

# Applied quals exam questions by topic\*

June 22, 2023

## Linear models

- 2005-5
- 2006-2 (Local linear regression and batch effects)
- 2006-5
- 2007-1
- 2007-4
- 2008-1 (Forward-stepwise regression and the QR decomposition)
- 2008-2 (Modeling soccer goals with linear regression)
- 2009-6 (Adaptivity in linear regression)
- 2011-4 (Assessing time series correlations)
- 2012-2 (Weighted least squares and optimal design)
- 2013-2 (Dropout in linear regression)
- 2013-4 (Dependence in time series regression)
- 2013-5 (Estimating out-of-sample error)

---

\*This document evolved throughout applied coaching by Nikos Ignatiadis, Isaac Gibbs, and Dan Kluger.

- 2014-2 (Big data linear regression with categorical features)
- 2014-3 (Paired versus unpaired regression models - power and validity)
- 2014-4 (Time series principal component regression)
- 2015-5 (Huber loss connection with  $L_1$  penalization)
- 2016-2 (Three angles and measurement error)
- 2016-5 (ANOVA – with interactions without)
- 2016-6 (Testing linearity with replicates)
- 2018-3 (Marginal linear models on the genome)
- 2019-4 (Judging the conclusions of a clinical trial)
- 2019-6 (Solving the population LASSO)
- 2021-6 (Cross-validation in the normal linear model)

Some problems emphasizing computational complexity:

- 2007-2
- 2008-1 (Forward-stepwise regression and the QR decomposition)
- 2010-4 (High-dimensional ridge regression computations)

Mixed models

- 2010-5 (PCA and technical bias correction)
- 2010-1 (Testing with batched data)

## Normal Theory

- 2009-2 (Multiple testing in a Gaussian setup)
- 2010-1 (Testing with batched data)
- 2011-5 (Entropy of Gaussians, (sparse) PCA and “ICA”)
- 2017-6 (A simple Gaussian Hierarchical Model)

## GLMs, exponential families

- 2005-4
- 2008-3 (Poisson regression and Lindsey's method)
- 2009-5 (Fitting a truncated gamma with Lindsey's method)
- 2010-6 (Estimating a truncated Poisson)
- 2012-1 (Deviance manipulations)
- 2012-4 (Separation in logistic regression)
- 2015-1 (Exponential family manipulations for the Gamma distribution)
- 2015-3 (Confidence intervals for the exponential distribution)
- 2015-6 (Life tables, survival curves and GLMs)
- 2016-4 (Separation in logistic regression without intercept)
- 2018-1 (A life table GLM with non-canonical link)
- 2020-3 (Poisson glm with offsets)
- 2021-1 (Modeling association between vaccination and death rates for Covid-19)

## Modeling questions

- 2006-2 (Local linear regression and batch effects)
- 2006-5
- 2008-6 (Modeling PCR data with censored regression)
- 2009-4 (Multiple testing and combining p-values)
- 2010-3 (Turtle diseases and log-rank test)
- 2011-2 (Poisson GLM modeling of call centers)

- 2013-1 (Finding the optimal line through a point cloud)
- 2014-1 (Modeling relative intensity of light sources)
- 2014-5 (A latent variable model for soccer)
- 2016-2 (Three angles and measurement error)
- 2017-3 (Disentangling Days of the Week)
- 2017-4 (Ranking Sports Teams)
- 2018-2 (Group Lasso, cross-validating the whole procedure)
- 2020-6 (Estimating the SARS-CoV-2 mutation rate)
- 2021-1 (Modeling association between vaccination and death rates for Covid-19)

## **Resampling: Bootstrap, Cross-validation, Permutations**

- 2005-6
- 2006-1
- 2008-1 (Forward-stepwise regression and the QR decomposition)
- 2013-5 (Estimating out-of-sample error)
- 2014-6 (Selective inference brainstorming)
- 2015-2 (Bias in cross-validation)
- 2015-3 (Confidence intervals for the exponential distribution)
- 2018-2
- 2019-5 (How to bootstrap swimmers)
- 2020-4 (How to bootstrap starfish data)
- 2021-6 (Cross-validation in the normal linear model)
- Can be used for many problems..

