SF 2930 Regression Analysis

Lecture 3

Multiple Linear Regression, Part 1: Introduction

Timo Koski

KTH Royal Institute of Technology

2023

Learning outcomes

- Ordinary multiple regression model, k regressors x_i, real valued response Y
- Normal multiple regression model, k regressors x_i, real valued response Y with normal distribution
- LSE $\hat{\beta}$ of the regression parameters
- Geometry of LSE , the hat matrix H, $H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$
- Properties of $\hat{\beta}$: mean (unbiasedness), covariance matrix, estimation of the variance $sigma^2$
- Properties of LSE residuals.

The material of the present lecture and the next lecture is covered with time 1:22:12 in the following item from MITOpenCourseWare

MIT 18.S096 Topics in Mathematics with Applications in Finance, Fall 2013
Peter Kempthorne Lecture 6: Regression Analysis
`https://www.youtube.com/watch?v=l1kLCrxL9Hk`

MIT
OCW

We have the training set

$$\mathcal{D}_{tr} := \left\{ \left(y_i, \, x_{i1}, \ldots, x_{ij} \right)_{i=1}^{n} \right\}_{j=1}^{k}$$

sampled n times from a source. The y_is are n outcomes/instantiations of the dependent response variable Y and are x_{ij} are corresponding instantiations of the k explanatory variables, or covariates, or, prediction variables $x_1, \ldots, x_k$.
A multiple linear regression model treats the relationship between the dependent response variable y and the k of expanatory variables as linear.
This relationship is enhanced with a statistical model through a disturbance term or error variable ε_i for each y_i — an *unobserved random variable* that adds "*noise*" to the linear relationship.

Measurements with Noise, A Case

In many situations we think first of

$$Y(t) = f(\boldsymbol{\beta}, t) + \varepsilon(t) \tag{1}$$

We take a finite set of basis functions $\{\phi_j(t)\}_{j=1}^{k}$ and write as our model

$$Y(t) = \beta_0 + \sum_{i=1}^{k} \beta_j \phi_j(t) + \varepsilon(t) \tag{2}$$

The observations: we sample n times the response and covariates at $t_1, \ldots, t_n$ and set $y_i = Y(t_i)$, $x_{ij} = \phi_j(t_i)$, $\varepsilon_i = \varepsilon(t_i)$ for $i = 1, \ldots, t_n$ $j = 1, \ldots, k$. Hence we obtain the ordinary multiple regression model equations:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

Multiple Linear Regression Model

$\mathcal{D}_{tr}$ is given. Multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $^{\mathsf{T}}$ is the transpose, and $\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}$ are the scalar products of the vectors $\mathbf{x}_i \in \mathbb{R}^{k+1}$ and $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. $n > k + 1$. These n equations are conveniently written in a compact matrix notation as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where (next slide)

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

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

X: Design Matrix

In general: X is a $p \times p$ square matrix whose entries are either $+1$ or -1 and whose rows are mutually orthogonal $\Rightarrow XX^T = p\mathbf{I}_p$. For example:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Observational X, Mechanical Model

The determination of the Earth's gravity field from highly accurate satellite measurements. The model[1] for gravitational potential is

$$V(r, \theta, \lambda) = \frac{GM}{R} \sum_{l=0}^{L} \left(\frac{r}{R}\right)^{l} \sum_{m=0}^{l} P_{lm}(\cos(\theta)) \left[C_{lm} \cos(m\lambda) + S_{lm}(m\lambda)\right]$$

where G is the gravitational constant, M is the Earth's mass, R is the Earth's reference radius, P_{lm} represents the fully normalized l-degree Legendre polynomials of order m, andC_{lm} and S_{lm} are the corresponding normalized harmonic coefficients. For the mission, the chosen value for L is 300.

[1] Duff, Iain S and Gratton, Serge: The parallel algorithms team at CERFACS, SIAM News, 39, 10, 2006

$$V(r,\theta,\lambda) = \frac{GM}{R}\sum_{l=0}^{L}\left(\frac{r}{R}\right)^{l}\sum_{m=0}^{l} P_{lm}(\cos(\theta))\left[C_{lm}\cos(m\lambda) + S_{lm}(m\lambda)\right]$$

We consider the following parameter estimation problem: Find the harmonic coefficients C_{lm} and S_{lm} as accurately as possible, using the satellite observations. This results in a linear least-squares problem involving millions of equations and 90,000 unknowns that engineers will need to solve on a daily basis on an eight-processor Power 5 IBM machine.

$n \geq k+1$, $n << k$

Remark

Assume $n \geq k+1$. What is the mathematical relationship between k and n? A formula or rule for this?

Remark

$n << k$ (meaning that n is much smaller than k). This is a situation of **big data***, i.e., an observational study with a huge number of possibly relevant explanatory/predictive factors observed. Multiple regression in this case is treated later in Lecture XXXX*

The (Ordinary) Multiple Linear Regression Model

$\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ and $n > k + 1$.

$$\mathbf{Y} = X\boldsymbol{\beta} + \varepsilon. \tag{3}$$

The following assumptions hold:

1) $E\left[\varepsilon\right] = \mathbf{0} \in \mathbb{R}^n$
2) $C_\varepsilon = E\left[\varepsilon\varepsilon^T\right] = \sigma^2 \mathbb{I}_n$ *(homoscedasticity)*
3) $X^T X$ *is invertible (to be discussed below)*

The model is called ordinary normal regression model, if additionally the following the following assumption holds:

4) $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$

The (Ordinary) Multiple Linear Regression Model

Remark

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

The assumption 4), i.e., $\varepsilon \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$ implies by the results in Lecture 2. that

$$\mathbf{Y} \sim N_n\left(X\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right).$$

Residuals

Let us now fix an arbitrary value of $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. Then we can compute the values of the **observed residuals**

$$e_i := y_i - \mathbf{x}_i^T \boldsymbol{\beta}, \quad i = 1, \ldots, n.$$

These are estimates of $\varepsilon_i \quad i = 1, \ldots, n$. Set

$$\mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

Now we can write

$$\mathbf{e} = \mathbf{y} - X\boldsymbol{\beta}$$

Least Squares Estimation (LSE)

We want to estimate β based on the training set $\mathcal{D}_{tr}$. One (respected and by Lecture 1 well known) way to do this is to minimize the squared norm (length) of the observed residuals:

$$\parallel \mathbf{e} \parallel^2 = \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$$

