Estimativa de σ^2 e Teorema de Gauss-Markov

Renato Assunção
ESRI and DCC/UFMG

Estimativa de σ^2

- Definimos $\sigma^2 = \text{Var}(\xi_i) = \text{Var}(y_i|\boldsymbol{x}_i) = \mathbb{E}(y_i - \mathbb{E}(y_i|\boldsymbol{x}_i))^2$.
- Se o modelo de regressão linear está correto: $\mathbb{E}(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} \approx \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$.
- Podemos olhar para o **resíduo**

$$r_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$$

como uma aproximação para o erro

$$\xi_i = y_i - \mathbb{E}(y_i|\boldsymbol{x}_i)\,.$$

Estimativa de σ^2

- Para obter um valor aproximado para

$$\sigma^2 = \mathbb{E}(y_i - \mathbb{E}(y_i|\boldsymbol{x}_i))^2 \,,$$

  podemos usar:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{(average of squared residuals)}$$

- Em geral, teremos $\hat{\sigma}^2 \neq \sigma^2$
- Mais que isto: $\hat{\sigma}^2$ é uma varável aleatória enquanto σ^2 é um valor fixo e desconhecido.
- Por quê?

Distribuição de $\hat{\sigma}^2$

- Se $\hat{\sigma}^2$ é uma variável aleatória, ela possui uma lista de valores possíveis e probabilidades associadas.
- Lista de valores possíveis: $(0, \infty) = \mathbb{R}^+$.
- Lista de probabilidades: vamos aprender em breve que ela será uma densidade de probabilidade que usualmente terá o seguinte formato:

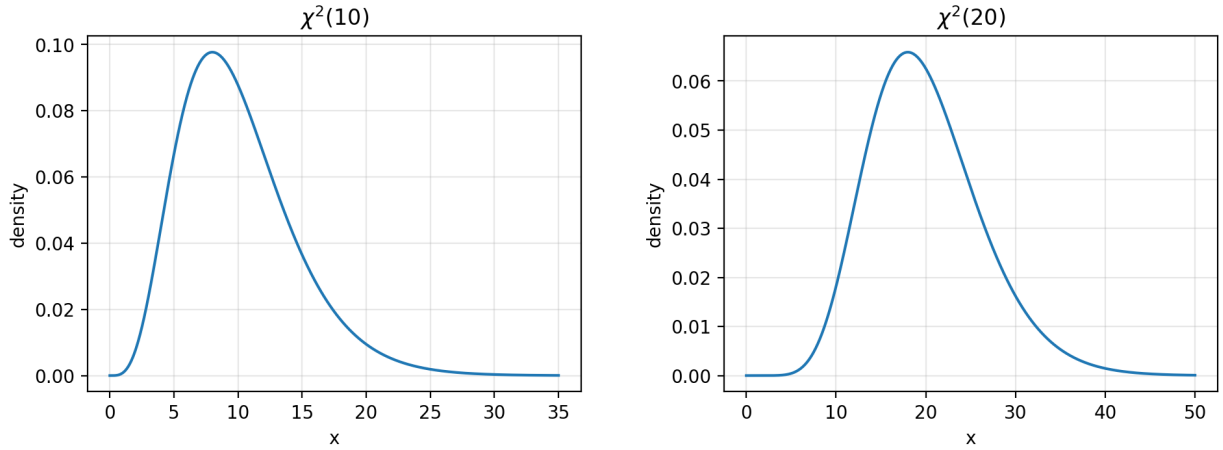

Figure: Densidade da distribuição de $\hat{\sigma}^2$ com baseada em $n = 15$ e $n = 25$ dados usando 4 features (e mais a coluna de 1's.

Distribuição de $\hat{\sigma}^2$

- Como $\hat{\sigma}^2$ é uma variável aleatória, ela possui esperança e variância.
- Na verdade, utilizamos uma leve modificação do estimador.
- Ao invés de dividir por n, nós dividimos por $n-(p+1)$:

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  Veremos a razão disso em breve.
- Se $n \gg p$, então $\frac{1}{n-(p+1)} \approx \frac{1}{n}$, o que significa que na prática não haverá nenhuma diferença significativa entre dvidir por n ou por $n-(p+1)$.
- **Exemplo (Cement Dataset):** $n = 1030$ e $p = 8$.

$$\frac{1}{1030-(8+1)} = 0,000979 \approx \frac{1}{1030} = 0,000970$$

Distribuição de $\hat{\sigma}^2$

Theorem

A soma dos quadrados dos resíduos

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \|\boldsymbol{r}\|^2$$

segue uma distribuição qui-quadrado com χ^2.

