Linear Models: An Integrated Approach by Debasis Sengupta and Sreenivasa Rao Jammalamadaka

Chapter 1: Introduction

- People want to make models to explain phenomena. But relevant factors may be left out, and measurements to validate come with some error. Statistical models are used here because they incorporate uncertainty.
- Linear model
  - How can an observed quality y be explained by a number of other quantities x1, …, xp?
  - Explain the response in terms of the explanatory variables
  - Betas are called parameters and xs are called regressors, predictors…
  - Y is called the response variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \qquad (1.1.1)$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are constants and ϵ is an error term that accounts for uncertainties. We shall refer to y as the *response variable*. It is also

  - Error term accounts for unaccounted factors, missing considerations, etc.
- Why Linear models?
  - Simple and easy to interpret - often used as the first step to other models
  - Useful even for certain non-linear models
  - Probability models with multivariate normal distributions
- Description of linear models and notations
  - If you have n observations, p predictors, and x_ij is the ith observation of the jth predictor variable and E_i is the unobservable error

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \qquad (1.3.2)$$

In this model,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

    - It's kinda like a linear transformation XB = y where X is and n x p matrix, B is a p x 1 matrix, and y is an n x 1 matrix
  - Errors are assumed to have zero mean with variances known to a scale factor

$$E(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad D(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{V}, \qquad (1.3.3)$$

where the notation E stands for expected value and D represents the dispersion (or variance-covariance) matrix. The vector $\boldsymbol{0}$ denotes a vector with zero elements (in this case n elements) and $\boldsymbol{V}$ is a known matrix of order $n \times n$. The parameter σ^2 is unspecified, along with the vector parameter $\boldsymbol{\beta}$. The elements of $\boldsymbol{\beta}$ are real-valued, while σ^2 is nonnegative.

- (y, XB, sigma^2V) is a shorthand triplet for a linear model (errors are uncorrelated and each has a variance sigma^2)
  - where V = I (the identity matrix), (y, XB, sigma^2 I) = homoscedastic linear model
  - More general structure with V is the general linear model
- If the explanatory variables are observed quantities, the x_ijs are assumed to be random, so the model is a conditional model of y given X

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}, \quad D(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2\boldsymbol{V}. \tag{1.3.4}$$

Thus, the error term in (1.3.2) is the difference $\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{X})$. The mean and dispersion of the error given in (1.3.3) should be interpreted as conditional on $\boldsymbol{X}$. The representation (1.3.4) is called the *linear regression model.* In this context, $\boldsymbol{\beta}$ is called the vector of regression parameters or regression coefficients (see Section 3.4 for a brief discussion on regression). An important aspect of the linear regression model is that the error $\boldsymbol{y} - E(\boldsymbol{y}|\boldsymbol{X})$ must be uncorrelated with $\boldsymbol{X}$ (see Exercise 1.6).

Suppose that the observations $(y_i, x_{i1}, \ldots, x_{ip})$ for $i = 1, 2, \ldots, n$ are statistically independent (in which case $\boldsymbol{V} = \boldsymbol{I}$). Then the conditional model (1.3.4) can be written in the simpler form

$$\begin{aligned} E(y|x_1, x_2, \ldots, x_p) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \\ Var(y|x_1, \ldots, x_p) &= \sigma^2, \end{aligned} \tag{1.3.5}$$

- Scope of the linear model
  - Polynomial regression
    - When there is a single explanatory variable, the mean response is sought to be explained as a polynomial function of the explanatory variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \epsilon,$$

where y and x are the response and the explanatory variable, respectively. This is known as the *polynomial regression model.* This model can be viewed as a special case of (1.1.1) with $x_j = x^j$, $j = 1, 2, \ldots, p$. Therefore, the methodology to be developed for the model (1.1.1) will be applicable to the polynomial regression model. □

    - Linearity in the parameters makes the model linear (as long as the right-hand side is linear in the parameters even if it is not linear in the explanatory variables. (parameters are your beta coefficients)
      - The nonlinearity in the explanatory variables can be removed with transformations (linearization)
  - Generalized linear model
    - A nonlinear function of expected response is model as a linear function of the regression parameters

$$\eta(E(y|x_1, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{1.4.4}$$

This model is known as the *generalized linear model*, and the function η is called the *link function*, which is assumed to be known. One can