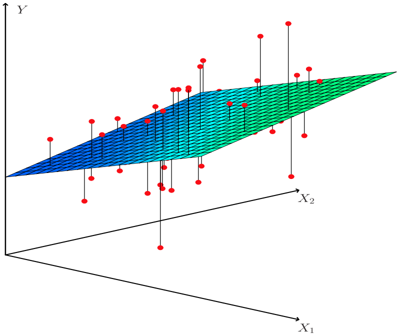
i.e.,

$$\widehat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{B}} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$$

But by the definition of the norm $\parallel \cdot \parallel$ on the Euclidean space $\mathbb{R}^n$ we find

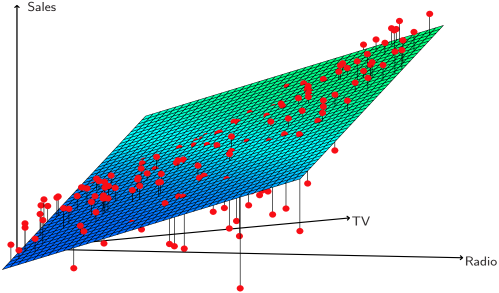
$$\parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right)^2 .$$

Hence this is nothing but an extension of the LSE in Lecture 1. (Set k=1 to regain the simple linear regression therein).



In three dimensional setting, with one response and two predictors $k = 2$, LSE fits a plane to the training data. [2]

[2] by Courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert *An introduction to statistical learning*, Chapter 3



In three dimensional setting, i.e. with one response and $k = 2$, LSE fits a plane to the training data [3]

[3] by Courtesy of James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert *An introduction to statistical learning*, Chapter 3

OLS Linear Regression is Adaline[4]

Consider our implementation of the ADAptive LInear NEuron (Adaline) from Chapter 2, Training Machine Learning Algorithms for Classification; we remember that the artificial neuron uses a linear activation function and we defined a cost function (=Q in this lecture T.K.), which we minimized to learn the weights via optimization algorithms, such as Gradient Descent (GD) and Stochastic Gradient Descent (SGD). This cost function in Adaline is the Sum of Squared Errors (SSE).

[4] p. 285 in Sebastian Raschka: *Python Machine Learning*. PACKT Publishing, Birmingham, 2015

OLS Linear Regression is Adaline..[5]

Essentially, OLS linear regression can be understood as Adaline without the unit step function so that we obtain continuous target values instead of the class labels -1 *and* 1. *To demonstrate the similarity, let's take the GD implementation of Adaline from Chapter 2, Training Machine Learning Algorithms for Classification, ...'*

[5] p. 285 in Sebastian Raschka: **Python Machine Learning**. PACKT Publishing, Birmingham, 2015

OLS Linear Regression is Adaline..[6]

As an alternative to using machine learning libraries, there is a closed-form solution for solving OLS involving a system of linear equations that can be found in most introductory statistics textbooks . . .

If you are interested in more information on how to obtain the normal equations, I recommend you take a look at

[6] p. 290 in Sebastian Raschka: **Python Machine Learning**. PACKT Publishing, Birmingham, 2015

Stanford Online

Andrew Ng (Adjunct Professor of Computer Science) lecturing **Stanford CS229: Machine Learning - Linear Regression and Gradient Descent** | Lecture 2 (Autumn 2018)

https://www.youtube.com/watch?v=4b4MUYve_U8&t=133s

Full Column Rank Implies that $X^T X$ is Positive Definite and Invertible

In order to assist machine learning libraries and to find the closed-form solution, we need an assumption.

Definition

An $n \times k$ matrix X has full column rank as soon as the k columns of X are linearly independent.

Lemma

Let A be any $n \times p$ matrix. Then

i) *$A^T A$ is symmetric positive semidefinite.*

ii) *If A has full column rank, then $A^T A$ is symmetric and positive definite.*

Full Column Rank Implies that $X^T X$ is Positive Definite and Invertible

The proof of *II*) is a short and clear one, and is recapitulated in the Appendix XXXX.

COROLLARY

If X has full column rank, then $X^T X$ is invertible.

Proof: By the preceding lemma, if X has full column rank, then $X^T X$ is positive definite. Hence $X^T X$ has positive determinant, and is therefore invertible. □

EXQ Full Column Rank for Simple Linear Regression

Consider X in simple linear regression, or, as seen in Lecture 2.:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} = (\mathbf{1}_n, \mathbf{x})$$

In which case shall X with $k+1=2$ not have full rank? Give your answer in terms of the $n \times 1$ vectors $\mathbf{1}_n$ and $\mathbf{x}$.

Least Squares Estimation

Hence we set, for simplicity of writing,

$$Q(\beta) := \| \mathbf{y} - X\beta \|^2 \tag{4}$$

and the LSE is the minimizer

$$\widehat{\beta} := \mathrm{argmin}_{\beta} Q(\beta) .$$

Proposition

If X has full column rank, then

$$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}. \tag{5}$$

$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$

Remark

Check that the matrix multiplications are compatible.

Remark

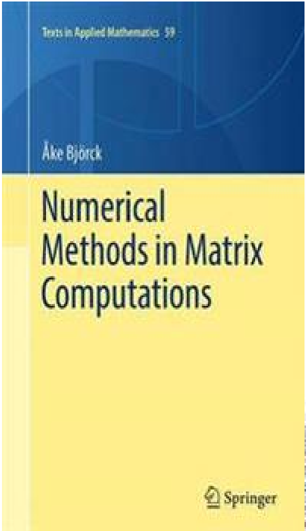
If k is large, how hard is it to invert $X^T X$ computationally? Computational maths & algorithmics reguested for.

$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ and Neural Engines (!?)

> However directly forming $X^T X$ is unstable for all but the most well-conditioned systems; in practice we would avoid forming $X^T X$ directly. A much more reliable and accurate method is based on QR factorization.

Citation from Zhang, Shaoshuai and Baharlouei, Elaheh and Wu, Panruo: **High accuracy matrix computations on neural engines: A study of QR factorization and its applications.** *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, pp. 17–28, June 23-26, Stockholm, 2020.
On QR : Chapter 2.3 in Åke Björck: *Numerical Methods in Matrix Computations. Springer 2015*

Computational Methods for $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$



Texts in Applied Mathematics 59

Åke Björck

Numerical Methods in Matrix Computations

Springer

See chapter 2 for matrix calculus in linear regression.