Proof.

Appendix □

- $\|\boldsymbol{r}\|^2$ é o comprimento ao quadrado do vetor de resíduos $\boldsymbol{R} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.
- $\hat{\sigma}^2$ é apenas a constante $1/(n-(p+1))$ vezes a variável aleatória $\|\boldsymbol{r}\|^2$ do teorema.
- Mas o que é a distribuição qui-quadrado com χ^2?

A Distribuição Qui-Quadrado (χ^2)

- O que é uma distribuição qui-quadrado?

Definition

Se $Z_1, Z_2, \ldots, Z_k$ são variáveis i.i.d. $N(0,1)$, então $Q = \sum Z_i^2 \sim \chi^2_k$.

- Seja $Z = (Z_1, \ldots, Z_k)$ a random vector of i.i.d. $N(0,1)$. It is a random points in $\mathbb{R}^k$ orbiting around the center $\mathbf{0} = (0, \ldots, 0)$.
- Then $Q = \sum Z_i^2$ is the squared length of this random vector.
- Sometimes it is close to the origin $\mathbf{0}$, and its length will be small, close to 0.
- Sometimes it is farther away from $\mathbf{0}$, with a rather long length.
- How often can it be close to $\mathbf{0}$? And how long can it be expected?
- Let us look at the density of $Q = \sum Z_i^2 \sim \chi^2_k$.

A Distribuição Qui-Quadrado (χ^2)

