Notes for 751-752
Sections 1-3

Martin Lindquist*

August 25, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

1 Introduction

Linear models are the cornerstone of statistical methodology. They allow us to model the relationship between a response variable (dependent variable) and one or more explanatory variables (covariate, predictor, factor, independent variable), and determine which of these variables are important. In addition, they are relatively easy to work with, and are applicable to many real-life settings. Properly constructed, linear models encompass many common statistical techniques, including t-tests, regression, analysis of variance, and analysis of covariance.

Let y_i be the response variable and $(x_{i1}, x_{i2}, \ldots x_{ip})$ a set of explanatory variables for observation $i = 1, \ldots n$. The relationship between the variables can be expressed using the model:

$$y_i = \mu(x_{i1}, x_{i2}, \ldots x_{ip}) + \epsilon_i \tag{1}$$

where $\mu(x_{i1}, x_{i2}, \ldots x_{ip})$ is a function of the explanatory variables, which are assumed to be known constants, and the error term ϵ. The function $\mu(.)$ is the deterministic part of the model, while y_i and ϵ_i are both random. The form of $\mu(.)$ is generally assumed to be known, but contains unknown parameters. Here the term *linear* implies that $\mu(.)$ is a linear function of some unknown parameters, e.g., $\mu(x_{i1}, x_{i2}) = \beta_1 x_{i1} + \beta_2 x_{i2}$. However, the original explanatory variables can be subjected to arbitrary transformations.

The variables $x_{i1}, x_{i2}, \ldots x_{ip}$ can be quantitative, transformations of quantitative variables, basis expansions, numeric, or interactions. In the next section we explore various models that can be created using these inputs.

Throughout the course we will make additional assumptions about the mean, variance, and distribution of the error, but for now we leave this information unspecified.

1.1 Examples

Below follow some specific cases of linear models. Many times it is convenient to write the models in matrix form so we provide details of how to do so in each case.

Example: 1.1 Simple linear regression.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example: 1.2 Polynomial regression.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad \text{for } i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example: 1.3 Multiple linear regression.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad \text{for } i = 1, \ldots n$$

or

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example: 1.4 Data transformations.

$$log(y_i) = \beta_0 + \beta_1 log(x_i) + \epsilon_i \qquad \text{for } i = 1, \ldots n$$

or

$$\begin{pmatrix} log(y_1) \\ log(y_2) \\ \vdots \\ log(y_n) \end{pmatrix} = \begin{pmatrix} 1 & log(x_1) \\ 1 & log(x_2) \\ \vdots & \vdots \\ 1 & log(x_n) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Example: 1.5 One-way analysis of variance.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad \text{for } i = 1, 2; \;\; j = 1, \ldots J$$

or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2J} \end{pmatrix}$$

Example: 1.6 Two-way analysis of variance.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \qquad \text{for } \; i, j = 1, 2; \;\; k = 1, \ldots K$$

or

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{11K} \\ \hline y_{121} \\ \vdots \\ y_{12K} \\ \hline y_{211} \\ \vdots \\ y_{21K} \\ \hline y_{221} \\ \vdots \\ y_{22K} \end{pmatrix} = \left(\begin{array}{ccccc} 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 \\ \hline 1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ \hline 1 & 0 & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 \\ \hline 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 \end{array}\right) \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{11K} \\ \hline \varepsilon_{121} \\ \vdots \\ \varepsilon_{12K} \\ \hline \varepsilon_{211} \\ \vdots \\ \varepsilon_{21K} \\ \hline \varepsilon_{221} \\ \vdots \\ \varepsilon_{22K} \end{pmatrix}$$

Example: 1.7 Analysis of covariance.