Proof of $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y} = \mathrm{argmin}_\beta Q(\beta)$

By the corollary above, $X^T X$ is invertible and $\widehat{\beta}$ is well defined. Take any β and write

$$\mathbf{y} - X\boldsymbol{\beta} = \mathbf{y} - X\widehat{\boldsymbol{\beta}} + X\widehat{\boldsymbol{\beta}} - X\boldsymbol{\beta}$$

Let us set for ease of writing $U := \mathbf{y} - X\widehat{\boldsymbol{\beta}}$ and $V := X\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. By definition of the norm

$$\parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta})$$

Hence we have

$$\begin{aligned} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 &= (U+V)^T(U+V) = (U^T + V^T)(U+V) \\ &= U^T U + U^T V + V^T U + V^T V \end{aligned}$$

Set $\mathbf{e}_{LSE} := \mathbf{y} - X\widehat{\boldsymbol{\beta}}$. Thus

$$U^T U = \left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right)^T \left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right) = \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}.$$

Proof of $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y} = \operatorname{argmin}_{\beta} Q(\beta)$

Next

$$V^T V = \left(X\left(\widehat{\beta}-\beta\right)\right)^T X\left(\widehat{\beta}-\beta\right) = \left(\widehat{\beta}-\beta\right)^T X^T X\left(\widehat{\beta}-\beta\right).$$

In the above, $U^T V = V^T U$, since these are scalar products. Let us expand $U^T V$.

$$U^T V = \left(\mathbf{y} - X\widehat{\beta}\right)^T X\left(\widehat{\beta}-\beta\right) = \left(X^T\left(\mathbf{y} - X\widehat{\beta}\right)\right)^T\left(\widehat{\beta}-\beta\right)$$

$$= \left(X^T\mathbf{y} - X^T X\widehat{\beta}\right)^T\left(\widehat{\beta}-\beta\right)$$

Proof of $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y} = \operatorname{argmin}_{\beta} Q(\beta)$

But $X^T X\widehat{\beta} = X^T\mathbf{y}$ and thus the last expression equals zero, since

$$\left(X^T\mathbf{y} - X^T X\widehat{\beta}\right)^T \left(\widehat{\beta} - \beta\right) = \mathbf{0}_k^T \left(\widehat{\beta} - \beta\right) = 0.$$

Hence we have established the following decomposition

$$Q(\beta) = \parallel \mathbf{y} - X\beta \parallel^2 = \mathbf{e}_{LSE}^T \mathbf{e}_{LSE} + \left(\widehat{\beta} - \beta\right)^T X^T X \left(\widehat{\beta} - \beta\right) \qquad (6)$$

In the right hand side we have two non-negative terms. The first term in the right hand side of (6) does not depend on β. Hence we can minimize the expression by minimizing the second term, which is a quadratic form.

Apply the lemma *II*) to X, which assumed to have full column rank $p = k + 1$. Then $X^T X$ is a positive definite matrix, hence the quadratic form is zero if and only if we choose $\beta - \widehat{\beta}$ to be the zero vector $\mathbf{0}_k$. Hence we have shown the proposition as claimed. *Q.E.D*.

An Additional Insight to $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$

Let the $k+1$ columns of the design matrix X be denoted by

$$\mathbf{x}_0^{(c)} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \quad \mathbf{x}_j^{(c)} = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad j = 1, \ldots, k, \quad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}.$$

Let $\mathrm{sp}\,(X) = \mathrm{sp}\{\mathbf{x}_0^{(c)}, \mathbf{x}_1^{(c)} \ldots, \mathbf{x}_k^{(c)}\}$ be the linear span of the columns of X. Then $X\widehat{\boldsymbol{\beta}} = \sum_{j=0}^{k} \mathbf{x}_j^{(c)} \widehat{\beta}_j \in \mathrm{sp}\,(X)$.

Hat Matrix

$$H := X(X^T X)^{-1} X^T. \tag{7}$$

Then the predicted values are $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y}$.

$$\widehat{\mathbf{y}} = H\mathbf{y} \in \mathrm{sp}\,(X)\,.$$

H is called[7] **hat matrix**

[7] MPV p. 73: Hat matrix and its properties play a central role in regression analysis. But MVP makes a suboptimal use of *H*

Properties of the Hat Matrix H

A) H is symmetric, i.e., $H^T = H$

B) H is idempotent, i.e., $H^2 = HH = H$

C) $\mathrm{Tr}(H) = k + 1$

Proofs are found in the Appendix XX. The rule C) will be manifest in Lecture 4. From linear algebra we recall by A) and B):

H is an orthogonal projection matrix in $\mathbb{R}^n$

Least Square Residuals

The $n \times 1$ vector $\mathbf{e}_{LSE}$ of observed residuals that correspond to the least square minimizer as computed by

$$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\beta} = \mathbf{y} - H\mathbf{y} \quad (= \mathbf{y} - \widehat{\mathbf{y}}).$$

We have now

$$\mathbf{e}_{LSE}^{T}\widehat{\mathbf{y}} = \mathbf{e}_{LSE}^{T}H\mathbf{y} = \mathbf{y}^{T}H\mathbf{y} - \mathbf{y}^{T}H^{T}H\mathbf{y}$$

and since H is symmetric, A) above,

$$= \mathbf{y}^{T}H\mathbf{y} - \mathbf{y}^{T}HH\mathbf{y}$$

and since H is idempotent, B) above,

$$= \mathbf{y}^{T}H\mathbf{y} - \mathbf{y}^{T}H\mathbf{y} = 0.$$

That is,

$$\mathbf{e}_{LSE}^{T}H\mathbf{y} = 0.$$

Fitted values $\widehat{\mathbf{y}}$ *are orthogonal to the LSE residuals* $\mathbf{e}_{LSE}$*.*

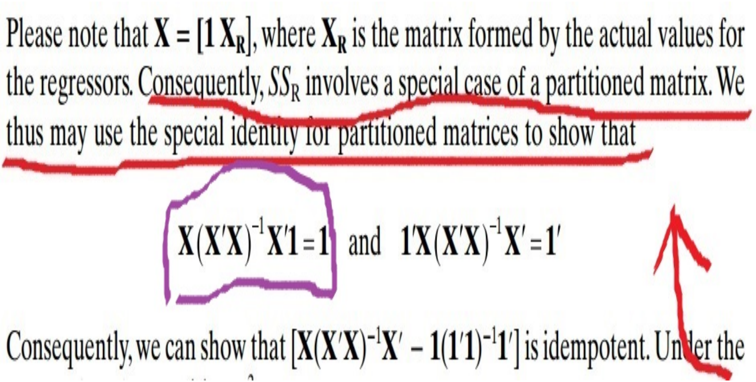
Orthogonal Projection

$$\mathbf{e}_{LSE} := \mathbf{y} - X\widehat{\beta} = \mathbf{y} - H\mathbf{y} \Leftrightarrow \mathbf{y} = H\mathbf{y} + \mathbf{e}_{LSE}$$
$$\mathbf{e}_{LSE}^{T} H\mathbf{y} = 0.$$

The LSE vector of residuals $\mathbf{e}_{LSE}$ is orthogonal to $\mathrm{sp}\,(X)$. $H\mathbf{y} \in \mathrm{sp}\,(X)$ is the orthogonal projection of $\mathbf{y}$ to $\mathrm{sp}\,(X)$.

