in y about $x_i^T\hat{\beta}$

… comparing Models

LM 3.1,2; SM 8.5

- Residual sum of squares RSS

- Decomposition of variance TSS

... comparing Models

LM 3.1,2; SM 8.5

- Residual sum of squares RSS

- Decomposition of variance TSS

- Corrected TSS $\sum(y_i - \bar{y})^2$

$$
\begin{aligned}
(y - \bar{y}\mathbf{1})^{\mathrm{T}}(y - \bar{y}\mathbf{1}) &= (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta} - n\bar{y}^2 \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - x_i^{\mathrm{T}}\hat{\beta})^2 + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}_2
\end{aligned}
$$

... comparing models

$\beta_2 = (\beta_1 \cdots \beta_{p-1})$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}_2$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

(corr'd)

(without β_0)

RSS, $SS(\hat{\beta})$

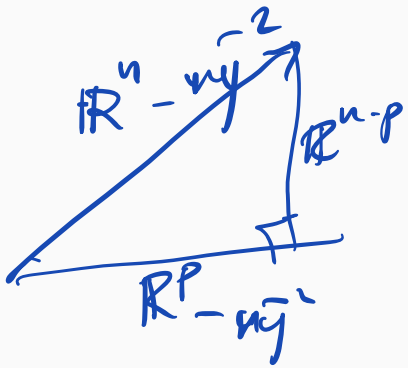
- LHS is
- comparison of LHS to $SS(\hat{\beta})$ reflects

$$R^2 = 1 - \frac{RSS}{TSS_{\text{corr'd}}}$$

coeff't of determination

if you leave out β_0 then ~~R^2~~

… comparing models



$$\boxed{R^2 = 1 - \frac{RSS}{TSS}}$$

$$\overset{n-1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \overset{n-p}{(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})} + \overset{p-1}{\hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}_2}$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

p-1 d.f.

explained by model

RSS, $SS(\hat{\beta})$

- LHS is
- comparison of LHS to $SS(\hat{\beta})$ reflects
- $$F = \frac{\overbrace{(TSS - RSS)}^{\text{reg. SS}}/(p-1)}{\underbrace{RSS}_{\text{residual SS}}/(n-p)} \sim F_{p-1,\,n-p} \quad \text{under } H_0$$
- here $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, but we don't care about β_1 $\quad (\cancel{\beta_0}, \beta_1, \ldots, \beta_{p-1})$

$H_0: \beta_1 \cdots \beta_{p-1} = 0$

… comparing models

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

RSS, $SS(\hat{\beta})$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to ~~$SS(\hat{\beta})$~~ reflects importance of other βs, i.e. importance of explanatory variables

reg SS

... comparing models

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

RSS, $SS(\hat{\beta})$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to $SS(\hat{\beta})$ reflects importance of other βs, i.e. importance of explanatory variables
- Reg SS

$$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}$$

… comparing models

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) + \hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}$$

$$\text{Total SS} = \text{Residual SS} + \text{Regression SS}$$

RSS, $SS(\hat{\beta})$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to $SS(\hat{\beta})$ reflects importance of other βs, i.e. importance of explanatory variables
- $$F = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}$$ $H_0: \beta_1 = \cdots \beta_{p-1} = 0$
- here $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, but we don't care about β_1 $(\beta_0, \beta_1, \ldots, \beta_p)$

... comparing models

- same argument can be derived for comparing submodels
- for example, testing $(\underline{\beta_2, \beta_3, \beta_4}) = (0, 0, 0)$

... comparing models

- same argument can be derived for comparing submodels
- for example, testing $(\beta_2, \beta_3, \beta_4) = (0, 0, 0)$ $= H_0$
- fit full model $\longrightarrow$ RSS_{full}; fit reduced model $\longrightarrow$ RSS_{red}

resid

residual SS

- $$F = \frac{(RSS_{red} - RSS_{full})/(p-q)}{RSS_{full}/(n-p)} \sim F_{p-q,\, n-p}$$

$H_0: \beta_2$

coeff's in full

e.g.