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}_{..}) + \epsilon_{ij} \qquad \text{for } \; i = 1, 2; \;\; j = 1, \ldots J$$

or

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1J} \\ y_{21} \\ \vdots \\ y_{2J} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & (x_{11} - \bar{x}_{..}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & (x_{1J} - \bar{x}_{..}) \\ 1 & 0 & 1 & (x_{21} - \bar{x}_{..}) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & (x_{2J} - \bar{x}_{..}) \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2J} \end{pmatrix}$$

1.2 The General Linear Model

The general linear model in matrix form:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Equivalent shorthand form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Here $\mathbf{y}$ $(n \times 1)$ is the response vector, $\mathbf{X}$ $(n \times p)$ is the design (or model or regression) matrix, $\boldsymbol{\beta}$ $(p \times 1)$ is the vector of regression coefficients, $\boldsymbol{\varepsilon}$ $(n \times 1)$ is the error vector (mean $\mathbf{0}$).

Note: 1.8 Usually $x_{i0} = 1$ for all i, i.e., there is an intercept β_0 in the model.

Note: 1.9 $x_{i0}, x_{i1}, \ldots, x_{i,p-1}$ are called the predictor variables or regressor variables. They are known constants.

Note: 1.10 The model is linear in the unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_{p-1}$, i.e.,

$$\mathbf{y} = \sum_{j=0}^{p-1} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon},$$

where $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$.

Note: 1.11 We will give assumptions on the distribution of $\boldsymbol{\varepsilon}$ when we discuss estimation methods.

1.3 Least Squares Estimation

Definition: 1.12 An estimate $\hat{\boldsymbol{\beta}}$ is a least-squares estimate of $\boldsymbol{\beta}$ if it minimizes $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$ over all $\boldsymbol{\beta}$.

Note: 1.13 We can re-express the least-squares criteria as follows:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

1.4 Goals

We have multiple goals when working with linear models. These include:

- Finding the best 'fit' between the explanatory variables and the response. The standard procedure for this is the method of least squares, but other methods exist.
- Obtaining good parameter estimates that allow us to determine the relationship between variables.
- Making predictions that allow us to determine what response can we expect to get under a new set of experimental conditions.
- Performing inference.

1.5 Software

R is a programming language and software environment for statistical computing and graphics that is highly extensible. In recent years it has become the most popular language among statisticians for developing new statistical methodology.

Linear models can be fit in R using using the `lm` command. Here we illustrate simple linear regression using the `mtcars` data set that is directly available in R. The data was extracted from the 1974 Motor Trend US magazine, and consists of gas consumption (`mpg`) and 10 other aspects of automobile design and performance for a total of 32 cars. Here we will fit a simple linear regression using gas consumption (`mpg`) as the response variable and weight (`wt`) as the explanatory variable.

```
> fit <- lm(mpg ~ wt, data=mtcars)
> fit

Call:
lm(formula = mpg ~ wt, data = mtcars)

Coefficients:
(Intercept)           wt
     37.285       -5.344
```

We will revisit R and the `lm` command throughout the course.

2 Review of Linear Algebra and Matrices

2.1 Matrix Notation and Elementary Properties

Definition: 2.1 Matrix: An $m \times n$ matrix with elements a_{ij} is denoted $\mathbf{A} = (a_{ij})_{m\times n}$.

Definition: 2.2 Vector: A vector of length n is denoted $\mathbf{a} = (a_i)_n$. If all elements equal 1 it is denoted $\mathbf{1}_n$.

Definition: 2.3 Diagonal Matrix:

$$\mathrm{diag}(a_1, \ldots, a_n) \equiv \begin{pmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & a_n \end{pmatrix}.$$

Definition: 2.4 Identity Matrix: $\mathbf{I}_{n\times n} \equiv \mathrm{diag}(\mathbf{1}_n)$.

Definition: 2.5 Matrix Transpose: If $\mathbf{A} = (a_{ij})_{m\times n}$, then $\mathbf{A}' \equiv (a'_{ij})_{n\times m}$ where $a'_{ij} = a_{ji}$.

Definition: 2.6 If $\mathbf{A} = \mathbf{A}'$, then $\mathbf{A}$ is symmetric.

Definition: 2.7 Matrix Sum: If $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{m\times n}$, then $\mathbf{A} + \mathbf{B} = (a_{ij} + b_{ij})_{m\times n}$.

Theorem: 2.8 Matrix sums satisfy $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ and $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.

Definition: 2.9 Matrix Product: If $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{n\times p}$, then $\mathbf{AB} = (c_{ij})_{m\times p}$, where $c_{ij} = \sum_k a_{ik}b_{kj}$.