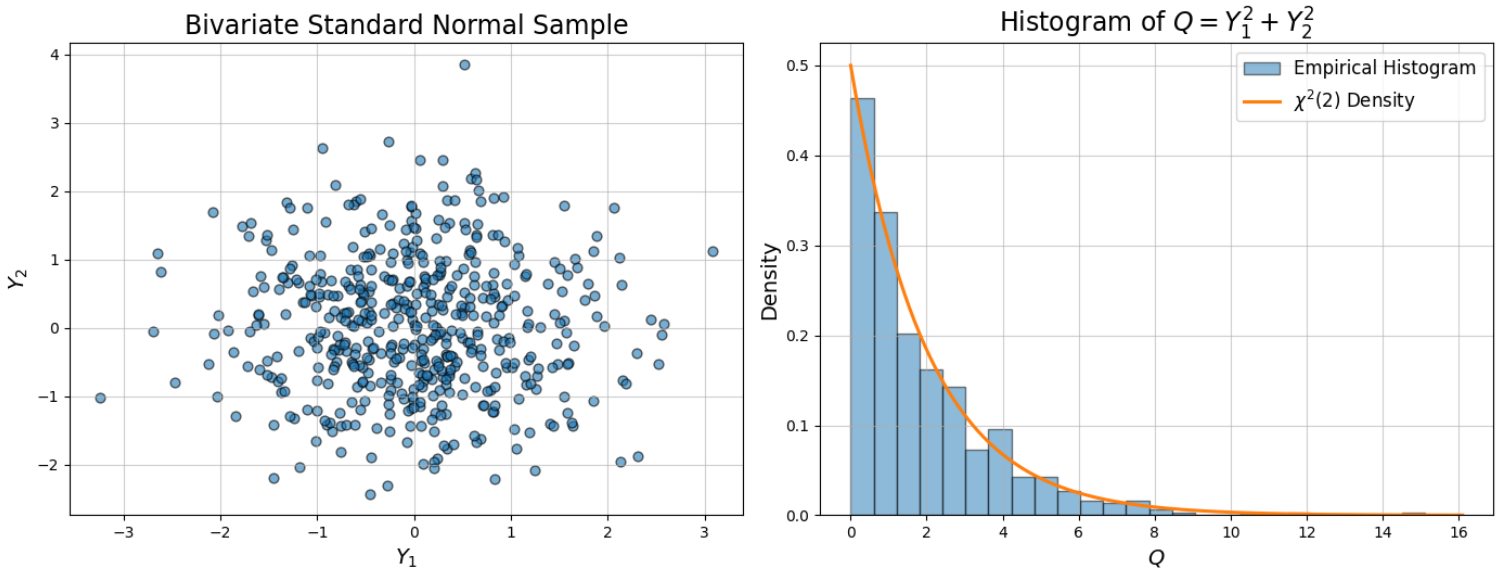

Figure: Sample of $\boldsymbol{Z} = (Z_1, Z_2)$ and its histogram together with the probability density of a Chi-square for $Q = Z_1^2 + Z_2^2$.

Note that the values of Q are spread around 2, the number of dimensions. Also, the squared lengths are hardly larger than 6.

A Distribuição Qui-Quadrado (χ^2)

Random points $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ are in 3-dimensional space.

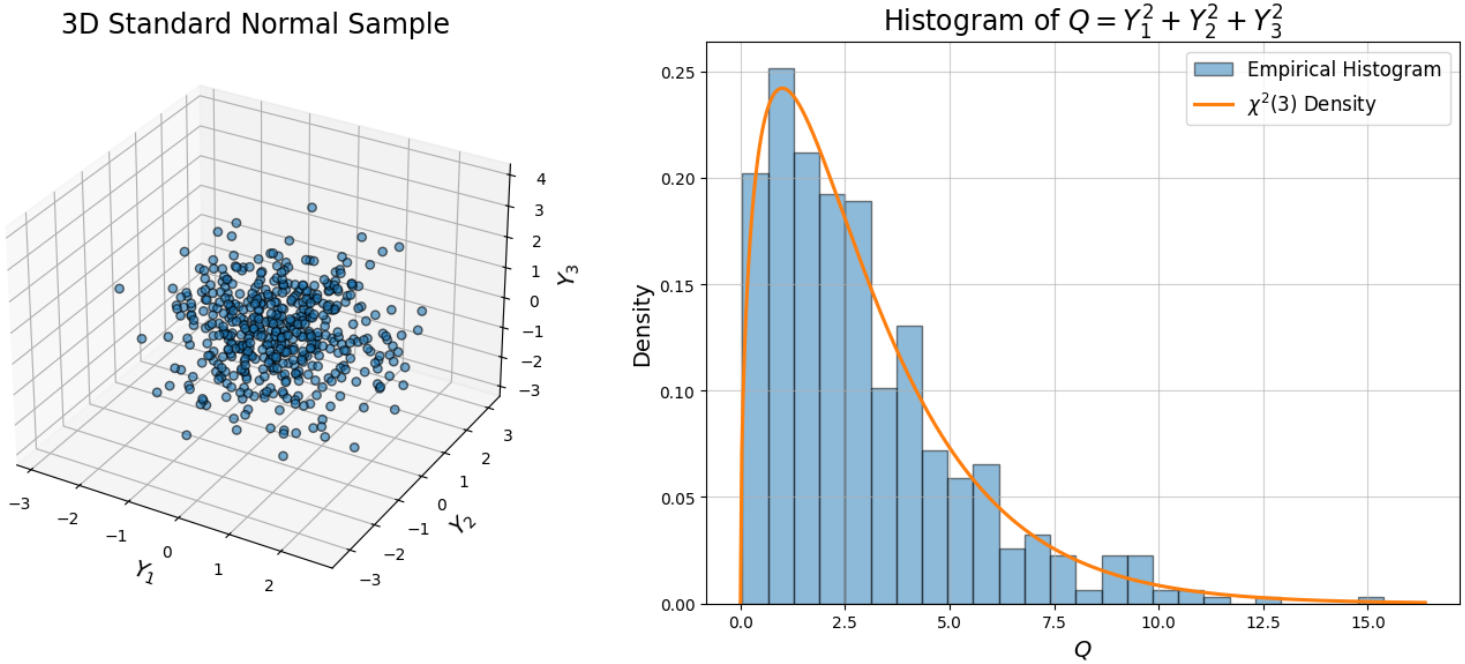

Figure: Sample of $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ and its histogram together with the probability density of a Chi-square for $Q = Z_1^2 + Z_2^2 + Z_3^2$.