    - You can redefine z = n(y) where z includes the E error term

        - Error due to measurement of ignored factors and linearization
- Related Models
    - Term linear model can be more general (any model which connects variables or their transformed versions through a linear relationship)
    - Autoregressive model
        - Where the explanatory variables are collections of past values of the response itself AR(p)
        - Serves as a vehicle for linear prediction (predicting future values using a linear combination of its past values)

A more general linear model for time series data is the autoregressive moving average model of order (p, q), also known as ARMA(p, q) and given by

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \theta_0 \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad (1.5.2)$$

    - Mixed-effects linear model
        - A combination of fixed and random parameters are involved
    - The simplest form of linear modeling is usually used as a validation benchmark
- Uses of the linear model
    - Regression analysis: average response is explained via regressors
    - Or used as a vehicle of analysis:
        - Relationship with a particular regressor (using your beta coefficient)
        - Examine statistically certain empirical beliefs: x2 effects y 3 times more than x1 (B2 >= 3B1) - this can be tested using available data
        - Area of prediction: unobserved values of response may be predicted on the basis of the fitted model and the values of the explanatory variables corresponding to the unobserved response -> calibration
        - Designed experiments: effects of certain variables or testing
        - Control: if one of the explanatory variables can be controlled, which value produces the desired level of response?
        - Optimization: shooing the right explanatory variable
            - Maximum or minimum of the estimated response surface within a certain range of the variables
        - Fill missing data using related variables like a prediction or detect bad

<u>Review of Linear Algebra</u>

- Matrices and Vectors
    - Matrix = rows and columns of numbers
        - M rows and n columns = order m x n
        - Product: number of rows in A = number of columns in B
            - Order matters
        - Diagonal matrix: all off-diagonal elements are 0
            - Sum of diagonals = trace = tr(A)

$$\operatorname{tr}(\boldsymbol{A}_{m\times n}\boldsymbol{B}_{n\times m}) = \operatorname{tr}(\boldsymbol{B}_{n\times m}\boldsymbol{A}_{m\times n}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}b_{j,i}.$$

- Transpose of A = A' = ((a_ j,i)
  - (AB)' = B'A'
  - Symmetric if A' = A

$\boldsymbol{A}+\boldsymbol{B}$ is defined as

$$\boldsymbol{A}+\boldsymbol{B} = ((a_{i,j}+b_{i,j})).$$

The scalar product of a matrix $\boldsymbol{A}$ with a real number c is defined as

$$c\boldsymbol{A} = ((ca_{i,j})).$$

The difference of $\boldsymbol{A}_{m\times n}$ and $\boldsymbol{B}_{m\times n}$ is defined as $\boldsymbol{A}+(-1)\boldsymbol{B}$. Thus,

$$\boldsymbol{A}-\boldsymbol{B} = ((a_{i,j}-b_{i,j})).$$

The product of the matrix $\boldsymbol{A}_{m\times n}$ with the matrix $\boldsymbol{B}_{n\times k}$, denoted by $\boldsymbol{AB}$, is defined as

$$\boldsymbol{AB} = \left(\left(\sum_{l=1}^{n} a_{i,l}b_{l,j}\right)\right).$$

- Matrix with single column = column vector
  - Synonymous to a *vector* = column vector
  - Order of a vector = n x 1 = order n for brevity
- Matrix with single row = row vector
- Non-trivial vector or matrix = not identically equal to 0
- **1** = vector of 1s
  - Identity matrix = all diagonal elements 1
    - AI = IA = A
- Kronecker product

The *Kronecker product* of two matrices $\boldsymbol{A}_{m\times n}$ and $\boldsymbol{B}_{p\times q}$, denoted by

$$\boldsymbol{A}\otimes\boldsymbol{B} = ((a_{ij}\boldsymbol{B})),$$

is a partitioned $mp \times nq$ matrix with $a_{ij}\boldsymbol{B}$ as its (i,j)th block. This product is found to be very useful in the manipulation of matrices with special block structure. It follows from the definition of the Kronecker product that

