Title:

Some basic statistics for Machine Learning problems

Present by:

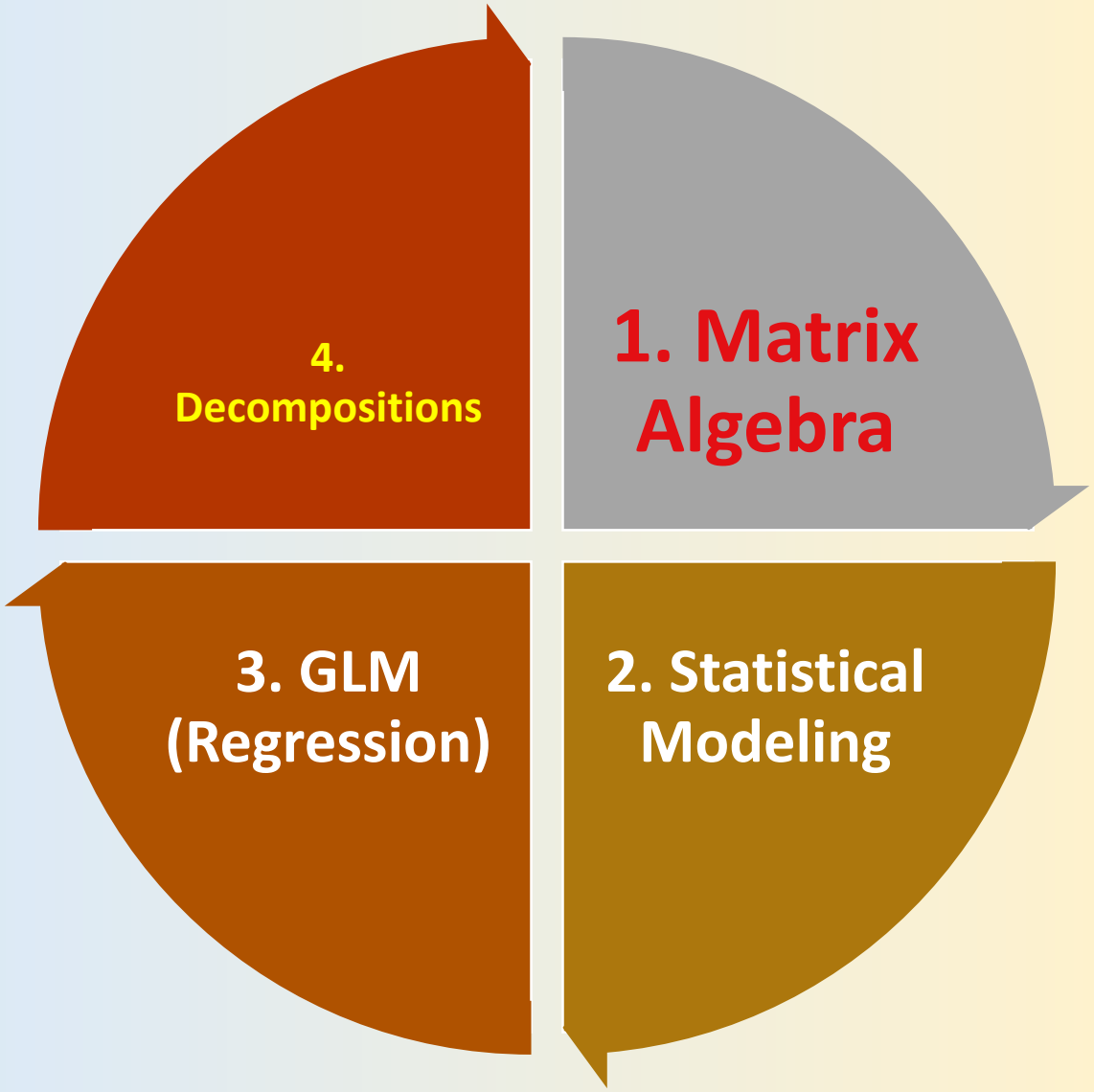
Shakib Panahbehagh

Februray 2018

The best thing about being a statistician is that you get to play in everyone's backyard!