8-3

q = # " " reduced

... comparing models

- same argument can be derived for comparing submodels
- for example, testing $(\beta_2, \beta_3, \beta_4) = (0, 0, 0)$

- fit full model $\longrightarrow RSS_{full}$; fit reduced model $\longrightarrow RSS_{red}$
- $$F = \frac{(RSS_{red} - RSS_{full})/(p-q)}{RSS_{full}/(n-p)}$$

- see LM 3.1, SM §8.2 (p.367) for connection to likelihood ratio test

- when would we want to do this?

... comparing models

```
head(prostate)
#     lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547

model1 <- lm(lpsa ~ ., data = prostate)
```

… comparing models

```
> summary(model1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Residual standard error: 0.7084 on 88 degrees of freedom

F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

out in reduced model

... comparing models

```
model3 <- lm(lpsa ~ lcavol + lweight  + age + lbph + svi, data = prostate)
anova(model3,model1)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight  + age + lbph + svi
Model 3: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     91 45.526
2     88 44.163  3   1.3625 0.905 0.4421
```

$p-q$

$= 8-5 = 3$

$F_{3,88}$

$\Pr F_{88} > 0.905$

$= 0.4421$

does this make sense?

not really

Factor variables

- F-tests are used when the columns to be removed form a group
- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix
- these variables should either all be in, or all be out

in most cases

$y \sim \underline{\underline{ns}}(x, 3)$

Factor variables

- F-tests are used when the columns to be removed form a group
- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix
- these variables should either all be in, or all be out

in most cases

-
```
  prostate$gleason_factor <- factor(prostate$gleason)
  levels(prostate$gleason_factor)
[1] "6" "7" "8" "9"
  model_fac <- lm(lpsa ~ .-gleason, data=prostate)
```

... factor variables

```
  model_fac <- lm(lpsa ~ .-gleason, data=prostate)
  sumary(model_fac)
  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.91328    0.84084    1.09   0.2804
lcavol           0.56999    0.09010    6.33  1.1e-08
lweight          0.46879    0.16961    2.76   0.0070
age             -0.02175    0.01136   -1.91   0.0589
lbph             0.09968    0.05898    1.69   0.0946
svi              0.74588    0.24740    3.01   0.0034
lcp             -0.12511    0.09559   -1.31   0.1941
pgg45            0.00499    0.00467    1.07   0.2885
gleason_factor7  0.26761    0.21942    1.22   0.2259
gleason_factor8  0.49682    0.76927    0.65   0.5201
gleason_factor9 -0.05621    0.50020   -0.11   0.9108

n = 97, p = 11, Residual SE = 0.70, R-Squared = 0.67
```

check `model.matrix`

... factor variables

```
model_nog <- lm(lpsa ~ . - gleason - gleason_factor, data = prostate)