Hence $H^2\mathbf{y} = H\mathbf{y}$ by idempotence is a natural property, why ?

MPV p. 581



Please note that $\mathbf{X} = [\mathbf{1}\ \mathbf{X}_R]$, where $\mathbf{X}_R$ is the matrix formed by the actual values for the regressors. Consequently, SS_R involves a special case of a partitioned matrix. We thus may use the special identity for partitioned matrices to show that

$$\boxed{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = \mathbf{1}} \quad \text{and} \quad \mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{1}'$$

Consequently, we can show that $[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']$ is idempotent. Under the

MPV P. 581

In our typing the claim of MPV in the lila 'box' is

$$X(X^TX)^{-1}X^T\mathbf{1}_n = \mathbf{1}_n \Leftrightarrow H\mathbf{1}_n = \mathbf{1}_n \tag{8}$$

where

$$\mathbf{1}_n := \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Note that $\mathbf{1}_n$ is the first column in X. Hence $\mathbf{1}_n \in \mathrm{sp}\,(X)$ and thus $H\mathbf{1}_n = \mathbf{1}_n$ follows, since H projects a vector in $\mathrm{sp}\,(X)$ to the vector itself (idempotence).

MPV P. 581

More formally: The orthogonal projector represented by the matrix H splits up $\mathbb{R}^n$ into a direct sum of two orthogonal subspaces, $\mathrm{sp}\,(X)$ and its orthogonal complement $\mathrm{sp}\,(X)^{\perp}$, i.e., any $\mathbf{e} \in \mathrm{sp}\,(X)^{\perp}$ is orthogonal to every $\mathbf{z} \in \mathrm{sp}\,(X)$, and every $\mathbf{y} \in \mathbb{R}^n$ is uniquely decomposed as

$$\mathbf{y} = H\mathbf{y} + \mathbf{e},$$

where $H\mathbf{y} \in \mathrm{sp}\,(X)$, and $\mathbf{e} \in \mathrm{sp}\,(X)^{\perp}$. As $\mathbf{1}_n \in \mathbb{R}^n$, we set $\mathbf{y} = \mathbf{1}_n$ and thus

$$\mathbf{1}_n = H\mathbf{1}_n + \mathbf{e}.$$

Hence $\mathbf{e}^T\mathbf{1}_n = \mathbf{e}^T H\mathbf{1}_n + \mathbf{e}^T\mathbf{e}$. Since $\mathbf{1}_n \in \mathrm{sp}\,(X)$, $\mathbf{e}^T\mathbf{1}_n = 0$. Since $H\mathbf{1}_n \in \mathrm{sp}\,(X)$, $\mathbf{e}^T H\mathbf{1}_n = 0$. Hence $\mathbf{e}^T\mathbf{e} = 0$, i.e. $\parallel \mathbf{e} \parallel^2 = 0$. But this means that $\mathbf{e} = \mathbf{0}_n$. Thus we have

$$\boxed{H\mathbf{1}_n = \mathbf{1}_n. \qquad (9)}$$

MPV p. 581

Now we may write the expression in MVP as

$$X(X^TX)^{-1}X^T - \mathbf{1}_n\left(\mathbf{1}_n\mathbf{1}_n^T\right)^{-1}\mathbf{1}_n^T = H - \frac{1}{n}\mathbf{1}_n\mathbf{1}^T. \tag{10}$$

In Lecture 5 the need arises to find, whether the matrix $H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is idempotent. Let us not keep ourselves in suspense.

$H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ IS IDEMPOTENT

$$\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}^T\right)\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) = HH - \frac{1}{n}H\mathbf{1}_n\mathbf{1}_n^T - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T H + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\mathbf{1}_n\mathbf{1}_n^T.$$

But $HH = H$, since H is idempotent, $H\mathbf{1}_n\mathbf{1}_n^T = \mathbf{1}_n\mathbf{1}_n^T$ by (9). Next, $\mathbf{1}_n\mathbf{1}_n^T H = \mathbf{1}_n\mathbf{1}_n^T$, since $\mathbf{1}_n^T H = \left(H^T\mathbf{1}_n\right)^T = \left(H\mathbf{1}_n\right)^T = \mathbf{1}_n^T$, as H is symmetric and by (9). As pointed out earlier, and is seen immediately, $\mathbf{1}_n^T\mathbf{1}_n = n$. Hence $\mathbf{1}_n\mathbf{1}^T\mathbf{1}_n\mathbf{1}_n^T = n\mathbf{1}_n\mathbf{1}_n^T$. Therefore and we have established the desired idempotence.

EXQ Pythagoras's Theorem

Show that

$$\| \mathbf{y} \|^2 = \| \hat{\mathbf{y}} \|^2 + \| \mathbf{e}_{LSE} \|^2$$

The data point $\mathbf{y}$ *is the hypotenuse of the right-angled triangle in* $\mathbb{R}^n$ *with the base of predicted/fitted values* $\hat{\mathbf{y}}$ *and the altitude of the LSE- residual* $\mathbf{e}_{LSE}$*. This is next illustrated in a Figure.*

By Courtesy of Puntanen, S. and Isotalo, J. and Styan, GPH: *Formulas Useful for Linear Regression Analysis and Related Matrix Theory*. In the Figure $\hat{\varepsilon} \leftrightarrow \mathbf{e}_{LSE}$

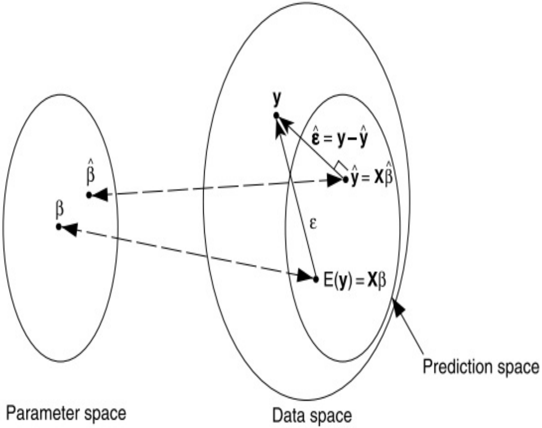

$\mathbf{y}$
$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$
$\hat{\beta}$
$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
β
ε
$E(\mathbf{y}) = \mathbf{X}\beta$
Prediction space
Parameter space
Data space

Figure 7.4 Geometric relationships of vectors associated with the multiple linear regression model.