(a) $(\boldsymbol{A}_1+\boldsymbol{A}_2)\otimes\boldsymbol{B} = \boldsymbol{A}_1\otimes\boldsymbol{B}+\boldsymbol{A}_2\otimes\boldsymbol{B}$;
(b) $\boldsymbol{A}\otimes(\boldsymbol{B}_1+\boldsymbol{B}_2) = \boldsymbol{A}\otimes\boldsymbol{B}_1+\boldsymbol{A}\otimes\boldsymbol{B}_2$;
(c) $(\boldsymbol{A}_1\boldsymbol{A}_2)\otimes(\boldsymbol{B}_1\boldsymbol{B}_2) = (\boldsymbol{A}_1\otimes\boldsymbol{B}_1)(\boldsymbol{A}_2\otimes\boldsymbol{B}_2)$;
(d) $(\boldsymbol{A}\otimes\boldsymbol{B})' = \boldsymbol{A}'\otimes\boldsymbol{B}'$;
(e) $\boldsymbol{A}_{m\times n}\boldsymbol{b} = (\boldsymbol{b}'\otimes\boldsymbol{I}_{m\times m})\operatorname{vec}(\boldsymbol{A})$, where $\operatorname{vec}(\boldsymbol{A})$ is the vector obtained by successively concatenating the columns of $\boldsymbol{A}$.

    - Column rank = maximum number of linearly independent columns of A
        - Row rank is the same for rows
        - Column rank of A_mxn is n = full rank (full row rank same)
        - Rank deficient = neither full row rank nor full column rank
        - An important result of matrix theory: row rank of any matrix = column rank
            - p(A) <= min{m, n}
            - A square matrix B_nxn is called full rank or non-singular if p(B) = n
                - p(B) < n = singular matrix -> rank deficient
    - Inner product: a generalization of the dot product
        - a'b = b'a
    - Norm of a vector: magnitude (length of the vector from origin to point)
        - ||a|| = (a'a)^½ = sqrt(x_1^2 + … + x_i^2)
            - If ||a|| = 1, vector is unit norm or unit vector
- Inverse and generalized Inverses
    - If AB = I, then B is the right-inverse of A and A is the left inverse of B
        - Right inverse of A = A^-R (only exists when A is full row rank)
        - Left inverse of B = B^-L (only when B is full column rank)
        - Both inverses only exist if A is a square matrix and full rank (RREF pivots = I, so all vector lin indep). Then A^-L and A^-R are unique and equal
            - The inverse of a non-singular matrix

              by $\boldsymbol{A}^{-1}$. By definition, the inverse exists and is unique if and only if $\boldsymbol{A}$ is nonsingular, and $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are both nonsingular with the same order, then $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

            - Non_singular matrix is an invertible matrix
    - If A is invertible and A^-1 = A', then A is an orthogonal matrix
- Vector space and projection
    - A vector u is orthogonal to another vector v if u'v = 0.
        - If a vector is orthogonal to all vectors in a vector space, orthogonal to S
        - Two vector spaces can be orthogonal to each other
    - Number of bases is a uniquely defined attribute of vector spaces
        - The number is called the dimension of the vector space: dim(S)
        - If S consists of n-component vectors, dim(S) <= n
    - Basis
        - Orthogonal bases = vectors in basis set are orthogonal to each other
        - Unit norm = basis is called orthonormal
        - You can always construct these special bases for a vector space using Gram-Schmidt Orthogonalization
    - Projection matrices = confusing (pg. 35)
- Column Space
    - Spanned by the columns of A = C(A)
    - Row space = C(A')

Let the matrix $\boldsymbol{A}$ have the columns $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$. If $\boldsymbol{x}$ is a vector having components $x_1, x_2, \ldots, x_n$, then the matrix-vector product

$$\boldsymbol{A}\boldsymbol{x} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n,$$

represents a linear combination of the columns of $\boldsymbol{A}$. The set of all vectors that may be expressed as linear combinations of the columns of $\boldsymbol{A}$ is a vector space and is called the *column space* of $\boldsymbol{A}$. We denote it

