Model Assessment and Selection in Multiple and Multivariate Regression

Ho Tu Bao
Japan Advance Institute of Science and Technology
John von Neumann Institute, VNU-HCM

Statistics and machine learning

Statistics

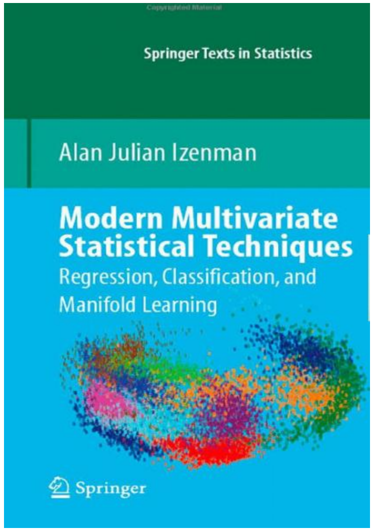
- Long history, fruitful
- Aims to analyze datasets
- Early focused on numerical data
- Multivariate analysis = linear methods on small to medium-sized data sets + batch processing.
- 1970s: interactive computing + exploratory data analysis (EDA)
- Computing power & data storage → machine learning and data mining (aka EDA extension).
- Statisticians interested in ML

Machine learning

- Newer, fast development
- Aims to exploit datasets to learn
- Early focused on symbolic data
- Tends closely to data mining (more practical exploitation of large datasets
- Increasing employs statistical methods
- More practical with computing power
- ML people: need to learn statistics!

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression



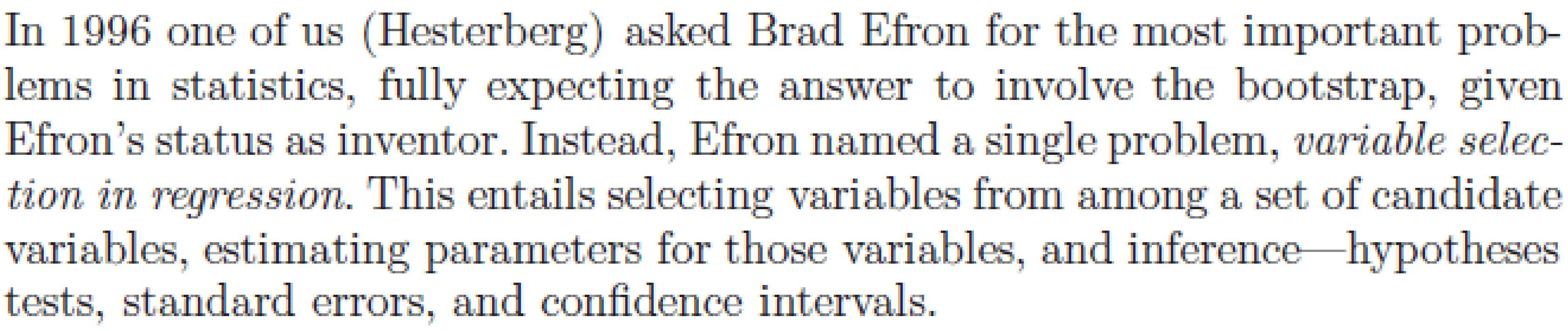
Springer Texts in Statistics

Alan Julian Izenman

Modern Multivariate Statistical Techniques

Regression, Classification, and Manifold Learning

Springer



In 1996 one of us (Hesterberg) asked Brad Efron for the most important problems in statistics, fully expecting the answer to involve the bootstrap, given Efron's status as inventor. Instead, Efron named a single problem, *variable selection in regression*. This entails selecting variables from among a set of candidate variables, estimating parameters for those variables, and inference—hypotheses tests, standard errors, and confidence intervals.

Hesterberg et al., LARS and 11 penalized regression

Introduction

Model and modeling

- Model:
  - Mô tả khái quát hoặc trừu tượng hóa của một thực thể (simplified description or abstraction of a reality).
- Modeling: Quá trình tạo ra một mô hình.
- Mathematical modeling: Description of a system using mathematical concepts and language
  - Linear vs. nonlinear; deterministic vs. probabilistic; static vs. dynamic; discrete vs. continuous; deductive, inductive, or floating.
  - A method for model assessment and selection
- Model selection: Select the most appropriate model
  - Given the problem target and the data → Choose appropriate methods and parameter settings for the most appropriate model.
  - No free lunch theorem.

Introduction

History

- The earliest form of regression was the method of least squares which was published by Legendre in 1805 and by Gauss in 1809.
- The term “regression” coined by Francis Galton in the 19th century to describe a biological phenomenon which was extended by Udny Yule and Karl Pearson to a more general statistical context (1897, 1903).
- In 1950s, 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.
- Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression in , time series, images, graphs, or other complex data objects, nonparametric regression, Bayesian methods for regression, etc.

Introduction

Regression and model

- Given $\{(\mathbf{X}_i, \mathbf{Y}_i), i = 1, \dots, n\}$ where each $\mathbf{X}_i$ is a vector of r random variables $\mathbf{X} = (X_1, \dots, X_r)^\tau$ in a space $\mathbb{X}$ and $\mathbf{Y}_\mathrm{i}$ is a vector of s random variables $\mathbf{Y} = (Y_1, \dots, Y_s)^\tau$ in a space $\mathbb{Y}$
- The problem is to learn a function $f\colon \mathbb{X} \longrightarrow \mathbb{Y}$ from $\{(\mathbf{X}_i, \mathbf{Y}_i), i = 1, \dots, n\}$ satisfies $f(\mathbf{X}_i) = \mathbf{Y}_i, i = 1, \dots, n$.
- When $\mathbb{Y}$ is discrete the problem is called classification and when $\mathbb{Y}$ is continuous the problem is called regression. For regression:
  - When r = 1 and s =1 the problem is called simple regression.
  - When r > 1 and s =1 the problem is called multiple regression.
  - When r > 1 and s > 1 the problem is called multivariate regression.

Introduction

Least square fit

- **Problem statement**
  Adjusting the parameters of a model function to best fit a data set
  - ❑ The model function has adjustable parameters, held in the vector $\boldsymbol{\beta}$
  - ❑ The goal is to find the parameter values for the model which "best" fits the data
  - ❑ The least square method finds its optimum when the sum, S, of squared residuals is a minimum.

Data set

$$\{(\mathbf{X}_i, \mathbf{Y}_i), i = 1, \dots, n\}$$

Model function

$$f(\mathbf{X}, \boldsymbol{\beta})$$

Sum of squared residuals

$$S = \sum_{i=1}^{n} r_i^{\,2}$$

$$r_i = Yi \; - f(\mathbf{X}_i, \boldsymbol{\beta})$$

E.g.

$$f(\mathbf{X}, \boldsymbol{\beta}) = \beta_0 + \beta_1 \mathbf{X}$$

β_0 : *intercept* (phần bị chắn)
β_1 : *slope* (độ dốc)

Introduction

Least square fit

- **Solving the problem**
  - ❑ Minimum of the sum of squared residuals is found by setting the gradient to zero.
  - ❑ The gradient equations apply to all least squares problems.
  - ❑ Each particular problem requires particular expression for the model and its partial derivatives.

Gradient on β

$$\frac{\partial S}{\partial \beta_j} = 2\sum_i r_i \frac{\partial ri}{\partial \beta_i} = 0,\ j = 1, \dots, m$$

$$r_i = Yi \ \_\ f(\mathbf{X}_i, \boldsymbol{\beta})$$

Gradient equation

$$-2\sum_i r_i \boxed{\frac{\partial f(\mathbf{X}_i, \boldsymbol{\beta})}{\partial \beta_j}} = 0,\ j = 1, \dots, m$$

Introduction

Least square fit

- Linear least squares
  - Coefficients φ_i are functions of X_i
- Non-linear least squares
  - There is no closed-form solution to a non-linear least squares problem.
  - Numerical algorithms are used to find the value of the parameter $\boldsymbol{\beta}$ which minimize the objective.
  - The parameters $\boldsymbol{\beta}$ are refined iteratively and the values are obtained by successive approximation.

Linear model function

$$f(\mathbf{X}_i, \boldsymbol{\beta}) = \sum_{j=1}^{m} \beta_j \, \varphi_j(\mathbf{X}_i)$$

$$\boldsymbol{X}_{ij} = \frac{\partial f(\mathbf{X}_i, \boldsymbol{\beta})}{\partial \beta_j} = \varphi_j(\mathbf{X}_i)$$

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

Iterative approximation

Shift vector

$$\beta_j^{k+1} = \beta_j^k + \Delta\beta_j$$

$$f(\mathbf{X}_i, \boldsymbol{\beta}) = f^k(\mathbf{X}_i, \boldsymbol{\beta}) + \sum_{j=1}^{m} \frac{\partial f(\mathbf{X}_i, \boldsymbol{\beta})}{\partial \beta_j} (\beta_j - \beta_j^k)$$

$$= f^k(\mathbf{X}_i, \boldsymbol{\beta}) + \sum_{j=1}^{m} J_{ij} \, \Delta\beta_j$$

Gradient equation

Gauss-Newton algorithm

$$-2 \sum_{i=1}^{n} J_{ij} \left(\Delta Y_i - \sum_{j=1}^{m} J_{ij} \, \Delta\beta_j \right) = 0$$

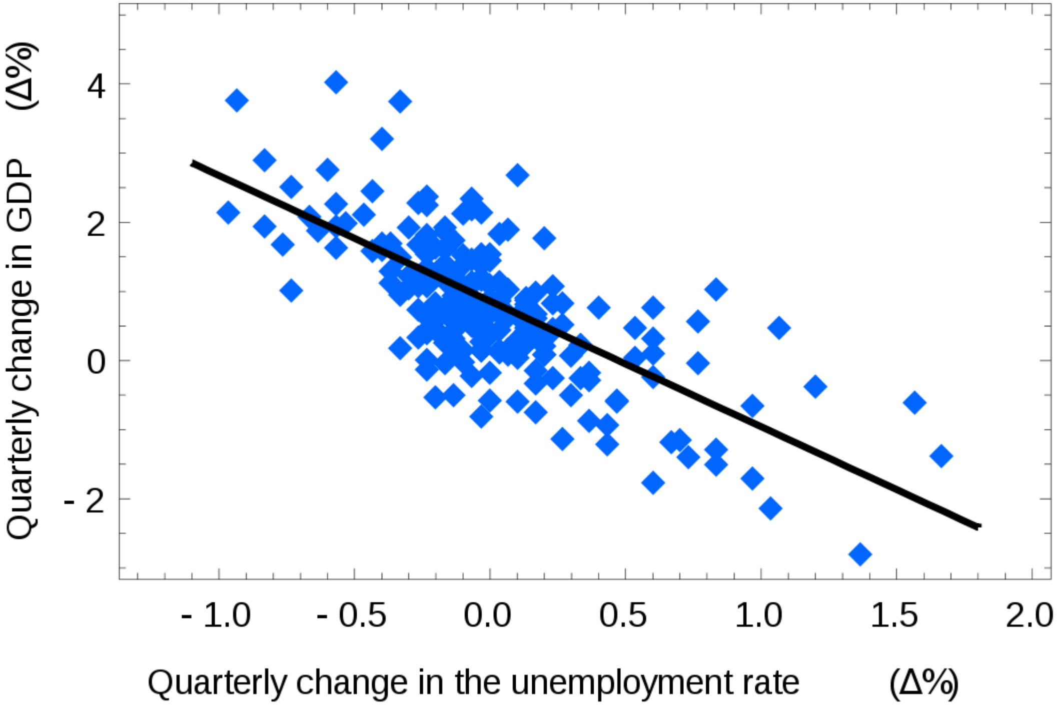
an expression is said to be a *closed-form expression* if it can be expressed analytically in terms of a *finite* number of certain "well-known" functions.