Projection Geometrically for Simple Linear Regression

In the next figure

$$\mathcal{C}\left(\mathbf{1};\mathbf{x}\right) \leftrightarrow sp(X)$$

$$\mathbf{J} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$$

Projection Geometrically for Simple Linear Regression

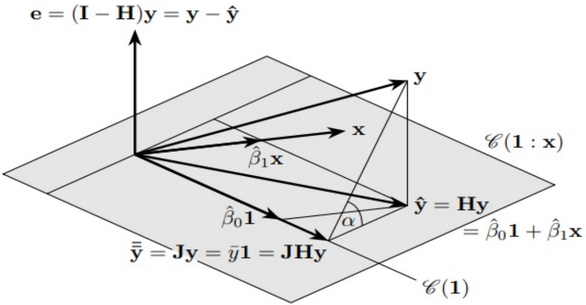

$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}}$
$\mathbf{y}$
$\mathbf{x}$
$\mathscr{C}(\mathbf{1} : \mathbf{x})$
$\hat{\beta}_1\mathbf{x}$
$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$
$= \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}$
$\hat{\beta}_0\mathbf{1}$
α
$\bar{\bar{\mathbf{y}}} = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1} = \mathbf{J}\mathbf{H}\mathbf{y}$
$\mathscr{C}(\mathbf{1})$

Figure 8.3 Projecting $\mathbf{y}$ onto $\mathscr{C}(\mathbf{1} : \mathbf{x})$.

From Puntanen S., Styan G.P.H., Isotalo J.: *Matrix Tricks for Linear Statistical Models*. Springer 2011.

The (Ordinary) Normal Multiple Linear Regression Model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{11}$$

where

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$$

Then by known properties of the multivariate normal distribution

$$\mathbf{Y} \sim N_n\left(X\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right) \tag{12}$$

and, since $\left(\sigma^2 \mathbb{I}_n\right)^{-1} = \sigma^{-2}\mathbb{I}_n$ and $\det \sigma^2 \mathbb{I}_n = \sigma^{2n}$

$$f_{\mathbf{Y}}\left(\mathbf{y}\right) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2} \|\mathbf{y} - X\boldsymbol{\beta}\|^2} \tag{13}$$

The Maximum Likelihood Estimator (MLE) for Normal Multiple Linear Regression Model 1

We have the $-1\cdot$ loglikelihood function

$$l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right) := -\ln f_{\mathbf{Y}}\left(\mathbf{y}\right) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$$

We choose $\boldsymbol{\beta}$ and σ^2 so that $l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right)$ is minimized. The maximum likelihood estimates $\widehat{\boldsymbol{\beta}}_{ML}$ and $\widehat{\sigma}_{ML}$ are found by computing the gradient

$\nabla l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2)\right) = \left(\frac{\partial}{\partial\beta_0} l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right),\ldots,\frac{\partial}{\partial\beta_k} l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right),\frac{\partial}{\partial\sigma^2} l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right)\right)^T$, and solving $\nabla l_{\mathbf{y}}\left(\widehat{\boldsymbol{\beta}}_{ML},\widehat{\sigma}_{ML}\right) = \mathbf{0}_{k+2}$.

MLE 2

$$l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right) := -\ln f_{\mathbf{Y}}\left(\mathbf{y}\right) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) + \frac{1}{2\sigma^2} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$$

For $j = 0, 1, \ldots, k$ it holds that

$$\frac{\partial}{\partial\beta_j} l_{\mathbf{y}}\left(\boldsymbol{\beta},\sigma^2\right) = \frac{1}{2\sigma^2}\frac{\partial}{\partial\beta_j} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2 = \frac{1}{2\sigma^2}\frac{\partial}{\partial\beta_j} Q\left(\boldsymbol{\beta}\right)$$

by (4). Hence $\widehat{\boldsymbol{\beta}}_{ML}$ can be found by solving first w.r.t. $\boldsymbol{\beta}$ the equations

$$\frac{1}{2}\nabla Q\left(\boldsymbol{\beta}\right) = \mathbf{0}_k$$

By expansion,

$$Q\left(\boldsymbol{\beta}\right) = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T\mathbf{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}. \tag{14}$$

MLE 2

Since $\mathbf{y}^T X \boldsymbol{\beta} = \boldsymbol{\beta}^T X^T \mathbf{y}$ we get

$$Q(\boldsymbol{\beta}) = \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y}. \tag{15}$$

Then

$$\frac{1}{2} \nabla Q(\boldsymbol{\beta}) = \nabla \left[\frac{1}{2} \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - \boldsymbol{\beta}^T X^T \mathbf{y} \right].$$

MLE 3

$$\nabla \frac{1}{2} Q(\boldsymbol{\beta}) = \nabla \left[\frac{1}{2} \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - \boldsymbol{\beta}^T X^T \mathbf{y} \right].$$

One checks easily that

$$\nabla \boldsymbol{\beta}^T X^T X \boldsymbol{\beta} = 2X^T X \boldsymbol{\beta}, \nabla \boldsymbol{\beta}^T X^T \mathbf{y} = X^T \mathbf{y}$$

Hence

$$\frac{1}{2} \nabla Q(\boldsymbol{\beta}) = \mathbf{0}_k \Leftrightarrow X^T X \boldsymbol{\beta} = X^T \mathbf{y}$$

and, if $X^T X$ is invertible, $\widehat{\boldsymbol{\beta}}_{ML} = \widehat{\boldsymbol{\beta}} =$ the least squares estimate.

The MLE for Normal Multiple Linear Regression Model 4

Now one inserts $\widehat{\boldsymbol{\beta}}$ to get

$$l_{\mathbf{y}}\left(\widehat{\boldsymbol{\beta}},\sigma^2\right)=\frac{n}{2}\ln(2\pi)+\frac{n}{2}\ln(\sigma^2)+\frac{1}{2\sigma^2}\mathbf{e}_{LSE}^T\mathbf{e}_{LSE}$$

$$\frac{\partial}{\partial\sigma^2}l_{\mathbf{y}}\left(\widehat{\boldsymbol{\beta}},\sigma^2\right)=0\Leftrightarrow\widehat{\sigma^2}_{\mathrm{MLE}}=\frac{\mathbf{e}_{LSE}^T\mathbf{e}_{LSE}}{n}.$$

The estimate $\widehat{\sigma^2}_{\mathrm{MLE}}=\frac{\mathbf{e}_{LSE}^T\mathbf{e}_{LSE}}{n}$ is not used, since it is biased. For this and other prpoperties of LSE in normal multiple linear regression we need the notion of the true value of $\boldsymbol{\beta}_*$

The **True** Normal Multiple Linear Regression Model

$$\varepsilon \in N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$$

Now we suppose that the exists an unknown $\boldsymbol{\beta}_*$ such that

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \varepsilon. \tag{16}$$

in the sense that $\boldsymbol{\beta}_*$ is the **true value** [8] in $\mathbb{R}$, i.e., it underlies via (16) the observed vector $\mathbf{y}$, an outcome of $\mathbf{Y}$.