The Q values are spread around 3, the number of dimensions, and hardly $Q > 8$.

A Distribuição Qui-Quadrado (χ^2)

Chi-square density with $k = 10$ and $k = 20$:

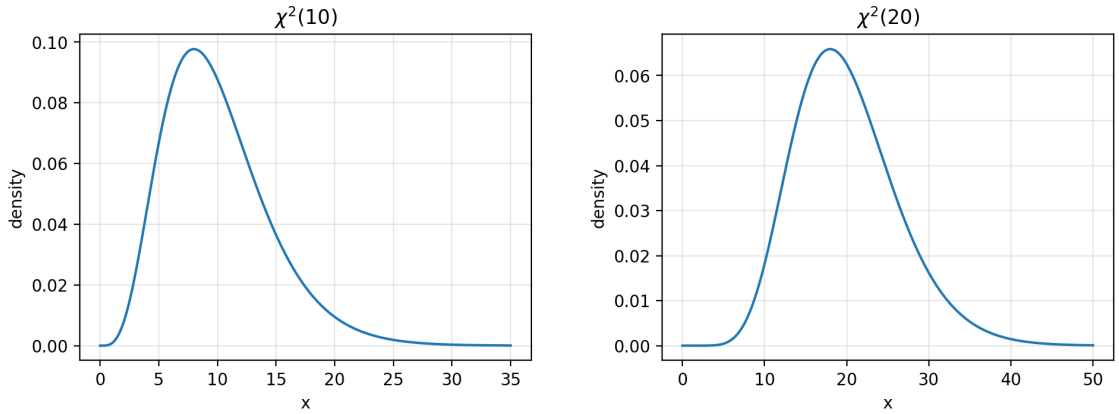

Figure: Densidade da distribuição de χ^2_k com $k = 10$ and $k = 20$

The Q values spread around k, and the SD also increases with k.

A Distribuição Qui-Quadrado (χ^2)

- Mathematical analysis finds the density function:

$$f_k(x) = (cte)x^{(k/2)-1}e^{-x/2} \quad \text{for } x > 0$$

- For example, with $k = 10$, we have $f_{10}(x) = (cte)x^4e^{-x/2}$.
- It is the product of a polynomial (x^4) times an exponentially decreasing function.
- Propriedades:
  1. $\mathbb{E}(Q) = k$, aumenta linearmente com a dimensão k.
  2. $\mathbb{V}(Q) = 2k \Rightarrow DP(Q) = \sqrt{2k} \approx 1.4\sqrt{k}$. SD também aumenta mas a uma taxa mais lenta que $\mathbb{E}(Q)$.
  3. Quando k aumenta, a densidade vai ficando simétrica e similar a uma Gaussiana.

A Distribuição Qui-Quadrado (χ^2)