-- John Tukey

Overview


4.
Decompositions
1. Matrix
Algebra
3. GLM
(Regression)
2. Statistical
Modeling

1. Rank of a matrix

The *rank* of any square or rectangular matrix $\mathbf{A}$ is defined as

$$\begin{aligned}\text{rank}(\mathbf{A}) &= \text{number of linearly independent columns of } \mathbf{A} \\ &= \text{number of linearly independent rows of } \mathbf{A}.\end{aligned}$$

It can be shown that the number of linearly independent columns of any matrix is always equal to the number of linearly independent rows.

If a matrix $\mathbf{A}$ has a single nonzero element, with all other elements equal to 0, then $\text{rank}(\mathbf{A}) = 1$. The vector $\mathbf{0}$ and the matrix $\mathbf{O}$ have rank 0.

Suppose that a rectangular matrix $\mathbf{A}$ is $n \times p$ of rank p, where $p < n$. (We typically shorten this statement to "$\mathbf{A}$ is $n \times p$ of rank $p < n$.") Then $\mathbf{A}$ has maximum possible rank and is said to be of *full rank*. In general, the maximum possible rank of an $n \times p$ matrix $\mathbf{A}$ is $\min(n, p)$. Thus, in a rectangular matrix, the rows or columns (or both) are linearly dependent. We illustrate this in the following example.

2. Quadratic forms

1.5 Quadratic forms

1. A **quadratic form** is a polynomial expression in which each term has degree 2. Thus, $y_1^2 + y_2^2$ and $2y_1^2 + y_2^2 + 3y_1y_2$ are quadratic forms in y_1 and y_2, but $y_1^2 + y_2^2 + 2y_1$ or $y_1^2 + 3y_2^2 + 2$ are not.
2. Let $\mathbf{A}$ be a symmetric matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix},$$

where $a_{ij} = a_{ji}$; then the expression $\mathbf{y}^T\mathbf{A}\mathbf{y} = \sum_i \sum_j a_{ij}y_iy_j$ is a quadratic form in the y_i's. The expression $(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is a quadratic form in the terms $(y_i - \mu_i)$ but not in the y_i's.

3. The quadratic form $\mathbf{y}^T\mathbf{A}\mathbf{y}$ and the matrix $\mathbf{A}$ are said to be **positive definite** if $\mathbf{y}^T\mathbf{A}\mathbf{y} > 0$ whenever the elements of $\mathbf{y}$ are not all zero. A necessary and sufficient condition for positive definiteness is that all the determinants

$$|A_1| = a_{11}, |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \ldots, \text{and}$$

$|A_n| = \det \mathbf{A}$ are positive. If a matrix is positive definite, then it can be inverted and also it has a square root matrix $\mathbf{A}^*$ such that $\mathbf{A}^*\mathbf{A} = \mathbf{A}$. These properties are useful for the derivation of several theoretical results related to estimation and the probability distributions of estimators.

4. The rank of the matrix $\mathbf{A}$ is also called the degrees of freedom of the quadratic form $Q = \mathbf{y}^T\mathbf{A}\mathbf{y}$.

3. Matrix Differentiation

Proposition 5 *Let*

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{y}$ *is* $\mathrm{m} \times 1$, $\mathbf{x}$ *is* $\mathrm{n} \times 1$, $\mathbf{A}$ *is* $\mathrm{m} \times \mathrm{n}$, *and* $\mathbf{A}$ *does not depend on* $\mathbf{x}$, *then*

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$$

Proposition 6 *Let*

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

where $\mathbf{y}$ *is* $\mathrm{m} \times 1$, $\mathbf{x}$ *is* $\mathrm{n} \times 1$, $\mathbf{A}$ *is* $\mathrm{m} \times \mathrm{n}$, *and* $\mathbf{A}$ *does not depend on* $\mathbf{x}$, *as in Proposition 5. Suppose that* $\mathbf{x}$ *is a function of the vector* $\mathbf{z}$, *while* $\mathbf{A}$ *is independent of* $\mathbf{z}$. *Then*

$$\frac{\partial \mathbf{y}}{\partial \mathbf{z}} = \mathbf{A}\frac{\partial \mathbf{x}}{\partial \mathbf{z}}$$