[8] MPV uses the phrase 'correct model'

PROPERTIES OF $\widehat{\boldsymbol{\beta}}$: $E\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}_*$

We compute the expectation of $\widehat{\boldsymbol{\beta}}$ w.r.t. the true model in (16).

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} = (X^T X)^{-1} X^T (X\boldsymbol{\beta}_* + \boldsymbol{\varepsilon})$$

$$= (X^T X)^{-1} X^T X \boldsymbol{\beta}_* + (X^T X)^{-1} X^T \boldsymbol{\varepsilon}$$

Thus

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_* + (X^T X)^{-1} X^T \boldsymbol{\varepsilon}. \tag{17}$$

Then by rules of computation with expectation vectors (see Lecture 2.)

$$E\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}_* + (X^T X)^{-1} X^T E\left[\boldsymbol{\varepsilon}\right] = \boldsymbol{\beta}_* + \mathbf{0} = \boldsymbol{\beta}_*.$$

$\widehat{\boldsymbol{\beta}}$ is unbiased

Properties of $\widehat{\beta}$: Covariance Matrix $\mathbf{C}_{\widehat{\beta}} = \sigma^2 (X^T X)^{-1}$

By definition:

$$\mathbf{C}_{\widehat{\beta}} = E\left[\left(\widehat{\beta} - E\left[\widehat{\beta}\right]\right)\left(\widehat{\beta} - E\left[\widehat{\beta}\right]\right)^T\right]$$

But by the true model and preceding computations

$$\widehat{\beta} - E\left[\widehat{\beta}\right] = \boldsymbol{\beta}_* + (X^T X)^{-1} X^T \boldsymbol{\varepsilon} - \boldsymbol{\beta}_* = (X^T X)^{-1} X^T \boldsymbol{\varepsilon}.$$

Hence $\mathbf{C}_{\widehat{\beta}} = E\left[(X^T X)^{-1} X^T \boldsymbol{\varepsilon}\left((X^T X)^{-1} X^T \boldsymbol{\varepsilon}\right)^T\right]$

$$= (X^T X)^{-1} X^T \underbrace{E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right]}_{=\sigma^2 \mathbb{I}_n} X (X^T X)^{-1}$$

$$= \sigma^2 \underbrace{(X^T X)^{-1} X^T X}_{=\mathbb{I}_{k+1}} (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

PROPERTIES OF $\widehat{\beta}$

Since $\widehat{\beta}$ is a linear transformation of the normal random vector **Y**:

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}_{*}, \sigma^2 (X^T X)^{-1}\right).$$

Bias of $\widehat{\sigma^2}_{\text{MLE}}$

The residual vector $\mathbf{e}_{LSE}$ is the outcome of the random vector

$$\widehat{\varepsilon} = Y - X\widehat{\boldsymbol{\beta}}$$

Then by (16) and (17)

$$\begin{aligned}
\widehat{\varepsilon} &= X\boldsymbol{\beta}_* + \varepsilon - X\widehat{\boldsymbol{\beta}} \\
&= X\boldsymbol{\beta}_* + \varepsilon - X\left(\boldsymbol{\beta}_* + (X^T X)^{-1} X^T \varepsilon\right) \\
&= \left(\mathbb{I}_n - X(X^T X)^{-1} X^T\right)\varepsilon \\
&= (\mathbb{I}_n - H)\,\varepsilon. \qquad (18)
\end{aligned}$$

Bias of $\widehat{\sigma^2}_{\mathrm{MLE}}$

Hence

$$E\left[\widehat{\sigma^2}_{\mathrm{MLE}}\right] = \frac{1}{n} E\left[\widehat{\boldsymbol{\epsilon}}^T \widehat{\boldsymbol{\epsilon}}\right] =$$

$$= \frac{1}{n} \mathrm{Tr}\, E\left[\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T\right]$$

where we evoked rule 2. in Appendix C for Traces. In view of (18)

$$E\left[\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T\right] = E\left[\left(\mathbb{I}_n - H\right)\varepsilon\left(\left(\mathbb{I}_n - H\right)\varepsilon\right)^T\right] =$$

$$= \left(\mathbb{I}_n - H\right)\underbrace{E\left[\varepsilon\varepsilon^T\right]}_{=\sigma^2\mathbb{I}_n}\left(\mathbb{I}_n - H\right)^T = \sigma^2\left(\mathbb{I}_n - H\right)\left(\mathbb{I}_n - H\right)^T$$

We have $\left(\mathbb{I}_n - H\right)^T = \mathbb{I}_n^T - H^T = \mathbb{I}_n - H$

BIAS OF $\widehat{\sigma^2}_{\text{MLE}}$

$$E\left[\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T\right] = \sigma^2 \left(\mathbb{I}_n - H\right)\left(\mathbb{I}_n - H\right)$$

$$\left(\mathbb{I}_n - H\right)\left(\mathbb{I}_n - H\right)^T = \mathbb{I}_n - H - H + HH = \mathbb{I}_n - 2H + H = \mathbb{I}_n - H$$

By property 3. in Appendix C and the rule C) above

$$\mathrm{Tr}\, E\left[\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}^T\right] = \sigma^2\, \mathrm{Tr}\left(\mathbb{I}_n - H\right) = \sigma^2\left(\mathrm{Tr}\,\mathbb{I}_n - \mathrm{Tr}\, H\right) = \sigma^2(n-k-1).$$

and

$$E\left[\widehat{\boldsymbol{\epsilon}}^T\widehat{\boldsymbol{\epsilon}}\right] = \sigma^2(n-k-1)$$

Hence

$$\widehat{\sigma^2} := \frac{1}{(n-k-1)}\mathbf{e}_{LSE}^T\mathbf{e}_{LSE}$$

is an unbiased estimator of σ^2.