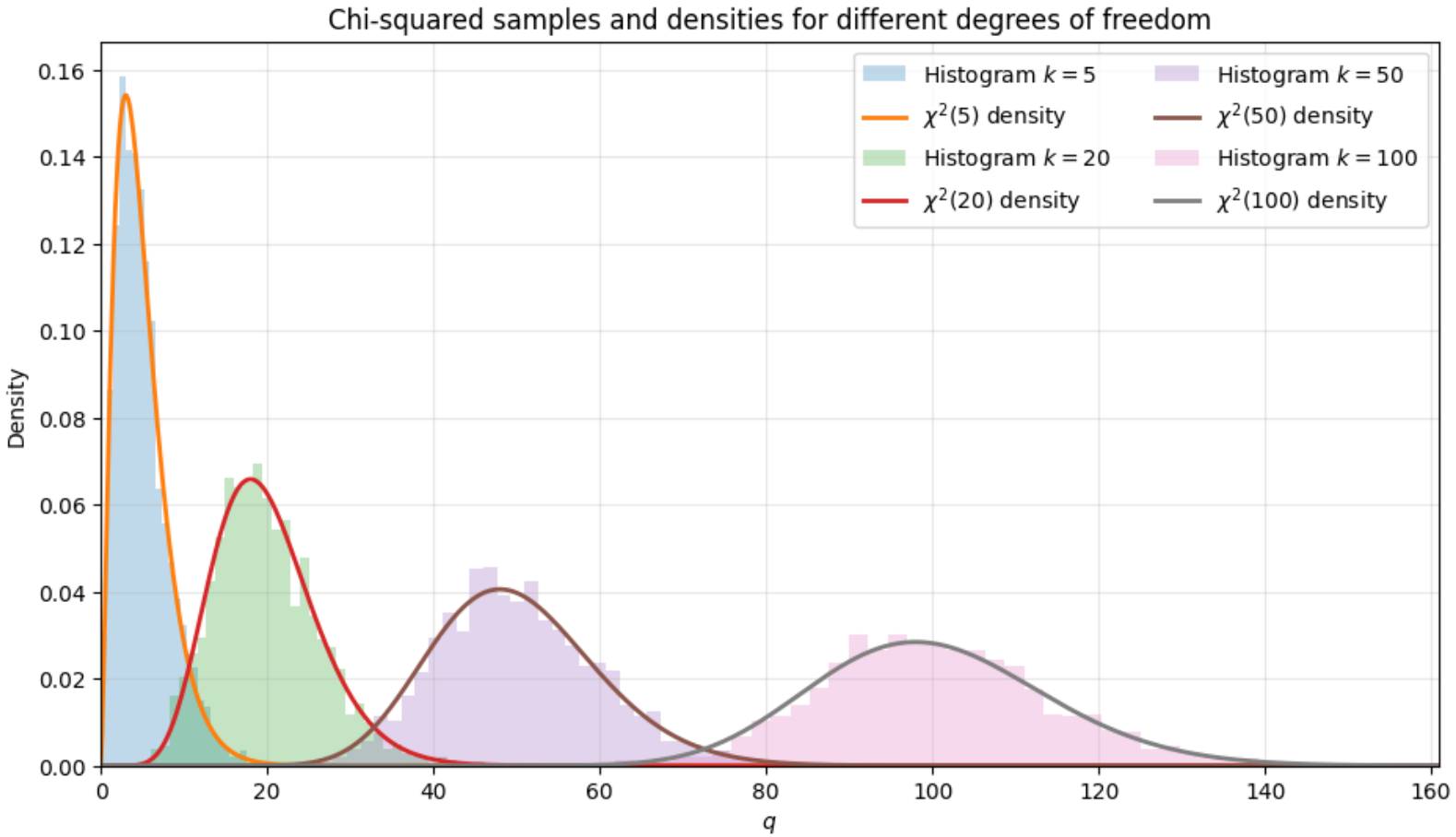

Figure: Chi-square density with different k and their samples.

Contraste: ξ and r

Let us connect the chi-square distribution with the estimator $\hat{\sigma}^2$.
We have two representations for the random vector $\boldsymbol{Y}$:

$$\begin{aligned} \textbf{Modelo Real} \quad & \boldsymbol{Y} = \boldsymbol{X}\beta + \xi \\ \textbf{Observado} \quad & \boldsymbol{Y} = \boldsymbol{X}\hat{\beta} + \boldsymbol{r} \end{aligned}$$

These are different representations

$\boldsymbol{Y}$ is the same vector in both rows. However:

- $\beta \neq \hat{\beta}$
- $\xi \neq \boldsymbol{r}$. That is, $\xi = \boldsymbol{Y} - \boldsymbol{X}\beta \neq \boldsymbol{Y} - \boldsymbol{X}\hat{\beta} = \boldsymbol{r}$

How is $\xi \neq \boldsymbol{r}$?