Introduction

Simple linear regression and correlation

Okun’s law (Macroeconomics): An example of the simple linear regression

The GDP growth is presumed to be in a linear relationship with the changes in the unemployment rate.



Introduction

Simple linear regression and correlation

- Correlation analysis (correlation coefficient) is for determining whether a relationship *exists.*
- Simple linear regression is for examining the relationship between two variables (if a linear relationship between them exists).
- Mathematical equations describing these relationships are models, and they fall into two types: *deterministic* or *probabilistic*.
  - ❑ Deterministic model: an equation or set of equations that allow us to fully determine the value of the dependent variable from the values of the independent variables.

  Contrast this with…

  - ❑ Probabilistic model: a method used to capture the randomness that is part of a real-life process.

Introduction
Simple linear regression

Example: Do all houses of the same size sell for exactly the same price?

- Models
  - Deterministic model: approximates the relationship we want to model and add a random term that measures the error of the deterministic component.

The cost of building a new house is about \$75 per square foot and most lots sell for about \$25,000. Hence the approximate selling price (Y) would be:

$$Y = \$25{,}000 + (75\$/\text{ft}^2)(X)$$

(where X is the size of the house in square feet)

Introduction

Background of model design

- *The facts*
  - ❑ Having too many input variables in the regression model ⇒ an overfitting regression function with an inflated variance
  - ❑ Having too few input variables in the regression model ⇒ an underfitting and high bias regression function with poor explanation of the data
- *The "importance" of a variable*

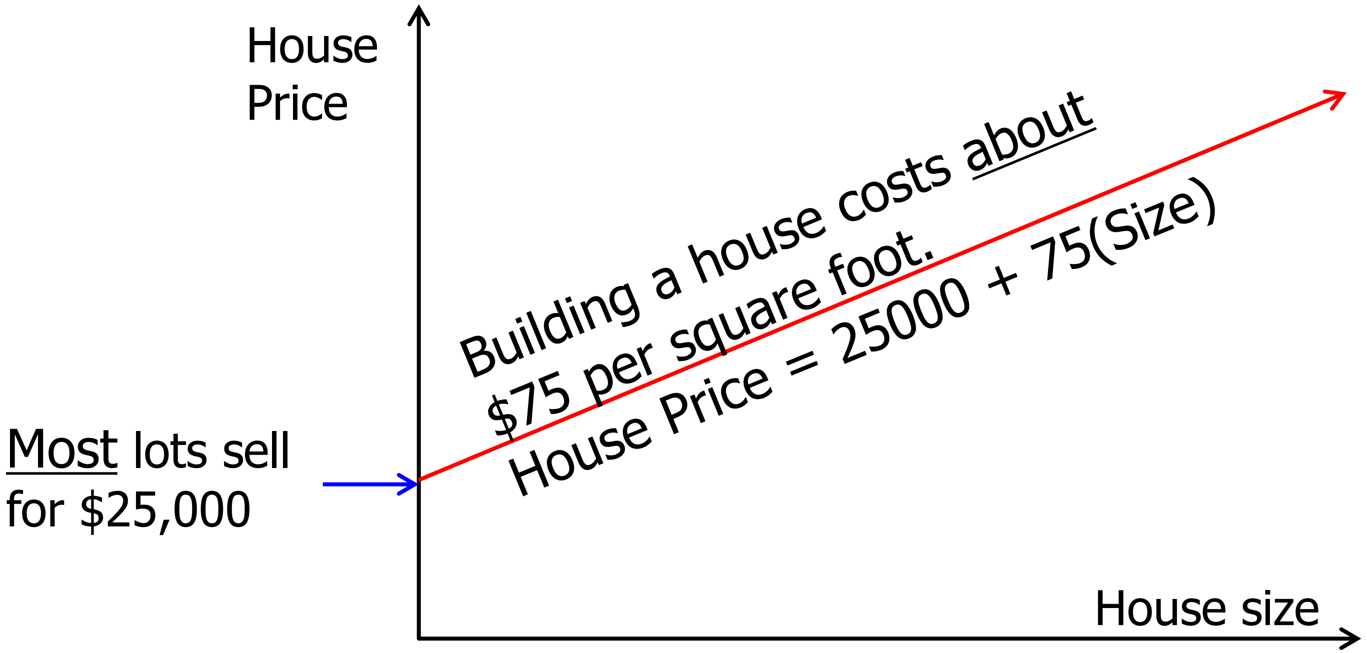
  Depends on how seriously it will affects prediction accuracy if it is dropped.
- *The behind driving force*

  The desire for a simpler and more easily interpretable regression model combined with a need for greater accuracy in prediction.

Introduction

Simple linear regression

A model of the relationship between house size (independent variable) and house price (dependent variable) would be:

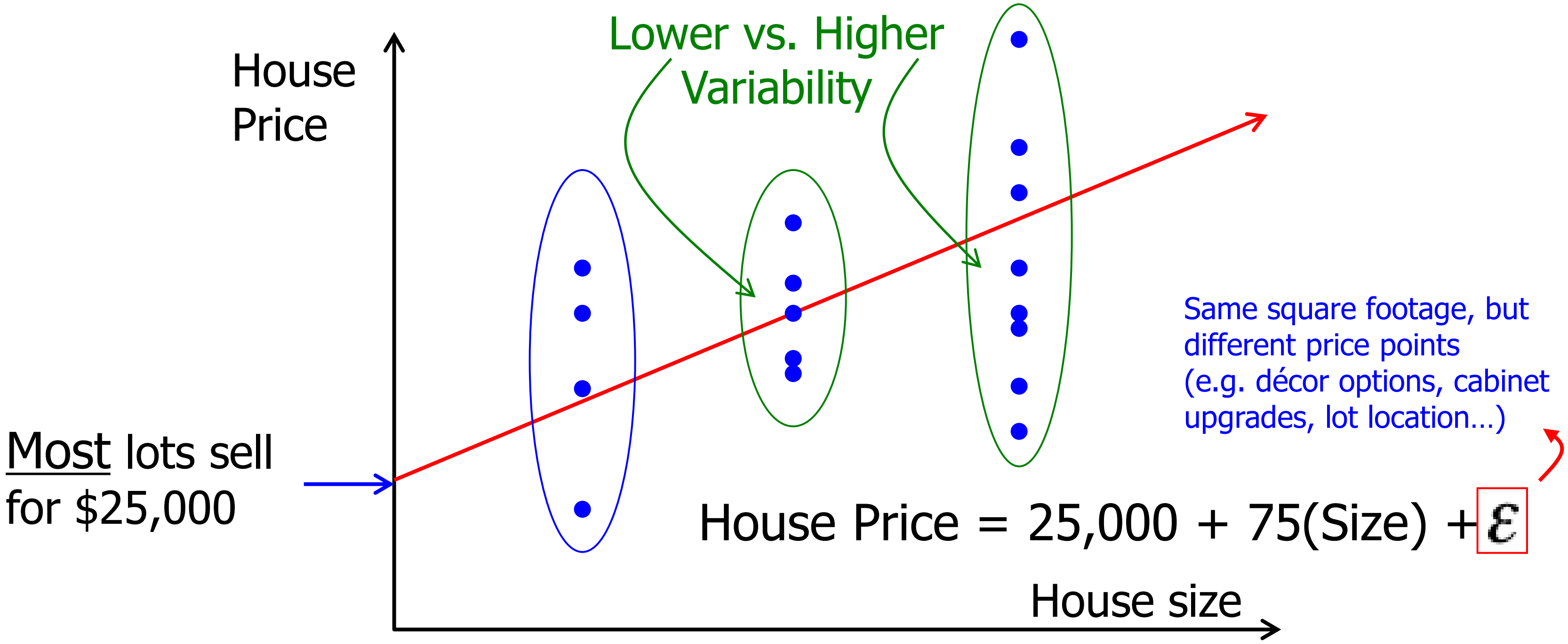


In this model, the price of the house is completely determined by the size

Introduction

Simple linear regression

A model of the relationship between house size (independent variable) and house price (dependent variable) would be:



Lower vs. Higher Variability

Same square footage, but different price points (e.g. décor options, cabinet upgrades, lot location…)

Most lots sell for $25,000

$$\text{House Price} = 25{,}000 + 75(\text{Size}) + \boxed{\varepsilon}$$

In this model, the price of the house is completely determined by the size

Introduction

Simple linear regression

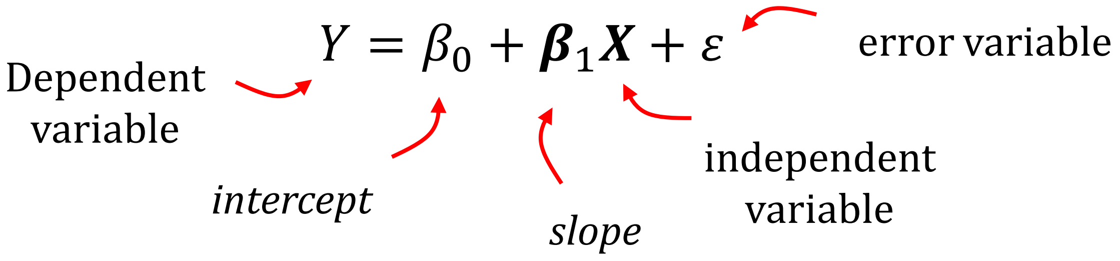
Example: Do all houses of the same size sell for exactly the same price?

- Probabilistic model:

$$Y = 25{,}000 + 75\,\mathbf{X} + \varepsilon$$

where ε is the random term (*error variable*). It is the difference between the actual selling price and the estimated price based on the size of the house. Its value will vary from house sale to house sale, even if the square footage ($\mathbf{X}$) remains the same.

- First order simple linear regression model:



$$Y = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{X} + \varepsilon$$

Introduction

Regression and model

- Techniques for modeling and analyzing *the relationship* between dependent variables and independent variables.