Proposition 2.4.1

(a) $\mathcal{C}(\boldsymbol{A} : \boldsymbol{B}) = \mathcal{C}(\boldsymbol{A}) + \mathcal{C}(\boldsymbol{B})$.
(b) $\mathcal{C}(\boldsymbol{A}\boldsymbol{B}) \subseteq \mathcal{C}(\boldsymbol{A})$.
(c) $\mathcal{C}(\boldsymbol{A}\boldsymbol{A}') = \mathcal{C}(\boldsymbol{A})$. *Consequently,* $\rho(\boldsymbol{A}\boldsymbol{A}') = \rho(\boldsymbol{A})$.
(d) $\mathcal{C}(\boldsymbol{C}) \subseteq \mathcal{C}(\boldsymbol{A})$ *only if* $\boldsymbol{C}$ *is of the form* $\boldsymbol{A}\boldsymbol{B}$ *for a suitable matrix* $\boldsymbol{B}$.
(e) If $\mathcal{C}(\boldsymbol{B}) \subseteq \mathcal{C}(\boldsymbol{A})$, *then* $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{B} = \boldsymbol{B}$, *irrespective of the choice of the g-inverse. Similarly,* $\mathcal{C}(\boldsymbol{B}') \subseteq \mathcal{C}(\boldsymbol{A}')$ *implies* $\boldsymbol{B}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{B}$.
(f) $\mathcal{C}(\boldsymbol{B}') \subseteq \mathcal{C}(\boldsymbol{A}')$ *and* $\mathcal{C}(\boldsymbol{C}) \subseteq \mathcal{C}(\boldsymbol{A})$ *if and only if* $\boldsymbol{B}\boldsymbol{A}^-\boldsymbol{C}$ *is invariant under the choice of the g-inverse.*
(g) $\boldsymbol{B}'\boldsymbol{A} = \boldsymbol{0}$ *if and only if* $\mathcal{C}(\boldsymbol{B}) \subseteq \mathcal{C}(\boldsymbol{A})^\perp$.
(h) $dim(\mathcal{C}(\boldsymbol{A})) = \rho(\boldsymbol{A})$.
(i) If $\boldsymbol{A}$ *has* n *rows, then* $dim(\mathcal{C}(\boldsymbol{A})^\perp) = n - \rho(\boldsymbol{A})$.
(j) If $\mathcal{C}(\boldsymbol{A}) \subseteq \mathcal{C}(\boldsymbol{B})$ *and* $\rho(\boldsymbol{A}) = \rho(\boldsymbol{B})$, *then* $\mathcal{C}(\boldsymbol{A}) = \mathcal{C}(\boldsymbol{B})$. *In particular,* $\mathcal{C}(\boldsymbol{I}_{n\times n}) = I\!R^n$.
(k) $\rho(\boldsymbol{A}\boldsymbol{B}) \leq \min\{\rho(\boldsymbol{A}), \rho(\boldsymbol{B})\}$.
(l) $\rho(\boldsymbol{A} + \boldsymbol{B}) \leq \rho(\boldsymbol{A}) + \rho(\boldsymbol{B})$.

- ^ confusing

- Matrix Decompositions
  - Rank-factorization

    Any non-null matrix $\boldsymbol{A}_{m\times n}$ of rank r can be written as $\boldsymbol{B}_{m\times r}\boldsymbol{C}_{r\times n}$, where $\boldsymbol{B}$ has full column rank and $\boldsymbol{C}$ has full row rank. This is called a *rank-factorization.*

  - Singular value decomposition
    - Non-zero elements of D are singular values
      - Columns of U and V corresponding to the singular values are called the left and right singular vectors of A

    Any matrix $\boldsymbol{A}_{m\times n}$ can be written as $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where $\boldsymbol{U}_{m\times m}$ and $\boldsymbol{V}_{n\times n}$ are orthogonal matrices and $\boldsymbol{D}_{m\times n}$ is a diagonal matrix with nonnegative diagonal elements. This is called a *singular value decomposition*

    - p(A) = p(D), so number of positive singular values = rank

Example 2.5.1 Consider the matrix

$$\boldsymbol{A} = \begin{pmatrix} 4 & 5 & 2 \\ 0 & 3 & 6 \\ 4 & 5 & 2 \\ 0 & 3 & 6 \end{pmatrix}.$$

The rank of $\boldsymbol{A}$ is 2. An SVD of $\boldsymbol{A}$ is $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$, where

$$\boldsymbol{U} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 12 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \boldsymbol{V} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} & -\frac{2}{3} \\ \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \end{pmatrix}.$$

- Spectral decomposition, eigenvalues, and eigenvectors are confusing (pg 42-45)
- Lowner order
  - Come back to this
- Solution of linear equations
  - Ax = b has a solution if and only if b is in C(A)