Proposition

1. $$\widehat{\mathbf{Y}} \sim N_n\left(X\boldsymbol{\beta}_*, \sigma^2 H\right). \tag{19}$$

2. $$\widehat{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\left(\mathbb{I}_n - H\right)\right) \tag{20}$$

Proof:

1. $\widehat{\mathbf{Y}} = H\mathbf{Y} = HX\boldsymbol{\beta}_* + H\varepsilon$. By definition of the hat matrix $HX\boldsymbol{\beta}_* = X\boldsymbol{\beta}_*$. Hence $E\left[\widehat{\mathbf{Y}}\right] = X\boldsymbol{\beta}_*$. By the known rules

   $$C_{\widehat{\mathbf{Y}}} = H\sigma^2\mathbb{I}_n H^T = \sigma^2 HH^T = \sigma^2 HH = \sigma^2 H.$$

   Since $\mathbf{Y}$ is a multivariate normal vector, the assertion in (19) follows.

2. It has been shown in (18) that

   $$\widehat{\varepsilon} = \left(\mathbb{I}_n - H\right)\varepsilon$$

Then the assertion about the distribution of the LSE residualerna follows as with the distribution of $\widehat{\mathbf{Y}}$ using the idempotency of $\mathbf{I}_n - H$.

□

We see also that $\widehat{\mathbf{Y}}$ is unbiased in the sense that

$$E\left[\widehat{\mathbf{Y}}\right] = E\left[\mathbf{Y}\right].$$

SUMMARY PART 1

$$\varepsilon \sim N_n\left(\mathbf{0}_n, \sigma^2 \mathbb{I}_n\right)$$

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \varepsilon.$$

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

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}_*, \sigma^2 (X^T X)^{-1}\right).$$

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_* + (X^T X)^{-1} X^T \varepsilon.$$

SUMMARY PART 2

$$\widehat{\mathbf{Y}} \sim N_n\left(X\boldsymbol{\beta}_*, \sigma^2 H\right). \tag{21}$$

$$\widehat{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\left(\mathbb{I}_n - H\right)\right) \tag{22}$$

$$\parallel \mathbf{y} \parallel^2 = \parallel \hat{\mathbf{y}} \parallel^2 + \parallel \mathbf{e}_{LSE} \parallel^2 \tag{23}$$

$$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\boldsymbol{\beta}} = \mathbf{y} - H\mathbf{y}$$

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)}\mathbf{e}_{LSE}^T\mathbf{e}_{LSE}$$

C.f. s^2 when $k = 1$ in Lecture 1

Appendix A: Inner Product Spaces

Appendix B: Check $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ in the case $k = 1$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

By Cramer's rule

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

Appendix B: Check $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ in the case $k = 1$

$$(X^T X)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

We deal first with the determinant

$$\det(X^T X) = n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 = n\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right]$$

$$= n\sum_{i=1}^n (x_i - \bar{x})^2 = nS_{xx}$$

by the algebra in Lecture 1.

Appendix B: Check $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$ in the case $k = 1$

Thus

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \qquad (24)$$

Next

$$X^T \mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Then with $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)^T = (X^T X)^{-1} X^T \mathbf{y}$

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n y_i \sum_{i=1}^n x_i + n \sum_{i=1}^n x_i y_i \end{pmatrix} \qquad (25)$$

Appendix B: Check

$$(X^TX)^{-1}X^T\mathbf{y} = \frac{1}{nS_{xx}}\begin{pmatrix} \sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n y_i \sum_{i=1}^n x_i + n\sum_{i=1}^n x_i y_i \end{pmatrix}$$

Here

$$\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i$$

$$= \sum_{i=1}^n y_i \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right] + \sum_{i=1}^n y_i n\bar{x}^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i$$

where by the identity from Lecture 1 noted above

$$= S_{xx}\sum_{i=1}^n y_i + +n\bar{x}^2\sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i$$

$$= S_{xx}\sum_{i=1}^n y_i + n\bar{x}^2\sum_{i=1}^n y_i - n\bar{x}\sum_{i=1}^n x_i y_i$$

Appendix B: Check

Contnd

$$S_{xx}\sum_{i=1}^{n} y_i + n\bar{x}^2\sum_{i=1}^{n} y_i - n\bar{x}\sum_{i=1}^{n} x_i y_i = S_{xx}\sum_{i=1}^{n} y_i + n\bar{x}\left[\bar{x}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i\right]$$

$$= S_{xx}\sum_{i=1}^{n} y_i + n\bar{x}\left[n\bar{x}\bar{y} - \sum_{i=1}^{n} x_i y_i\right]$$

$$= S_{xx}\sum_{i=1}^{n} y_i - n\bar{x}S_{xy}$$

by an identity established in Lecture 1. i.e.,

$$\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = S_{xy}$$

Appendix B: Check

W.r.t (25) we have at this point established

$$\widehat{\beta}_0 = \frac{1}{nS_{xx}}\left[S_{xx}\sum_{i=1}^{n} y_i - n\bar{x}S_{xy}\right] = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{S_{xy}}{S_{xx}}\bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

that is

$$\widehat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}. \tag{26}$$

Next, from (25)

$$\widehat{\beta}_1 = \frac{1}{nS_{xx}}\left[-\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + n\sum_{i=1}^{n} x_i y_i\right] = \frac{1}{nS_{xx}}\left[n\left(\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}\right)\right]$$

Appendix B: Check

Contnd, by an identity in Lecture 1.

$$\widehat{\beta}_1 = \frac{1}{nS_{xx}}\left[n\left(\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}\right)\right] = \frac{1}{S_{xx}}\left[\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}\right] = \frac{S_{xy}}{S_{xx}}.$$

Hence we have

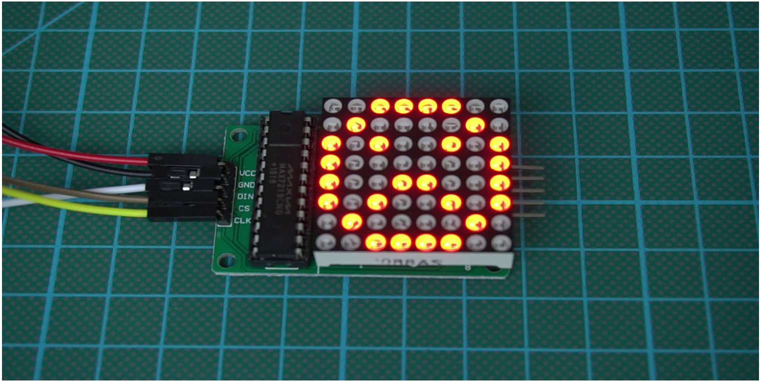
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{27}$$

which is the LSE for the slope in simple linear regression. When we insert (27) in (26) we get

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}, \tag{28}$$

which is the LSE for the intercept in simple linear regression.