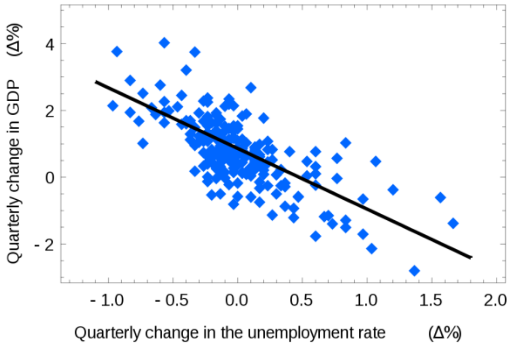
Input (independent, predictor, explanatory) Output (dependent, predicted, response)

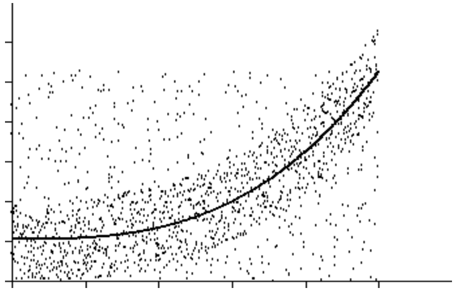
$$\boldsymbol{X}_i = \begin{pmatrix} X_1^i \\ \vdots \\ X_r^i \end{pmatrix} \xrightarrow[\textit{The relationship = Regression model}]{} \boldsymbol{Y}_i = \begin{pmatrix} Y_1^i \\ \vdots \\ Y_s^i \end{pmatrix}$$

- Different forms of

$$f : \mathbb{X} \longrightarrow \mathbb{Y}$$

  - ❑ linear vs. nonlinear,
  - ❑ parametric vs. nonparametric.

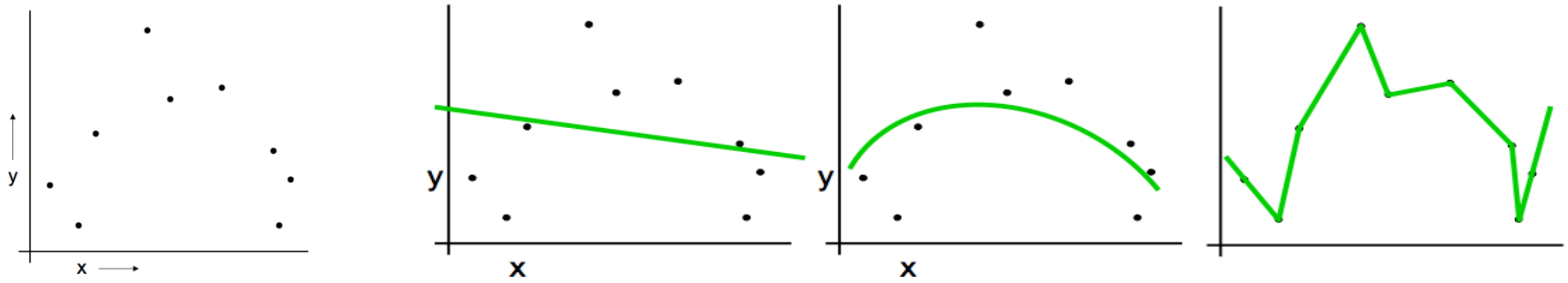




Linear combination of the parameters (but need not be linear in the independent variables)

Introduction

Model selection and model assessment



- Model Selection: Estimating performances of different models to choose the best one (produces the minimum of the *test error*).
- Model Assessment: Having chosen a model, estimating the *prediction error* on new data.

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression

Regression function and least squares

Consider the problem of predicting Y by a function, $f(\mathbf{X})$, of $\mathbf{X}$

- Loss function

$$L\big(Y, f(\mathbf{X})\big)$$

measures the prediction accuracy gives the loss incurred if Y is predicted by $f(\mathbf{X})$.

- Risk function

$$R(f) = E\big\{L\big(Y, f(\mathbf{X})\big)\big\}$$

(expected loss) measures the quality of $f(\mathbf{X})$ as a predictor.

- Bayes rule and the Bayes risk

The Bayes rule is the function f^* which minimizes $R(f)$, and the Bayes risk is $R(f^*)$.

Regression function and least squares

Regression function

- Squared-error loss function

$$L\big(Y, f(\mathbf{X})\big) = \big(Y - f(\mathbf{X})\big)^2$$

- Mean square error criterion

$$R(f) = E\left\{\big(Y - f(\mathbf{X})\big)^2\right\} = E_{\mathbf{X}}\left[E_{Y|\mathbf{X}}\left\{\big(Y - f(\mathbf{X})\big)^2 | \mathbf{X}\right\}\right]$$

- We can write $Y - f(\mathbf{x}) = (Y - \mu(\mathbf{x})) + (\mu(\mathbf{x}) - f(\mathbf{x}))$ where

$$\boxed{\mu(\mathbf{x}) = E_{Y|\mathbf{X}}\{Y|\mathbf{X} = \mathbf{x}\}}$$

is called regression function of Y on $\mathbf{X}$. Squaring both sides and taking the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$, as $E_{Y|\mathbf{X}}\{Y - \mu(x)\}|\mathbf{X} = \mathbf{x}\} = 0$, we have

$$E_{Y|\mathbf{X}}\left\{\big(Y - f(\mathbf{X})\big)^2 | \mathbf{X} = \mathbf{x}\right\} = E_{Y|\mathbf{X}}\left\{\big(Y - \mu(\mathbf{x})\big)^2 | \mathbf{X} = \mathbf{x}\right\} + \big(\mu(\mathbf{x}) - f(\mathbf{x})\big)^{\mathbf{2}}$$

Regression function and least squares

Least squares

- Taking $f^*(\mathbf{x}) = \mu(\mathbf{x}) = E_{Y|X}\{Y|\mathbf{X} = \mathbf{x}\}$, the previous equation is minimized

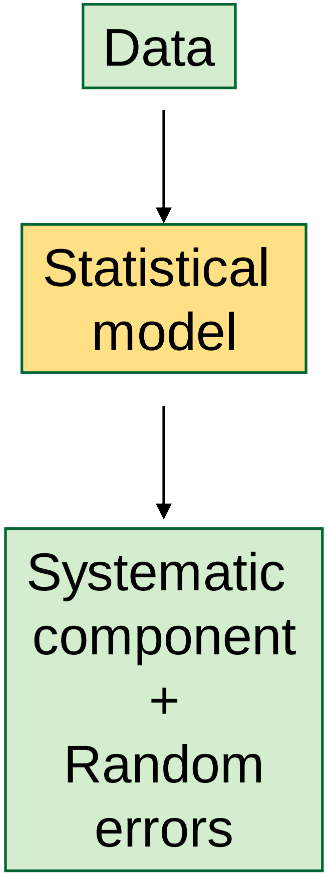
$$E_{Y|\mathbf{x}}\{(Y - f^*(\mathbf{x}))^2|\mathbf{X} = \mathbf{x}\}$$

$$= E_{Y|\mathbf{x}}\{(Y - \mu(\mathbf{x}))^2|\mathbf{X} = \mathbf{x}\}$$

- Taking expectations of both sides, we have Bayes risk

$$R(f^*) = \min_f R(f) = E\left\{\left(Y - \mu(\mathbf{X})\right)^2\right\}$$

The regression function $\mu(\mathbf{X})$ of Y on $\mathbf{X}$, evaluated at $\mathbf{X} = \mathbf{x}$, is the “best” predictor (defined by using minimum mean squared error).



Data

Statistical model

Systematic component + Random errors

Regression function and least squares

- Assumption

  The output variables $\boldsymbol{Y}$ are linearly related to the input variables $\boldsymbol{X}$

- The model

$$\mu(\mathbf{X}) = \beta_0 + \sum_{i=1}^{r} \beta_i\, \boldsymbol{X}_i \quad \Longrightarrow \quad Y = \beta_0 + \sum_{i=1}^{r} \beta_i\, \boldsymbol{X}_i + e$$

  is treated depending on assumptions on how $X_1, \dots, X_r$ were generated.

  - $\boldsymbol{X}_i$: the input (or independent, predictor) variables
  - Y: the output (or dependent, response) variable
  - e : (error) the unobservable random variable with mean 0, variance σ^2

- The tasks
  - To estimate the true values of $\beta_0, \beta_1, \dots, \beta_r$, and σ^2
  - To assess the impact of each input variable on the behavior of Y
  - To predict future values of Y
  - To measure the accuracy of the predictions

Regression function and least squares

Random-X case vs. Fixed-X case

Random-X case

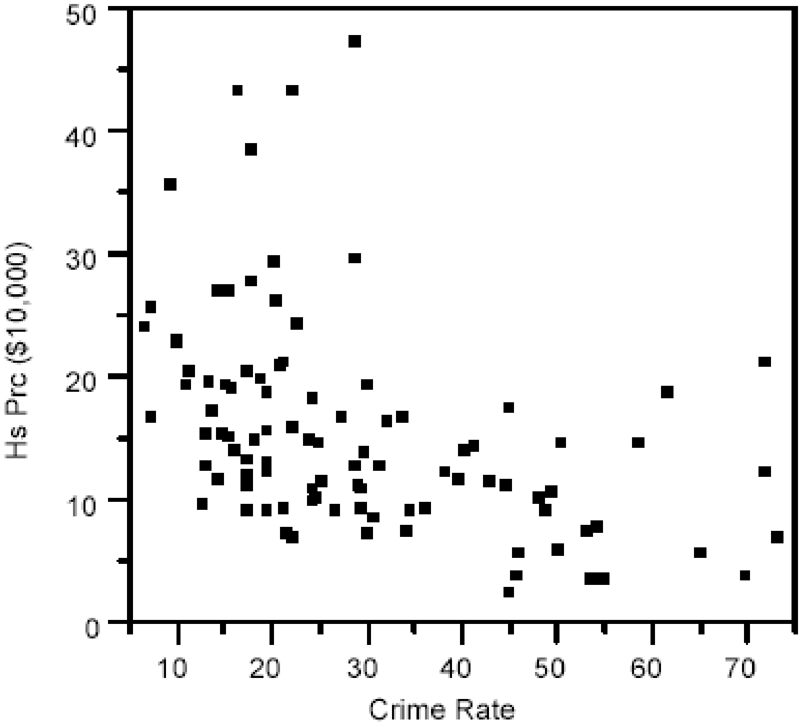
- **X** is a random variable, also known as the *structural model* or *structural relationship.*
- Methods for the structural model require some estimate of the variability of the variable **X.**
- The least squares fit will still give the best linear predictor of Y, but the estimates of the slope and intercept will be biased.
- $\mathbf{E(Y|X) = \beta_0 + \beta_1 X}$