anova(model_fac, model_nog) # compare two models

Analysis of Variance Table

Model 1: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45 + gleason_factor) - gleason - gleason_factor
Model 2: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45 + gleason_factor) - gleason
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     89 44.2
2     86 42.7  3      1.48 0.99    0.4
```

... factor variables

- with designed experiments, covariates are often factors set at pre-determined levels
- see Ch 14 LM-2 (Ch 13 LM-1)

Example 8.4 in SM

... factor variables

- with designed experiments, covariates are often factors set at pre-determined levels
- see Ch 14 LM-2 (Ch 13 LM-1)

Example 8.4 in SM

- if the design is perfectly balanced, then X has orthogonal columns, and $X^{\mathrm{T}}X$ is diagonal

- so $\hat{\beta}_j$'s are uncorrelated, and hence independent (under normality assumption)

... factor variables

- with designed experiments, covariates are often factors set at pre-determined levels
- see Ch 14 LM-2 (Ch 13 LM-1)

Example 8.4 in SM

- if the design is perfectly balanced, then X has orthogonal columns, and $X^{\mathrm{T}}X$ is diagonal
- so $\hat{\beta}_j$'s are uncorrelated, and hence independent (under normality assumption)
- more generally we might have $X^{\mathrm{T}}X$ block diagonal, e.g.

importance?

read Ch 3.5

Ch 3.3, 3.4, 3.6 more specialized

In the News

New Brunswick

'Nice snake, shame about the legs' and other science humour

Does humour in a scholarly paper's title make scientists want to read more?

Mia Urquhart · CBC News · Posted: Sep 20, 2022 7:00 AM AT | Last Updated: 9 hours ago

In the News

If this title is funny, will you cite me?

Citation impacts of humour and other features of article titles in ecology and evolution

Stephen B. Heard[1,2], Chloe A. Cull[1,3], and Easton R. White[4]

[1]Dept. of Biology, University of New Brunswick, Fredericton, NB Canada E3B 5A3

[2]Corresponding author. sheard@unb.ca; Dept. of Biology, University of New Brunswick, PO Box 4400, Fredericton, NB Canada E3B 5A3. Phone: 506-452-6047. FAX: 506-435-3570.

see humour-science.pdf

In the News

Tea drinkers enjoy possible health benefits, study suggests

A cup of tea just got a bit more relaxing.

Tea can be part of a healthy diet and people who drink tea may even be a little more likely to live longer than those who don't, according to a large study.

Tea contains helpful substances known to reduce inflammation. Past studies in China and Japan, where green tea is popular, suggested health benefits. The new study extends the good news to the U.K.'s favourite drink: black tea.

Scientists from the U.S. National Cancer Institute asked about the tea habits of nearly a half million adults in the United Kingdom, then followed them for up to 14 years. They adjusted for risk factors such as health, socioeconomics, smoking, alcohol intake, diet, age, race and gender.

Higher tea intake – two or more cups daily – was linked to a modest benefit: A nine to 13 per cent lower risk of death from any cause versus non-tea drinkers. Tea temperature, or adding milk or sugar, didn't change the results.

The study, published Monday in Annals of Internal Medicine, found the association held up for heart-disease deaths, but there was no clear trend for cancer deaths. Researchers weren't sure why, but it's possible there weren't enough cancer deaths for any effect to show up, said Maki Inoue-Choi, who led the study.

A study like this, based on observing people's habits and health, can't prove cause and effect.

"Observational studies like this always raise the question: Is there something else about tea drinkers that makes them healthier?" said Marion Nestle, a professor of food studies at New York University. "I like tea. It's great to drink. But a cautious interpretation seems like a good idea."

There's not enough evidence to advise changing tea habits, said Inoue-Choi.

"If you drink one cup a day already, I think that is good," she said. "And please enjoy your cup of tea."

ASSOCIATED PRESS

Associated Press Health and Science Department receives support from the Howard Hughes Medical Institute's Department of Science Education.

In the News

Home » News & Events » News Releases

NEWS RELEASES

Media Advisory Monday, August 29, 2022

NIH study of tea drinkers in the UK suggests health benefits for black tea

What

A prospective study of half a million tea drinkers in the United Kingdom has shown that higher tea intake was associated with a modestly lowered risk of death. The study, led by researchers at the National Cancer Institute, part of the National Institutes of Health, is a large and comprehensive analysis of the potential mortality benefits of drinking black tea, which is the most common type of tea consumed in the U.K.