Proposition 8 *For the special case in which the scalar* α *is given by the quadratic form*

$$\alpha = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$$

where $\mathbf{x}$ *is* $n \times 1$, $\mathbf{A}$ *is* $n \times n$, *and* $\mathbf{A}$ *does not depend on* $\mathbf{x}$, *then*

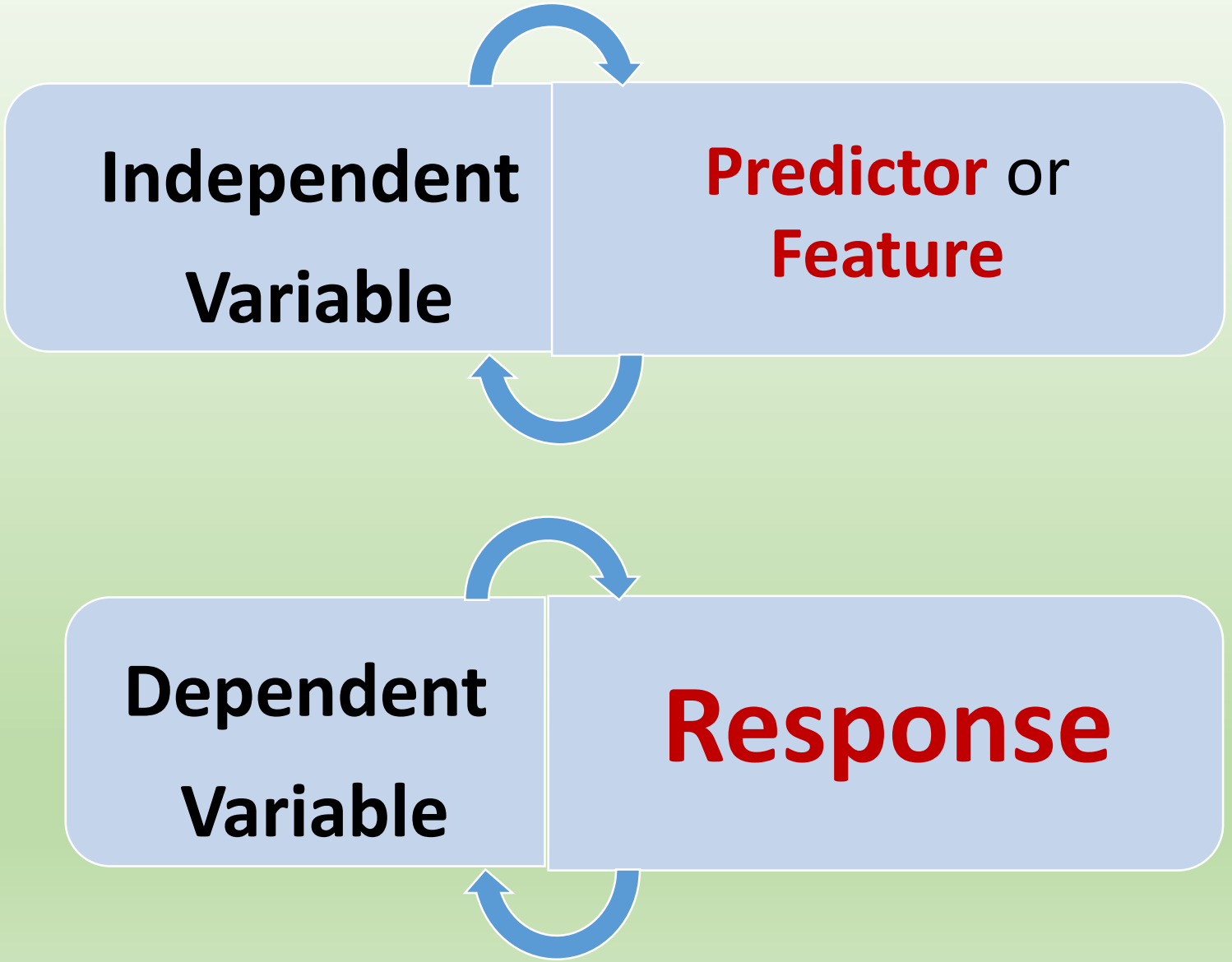
$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{x}^{\mathsf{T}}\left(\mathbf{A} + \mathbf{A}^{\mathsf{T}}\right)$$