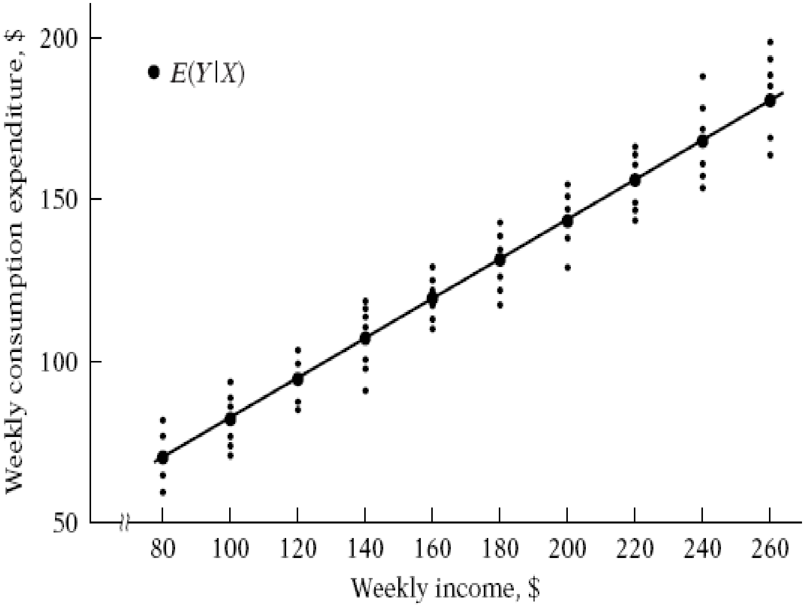
Fixed-X case (Fisher, 1922)

- **X** is fixed, but measured with noise, is known as the *functional model* or *functional relationship.*
- The fixed-X assumption is that the explanatory variable is measured without error.
- Distribution of the regression coefficient is unaffected by the distribution of **X.**
- $\mathbf{E(Y|X=x) = \beta_0 + \beta_1 x}$

Regression function and least squares

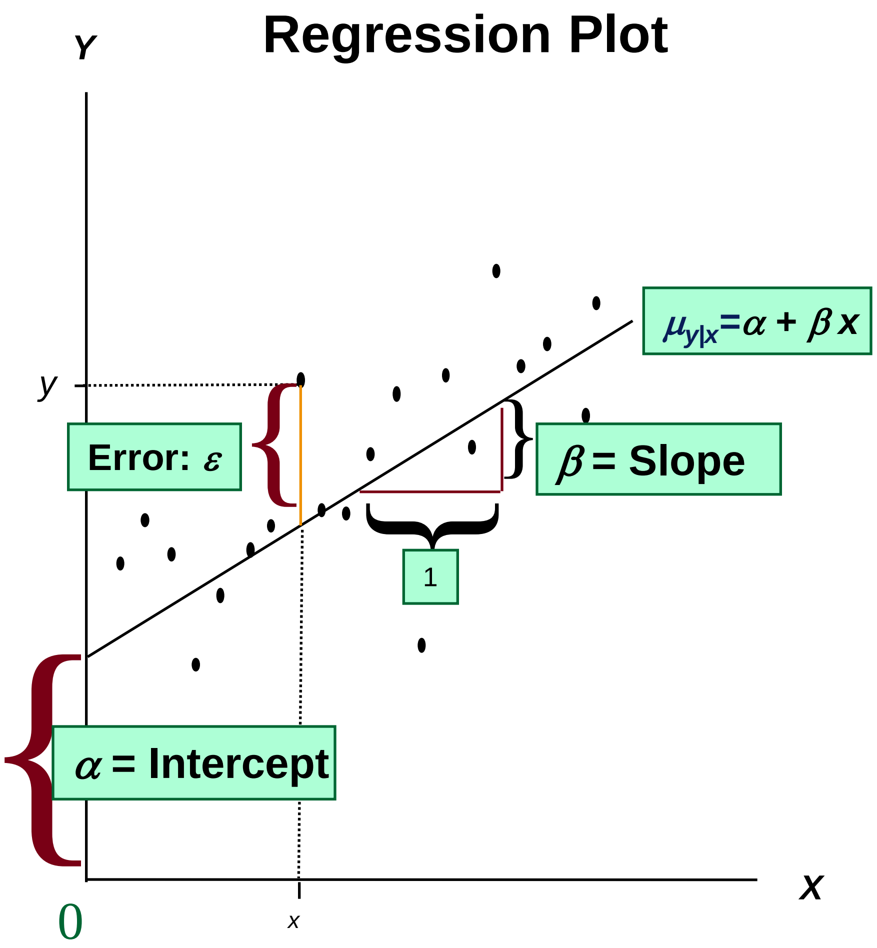
Random-X case vs. Fixed-X case



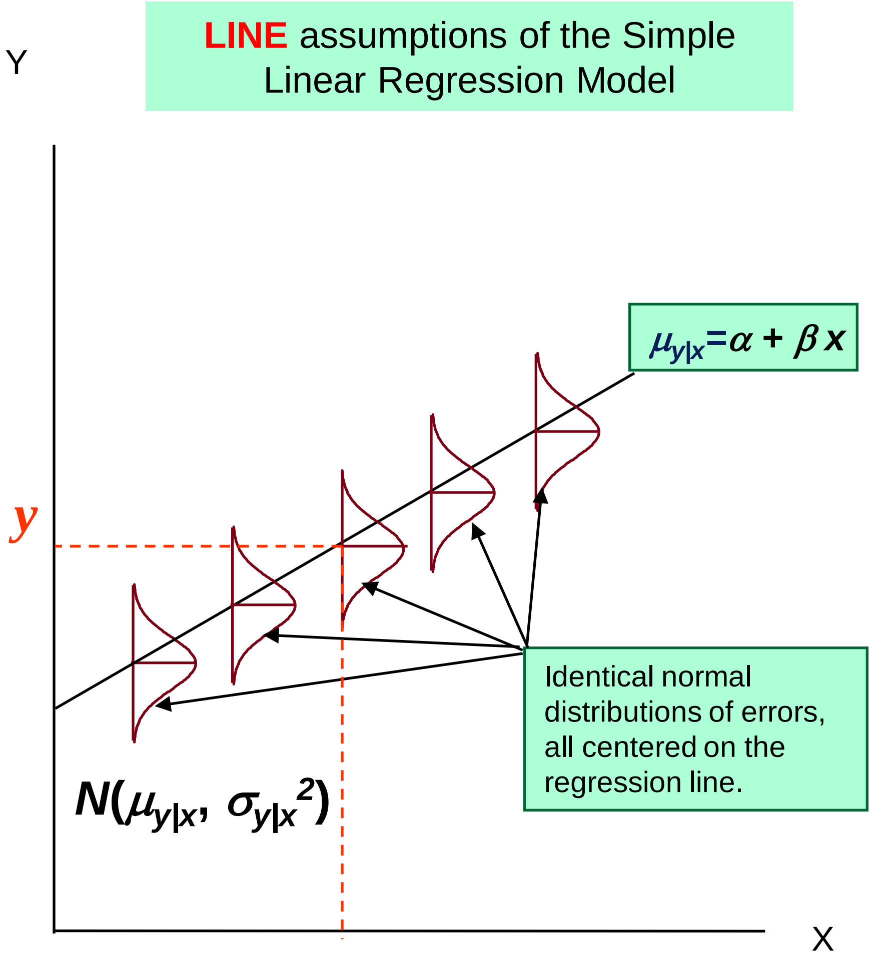


Regression function and least squares

Random-X case vs. Fixed-X case



Regression Plot



LINE assumptions of the Simple Linear Regression Model

Identical normal distributions of errors, all centered on the regression line.

Y
$\mu_{y|x} = \alpha + \beta x$
y
$N(\mu_{y|x}, \sigma_{y|x}^2)$
X

Regression function and least squares

Random-X case

- Regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_r \end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 \\ \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_r \end{pmatrix}$$

- Regression function

$$\mu(\mathbf{X}) = \mathbf{Z}^\tau \boldsymbol{\alpha} = \beta_0 + \boldsymbol{\beta}\mathbf{X} = \beta_0 + \sum_{j=1}^{r} \beta_j \boldsymbol{X}_j$$

- Let

$$S(\boldsymbol{\alpha}) = E\{(Y - \mathbf{Z}^\tau \boldsymbol{\alpha})^2\}$$

and define $\boldsymbol{\alpha}^* = \arg\min_{\alpha} S(\boldsymbol{\alpha})$

Regression function and least squares

Random-X case

- Setting differential of $S(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ to 0

$$\frac{\partial S(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = -2E(\mathbf{Z}Y - \mathbf{Z}\mathbf{Z}^{\tau}\boldsymbol{\alpha}) = 0$$

- We get $\boldsymbol{\alpha}^{*} = [\mathrm{E}(\mathbf{Z}\mathbf{Z}^{\tau})]^{-1}\mathrm{E}(\mathbf{Z}Y)$

$$\boldsymbol{\beta}^{*} = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}, \quad \beta_0^{*} = \mu_Y - \boldsymbol{\mu}_X^{\tau}\boldsymbol{\beta}^{*}$$

- In practice, $\boldsymbol{\mu}_X, \mu_Y, \boldsymbol{\Sigma}_{XX}$, and $\boldsymbol{\Sigma}_{XY}$ are unknown, we estimate them by ML using data generated by the joint distribution of $(\mathbf{X}, Y)$. Let

$$\mathcal{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^{\tau}, \mathcal{Y} = (Y_1, \cdots, Y_n)^{\tau}, \bar{\mathbf{X}} = \frac{1}{n}\sum_1^{\mathrm{n}} \mathbf{X}_j, \bar{Y} = \frac{1}{n}\sum_1^{n} Y_j,$$

$$\bar{\mathcal{X}} = (\bar{\mathbf{X}}, \dots, \bar{\mathbf{X}})^{\tau}, \bar{\mathcal{Y}} = (\bar{Y}, \dots, \bar{Y})^{\tau}, \mathcal{X}_c = \mathcal{X} - \bar{\mathcal{X}}, \mathcal{Y}_c = \mathcal{Y} - \bar{\mathcal{Y}},$$

$$\hat{\boldsymbol{\beta}}^{*} = (\mathcal{X}_c^{\tau}\mathcal{X}_c)^{-1}\mathcal{X}_c^{\tau}\mathcal{Y}_c, \qquad \hat{\beta}_0^{\tau} = \bar{Y} - \bar{\mathbf{X}}^{\tau}\hat{\boldsymbol{\beta}}^{*}$$

Covariance matrix: $\boldsymbol{\Sigma}_{XX} = \mathrm{cov}(\mathbf{X},\mathbf{X}) = \mathrm{E}\{(\mathbf{X} - \boldsymbol{\mu}_X)(\mathbf{X} - \boldsymbol{\mu}_X)^{\tau}\}$

Regression function and least squares

Fixed-X case

- $X_1, \dots, X_r$ are fixed in repeated sampling and Y may be selected in the designed experiment or Y may be observed conditional on the $X_1, \dots, X_r$

$$\mathcal{Z} = \begin{pmatrix} 1 & X_1^1 & \cdots & X_r^1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_1^n & \cdots & X_r^n \end{pmatrix}\text{- input variables,} \quad \mathcal{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_r \end{pmatrix}\text{- output variables}$$

