📖

Statistics Unit 2

Topic: Linear Regression

- Date: @March 5, 2024
- Lane: *Optimization, Statistics*

What it covers

- What is linear regression?
- An optimization problem: Minimizing mean squared error
- Variance, errors & diagnostics
- Bias- variance tradeoffs.

Requires

Random variables, expectation, variance, projection operators.

Required by

Testing, Estimation

Density estimation

What is linear regression?

A basic problem in science - recover from noisy data a function (an unobservable) of a set of *explanatory* (”independent”) variables , ***X,*** for an *outcome* (”dependent”) variable, ***Y.*** The function can then be applied to a point, ***x*** not in the set of observed explanatory variables, either an “interpolation” or an “extrapolation.” As notation the estimated objects, such as the regression function $\hat{f}(\boldsymbol{x})$ are shown with a “hat” to distinguish them from the actual, unobserved item, e.g. $f(\boldsymbol{x})$. The regression function can be defined $y = E[f(\boldsymbol{x})|\boldsymbol{X}]$, where the *conditional expectation* makes it clear that this applies point-wise over ***X.***

We solve linear regression by estimating the β_i “beta coefficients” in this linear combination, where the explanatory variables $\boldsymbol{x}_i$are the columns of the design matrix (what in Linear Algebra we called ***A),*** and the ***y*** is a column vector - the outcome to be fitted to the estimator

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

This is typically be solved as an optimization problem by minimizing the *mean squared error* $\epsilon = ||\hat{y} - y||^2$.

Note that the first two terms resemble the first two polynomial regression terms. Although $f(\boldsymbol{x})$ is linear the properties of linear regression generalize to a wide range of regression methods — as our brief introduction to polynomial regression shows. Note that each column is typically an “independent” variable - hence the term “multivariate linear regression.”

Minimizing mean squared error

One could use different criteria to solve for the equation’s beta coefficients, however “ordinary least squares” (OLS), minimizes the sum of the squared error between the actual and predicted values. $\sum_i^n (y_i - \hat{y}_i)^2$ ML practitioners call the the “mean squared error”. In statistics it is called the “residual sum of squares” (RSS). In matrix form, taking the derivative of RSS and setting it to zero gives:

$$0 = \frac{\partial \mathrm{RSS}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T(y - \boldsymbol{X}\beta)$$

This gives us an equation for each beta. Solving for the beta

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

So the prediction becomes $\hat{y} = \boldsymbol{X}\hat{\beta}$. Putting these two together we see that the normal equation optimization simply rediscovers the projection operator into the column space of the design matrix ***X.*** This is because the optimization results in the error $y - \hat{y}$ being orthogonal to the column space.

The betas are simply a measure of the sensitivity of the outcome to a change in that feature. The linearity of the regression equation makes this clear

$$\beta_i = \frac{\partial f(x)}{\partial x}$$

The error term - residual variance and standard errors.

For inference we are interested in the variation in the beta coefficients, as shown by their *sampling properties.* To estimate their variance comes from the assumption that ϵ has zero mean, independent of the value of *x,* and from applying the definition of variance to the solution for beta, which is a function of the familiar Gram matrix

$$\sigma^2(\beta) = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2(\mathrm{RSS})$$

where $\sigma^2(\mathrm{RSS})$ is estimated by computing as the mean squared residuals. Given the variance of the individual beta coefficients we can compute error intervals for each beta - called the *standard errors.* Note that these are linear in the residual variance. From the standard errors we can form *interval estimates* of the beta. The convention in statistics is to consider those beta-s whose intervals include zero to have dubious sign, and to disregard them. This is not a strong result, but merely an approximation for ordering the beta-s by importance, or deciding which features are not relevant and can be dropped from the regression.

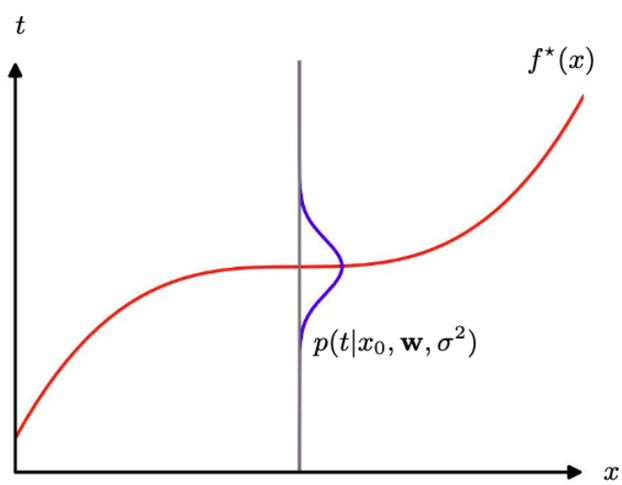
(assumptions)

Prediction

Given the estimated regression function, we can apply it to values of *x* that are not in the original data set, to predict values, as assumed by the model. Since we also have an estimate of the error variance, we can apply that to the prediction to obtain an interval prediction at *x.* The interval tends to increase at the extremes of the range of x.