Theorem: 2.10 Matrix products satisfy $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Definition: 2.11 Matrix Trace: The sum of the diagonal elements, $\mathrm{tr}(\mathbf{A}) \equiv \sum_i a_{ii}$.

Theorem: 2.12 The trace satisfies $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$ if $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{m\times n}$, and $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$ if $\mathbf{A}$ and $\mathbf{B}$ are square matrices.

2.2 Vector Spaces

Definition: 2.13 A (real) vector space consists of a non empty set $\mathbf{V}$ of vectors, and two operations:

(1) Addition is defined for pairs of elements in $\mathbf{V}$, $\mathbf{x}$ and $\mathbf{y}$, and yields an element in $\mathbf{V}$, denoted by $\mathbf{x}+\mathbf{y}$.

(2) Scalar multiplication, is defined for the pair α, a real number, and an element $\mathbf{x} \in \mathbf{V}$, and yields an element in $\mathbf{V}$ denoted by $\alpha\mathbf{x}$.

Eight properties are assumed to hold for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{V}, \alpha, \beta, 1 \in \mathbb{R}$:

(1) $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$

(2) $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$

(3) There is an element in $\mathbf{V}$ denoted $\mathbf{0}$ such that $\mathbf{0} + \mathbf{x} = \mathbf{x} + \mathbf{0} = \mathbf{x}$

(4) For each $\mathbf{x} \in \mathbf{V}$ there is an element in $\mathbf{V}$ denoted $-\mathbf{x}$ such that $\mathbf{x} + (-\mathbf{x}) = (-\mathbf{x}) + \mathbf{x} = \mathbf{0}$

(5) $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$ for all α

(6) $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$ for all α, β

(7) $1\mathbf{x} = \mathbf{x}$

(8) $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$ for all α, β

Definition: 2.14 Vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ are linearly independent if $\sum_i c_i\mathbf{a}_i \neq 0$ unless $c_i = 0$ for all i.

Definition: 2.15 A linear basis or coordinate system in a vector space $\mathbf{V}$ is a set $\mathbf{B}$ of linearly independent vectors in $\mathbf{V}$ such that each vector in $\mathbf{V}$ can be written as a linear combination of the vectors in $\mathbf{B}$.

Definition: 2.16 The dimension of a vector space is the number of vectors in any basis of the vector space.

Definition: 2.17 Let $\mathbf{V}$ be a p dimensional vector space and let $\mathbf{W}$ be an n dimensional vector space. A linear transformation $\mathbf{L}$ from $\mathbf{V}$ to $\mathbf{W}$ is a mapping (function) from $\mathbf{V}$ to $\mathbf{W}$ such that

$$\mathbf{L}(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha\mathbf{L}(\mathbf{x}) + \beta\mathbf{L}(\mathbf{y}) \text{ for every } \mathbf{x}, \mathbf{y} \in \mathbf{V} \text{ and all } \alpha, \beta \in \mathbb{R}.$$

2.3 Range, Rank, and Null Space

Definition: 2.18 Range (Column Space): $\mathcal{R}(\mathbf{A}) \equiv$ the linear space spanned by the columns of $\mathbf{A}$.

Definition: 2.19 Rank: $\text{rank}(\mathbf{A}) \equiv \text{r}(\mathbf{A}) \equiv$ the number of linearly independent columns of $\mathbf{A}$ (i.e., the dimension of $\mathcal{R}(\mathbf{A})$), or equivalently, the number of linearly independent rows of $\mathbf{A}$.

Theorem: 2.20 Decreasing property of rank: $\text{rank}(\mathbf{AB}) \leq \min\{\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B})\}$.

Definition: 2.21 Null Space: $\mathcal{N}(\mathbf{A}) \equiv \{\mathbf{x} : \mathbf{Ax} = \mathbf{0}\}$. The nullity of $\mathbf{A}$ is the dimension of $\mathcal{N}(\mathbf{A})$.

Theorem: 2.22 $\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = n$, the number of columns of $\mathbf{A}$.