- The regression function

$$Y_i = \beta_0 + \sum_{j=1}^{r} \beta_j X_{ij} + e_i\,, \qquad i = 1,2,\dots,n$$

$$\mathcal{Y} = \mathcal{Z}\boldsymbol{\beta} + \boldsymbol{e}$$

$\boldsymbol{e}$: random n-vector of unobservable errors with $E(\boldsymbol{e}) = 0$, $\mathrm{var}(\boldsymbol{e}) = \sigma^2 \mathbf{I}_{n.}$

- Error sum of squares

$$ESS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{e}_i^2 = \boldsymbol{e}^\tau \boldsymbol{e} = (\mathcal{Y} - \mathcal{Z}\boldsymbol{\beta})^\tau\,(\mathcal{Y} - \mathcal{Z}\boldsymbol{\beta})$$

Regression function and least squares

Fixed-X case

- Estimate $\boldsymbol{\beta}$ by minimizing $ESS(\boldsymbol{\beta})$ w.r.t. $\boldsymbol{\beta}$. Set differential w.r.t. $\boldsymbol{\beta}$ to 0

$$\frac{\partial ESS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathcal{Z}^{\tau}(\mathcal{Y} - \mathcal{Z}\boldsymbol{\beta}) = 0$$

The unique ordinary least-squares (OLS) estimator of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}_{ols} = (\mathcal{Z}^{\tau}\mathcal{Z})^{-1}\mathcal{Z}^{\tau}\mathcal{Y}$$

- Even though the descriptions differ as to how the input data are generated, the ordinary least-squares estimator estimates turn out to be the same for the random-X case and the fixed-X case:

$$\widehat{\boldsymbol{\beta}}^{*} = (\mathcal{X}_c^{\tau}\mathcal{X}_c)^{-1}\mathcal{X}_c^{\tau}\mathcal{Y}_c, \qquad \hat{\beta}_0^{\tau} = \bar{Y} - \bar{\mathbf{X}}^{\tau}\widehat{\boldsymbol{\beta}}^{*}$$

- The components of the n-vector of OLS *fitted values* are the vertical projections of the n points onto the LS regression surface (or hyperplane)

$$\hat{Y}_i = \hat{\mu}(\mathbf{X}_i) = \mathbf{X}_i^{\tau}\widehat{\boldsymbol{\beta}}_{ols}, i = 1, \dots, n.$$

Regression function and least squares

Fixed-X case

- The variance of $\hat{Y}_i$ for fixed $\mathbf{X}_i$ is given by

$$var(\hat{Y}_i|\mathbf{X}_i) = \mathbf{X}_i^{\tau}\{var(\hat{\beta}_{ols})\}\mathbf{X}_i = \sigma^2\mathbf{X}_i^{\tau}(\mathcal{Z}^{\tau}\mathcal{Z})^{-1}\mathbf{X}_i$$

- The *n*-vector of fitted values $\hat{\mathcal{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^{\tau}$ is

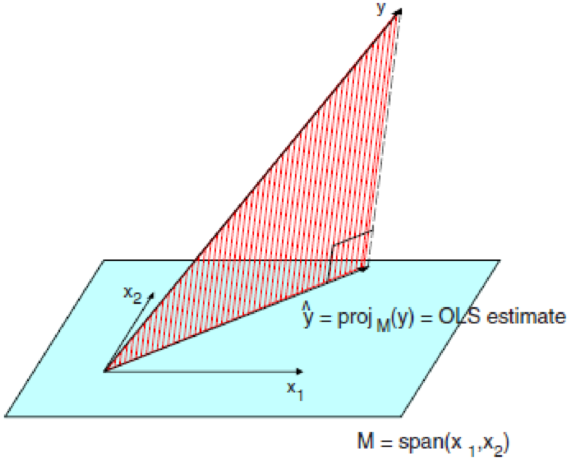
$$\hat{\mathcal{Y}} = \mathcal{Z}\widehat{\boldsymbol{\beta}}_{ols} = \mathcal{Z}(\mathcal{Z}^{\tau}\mathcal{Z})^{-1}\mathcal{Z}^{\tau}\mathcal{Y} = \mathbf{H}\mathcal{Y}$$

  where the $(n \times n)$-matrix $\mathbf{H} = \mathcal{Z}(\mathcal{Z}^{\tau}\mathcal{Z})^{-1}\mathcal{Z}^{\tau}$ is often called the *hat matrix*.

- The variance of Y is given by

$$var(\hat{\mathcal{Y}}|\mathbf{X}) = H\{var(\mathcal{Y})\}\mathbf{H}^{\tau} = \sigma^2\mathbf{H}$$

- The *residuals*, $\hat{\boldsymbol{e}} = \mathcal{Y} - \hat{\mathcal{Y}} = (\mathbf{I}_n - \mathbf{H})\mathcal{Y}$ are the OLS estimates of the unobservable errors **e**, and can be written as $\hat{\boldsymbol{e}} = (\mathbf{I}_n - \mathbf{H})\boldsymbol{e}$.



Regression function and least squares

ANOVA table for multiple regression model and F-statistic

Residual variance	$\hat{\sigma}^2 = \frac{RSS}{n-r-1}$
Total sum of squares	$S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (\mathcal{Y} - \bar{\mathcal{Y}})^\tau(\mathcal{Y} - \bar{\mathcal{Y}})$
Regression sum of of squares	$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_i)^2 = \hat{\beta}_{ols}^\tau(\mathcal{Z}^\tau\mathcal{Z})\hat{\beta}_{ols}$
Residual sum of squares	$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathcal{Y} - \mathcal{Z}\hat{\beta}_{ols})^\tau(\mathcal{Y} - \mathcal{Z}\hat{\beta}_{ols})$

- Use *F-statistic*, $F = \frac{SS_{reg}/r}{RSS/(n-r-1)}$, to see if there is a linear relationship between *Y* and the *X*s: F small → not reject $\beta = 0$, F large → $\exists j, \beta_j \neq 0$.

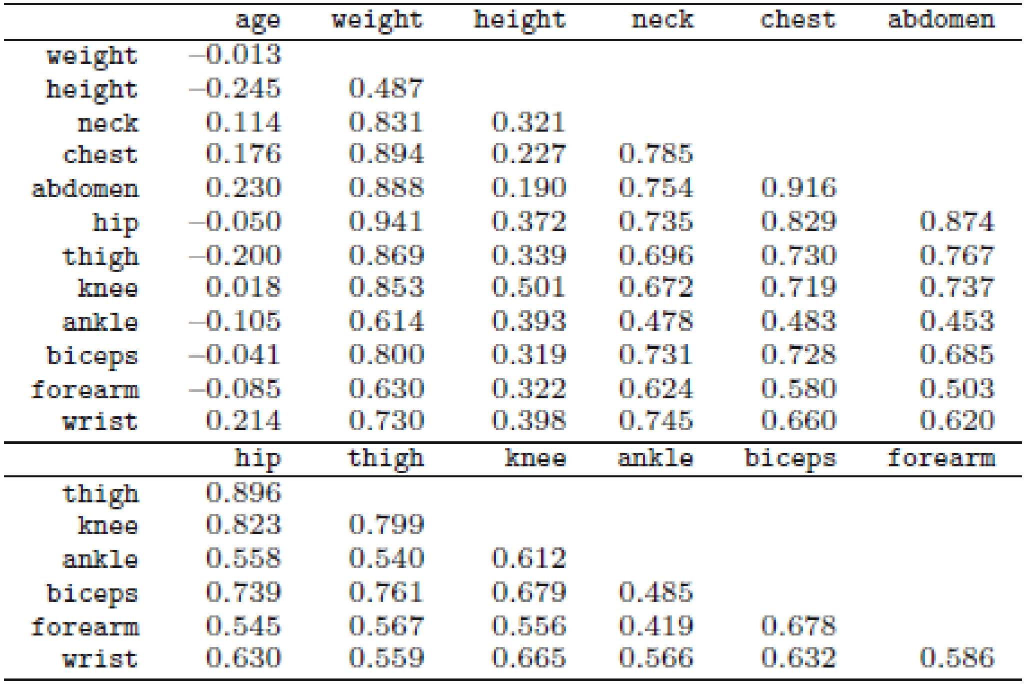
- If β_j= 0, use *t*-statistic, $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_{jj}}}$, where v_{jj} is the *j*th diagonal entry of $(\mathcal{Z}^\tau\mathcal{Z})^{-1}$. If $|t_j|$ large large → $\beta_j \neq 0$, else (near zero) → $\beta_j = 0$.

Regression function and least squares

Bodyfat data

- n = 252 men, to relate the percentage of bodyfat to age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, bicept, foream, wrist (13).

$$
\begin{aligned}
\text{bodyfat} = \beta_0 &+ \beta_1(\text{age}) \\
&+ \beta_2(\text{weight}) + \beta_3(\text{height}) \\
&+ \beta_4(\text{neck}) + \beta_5(\text{chest}) \\
&+ \beta_6(\text{abdomen}) + \beta_7(\text{hip}) \\
&+ \beta_8(\text{thigh}) + \beta_9(\text{knee}) \\
&+ \beta_{10}(\text{ankle}) \\
&+ \beta_{11}(\text{biceps}) \\
&+ \beta_{12}(\text{forearm}) \\
&+ \beta_{13}(\text{wrist}) + e
\end{aligned}
$$

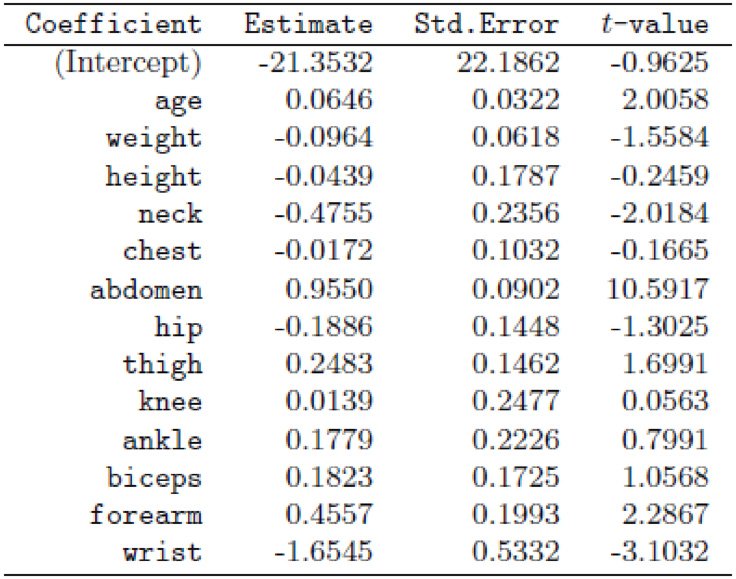


	age	weight	height	neck	chest	abdomen
weight	−0.013					
height	−0.245	0.487				
neck	0.114	0.831	0.321			
chest	0.176	0.894	0.227	0.785		
abdomen	0.230	0.888	0.190	0.754	0.916	
hip	−0.050	0.941	0.372	0.735	0.829	0.874
thigh	−0.200	0.869	0.339	0.696	0.730	0.767
knee	0.018	0.853	0.501	0.672	0.719	0.737
ankle	−0.105	0.614	0.393	0.478	0.483	0.453
biceps	−0.041	0.800	0.319	0.731	0.728	0.685
forearm	−0.085	0.630	0.322	0.624	0.580	0.503
wrist	0.214	0.730	0.398	0.745	0.660	0.620

	hip	thigh	knee	ankle	biceps	forearm
thigh	0.896					
knee	0.823	0.799				
ankle	0.558	0.540	0.612			
biceps	0.739	0.761	0.679	0.485		
forearm	0.545	0.567	0.556	0.419	0.678	
wrist	0.630	0.559	0.665	0.566	0.632	0.586