Accuracy of a prediction depends on the beta estimates, however predictions are less sensitive to variation in individual betas, and the errors are not “one-to-one” — there may be a range of betas that give accurate predictions.

The regression function $f^\star(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$.



(Illustration from C. Bishop & H. Bishop, “Deep Learning” (2023) Springer.

“Bias - variance” tradeoff.

Conventional statistics makes use of the independence of the variance of the error to the variance “explained” by the regression, to decompose the total variation in the data into two terms:

💡 TOTAL SUM of SQUARES = EXPLAINED SUM of SQUARES +RESIDUAL SUM of SQUARES

The ratio of ESS / TSS is called the “r-squared”, a rough approximation of how appropriate the model is to the data it has been applied to.

This decomposition is a kind of analysis of variance, making use of the probabilistic independence of the variables in the decomposition.

A further decomposition of variance splits the explained variance into a bias term and a variance term.

$$\sigma^2(x) = \sigma^2(\epsilon) + \text{Bias}^2(\hat{f}(x)) + \sigma^2(\hat{f}(x))$$

Different choice of models will make different tradeoffs between bias and variance, with more complicated models tending to reduce bias and the expense of greater variance. Thus there are reasons why bias is not always a bad thing. “Good” bias originating from informed judgment is sometimes called “inductive bias” — a term borrowed from early AI.

How can linear regression be extended?

... Linear regression re-appears in many guises ...

Readings

Each textbook spends a chapter on Linear Regression and related topics. These sections apply to the lectures in the rest of the course:

Boyd Ch. 12

Evans Chapter 10.3

ISLP_python, Ch.3

References

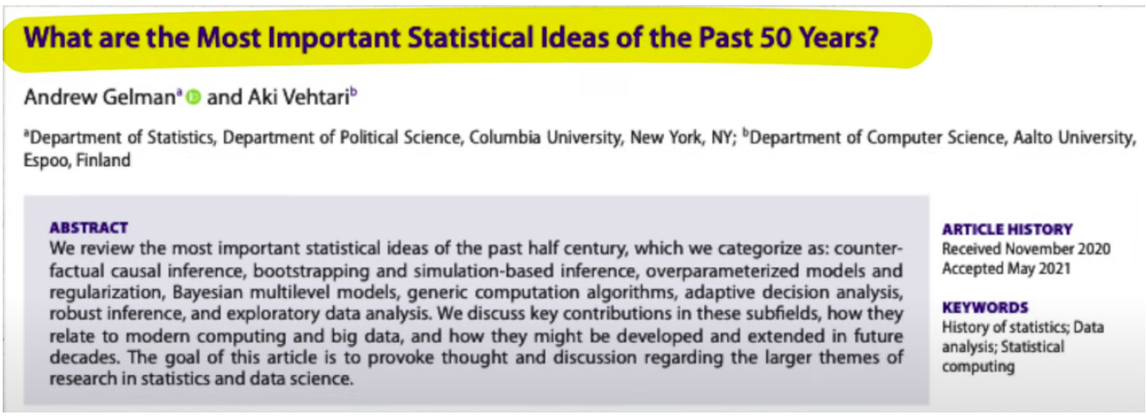
Video on Sampling Variation - Good explanation on re-sampling to estimate sample errors

Simplest Explanation of the Standard Errors of Regression Coefficients - Statistics Help

A simple tutorial explaining the standard errors of regression coefficients. This is a step-by-step explanation of the meaning and importance of the standard error.

https://www.youtube.com/watch?v=1oHe1a3JqHw

Some background -



What are the Most Important Statistical Ideas of the Past 50 Years?

Andrew Gelman[a] and Aki Vehtari[b]

[a]Department of Statistics, Department of Political Science, Columbia University, New York, NY; [b]Department of Computer Science, Aalto University, Espoo, Finland

ABSTRACT

We review the most important statistical ideas of the past half century, which we categorize as: counterfactual causal inference, bootstrapping and simulation-based inference, overparameterized models and regularization, Bayesian multilevel models, generic computation algorithms, adaptive decision analysis, robust inference, and exploratory data analysis. We discuss key contributions in these subfields, how they relate to modern computing and big data, and how they might be developed and extended in future decades. The goal of this article is to provoke thought and discussion regarding the larger themes of research in statistics and data science.

ARTICLE HISTORY
Received November 2020
Accepted May 2021

KEYWORDS
History of statistics; Data analysis; Statistical computing

References

- Searle “Linear Models” (1971).
- Gareth James , Daniela Witten , Trevor Hastie, Robert Tibshirani Introduction to Statistical Learning (2023)https://www.statlearning.com/resources-python (”ISLP” 2023)

- Bradley Efron, Trevor Hastie “Computer Age Statistical Inference” (”CASI” 2021) An advanced, readable text addressed to those already conversant in the advances in statistics brought on by computer science and machine learning.