Proposition 9 *For the special case where* $\mathbf{A}$ *is a symmetric matrix and*

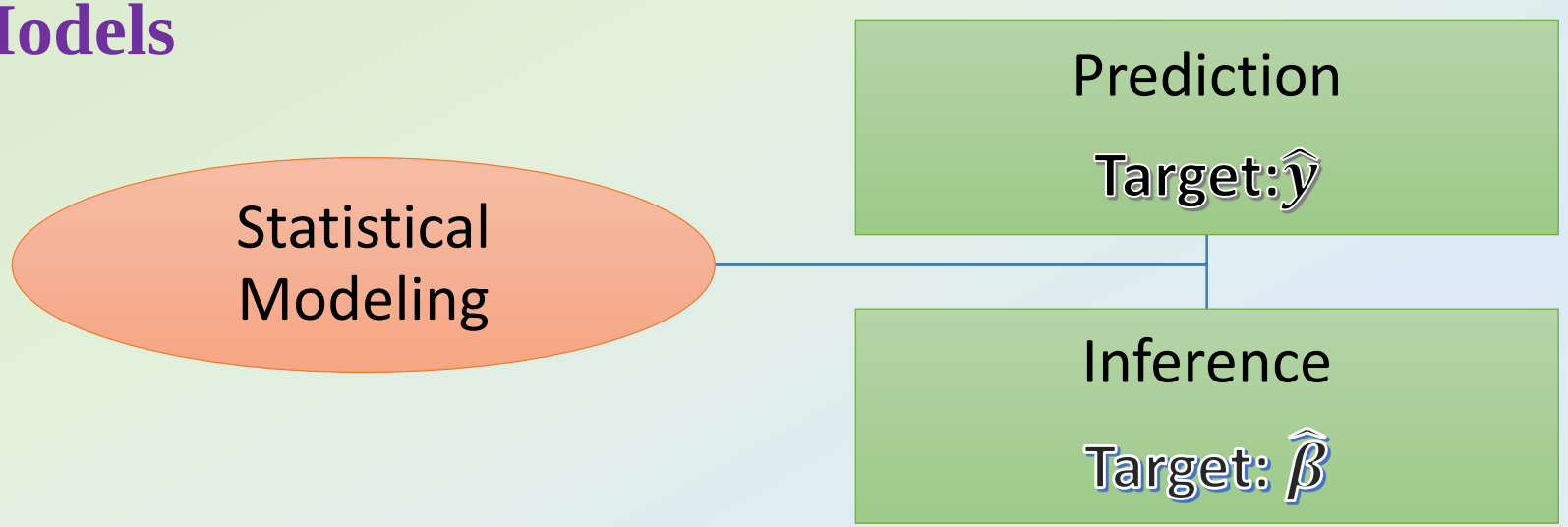
$$\alpha = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$$