Theorem: 2.23 $\text{r}(\mathbf{A}) = \text{r}(\mathbf{A}') = \text{r}(\mathbf{A}'\mathbf{A}) = \text{r}(\mathbf{A}\mathbf{A}')$.

2.4 Inverse

Definition: 2.24 An $n \times n$ matrix $\mathbf{A}$ is invertible (or non-singular) if there is a matrix $\mathbf{A}^{-1}$ such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_{n\times n}$. Equivalently, $\mathbf{A}$ $(n \times n)$ is invertible if and only if $\text{rank}(\mathbf{A}) = n$.

Theorem: 2.25 If $\mathbf{A}$ is invertible, then $\mathbf{A}'$ is invertible and $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.

Theorem: 2.26 Inverse of Product: $(\mathbf{AB})^{-\mathbf{1}} = \mathbf{B}^{-\mathbf{1}}\mathbf{A}^{-\mathbf{1}}$ if $\mathbf{A}$ and $\mathbf{B}$ are invertible.

2.5 Inner Product, Length, and Orthogonality

Definition: 2.27 Inner product: $\mathbf{a}'\mathbf{b} = \sum_i a_i b_i$, where $\mathbf{a} = (a_i)$, $\mathbf{b} = (b_i)$.

Definition: 2.28 Vector norm (length): $||\mathbf{a}|| = \sqrt{\mathbf{a}'\mathbf{a}}$.

Definition: 2.29 Orthogonal vectors: $\mathbf{a} = (a_i)$ and $\mathbf{b} = (b_i)$ are orthogonal if $\mathbf{a}'\mathbf{b} = 0$.

Definition: 2.30 Orthogonal matrix: $\mathbf{A}$ is orthogonal if its columns are orthogonal vectors of length 1, or equivalently, if $\mathbf{A}^{-1} = \mathbf{A}'$.

2.6 Determinants

Definition: 2.31 For a square matrix $\mathbf{A}$, $|\mathbf{A}| \equiv \sum_i a_{ij} A_{ij}$, where the cofactor $A_{ij} = (-1)^{i+j}|\mathbf{M}_{ij}|$, and $\mathbf{M}_{ij}$ is the matrix obtained by deleting the ith row and jth column from $\mathbf{A}$.

Theorem: 2.32 $\left| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right| = ad - bc.$

Theorem: 2.33 $|\mathbf{A}| = 0,$ if and only if $\mathbf{A}$ is singular.

Theorem: 2.34 $|\mathrm{diag}(a_1, \ldots, a_n)| = \prod_i a_i$.

Theorem: 2.35 $|\mathbf{AB}| = |\mathbf{A}| \cdot |\mathbf{B}|$.

Theorem: 2.36 $\left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} \right| = |\mathbf{A}| \cdot |\mathbf{C}|$.

Theorem: 2.37 If $|\mathbf{A}|$ is invertible, $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$.

2.7 Eigenvalues

Definition: 2.38 If $\mathbf{Ax} = \lambda\mathbf{x}$ where $\mathbf{x} \neq 0$, then λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is a corresponding eigenvector.

Let $\mathbf{A}$ be a symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$.

Theorem: 2.39 (Spectral Theorem, a.k.a. Principal Axis Theorem) For any symmetric matrix $\mathbf{A}$ there exists an orthogonal matrix $\mathbf{T}$ such that: $\mathbf{T'AT} = \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Theorem: 2.40 $\mathrm{r}(\mathbf{A}) =$ the number of non-zero λ_i

Theorem: 2.41 $\mathrm{tr}(\mathbf{A}) = \sum_i \lambda_i$.

Theorem: 2.42 $|\mathbf{A}| = \prod_i \lambda_i$.

2.8 Positive Definite and Semidefinite Matrices

Definition: 2.43 A symmetric matrix $\mathbf{A}$ is positive semidefinite (p.s.d.) if $\mathbf{x'Ax} \geq 0$ for all $\mathbf{x}$.

Theorem: 2.44 If $\mathbf{A}$ is aa p.s.d matrix, then

(a) The diagonal elements a_{ii} are all non-negative.