Contraste: $\boldsymbol{\xi}$ and $\boldsymbol{r}$

- $\boldsymbol{\xi} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.
- As the components of $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_n)$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, we have immediately that

$$\frac{1}{\sigma}\boldsymbol{\xi} = (\frac{\xi_1}{\sigma}, \frac{\xi_2}{\sigma}, \ldots, \frac{\xi_n}{\sigma})$$

  are i.i.d. $\mathcal{N}(0, 1)$.
- Therefore, by definition, their squared length follows a chi-squared probability distribution with n degrees of freedom:

$$\|\frac{\boldsymbol{\xi}}{\sigma}\|^2 = \left(\frac{\xi_1}{\sigma}\right)^2 + \ldots + \left(\frac{\xi_n}{\sigma}\right)^2 \sim \chi^2_n$$

- How about the residuals? Their squared length is also a chi-squared distribution but not with n degrees of freedom:

Contraste: ξ and r

- Considering the residual vector, we have:

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} \\ &= \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \end{aligned}$$

- There are two crucial properties of the matrix $\boldsymbol{H}$ that are very easy to prove using that $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. (lista de exercicios):
  - $\boldsymbol{H}$ is symmetric.
  - $\boldsymbol{H}$ is idempotent (that is, $\boldsymbol{H}^2 = \boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$
- Show that $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\xi$.
- As $\boldsymbol{Y}$ is a multivariate Gaussian $\mathcal{N}_n(\boldsymbol{X}\beta, \sigma^2\boldsymbol{I}_n)$, we use the property of a Gaussian multiplied by the constant matrix $(\boldsymbol{I} - \boldsymbol{H})$:

$$\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \mathcal{N}_n((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\beta, (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}_n(\boldsymbol{I} - \boldsymbol{H})'$$

- It is a very simple (lista de exercicios) to show that this expression reduces to

$$\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$$

Contraste: $\boldsymbol{\xi}$ and $\boldsymbol{r}$

- Therefore, we found that $\boldsymbol{\xi} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ and $\|\boldsymbol{\xi}\|^2/\sigma^2 \sim \chi^2_n$
- We also found that $\boldsymbol{r} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})$.
- What can we say about $\|\boldsymbol{r}\|^2/\sigma^2$, the squared length of the residual vector?
- The only difference between the distributions of $\boldsymbol{\xi}$ and $\boldsymbol{r}$ is the covariance matrix.
- The coordinates of the vector $\boldsymbol{\xi}$ are i.i.d. $N(0, \sigma^2)$.
- The coordinates of the vector $\boldsymbol{r}$ are all Gaussian with zero mean.
- However, they have different variances (the diagonal of $(\boldsymbol{I} - \boldsymbol{H})$ is usually not constant).
- Furthermore, the residuals $r_i = y_i - \hat{y}_i$ and $r_j = y_j - \hat{y}_j$ are not independent because their covariance is $(\boldsymbol{I} - \boldsymbol{H})[i, j] \neq 0$, usually.
- They can have negative or positive correlations.

Quadratic forms in mathematics

Definition (Quadratic forms)

In mathematics, given a $n \times n$ symmetric matrix $\boldsymbol{A}$, it determines a *quadratic form* defined as

$$\begin{aligned} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} &= \sum_{ij} \boldsymbol{A}_{ij} x_i x_j \\ &= \sum_{i=1}^{n} \boldsymbol{A}_{ii} x_i^2 + \sum_{i \neq j} \boldsymbol{A}_{ij} x_i x_j \\ &= \sum_{i=1}^{n} \boldsymbol{A}_{ii} x_i^2 + 2 \sum_{i<j} \boldsymbol{A}_{ij} x_i x_j \end{aligned}$$

The last equality is due to the $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}$ by the symmetry of $\boldsymbol{A}$.

Quadratic form of a Gaussian vector

- When we build a quadratic form using a random vector $\boldsymbol{X}$ we obtain a random variable.
- We studied these random quadratic forms in FECD-A.
- They are extremely important, with the notion of statistical (or probabilistic) distance based on them (rather than on the Euclidean distance).
- The Mahalanobis distance appeared naturally as a random quadratic form by taking $\boldsymbol{A} = \boldsymbol{\Sigma}^{-1}$, the inverse covariance matrix.
- First, we will learn a probability theorem that we will not prove.
- Next, we use this theorem to find the distribution of $\|\boldsymbol{r}\|^2$.