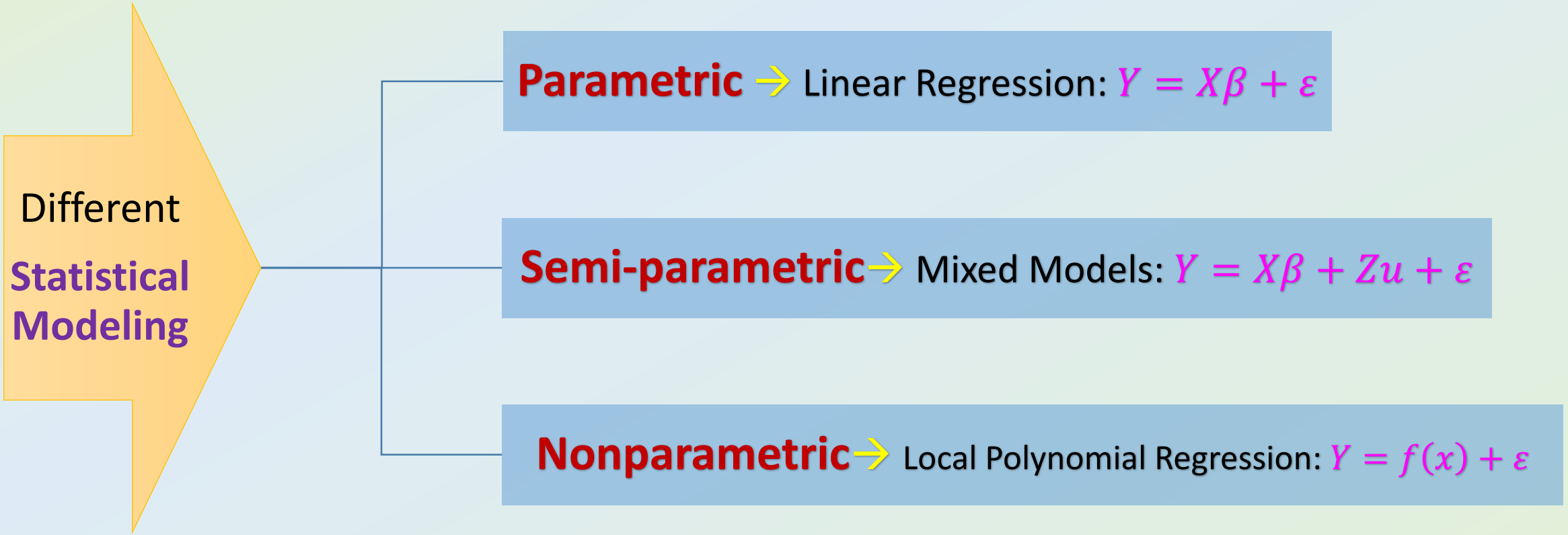
where $\mathbf{x}$ *is* $n \times 1$, $\mathbf{A}$ *is* $n \times n$, *and* $\mathbf{A}$ *does not depend on* $\mathbf{x}$, *then*

$$\frac{\partial \alpha}{\partial \mathbf{x}} = 2\mathbf{x}^{\mathsf{T}}\mathbf{A}$$


Independent Variable
Predictor or Feature
Dependent Variable
Response

4. Statistical Models





5. Simple Regression

the *simple linear regression* model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{6.1}$$

The designation *simple* indicates that there is only one x to predict the response y, and *linear* means that the model (6.1) is linear in β_0 and β_1. [Actually, it is the assumption $E(y_i) = \beta_0 + \beta_1 x_i$ that is linear; see assumption 1 below.] For example, a model such as $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is linear in β_0 and β_1, whereas the model $y_i = \beta_0 + e^{\beta_1 x_i} + \varepsilon_i$ is not linear.

In this chapter, we assume that y_i and ε_i are random variables and that the values of x_i are known constants, which means that the same values of $x_1, x_2, \ldots, x_n$ would be used in repeated sampling.

6. Basic assumptions:

1. $E(\varepsilon_i) = 0$ for all $i = 1, 2, \ldots, n$, or, equivalently, $E(y_i) = \beta_0 + \beta_1 x_i$.
2. $\text{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$, or, equivalently, $\text{var}(y_i) = \sigma^2$.
3. $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, or, equivalently, $\text{cov}(y_i, y_j) = 0$.

Assumption 1 states that the model (6.1) is correct, implying that y_i depends only on x_i and that all other variation in y_i is random. Assumption 2 asserts that the variance of ε or y does not depend on the values of x_i. (Assumption 2 is also known as the assumption of *homoscedasticity*, *homogeneous variance* or *constant variance*.) Under assumption 3, the ε variables (or the y variables) are uncorrelated with each other. In Section 6.3, we will add a normality assumption, and the y (or the ε) variables will thereby be independent as well as uncorrelated. Each assumption has been stated in terms of the ε's or the y's. For example, if $\text{var}(\varepsilon_i) = \sigma^2$, then $\text{var}(y_i) = E[y_i - E(y_i)]^2 = E(y_i - \beta_0 - \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$.