(b) All eigenvalues of $\mathbf{A}$ are nonnegative.

(c) $\mathrm{tr}(\mathbf{A}) \geq 0$.

Definition: 2.45 A symmetric matrix $\mathbf{A}$ is called positive definite (p.d.) if $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all non-zero $\mathbf{x}$.

Theorem: 2.46 If $\mathbf{A}$ is aa p.d matrix, then

(a) All diagonal elements and all eigenvalues of $\mathbf{A}$ are positive.

(b) tr$(\mathbf{A}) > 0$.

(c) $|\mathbf{A}| > 0$.

(d) There is a nonsingular $\mathbf{R}$ such that $\mathbf{A} = \mathbf{R}\mathbf{R}'$ (necessary and sufficient for $\mathbf{A}$ to be p.d.).

(e) $\mathbf{A}^{-1}$ is p.d.

2.9 Idempotent and Projection Matrices

Definition: 2.47 A matrix $\mathbf{P}$ is idempotent if $\mathbf{P}^2 = \mathbf{P}$. A symmetric idempotent matrix is called a projection matrix.

Properties of a projection matrix $\mathbf{P}$:

Theorem: 2.48 If $\mathbf{P}$ is an $n \times n$ matrix and rank$(\mathbf{P}) = r$, then $\mathbf{P}$ has r eigenvalues equal to 1 and $n - r$ eigenvalues equal to 0.

Theorem: 2.49 tr$(\mathbf{P})$ = rank$(\mathbf{P})$.

Theorem: 2.50 $\mathbf{P}$ is positive semidefinite.

2.10 Projections

Definition: 2.51 For two vectors **x** and **y**, the projection of **y** onto **x** is

$$\mathrm{Proj}_{\mathbf{x}}(\mathbf{y}) = \frac{\mathbf{x}'\mathbf{y}}{\mathbf{x}'\mathbf{x}}\mathbf{x}.$$

Theorem: 2.52 If V is a vector space and Ω is a subspace of V, then $\exists$ two vectors, $\mathbf{w}_1$, $\mathbf{w}_2 \in V$ such that

1. $\mathbf{y} = \mathbf{w}_1 + \mathbf{w}_2 \quad \forall\, \mathbf{y} \in V$,
2. $\mathbf{w}_1 \in \Omega$ and $\mathbf{w}_2 \in \Omega^{\perp}$.

Theorem: 2.53 $\|\mathbf{y} - \mathbf{w}_1\| \leq \|\mathbf{y} - \mathbf{x}\|$ for any $\mathbf{x} \in \Omega$. $\mathbf{w}_1$ is called the projection of $\mathbf{y}$ onto Ω.

Definition: 2.54 The matrix **P** that takes **y** onto $\mathbf{w}_1$ (i.e., $\mathbf{Py} = \mathbf{w}_1$) is called a projection matrix.

Theorem: 2.55 **P** projects $\mathbf{y}$ onto the space spanned by the column vectors of **P**.

Theorem: 2.56 **P** is a linear transformation.

Theorem: 2.57 $\mathbf{I} - \mathbf{P}$ is a projection operator onto $\Omega^{\perp}$.

2.11 Vector and Matrix Calculus

Definition: 2.58 Let $u = f(\mathbf{x})$ be a function of the variables $x_1, x_2, \ldots, x_p$ in $\mathbf{x} = (x_1, x_2, \ldots x_p)'$, and let $\partial u/\partial x_1$, $\partial u/\partial x_2$, ..., $\partial u/\partial x_p$ be the partial derivatives. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial u}{\partial x_1} \\ \frac{\partial u}{\partial x_2} \\ \vdots \\ \frac{\partial u}{\partial x_p} \end{pmatrix}.$$

Theorem: 2.59 Let $u = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$, where $\mathbf{a} = (a_1, a_2, \ldots a_p)'$ is a vector of constants. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}.$$

Theorem: 2.60 Let $u = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric matrix of constants. Then,

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

Definition: 2.61 Let $\mathbf{A}$ be an $n \times n$ nonsingular matrix with elements a_{ij} that are functions of a scalar x. We define $\partial\mathbf{A}/\partial x$ as the $n \times n$ matrix with elements $\partial a_{ij}/\partial x$.