Regression function and least squares

Fixed-X case



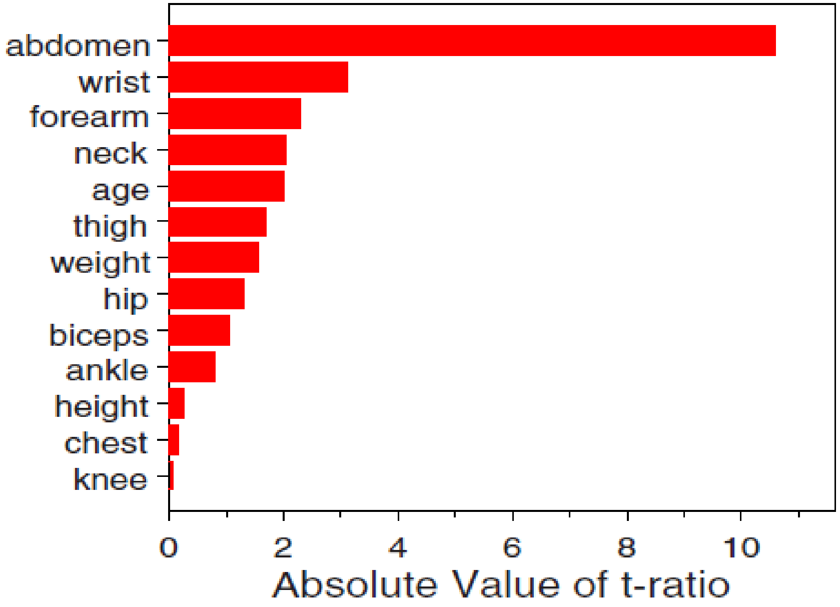
Coefficient	Estimate	Std.Error	*t*-value
(Intercept)	-21.3532	22.1862	-0.9625
age	0.0646	0.0322	2.0058
weight	-0.0964	0.0618	-1.5584
height	-0.0439	0.1787	-0.2459
neck	-0.4755	0.2356	-2.0184
chest	-0.0172	0.1032	-0.1665
abdomen	0.9550	0.0902	10.5917
hip	-0.1886	0.1448	-1.3025
thigh	0.2483	0.1462	1.6991
knee	0.0139	0.2477	0.0563
ankle	0.1779	0.2226	0.7991
biceps	0.1823	0.1725	1.0568
forearm	0.4557	0.1993	2.2867
wrist	-1.6545	0.5332	-3.1032

OLS estimation of coefficients :

- multiple R2 is 0.749
- residual sum of squares is 4420.1
- F-statistic is 54.5 on 13 and 238 degrees of freedom.

A multiple regression using variables having |t| > 2

- residual sum of squares 4724.9,
- R2 = 0.731,
- F-statistic of 133.85 on 5 and 246 degrees of freedom.



Multiple regression results for the bodyfat data.

The variable names are given on the vertical axis

(listed in descending order of their absolute t-ratios)

and the absolute value of the t-ratio for each variable

on the horizontal axis.

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression

Prediction accuracy and model assessment

- **The aims**
  - Prediction is the art of making accurate guesses about new response values that are independent of the current data.
  - Good predictive ability is often recognized as the most useful way of assessing the fit of a model to data.

- **Practice**
  - Learning data $\mathcal{L} = \{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$ for regression of Y on $\boldsymbol{X}$.
  - Prediction of a new Y^{new} by applying the fitted model to a brand-new $\boldsymbol{X}^{new}$, from the test set T.
  - Predicted Y^{new} is compared with the actual response value. The predictive ability of the regression model is assessed by its *prediction error* (or *generalization error*), an overall measure of the quality of the prediction, usually taken to be mean squared error.

Prediction accuracy and model assessment

Random-X case

- Learning data set $L = \{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$ are observations from the joint distribution of $(\mathbf{X}, Y)$ and

$$Y = \beta_0 + \sum_{j=1}^{r} \beta_j X_j + e = \mu(\mathbf{X}) + e$$

where $\mu(\mathbf{X}) = E(Y|\mathbf{X}), E(e|\mathbf{X}) = 0, var(e|\mathbf{X}) = \sigma^2$

- Given the test set $T = \{(\mathbf{X}^{new}, Y^{new})\}$, if the estimated OLS regression function at $\mathbf{X}$ is $\hat{\mu}(\mathbf{X}) = \hat{\beta}_0 + \mathbf{X}^\tau \widehat{\boldsymbol{\beta}}_{ols}$ then the predicted value of Y at $\mathbf{X}^{\text{new}}$ is $\hat{Y} = \hat{\mu}(\mathbf{X}^{new})$.

Prediction accuracy and model assessment

Random-X case

- Prediction Error

$$PE_R = E\{Y^{new} - \hat{\mu}(\mathbf{X}^{new})\}^2 = \sigma^2 + ME_R$$

- Model Error (also called the “expected bias-squared”)

$$ME_R = E\{\mu(\mathbf{X}^{new}) - \hat{\mu}(\mathbf{X}^{new})\}^2 = \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{ols}\right)^{\tau} \boldsymbol{\Sigma}_{XX}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{ols}\right)$$

Prediction accuracy and model assessment

Fixed-X case

- In $L = \{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$, $\{\boldsymbol{X}_i\}$ are fixed and only Y is random. Assume that

$$Y_i = \beta_0 + \sum_{j=1}^{r} \beta_j X_j^i + e_i = \mu(\mathbf{X}_i) + e_i$$

where $\mu(\mathbf{X}_i) = \beta_0 + \sum_{j=1}^{r} \beta_j X_j^i$ is the regression function evaluated at $\mathbf{X}_i$, and the errors e_i are iid with mean 0 and variance σ^2 and uncorrelated with $\{\boldsymbol{X}_i\}$.

- Assume the test data set generated by "future-fixed" $\{\mathbf{X}^{new}\}$ and $T = \{(\mathbf{X}_i, Y_i^{new}), i = 1, \ldots, m\}$, where $Y_i^{new} = \mu(X_i) + e_i^{new}$. The *predicted value* of Y^{new} at $\mathbf{X}$ is $\hat{\mu}(\mathbf{X}) = \hat{\beta}_0 + \mathbf{X}^{\tau}\widehat{\boldsymbol{\beta}}_{ols}$.

Prediction accuracy and model assessment

Fixed-X case

- Prediction Error

$$PE_F = E\left(\left\{\frac{1}{m}\sum_{i=1}^{m}\left(Y_i^{new} - \hat{\mu}(\mathbf{X}_i)\right)^2\right\}\right) = \sigma^2 + ME_F$$

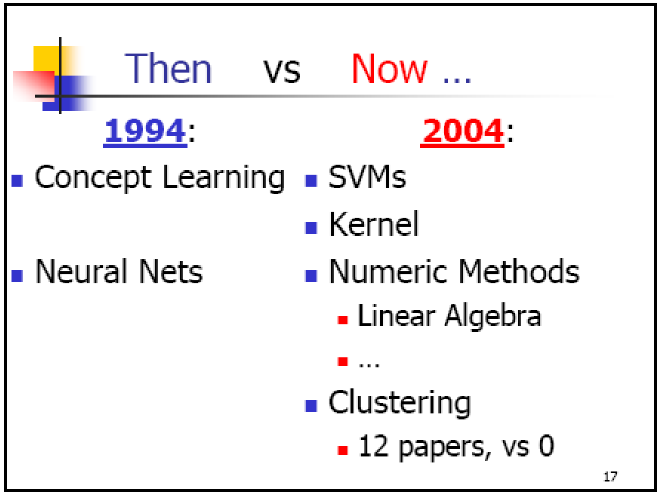
- Model Error

$$ME_F = \frac{1}{m}\sum_{i=1}^{m}\left(Y_i^{new} - \hat{\mu}(\mathbf{X}_i)\right)^2 = \left(\beta - \hat{\beta}_{ols}\right)^\tau \left(\frac{1}{m}\chi^\tau\chi\right)\Sigma_{XX}\left(\beta - \hat{\beta}_{OLS}\right)$$

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression

Observed from ICML 2004



Then vs Now ...

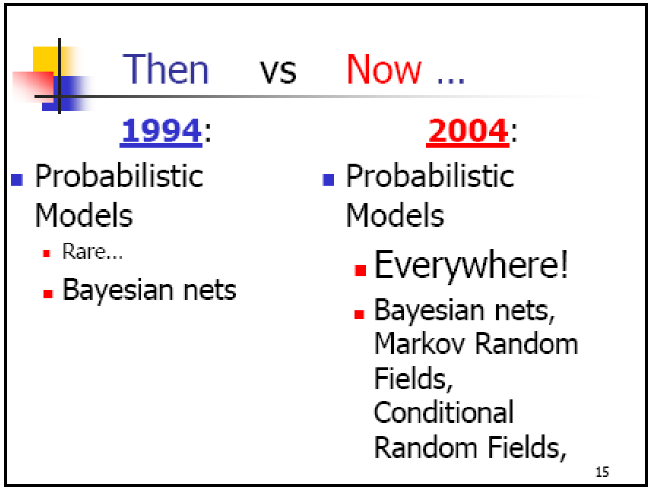
1994:

- Concept Learning
- Neural Nets

2004:

- SVMs
- Kernel
- Numeric Methods
  - Linear Algebra
  - ...
- Clustering
  - 12 papers, vs 0

17



Then vs Now ...

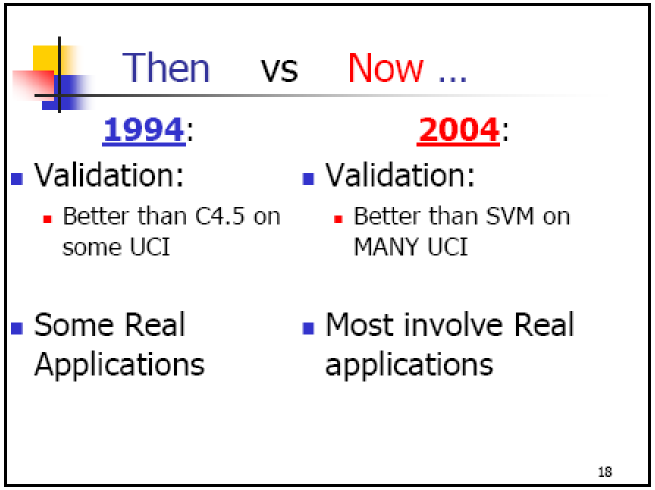
1994:

- Probabilistic Models
  - Rare...
  - Bayesian nets

2004:

- Probabilistic Models
  - Everywhere!
  - Bayesian nets, Markov Random Fields, Conditional Random Fields,

15



Then vs Now ...