Past studies finding a modest association between higher tea intake and lower risk of death have mainly focused on Asian populations, who commonly drink green tea. Studies on black tea have yielded mixed results.

In the new study, the researchers found that people who consumed two or more cups of tea per day had a 9% to 13% lower risk of death from any cause than people who did not drink tea. Higher tea consumption was also associated with a lower risk of death from cardiovascular disease, ischemic heart disease, and stroke. The association was seen regardless of preferred tea temperature, the addition of milk or sugar, and genetic variations affecting the rate at which people metabolize caffeine.

The findings, which appear Aug. 30, 2022, in the *Annals of Internal Medicine*, suggest that black tea, even at higher levels of intake, can be part of a healthy diet, the researchers wrote.

The study involved 498,043 men and women between ages 40 and 69 who participated in a large cohort study called UK Biobank. The participants were followed for about 11 years, and death information came from a linked database from the UK National Health Service.

Institute/C

National Canc

Contact

NCI Press Offi
240-760-6600

Connect w

Subscribe

RSS Feed

In the News

Annals of Internal Medicine

ORIGINAL RESEARCH

Tea Consumption and All-Cause and Cause-Specific Mortality in the UK Biobank

A Prospective Cohort Study

Maki Inoue-Choi, PhD; Yesenia Ramirez, MPH; Marilyn C. Cornelis, PhD; Amy Berrington de González, DPhil; Neal D. Freedman, PhD; and Erikka Loftfield, PhD

Background: Tea is frequently consumed worldwide, but the association of tea drinking with mortality risk remains inconclusive in populations where black tea is the main type consumed.

Objective: To evaluate the associations of tea consumption with all-cause and cause-specific mortality and potential effect modification by genetic variation in caffeine metabolism.

Design: Prospective cohort study.

Setting: The UK Biobank.

Participants: 498 043 men and women aged 40 to 69 years who completed the baseline touchscreen questionnaire from 2006 to 2010.

Measurements: Self-reported tea intake and mortality from all causes and leading causes of death, including cancer, all cardiovascular disease (CVD), ischemic heart disease, stroke

participants drinking 1 or fewer, 2 to 3, 4 to 5, 6 to 7, 8 to 9, and 10 or more cups per day were 0.95 (95% CI, 0.91 to 1.00), 0.87 (CI, 0.84 to 0.91), 0.88 (CI, 0.84 to 0.91), 0.88 (CI, 0.84 to 0.92), 0.91 (CI, 0.86 to 0.97), and 0.89 (CI, 0.84 to 0.95), respectively. Inverse associations were seen for mortality from all CVD, ischemic heart disease, and stroke. Findings were similar regardless of whether participants also drank coffee or not or of genetic score for caffeine metabolism.

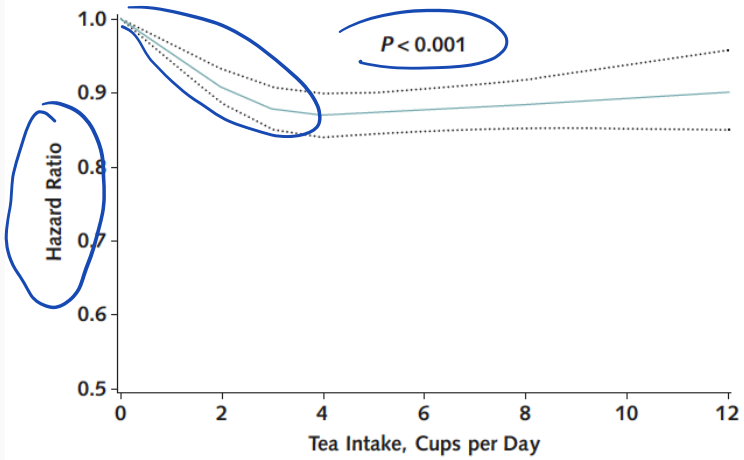
Limitation: Potentially important aspects of tea intake (for example, portion size and tea strength) were not assessed.

Conclusion: Higher tea intake was associated with lower mortality risk among those drinking 2 or more cups per day, regardless of genetic variation in caffeine metabolism. These findings suggest that tea, even at higher levels of intake, can be part of a healthy diet.

see tea-study.pdf

In the News

Figure 1. Dose-response association of tea consumption and all-cause mortality* in the UK Biobank.



* Hazard ratio was adjusted for age; sex; race and ethnicity (White, Black, Asian, mixed, or other race), assessment center, Townsend deprivation score, general health status (excellent, good, fair, or poor), cancer (yes or no), cardiovascular disease (yes or no), diabetes (yes or no), BMI (kg/m^2), tobacco smoking (25-level variable including current smoking status, smoking intensity [current and former smokers], time since quitting [former smokers], and cigar and pipe use [current and former smokers]); physical activity (>10 minutes of moderate or vigorous activity; days per week); alcohol intake (never drinker, former drinker, infrequent drinker [<1 drink per week], occasional drinker [>1 drink per week but <1 drink per day], moderate daily drinker [1 to 3 drinks per day]), or heavy drinker [>3 drinks per day]; coffee intake (cups per day); and dietary intake including vegetables (tablespoons per day), fruits (pieces per day), red meat (beef, lamb, and pork; 0 to 1, 1.5, 2, 2.5, 3 to 21 times per week as quintiles), and processed meat (0, <1, 1, 2 to 4, 5 to 6, and ≥7 times per week). The solid line represents hazard ratio; the dotted line represents 95% CI.

Annals of Internal Medicine • Vol. 175 No. 9 • September 2022 **1205**