> *(b) The equations* $\boldsymbol{Ax} = \boldsymbol{b}$ *have a unique solution if and only if* $\boldsymbol{b} \in \mathcal{C}(\boldsymbol{A})$ *and* $\rho(\boldsymbol{A}) = n$.
>
> *(c) If* $\boldsymbol{b} \in \mathcal{C}(\boldsymbol{A})$, *every solution to the equations* $\boldsymbol{Ax} = \boldsymbol{b}$ *is of the form* $\boldsymbol{A}^{-}\boldsymbol{b} + (\boldsymbol{I} - \boldsymbol{A}^{-}\boldsymbol{A})\boldsymbol{c}$ *where* $\boldsymbol{A}^{-}$ *is any fixed g-inverse of* $\boldsymbol{A}$ *and* $\boldsymbol{c}$ *is an arbitrary vector.*

- Optimization of quadratic forms and functions
  - Consider q(x) = x'Ax + b'x + c, where A is a symmetric without loss of generality
  - To minimize or maximize q(x) with respect to x, differentiate q(x) w.r.t components of x one at a time and set derivatives equal to 0. Check where the gradient is equal to 0.

Let $\boldsymbol{x} = (x_1, \ldots, x_n)'$. The gradient of a function $f(\boldsymbol{x})$ is defined as

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}.$$

  - Come back to this section

Chapter 3: Review of Statistical Results

- Covariance Adjustment

$E(\boldsymbol{u})$ and $D(\boldsymbol{u})$, respectively. For an ordered pair of random vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, we denote the matrix of covariances by $Cov(\boldsymbol{u}, \boldsymbol{v})$, that is,

$$Cov(\boldsymbol{u}, \boldsymbol{v}) = E[\{\boldsymbol{u} - E(\boldsymbol{u})\}\{\boldsymbol{v} - E(\boldsymbol{v})\}'].$$

According to this notation, $D(\boldsymbol{u}) = Cov(\boldsymbol{u}, \boldsymbol{u})$, and $Cov(\boldsymbol{v}, \boldsymbol{u}) = Cov(\boldsymbol{u}, \boldsymbol{v})'$. It is easy to see that $Cov(\boldsymbol{Au}, \boldsymbol{Bv}) = \boldsymbol{A}Cov(\boldsymbol{u}, \boldsymbol{v})\boldsymbol{B}'$ and $D(\boldsymbol{Au}) = \boldsymbol{A}D(\boldsymbol{u})\boldsymbol{A}'$.

  - Variance-covariance matrix = dispersion matrix D(u)
    - For a single dispersion matrix, column space does not contain all the vectors having the same order of its columns
    - So it's important to know which vectors are contained
- Basic distributions
  - Multivariate normal distribution: if for every fixed vector l with the same order as y, the RV l'y has the univariate normal distribution

The multivariate normal distribution is completely characterized by the mean vector and the dispersion matrix. If $E(\boldsymbol{y}) = \boldsymbol{\mu}$ and $D(\boldsymbol{y}) = \boldsymbol{V}$, the notation $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{V})$ indicates that $\boldsymbol{y}$ has the multivariate normal distribution with mean $\boldsymbol{\mu}$ and dispersion matrix $\boldsymbol{V}$. The joint probability density of such a random vector with n components is

$$f(\boldsymbol{y}) = (2\pi)^{-n/2}|\boldsymbol{V}|^{-1/2}\exp[-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})],$$

    - Come back to this section
- Distribution of quadratic forms
    - Come back to this section
- Regression
    - Approximate a random vector y by a suitable function of a random vector x

      the vector function $\boldsymbol{g}(\boldsymbol{x})$ that minimizes $E[(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{x}))'\boldsymbol{W}(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{x})]$, where $\boldsymbol{W}$ is an arbitrary, positive definite matrix that may depend on $\boldsymbol{x}$, is $\boldsymbol{g}(\boldsymbol{x}) = E(\boldsymbol{y}|\boldsymbol{x})$ (Exercise 3.7). This conditional expectation is called the *regression of* $\boldsymbol{y}$ *on* $\boldsymbol{x}$. It is sometimes referred to as the *regression function*, when viewed as a function of $\boldsymbol{x}$. If $\boldsymbol{y}$ is a scalar (denoted as