1994:

- Validation:
  - Better than C4.5 on some UCI
- Some Real Applications

2004:

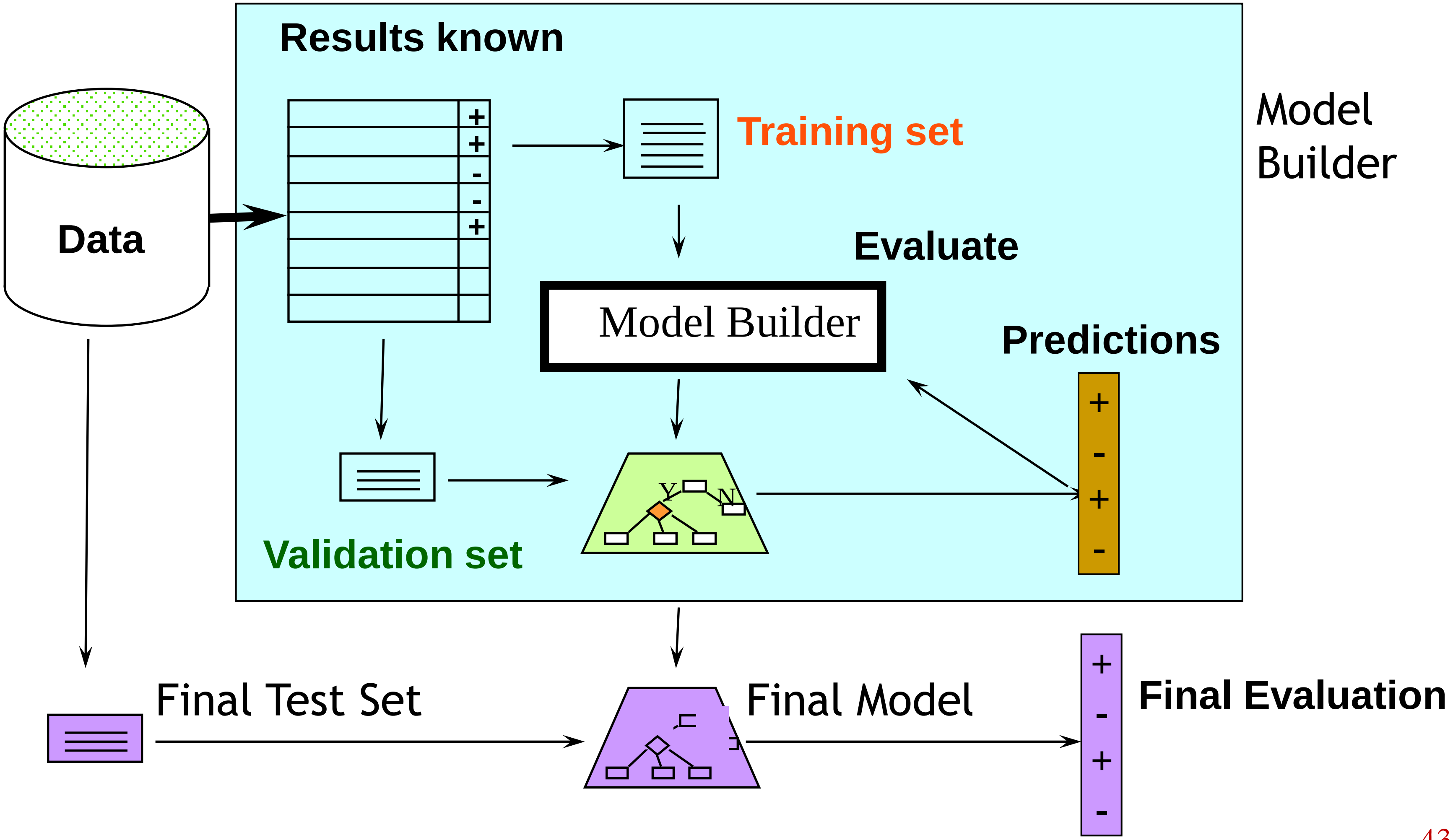
- Validation:
  - Better than SVM on MANY UCI
- Most involve Real applications

18

Russ Greiner, ICML'04 PC co-chair

Estimating prediction error

Training, validation, and testing data



Results known

Data

Training set

Model Builder

Evaluate

Model Builder

Predictions

Validation set

Final Test Set

Final Model

Final Evaluation

Estimating prediction error

Background

In cases the entire data set is not large enough and a division of the data into learning, validation, and test sets is not practical, we have to use alternative methods.

Apparent Error Rate

Applying the regression function obtained from OLS to the original sample data to see how well it predicts those same members

Cross-Validation

Split data set into two subsets, treating one subset as the learning set, and the other as the test set. Fit a model using this learning set and compute its prediction error. The learning set and the test set are then switched, and average all the prediction errors to estimate the test error.

Estimating prediction error

Background

Bootstrap

Drawing a random sample with replacement having the same size as the parent data set. Fit a model using this bootstrap sample and compute its prediction error. Repeat the procedure, and average all the prediction errors to estimate the test error.

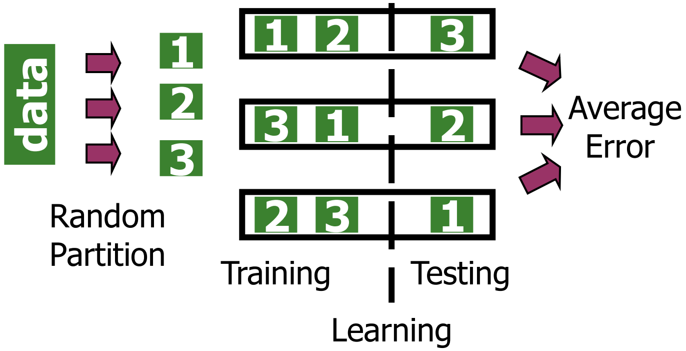
- ❑ Random-X case: Cross-validation and "unconditional bootstrap" are appropriate;
- ❑ Fixed-X case: "Conditional bootstrap" are appropriate but cross-validation is not appropriate for estimating prediction error.

Estimating prediction error

- Apparent error rate
  (*resubstitution error rate*)

$$\widehat{PE}(\hat{\mu}, D) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{new} - \hat{\mu}(\mathbf{X}_i)\right)^2 = \frac{RSS}{n}$$

Misleadingly optimistic, RSS/n will be PE too optimistic estimation with $\widehat{PE}(\hat{\mu}, D) < PE$



- Cross-Validation (*V-fold*)

$$D \Longrightarrow \{T_1, \dots, T_V\},\; D = \cup_{v=1}^{V} T_v\,,\; T_v \cap T_{v'} = \emptyset$$

$$L_v = D - T_v, \widehat{PE}_{CV/V} = \frac{1}{V}\sum_{v=1}^{V}\sum_{(\mathbf{X}_i Y_i)\in T_v}(Y_i - \hat{\mu}_{-v}(\mathbf{X}_i))^2$$

subtract σ^2 (obtain from the full data set) from $\widehat{PE}$ to get $\widehat{ME}$

- ❑ *Leave-one-out* rule: $V = n$, the most computationally intensive, but usually worse at model assessment than 10-fold (even 5-fold) CV.

Estimating prediction error

- Bootstrap (Efron, 1979)
  - **Unconditional Bootstrap**

    Random-X bootstrap sample (with replacement)

$$D_R^{*b} = \left\{\left(\mathbf{X}_i^{*b}, Y_i^{*b}\right), i = 1, \dots, n\right\}$$

$$\widehat{PE}\left(\hat{\mu}_R^{*b}, D\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_R^{*b}(\mathbf{X}_i)\right)^2$$

Simple bootstrap estimator of PE

$$\widehat{PE}_R(D) = \frac{1}{B}\sum_{b=1}^{B}\widehat{PE}\left(\hat{\mu}_R^{*b}, D\right) = \frac{1}{Bn}\sum_{b=1}^{B}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_R^{*b}(\mathbf{X}_i)\right)^2$$

Simple bootstrap estimator of PE using apparent error rate for D_R^{*b}

$$\widehat{PE}\left(D_R^{*b}\right) = \frac{1}{B}\sum_{b=1}^{B}\widehat{PE}\left(\hat{\mu}_R^{*b}, D_R^{*b}\right) = \frac{1}{Bn}\sum_{b=1}^{B}\sum_{i=1}^{n}\left(Y_i^{*b} - \hat{\mu}_R^{*b}\left(\mathbf{X}_i^{*b}\right)\right)^2$$

Estimating prediction error

- Bootstrap
  - ❑ **Unconditional Bootstrap**

    Simple estimators of PE are overly optimistic because there are observations common to the bootstrap samples $\{D_R^{*b}\}$ that determined $\{\hat{\mu}_R^{*b}\}$

    The *optimism* (improvement of PE by estimating the bias for D_R^{*b} using RSS/n as an estimate of PE and then correcting RSS/n by subtracting its estimated bias)

$$\widehat{opt}_R^b = \widehat{PE}\left(\hat{\mu}_R^{*b}, D\right) - \widehat{PE}\left(\hat{\mu}_R^{*b}, D_R^{*b}\right)$$

$$\widehat{opt}_R = \frac{1}{B}\sum_{b=1}^{B} \widehat{opt}_R^b = \widehat{PE}_R(D) - \widehat{PE}\left(D_R^{*b}\right)$$

$$\widehat{PE}_R = \frac{RSS}{n} + \widehat{opt}_R$$

Estimating prediction error

- Bootstrap
  - **Unconditional Bootstrap**

    The *optimism* (improvement of PE by estimating the bias for D_R^{*b} using RSS/n as an estimate of PE and then correcting RSS/n by subtracting its estimated bias)

    $$\widehat{PE}_R = \frac{RSS}{n} + \widehat{opt}_R$$

    - Computationally more expensive than cross-validation
    - Low bias, slightly better for model assessment than 10-fold cross-validation

    About 37% of the observations in $\mathcal{D}$ are left out of bootstrap sample

    $$\text{Prob}((X_i, Y_i) \in \mathcal{D}_R^{*b} = 1 - \left(1 - \frac{1}{n}\right)^n \longrightarrow 1 - e^{-1} \approx 0.632 \text{ as } n \rightarrow \infty$$

Estimating prediction error

- Bootstrap
  - **Conditional Bootstrap**

    Coefficients determination

    Estimate $\boldsymbol{\alpha}$ by minimizing $ESS(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$.

$$\widehat{\boldsymbol{\alpha}}_{OLS} = (\boldsymbol{Z}^{\tau}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\tau}\boldsymbol{Y}$$