## Multiple testing, Hypothesis testing, and Selective inference

- 2006-3
- 2006-6
- 2007-1
- 2007-4
- 2008-1 (Forward-stepwise regression and the QR decomposition)
- 2009-2 (Multiple testing in a Gaussian setup)
- 2009-4 (Multiple testing and combining p-values)
- 2011-1 (Multiple-testing across genes)
- 2012-3 (Multiple testing with gene expression data)
- 2014-6 (Selective inference brainstorming)
- 2017-2 (Combining p-values)
- 2017-5 (Do I Need to Use Bonferroni?)
- 2018-3 (Marginal linear models on the genome)
- 2021-5 (Testing and inference on a censored Gaussian draw)

## Categorical data analysis and count data

- 2006-4
- 2007-5
- 2011-2 (Poisson GLM modeling of call centers)
- 2012-6 (Testing for a discrete distribution)
- 2011-3 (Pearson  $\chi^2$  puzzle)

- 2016-3 (A missing species problem)
- 2017-4 (Ranking Sport Team's)

## **Classification**

- 2008-5 (LDA and QDA to model malt whiskey)
- 2009-1 (Linear discriminant analysis and QDA)
- 2011-6 (Logistic regression modeling with spatial data)
- 2012-4 (Separation in logistic regression)
- 2012-5 (PCA and sparse PCA before image classification)
- 2015-2 (Bias in cross-validation)
- 2016-1 (Quadratic discriminant analysis and EM)
- 2017-1 (Bayesian Logistic Regression)

## **Multivariate analysis**

- 2006-6
- 2007-3
- 2010-5 (PCA and technical bias correction)
- 2011-5 (Entropy of Gaussians, (sparse) PCA and “ICA”)
- 2012-5 (PCA and sparse PCA before image classification)
- 2013-1 (Finding the optimal line through a point cloud)
- 2014-4 (Time series principal component regression)
- 2018-4 (Low- rank covariances, missing data, factor models)
- 2018-5 (Gaussian Mixture model, Graphical Lasso, EM, NMF)

- 2019-2 (Estimating covariance of truncated Gaussian, EM, factor analysis)
- 2020-2 (Low rank matrix decomposition, least squares)
- 2021-6 (Finding an essential subset)
- 2021-4 (PCA versus k-means)

## **Survival analysis**

- 2005-1
- 2010-3 (Turtle diseases and log-rank test)
- 2015-6 (Life tables, survival curves and GLMs)
- 2018-1 (A life table GLM with non-canonical link)
- 2019-1 (Analysing clinical trial results using Cox model and log-rank test)

## **EM algorithm, Mixtures and Hierarchical models**

- 2005-2
- 2007-6
- 2009-1 (Linear discriminant analysis and QDA)
- 2009-3 (Evaluating accuracy using noisy ground truth data)
- 2010-2 (Modeling contaminated data as a mixture)
- 2013-6 (Estimating a Poisson mixture)
- 2014-5 (A latent variable model for soccer)
- 2016-1 (Quadratic discriminant analysis and EM)

- 2016-3 (A missing species problem)
- 2017-1 (Bayesian Logistic Regression)
- 2017-6 (A simple Gaussian Hierarchical Model)
- 2018-5 (Gaussian Mixture model, Graphical Lasso, EM, NMF)
- 2019-3 (Mixture model priors)
- 2020-5 (Mixture of regressions)

## Other topics

- 2005-3
- 2008-4 (Estimation and testing with a complex sampling scheme)
- 2013-3 (Combining quantile estimators)
- 2015-4 (Causal inferences based on structural assumptions)
- 2018-6 (Privacy-preserving data releases)
- 2020-1 (Maximum likelihood with truncated Gaussian data)
- 2021-3 (Constructing Conformal Prediction Intervals)