Appendix C: Matrices



- A and B conformal, $(AB)^T = B^T A^T$. A and B invertible, $(AB)^{-1} = B^{-1}A^{-1}$.$(A^T)^T = A$.
- A and B conformal, $(A+B)^T = A^T + B^T$

A is $n \times n$ and invertible.

$$\left(A^T\right)^{-1} = \left(A^{-1}\right)^T$$

- *Proof*: $A^T\left(A^{-1}\right)^T = \left(A^{-1}A\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$ and $\left(A^{-1}\right)^T A^T = \left(AA^{-1}\right)^T = \mathbb{I}_n^T = \mathbb{I}_n$ □
  Hence, as X^TX is $k \times k$, and symmetric

$$\left((X^TX)^{-1}\right)^T = (X^TX)^{-1} \qquad (29)$$

Full Column Rank Implies that $X^T X$ is Positive Definite and Invertible

Lemma

Let A be any $n \times (k+1)$ matrix. If A has full column rank, then $A^T A$ is symmetric positive definite.

Proof: Take any $\mathbf{x} \in \mathbb{R}^{k+1}$. Then $\mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T A\mathbf{x} = \parallel A\mathbf{x} \parallel^2$. The squared norm $\parallel A\mathbf{x} \parallel^2$ is $= 0$ if and only if $A\mathbf{x} = \mathbf{0}_n$. We need to show that $A\mathbf{x} = \mathbf{0}_n$ if and only if $\mathbf{x} = \mathbf{0}_{k+1}$.

Let the $k+1$ column vectors of the matrix A be denoted by

$$\mathbf{a}_j^{(c)} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix} \quad j = 1, \ldots, k+1$$

Full Column Rank Implies that $X^T X$ is Positive Definite and Invertible

By definition of a matrix multiplying a vector, $A\mathbf{x}$ is a linear combination of its column vectors, i.e.,

$$A\mathbf{x} = x_1\mathbf{a}_1^{(c)} + x_2\mathbf{a}_2^{(c)} + \ldots + x_{k+1}\mathbf{a}_{k+1}^{(c)}$$

But since A has full column rank, the column vectors are linearly independent, and therefore a linear combination of them is $\mathbf{0}_n$ if and only if $x_1 = x_2 = \ldots = x_{k+1} = 0$. □

Trace of a Square Matrix

Let A be a square matrix. The **trace** $\mathrm{Tr}\, A$ of A is the sum of the entries in main diagonal:

$$\mathrm{Tr}\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} = \sum_{j=1}^{k} a_{jj}$$

The following facts are easily established; the proofs are left as exercises:

- 1.If A is a $k \times n$-matrix, and B an $n \times k$-matrix, then $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$
- 2. In particular, if $\mathbf{a}$ is a column-vector, then $\mathbf{a}^T\mathbf{a} = \mathrm{Tr}\left(\mathbf{a}\mathbf{a}^T\right)$.
- 3. a and b are real numbers, $\mathrm{Tr}(aC + bD) = a\,\mathrm{Tr}\, C + b\,\mathrm{Tr}\, D$

Appendix C: Proofs of the statements about H

The rule (29), $\left((X^TX)^{-1}\right)^T = (X^TX)^{-1}$ is applied.

A) $H^T = (X(X^TX)^{-1}X^T)^T = (X^T)^T(X(X^TX)^{-1})^T$
$= X(X^TX)^{-1})^T X^T = X(X^TX)^{-1}X^T = H$

B) $H^2 = HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T =$
$X\underbrace{(X^TX)^{-1}\left(X^TX\right)}_{=\mathbb{I}_{k+1}}(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H.$

Note that X^TX is $(k+1)\times(k+1)$.

C) Set $A = X(X^TX)^{-1}$, $B = X^T$. By 1. above in this Appendix,

$$\mathrm{Tr}\, H = \mathrm{Tr}\, AB = \mathrm{Tr}\, BA = \mathrm{Tr}\, X^TX(X^TX)^{-1} = \mathrm{Tr}\, \mathbb{I}_{k+1} = k+1.$$

Appendix C: General Properties of Idempotent Matrices

The only invertible idempotent matrix is the identity matrix $\mathbb{I}$.

- *Proof*: Let A be an idempotent matrix. Assume that A^{-1} exists. By idempotency $A^2 = A$. We multiply this by A^{-1} to get

$$A^{-1}A^2 = A^{-1}A = \mathbb{I}.$$

But in the left hand side

$$A^{-1}A^2 = A^{-1}AA = A.$$

Hence $A = \mathbb{I}$. □

Appendix D: An Expression for SS_{Res}

Let us recall from Lecture 1 SS_{Res}, the Residual Sum of Squares defined for $k = 1$ as

$$SS_{\text{Res}} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The same definition is valid for any $k \geq 1$, and now we can write

$$SS_{\text{Res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \| \mathbf{y} - H\mathbf{y} \|^2 = \| \mathbf{y} - X\widehat{\beta} \|^2 = Q\left(\widehat{\beta}\right)$$

In (15) we have

$$Q(\beta) = \mathbf{y}^T\mathbf{y} - 2\beta^T X^T \mathbf{y} + \beta^T X^T X \beta.$$

Appendix D: An Expression for SS_{Res}

Hence

$$SS_{\mathrm{Res}} = Q\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}^T\mathbf{y} - 2\,\widehat{\boldsymbol{\beta}}^T X^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T X^T X\widehat{\boldsymbol{\beta}} \tag{30}$$

We rewrite the third term in the right hand side. As $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$ we have

$$X^TX\widehat{\boldsymbol{\beta}} = \underbrace{X^TX(X^TX)^{-1}}_{=\mathbb{I}_{k+1}}X^T\mathbf{y} = X^T\mathbf{y}.$$

Hence $\widehat{\boldsymbol{\beta}}^T X^TX\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^T X^T\mathbf{y}$ and we have in (30)

$$SS_{\mathrm{Res}} = Q\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}^T\mathbf{y} - \widehat{\boldsymbol{\beta}}^T X^T\mathbf{y}$$

Appendix D: An Expression for SS_{Res}

$$\boxed{SS_{\text{Res}} = Q\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}^T\mathbf{y} - \widehat{\boldsymbol{\beta}}^T X^T \mathbf{y} \qquad (31)}$$

This expression will turn out to be very useful in the sequel, e.g., in Lecture 6.