7. OLS in Simple Regression

Using a random sample of n observations $y_1, y_2, \ldots, y_n$ and the accompanying fixed values $x_1, x_2, \ldots, x_n$, we can estimate the parameters β_0, β_1, and σ^2. To obtain the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the method of least squares, which does not require any distributional assumptions (for maximum likelihood estimators based on normality, see Section 7.6.2).

In the *least-squares* approach, we seek estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{6.2}$$

Note that the predicted value $\hat{y}_i$ estimates $E(y_i)$, not y_i; that is, $\hat{\beta}_0 + \hat{\beta}_1 x_i$ estimates $\beta_0 + \beta_1 x_i$, not $\beta_0 + \beta_1 x_i + \varepsilon_i$. A better notation would be $\widehat{E(y_i)}$, but $\hat{y}_i$ is commonly used.

To find the values of β_0 and β_1 that minimize $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ in (6.2), we differentiate with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set the results equal to 0:

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0.$$

The solution to (6.3) and (6.4) is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- **Note that we need to check second derivation to be sure about minimum points.**

8. Mean and Variance of LS-estimators

$$E(\hat{\beta}_1) = \beta_1$$

$$E(\hat{\beta}_0) = \beta_0$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right].$$

- **What is this equations means?**

9. General Linear Model

The general multiple regression model with k predictor variables is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i. \qquad (2.4)$$

Here, for $j = 1, \ldots, k$, we have x_{ij} as the value of the jth predictor variable for the ith case, $i = 1, \ldots, n$.

This model can be written in matrix notation as

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

The ordinary least-squares (OLS) estimator minimizes

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

where $\|\mathbf{v}\| = \sqrt{\mathbf{v}^{\mathsf{T}}\mathbf{v}}$ is the "length" of the vector $\mathbf{v}$. As seen in the preceding examples, the least-squares estimator of $\boldsymbol{\beta}$ is

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

The error vector $\boldsymbol{\varepsilon}$ has zero mean: $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$. Additional assumptions about $\boldsymbol{\varepsilon}$ are often made. The first, often called *homoscedasticity*, is that

$$\text{var}(\varepsilon_i) = \sigma^2 \quad \text{for all } 1 \leq i \leq n.$$

It is also usual to assume that the errors are uncorrelated:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j.$$

Conditions (2.5) and (2.6) can be summarized by the expression

$$\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}.$$

Finally, there is the *normality assumption*; given (2.7), this translates to

$$\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

10. Regression Diagnostics

The term *regression diagnostics* refers to a large collection of techniques used to check the quality of the data and the adequacy of a regression model. Often data are misrecorded or do not come from the population of interest.

The two basic components of many diagnostics are the *fitted values* and the *residuals*. The ith fitted value is the estimate of $\mathsf{E}(y_i)$ from the model; that is, $\hat{y}_i = \mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}}$, where as before $\mathbf{x}_i^\mathsf{T}$ is the ith row of $\mathbf{X}$. The vector of all n fitted values is

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}, \tag{2.9}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}.$$

The matrix $\mathbf{H}$ is called the *hat matrix* since multiplication by $\mathbf{H}$ converts $\mathbf{y}$ to $\hat{\mathbf{y}}$. As will be seen in the following chapters, the hat matrix plays an extremely important role in regression theory and practice.

The ith residual is defined to be

$$e_i = y_i - \hat{y}_i,$$

Most of the information for determining the adequacy of a linear regression model is contained in the residuals, since these estimate that part of the model that was assumed to be random. Therefore, any patterns in the residuals reflect extra structure that is not accommodated by the model. Residual analysis for diagnosis of linear regression models is a very large topic

Noting that the vector of residuals is

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

and using

$$\mathbf{H} = \mathbf{H}^{\mathsf{T}} = \mathbf{H}^2,$$

we have

$$\mathrm{Cov}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\,\mathrm{Cov}(\mathbf{y})(\mathbf{I} - \mathbf{H})^{\mathsf{T}} = \sigma^2(\mathbf{I} - \mathbf{H}).$$

A natural question is: How much influence do these possible outliers have on the fitted line? Two diagnostics that are in common use to assessing influence are:

- the *hat diagonals* or *leverages*, which measure the potential of outliers in the predictors to influence the fit; and
- *Cook's distance* (also called *Cook's D*), which measures actual influence of an observation on the fit.