        - Minimum mean squared error criterion is to minimize E[y - g(x)]^2 with respect to function g
            - This occurs when g(x) = E(y|x), the regression of y on x
            - The approximation error y - E(y|x) is uncorrelated

              through a *linear* function of $\boldsymbol{x}$, including a constant. The solution to the minimization problem

              $$\min_{\boldsymbol{l}, c} E[y - \boldsymbol{l}'\boldsymbol{x} - c]^2$$

              may be called the *linear* regression of y on $\boldsymbol{x}$. It is known more commonly as the *best linear predictor* (BLP) of y given $\boldsymbol{x}$. We denote it by $\hat{E}(y|\boldsymbol{x})$.

    - Very complicated
- Basic concepts of inference
    - Sufficiency: a statistic t(y) is called sufficient for the parameter theta if the conditional distribution of y given t(y) = t0
        - You are summarizing information to make the summary brief
        - t(y) is called minimal if it is almost surely equal to a function of any other sufficient statistic almost surely for all theta - you have no more than what you need to know from y about theta
    - Aciliarry: z(y) for all theta if marginal distribution of z(y) does not depend on theta
    - Confusing
- Point estimation
    - Come back to this section
- Bayesian estimation
    - Making use of prior knowledge that is often expressed in terms of a prior distribution theta, denoted by pi(theta) -> called the prior

        - So you get the conditional distribution of y given theta
    - Average risk

$$r(\boldsymbol{t}, \pi) = \int R(\boldsymbol{\theta}, \boldsymbol{t})\, d\pi(\boldsymbol{\theta}),$$

    - Come back to this section

Chapter 4: Estimation in the Linear Model

> Consider the homoscedastic linear model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. This model is a special case of (1.3.2)–(1.3.3) where the model errors have the same variance and are uncorrelated. The unknown parameters of this model are the coefficient vector $\boldsymbol{\beta}$ and the error variance σ^2. In this chapter we deal with the problem of estimation of these parameters from the observables $\boldsymbol{y}$ and $\boldsymbol{X}$. We assume that $\boldsymbol{y}$ is a vector of n elements, $\boldsymbol{X}$ is an $n \times k$ matrix and $\boldsymbol{\beta}$ is a vector of k elements.

- Zero linear functions
    - An important tool in the theory of best linear unbiased estimation
    - Linear functions of the response which have zero expectation
- Linear estimation: some basic facts

> Much of the classical inference problems related to the linear model $(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ concern a *linear parametric function* (LPF), $\boldsymbol{p}'\boldsymbol{\beta}$. We often estimate it by a linear function of the response, $\boldsymbol{l}'\boldsymbol{y}$. Since $\boldsymbol{y}$ itself is modelled as a linear function of the parameter $\boldsymbol{\beta}$ plus error, it is reasonable to expect that one may be able to estimate $\boldsymbol{\beta}$ by some kind of a linear transformation in the reverse direction. This is why we try to estimate LPFs by linear estimators, that is, as linear functions of $\boldsymbol{y}$.

- Linear unbiased estimator and linear zero function
    - LZFs are important because they contain information about the error in the model (useful for estimating sigma^2), and the mean and variance of LZFs do not depend on beta (decoupled from XB)
    - Can be used to isolate noise
    - Ex:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

> The observations y_2 and y_4, being direct measurements of error, may be used to estimate the error variance. These are LZFs. The other two observations carry information about the two parameters. There are several unbiased estimators of α_1, such as y_1, $y_1 + y_2$ and $y_1 + y_4$. It appears that y_1 would be a natural estimator of α_1 since it is free from the baggage of any LZF. We shall formalize this heuristic argument later. □

        - But we usually don't get this nice of a split of errors in any real linear model—our goal is to achieve it
        - LZFs are sometimes referred to as linear error functions

> **Proposition 4.1.10** *A necessary and sufficient condition for the estimability of an LPF $(\boldsymbol{p}'\boldsymbol{\beta})$ is that $\boldsymbol{p} \in \mathcal{C}(\boldsymbol{X}')$.* □

    - Come back to this section
- Least squares estimation
    - Error vector can be written as (y - XB) so you estimate B to minimize the sum of sqaured elements from this error vector

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \quad \widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

    - X’X has to be singular (X has full column rank so B is estimable)
- Best linear unbiased estimation