Suppose $\widehat{\boldsymbol{\alpha}}_{OLS}$ to be the true value of the regression parameter, for the b th bootstrap sample, we sample with replacement from the residuals to get the bootstrapped residuals, $\hat{e}_i^{*b}$, and then compute the new set of responses

$$D_F^{*b} = \left\{\left(\mathbf{X}_i, Y_i^{*b} = \hat{\mu}(\boldsymbol{X}_i) + \hat{e}_i^{*b}\right), i = 1, 2, \dots n\right\}$$
$$\widehat{\boldsymbol{\alpha}}^{*b} = (\boldsymbol{Z}^{\tau}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\tau}\boldsymbol{Y}^{*b}$$

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression

Instability of least square estimates

If $\mathcal{X}_c^{\tau}\mathcal{X}_c$ is *singular* (as $\mathcal{X}_c$ has not less than full rank caused by columns of $\boldsymbol{Z}$ are collinear, or when $r > n$ or the data is ill-conditioned) then the OLS estimate of $\boldsymbol{\alpha}$ will not be unique

Ill-conditioned data:

- When the quantities to be computed are sensitive to small changes in the data, the computational results are likely to be numerically unstable.
- Too many highly correlated variables (*near collinearity*)
- The standard error of the estimated regression coefficients may be dramatically inflated (thổi phồng, khoa trương).
- The most popular measure of the ill-conditioning is the *condition number.*

Biased regression method

- As OLS estimates depend on $(\mathcal{Z}^\tau\mathcal{Z})^{-1}$ we would experience numerical complications in computing $\hat{\boldsymbol{\beta}}_{ols}$ if $\mathcal{Z}^\tau\mathcal{Z}$ were singular or nearly singular.
- If $\mathcal{Z}$ is ill-conditioned, small changes to $\mathcal{Z}$ lead to large changes in $(\mathcal{Z}^\tau\mathcal{Z})^{-1}$, and $\hat{\boldsymbol{\beta}}_{ols}$ becomes computationally unstable.
- One way: to abandon the requirement of an unbiased estimator of $\boldsymbol{\beta}$ and, instead, consider the possibility of using a *biased estimator* of $\boldsymbol{\beta}$.
  - ❑ Principal Components Regression
    Use the scores of the first t principal component of $\boldsymbol{Z}$.
  - ❑ Partial Least-Square Regression
    Construct *latent* variables from $\boldsymbol{Z}$ to retain most of the information that helps predict Y (reducing the dimensionality of the regression.)
  - ❑ Ridge Regression (ridge: chóp, dải đất hẹp dài trên đỉnh, luống, ...)
    Add a small constant k to the diagonal entries of the matrix before taking its inverse

$$\hat{\beta}_{rr}(k) = (\mathcal{X}^\tau\mathcal{X} + k\mathbf{I}_r)^{-1}\mathcal{X}^\tau\mathcal{Y}$$

Variable selection

- Motivation
  - Having too many input variables in the regression model
    ⇒ an overfitting regression function with an inflated variance
  - Having too few input variables in the regression model
    ⇒ an underfitting and high bias regression function with poor explanation of the data
- The “importance” of a variable

  Depends on how seriously it will affects prediction accuracy if it is dropped
- The behind driving force

  The desire for a simpler and more easily interpretable regression model combined with a need for greater accuracy in prediction.

Regularized regression

- A hybrid of these two ideas of Ridge Regression and Variable Selection.
- General penalized least-squares criterion

$$\phi(\boldsymbol{\beta}) = (\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})^{\tau}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}) + \lambda p(\boldsymbol{\beta})$$

for a given penalty function $p(\cdot)$ and *regularization parameter* λ.

- Define a family (indexed by $q > 0$) of penalized least-squares estimators in which the penalty function,

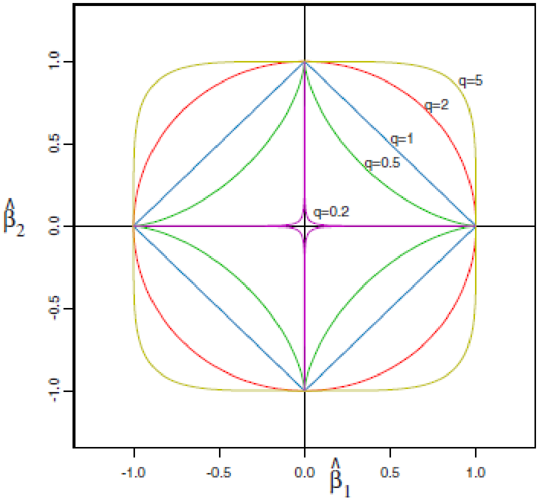
$$p_q(\boldsymbol{\beta}) = \sum_{j=1}^{r} |\beta_j|^q \sum_j |\beta_j|^q \leq c$$

bounds the L_q-norm (Frank and Friedman, 1993)

$$\sum_j |\beta_j|^q \leq c$$

Regularized regression

- q =2: *ridge regression*. The penalty function is rotationally invariant hypersphere centered at the origin, circular disk (r = 2) or sphere (r = 3).
- $q \neq 2$, the penalty is no longer invariant.
  - $q < 2$ (most interesting): penalty function collapses toward the coordinate axes →ridge regression and variable selection.
  - $q \approx 0$ penalty function places all its mass along the coordinate axes, and the contours of the elliptical region of $ESS(\boldsymbol{\beta})$ touch an undetermined number of axes, the result is variable selection.
  - q = 1 produces the lasso method having a diamond-shaped penalty function with the corners of the diamond on the coordinate axes.



Two-dimensional contours of the symmetric penalty function $p_q(\boldsymbol{\beta}) = |\beta_1|^q + |\beta_2|^q = 1$ for q = 0.2, 0.5, 1, 2, 5. The case q = 1 (blue diamond) yields the lasso and q = 2 (red circle) yields ridge regression.

Regularized regression

The Lasso

- The *Lasso* (least absolute shrinkage and selection operator) is a constrained OLS minimization problem in which

$$\phi(\boldsymbol{\beta}) = (\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})^{\tau}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}) + \lambda p(\boldsymbol{\beta})$$

is minimized for $\boldsymbol{\beta} = (\beta_j)$ subject to the diamond-shaped condition that $\sum_{j=1}^{r}|\beta_j| \leq c$ (Tibshirani, 1996). The regularization form of the problem is to find $\boldsymbol{\beta}$ to minimize

$$\phi(\boldsymbol{\beta}) = (\mathcal{Y} - \mathcal{X}\boldsymbol{\beta})^{\tau}(\mathcal{Y} - \mathcal{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{r} |\beta_j|$$

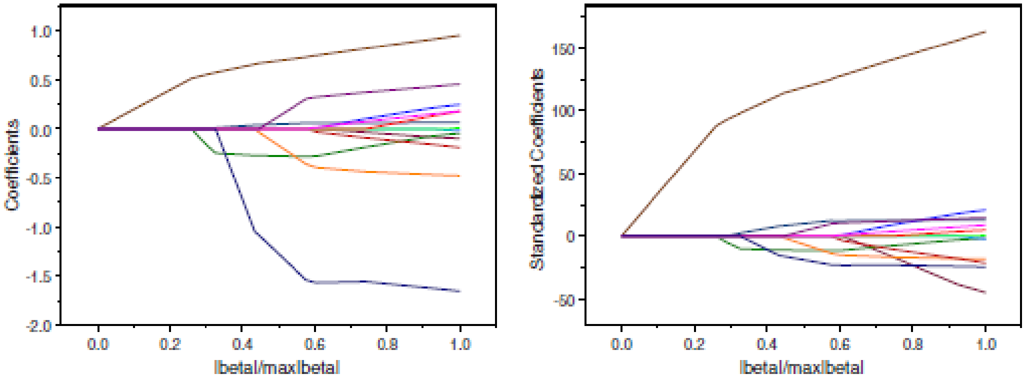
- This problem can be solved using complicated quadratic programming methods subject to linear inequality constraints.
- The Lasso has a number of desirable features that have made it a popular regression algorithm.

Lasso: toán tử chọn và co tuyệt đối tối thiểu

Regularized regression

The Lasso

- Like ridge regression, the Lasso is a shrinkage estimator of $\boldsymbol{\beta}$, where the OLS regression coefficients are shrunk toward the origin, the value of c controlling the amount of shrinkage.
- It behaves as a variable selection technique: for a given value of c, only a subset of the coefficient estimates, $\hat{\beta}_j$, will have nonzero values, and reducing the value of c reduces the size of that subset.



Lasso paths for the bodyfat data. The paths are plots of the coefficients $\{\hat{\beta}_j\}$ (left panel) and the standardized coefficients, $\{\hat{\beta}_j \parallel \mathcal{X}_j \parallel^2\}$ (right panel) plotted against. The variables are added to the regression model in the order: 6, 3, 1, 13, 4, 12, 7, 11, 8, 2, 10, 5, 9.

Regularized regression

The Garotte

- A different type of penalized least-squares estimator (Breiman, 1995).
- Let $\widehat{\boldsymbol{\beta}}_{ols}$ be the OLS estimator and let **W**= diag{**w**} be a diagonal matrix with nonnegative weights **w** = (w_j) along the diagonal. The problem is to find the weights **w** that minimize

$$\phi(\boldsymbol{w}) = (\mathcal{Y} - \mathcal{X}\mathbf{W}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}})^{\tau}(\mathcal{Y} - \mathcal{X}\mathbf{W}\widehat{\boldsymbol{\beta}}_{\mathrm{ols}})$$

subject to one of the following two constraints,

1. $\mathbf{w} \geq \mathbf{0}, \mathbf{1}_r^{\tau}\mathbf{w} = \sum_{j=1}^{r} \mathrm{w}_j \leq c$ (nonnegative garotte, thắt cổ)
2. $\mathbf{w}^{\tau}\mathbf{w} = \sum_{j=1}^{r} \mathrm{w}_i^2 \leq c$ (garotte)

- As c is decreased, more of the w_j become 0 (thus eliminating those particular variables from the regression function), while the nonzero $\hat{\beta}_{ols,j}$ shrink toward 0.

Outline

1. Introduction
2. The Regression Function and Least Squares
3. Prediction Accuracy and Model Assessment
4. Estimating Predictor Error
5. Other Issues
6. Multivariate Regression

Multivariate regression

- Multivariate regression has s output variables $\boldsymbol{Y} = (Y_1,\cdot\cdot\cdot,Y_s)^\tau$, each of whose behavior may be influenced by exactly the same set of inputs $\boldsymbol{X} = (X_1,\cdot\cdot\cdot,X_r)^\tau$.

- Not only are the components of **X** correlated with each other, but in multivariate regression, the components of **Y** are also correlated with each other (and with the components of **X**).

- Interested in estimating the regression relationship between **Y** and **X**, taking into account the various dependencies between the r-vector **X** and the s-vector **Y** and the dependencies within **X** and within **Y**.