Theorem: 2.62 Then,

$$\frac{\partial \mathbf{A}^{-1}}{\partial x} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}\mathbf{A}^{-1}.$$

Theorem: 2.63 Then,

$$\frac{\partial \log|\mathbf{A}|}{\partial x} = \mathrm{tr}(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial x}).$$

3 Single Parameter Regression

Here we will consider two simple linear models consisting of a single variable: *mean only regression* and *regression through the origin*. Throughout we will seek a particular estimate of these models, namely the least squares estimate.

3.1 Mean Only Regression

Consider the following model:

$$y_i = \mu + \epsilon_i \qquad \text{for } i = 1, \dots n,$$

which we can alternatively write as $\mathbf{y} = \mathbf{J}_n\mu + \epsilon$, where $\mathbf{J}_n$ is a vector of ones.

We seek to minimize $f(\mu) = ||\mathbf{y} - \mathbf{J}_n\mu||^2$ with respect to μ. To do so, we begin by rewriting f as follows:

$$\begin{aligned} f(\mu) &= \mathbf{y}'\mathbf{y} - 2\mathbf{J}_n'\mathbf{y}\mu + \mathbf{J}_n'\mathbf{J}_n\mu^2 \\ &= \mathbf{y}'\mathbf{y} - 2n\bar{\mathbf{y}}\mu + n\mu^2 \end{aligned}$$

Taking derivatives of f with respect to μ we obtain:

$$\frac{df}{d\mu} = -2n\bar{y} + 2n\mu.$$

This has a root at $\hat{\mu} = \bar{y}$. Note that the second derivative is $2n > 0$. Thus, the average is the least squares estimate in the sense of minimizing the Euclidean distance between the observed data and a constant vector.

Note we can think of this as projecting our n dimensional onto the best one dimensional subspace spanned by the vector $\mathbf{J}_n$.

3.1.1 R Code

Let's use the `diamond` dataset

```
> library(UsingR); data(diamond)
> y = diamond$price; x = diamond$carat
> mean(y)
[1] 500.0833
> #using least squares
> coef(lm(y ~ 1))
[1] 500.0833
```

Thus, in this example the mean only least squares estimate obtained via `lm` is the empirical mean.

3.2 Regression Through the Origin

We now consider the regression through the origin problem. Let $\mathbf{x} = (x_1, \ldots x_n)'$ be a vector. Here we seek to minimize $f(\mu) = ||\mathbf{y} - \mathbf{x}\beta||^2$ with respect to β. Note that the pairs, (x_i, y_i) form a scatterplot. Least squares is then finding the best multiple of the $\mathbf{x}$ vector to approximate $\mathbf{y}$. That is, finding the best line of the form $\mathbf{y} = \mathbf{x}\beta$ to fit the scatter plot. Thus we are considering lines through the origin hence the name.

Note that $f(\beta) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{x}\beta + \mathbf{x}'\mathbf{x}\beta^2$. Taking derivatives with respect to μ we obtain:

$$\frac{df}{d\beta} = -2\mathbf{y}'\mathbf{x} + 2\mathbf{x}'\mathbf{x}\beta.$$

Setting this equal to zero we obtain the equation:

$$\beta = \frac{\mathbf{y}'\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{< \mathbf{y}, \mathbf{x} >}{< \mathbf{x}, \mathbf{x} >}$$

Note that the second derivative is $2\mathbf{x}'\mathbf{x} > 0$. Thus, f is convex and $\hat{\beta}$ minimizes the least-squares criteria.

Note that we can think of this as projecting our n dimensional data onto the one dimensional subspace spanned by the single vector $\mathbf{x}$, i.e. $\{\beta\mathbf{x} | \beta \in \mathbb{R}\}$.

3.2.1 R Code

Let's continue with the diamond example. We'll center the variables first.

```
> yc = y - mean(y);
> xc = x - mean(x)
> sum(yc * xc) / sum(xc * xc)
[1] 3721.025
> coef(lm(yc ~ xc - 1))
      xc
3721.025
```