The ith *leverage* value is the ith diagonal of the hat matrix, H_{ii}. We know from (2.9) that the ith fitted value is

$$\hat{y}_i = \sum_{j=1}^{n} H_{ij} y_j = H_{i1} y_1 + \cdots + H_{ii} y_i + \cdots + H_{in} y_n, \qquad (2.12)$$

so that H_{ii} is the weight of y_i in the expression for $\hat{y}_i$, that is, the influence of y_i on its own fitted value. It should be appreciated that H_{ii} depends only on the predictors, not on the ys, so that H_{ii} measures only the *potential* for being influential and not actual influence.

11. Decompositions

Algorithm (Cholesky Least Squares)

(0) Set up the problem by computing A^*A and $A^*\boldsymbol{b}$.

(1) Compute the Cholesky factorization $A^*A = R^*R$.

(2) Solve the lower triangular system $R^*\boldsymbol{w} = A^*\boldsymbol{b}$ for $\boldsymbol{w}$.

(3) Solve the upper triangular system $R\boldsymbol{x} = \boldsymbol{w}$ for $\boldsymbol{x}$.

The operations count for this algorithm turns out to be $\mathcal{O}(mn^2 + \frac{1}{3}n^3)$.

Remark The solution of the normal equations is likely to be unstable. Therefore this method is *not recommended* in general. For small problems it is usually safe to use.

Solving linear equations by LU factorization

solve $Ax = b$ with A nonsingular of order n

factor-solve method using LU factorization

1. factor A as $A = PLU$ ($(2/3)n^3$ flops)
2. solve $(PLU)x = b$ in three steps
   - permutation: $z_1 = P^T b$ (0 flops)
   - forward substitution: solve $Lz_2 = z_1$ (n^2 flops)
   - back substitution: solve $Ux = z_2$ (n^2 flops)

total cost: $(2/3)n^3 + 2n^2$ flops, or roughly $(2/3)n^3$

this is the standard method for solving $Ax = b$

Algorithm (QR Least Squares)

(0) Set up the problem by computing A^*A and $A^*\boldsymbol{b}$.

(1) Compute the reduced QR factorization $A = \hat{Q}\hat{R}$.

(2) Compute $\hat{Q}^*\boldsymbol{b}$.

(3) Solve the upper triangular system $\hat{R}\boldsymbol{x} = \hat{Q}^*\boldsymbol{b}$ for $\boldsymbol{x}$.

This is well-defined since $\hat{R}^{-1}$ exists because A has full rank.

The operations count (using Householder reflectors to compute the QR factorization) is $\mathcal{O}(2mn^2 - \frac{2}{3}n^3)$.

Remark This approach is more stable than the Cholesky approach and is considered the *standard method for least squares problems*.

Algorithm (SVD Least Squares)

(1) Compute the reduced SVD $A = \hat{U}\hat{\Sigma}V^*$.

(2) Compute $\hat{U}^*\boldsymbol{b}$.

(3) Solve the diagonal system $\hat{\Sigma}\boldsymbol{w} = \hat{U}^*\boldsymbol{b}$ for $\boldsymbol{w}$.

(4) Compute $\boldsymbol{x} = V\boldsymbol{w}$.

This time the operations count is $\mathcal{O}(2mn^2 + 11n^3)$ which is comparable to that of the QR factorization provided $m \gg n$. Otherwise this algorithm is more expensive, but also more stable.

If $\operatorname{rank}(A) < n$ (which is possible even if $m < n$, i.e., if we have an *underdetermined* problem), then infinitely many solutions exist.

A common approach to obtain a well-defined solution in this case is to add an additional constraint of the form

$$\|\boldsymbol{x}\| \longrightarrow \min,$$

i.e., we seek the *minimum norm solution*.

12. Important considerations

Based on the *Orders* we saw:

- **Cholesky** is faster than **LU**
- **LU** is faster than **QR**
- **QR** is faster than SVD

Note: direction of stability is totally in the opposite direction. In the other hands the most stable decomposition is ***SVD***.

The best thing about being a statistician is that you get to play in everyone's backyard!

-- John Tukey

Thank You,
Have fun…