Quadratic form of a Gaussian vector

Theorem

Let $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$ and let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a symmetric idempotent matrix, that is, $\boldsymbol{A}' = \boldsymbol{A}$ and $\boldsymbol{A}^2 = \boldsymbol{A}$. Then

$$\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r, \qquad \text{where} \qquad r = \text{rank}(\boldsymbol{A}).$$

Quadratic form of a Gaussian vector

Proof: (optional)

- Since $\boldsymbol{A}$ is symmetric, the spectral theorem implies that there exists an orthogonal matrix $\boldsymbol{U}$ such that

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'.$$

- Since $\boldsymbol{A}$ is idempotent, its eigenvalues satisfy $\lambda^2 = \lambda$, so each eigenvalue is either 0 or 1.
- Therefore, if $r = \text{rank}(\boldsymbol{A})$, we may write

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{I}_r & 0 \\ 0 & 0 \end{pmatrix}.$$

- Define

$$\boldsymbol{W} = \boldsymbol{U}'\boldsymbol{Z}.$$

Since $\boldsymbol{U}$ is orthogonal and $\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, we also have

$$\boldsymbol{W} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n).$$

Quadratic form of a Gaussian vector

- Thus, the coordinates $W_1, \ldots, W_n$ are independent $N(0,1)$ variables, and

$$\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{Z} = \boldsymbol{W}'\boldsymbol{\Lambda}\boldsymbol{W} = \sum_{i=1}^{r} W_i^2.$$

- Since the sum of squares of r independent standard normal variables has a χ^2_r distribution, we conclude that

$$\boldsymbol{Z}'\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r.$$

Applying the theorem to $\|\boldsymbol{r}\|^2$

- As $\boldsymbol{\xi} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, then $\boldsymbol{\xi}/\sigma \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$.
- Simple matrix manipulation allows us to express the squared length of $\boldsymbol{r}$ as a quadratic form:

$$\|\boldsymbol{r}\|^2 = \boldsymbol{r}'\boldsymbol{r} = ((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi} = (\boldsymbol{\xi})'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}$$

- Dividing both sides by σ^2, we have

$$\|\boldsymbol{r}\|^2/\sigma^2 = (\boldsymbol{\xi}/\sigma)'(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{\xi}/\sigma)$$

- Hence, by the Gaussian quadratic form theorem,

$$\frac{\|\boldsymbol{r}\|^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Geometric interpretation

The degrees of freedom are exactly the dimension $n - (p + 1)$ of the residual subspace $\mathcal{C}(\boldsymbol{X})^{\perp}$ where the residual vector $\boldsymbol{r}$ lives.

Applying the theorem to $\|\boldsymbol{r}\|^2$

- As the expected value of a chi-square random variable is its degrees of freedom, we have that

$$\mathbb{E}\left[\frac{\|\boldsymbol{r}\|^2}{\sigma^2}\right] = n-p-1 \text{ or, exchanging the constants, } \mathbb{E}\left[\frac{\|\boldsymbol{r}\|^2}{n-p-1}\right] = \sigma^2$$

- That is, the expected value of the "average" of squared residuals is an unbiased estimator of σ^2:

$$\mathbb{E}\left[\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] = \sigma^2$$

- From now on, we adopt the following unbiased estimator of σ^2:

$$\widehat{\sigma^2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-p-1}$$

- We sum n residuals^2 but divide by $n-p-1$.

Degrees of freedom in the statsmodel output

- For the concrete compressive strength example, the statsmodel output for linear regression shows: the number of observation $n = 1030$, the number of degrees of freedom $n - p - 1 = 1030 - 8 - 1 = 1021$.
- Unfortunately, it did not show explicitly $\widehat{\sigma^2}$ or $\|\boldsymbol{r}\|^2$.

OLS Regression Results

Dep. Variable:	Compressive Strength	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.612
Method:	Least Squares	F-statistic:	204.3
Date:	Mon, 21 Apr 2025	Prob (F-statistic):	6.76e-206
Time:	12:12:55	Log-Likelihood:	-3869.0
No. Observations:	1030	AIC:	7756.
Df Residuals:	1021	BIC:	7800.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.1638	26.588	-0.871	0.384	-75.338	29.010
Cement	0.1198	0.008	14.110	0.000	0.103	0.136
Blast Furnace Slag	0.1038	0.010	10.245	0.000	0.084	0.124
Fly Ash	0.0879	0.013	6.988	0.000	0.063	0.113
Water	-0.1503	0.040	-3.741	0.000	-0.229	-0.071
Superplasticizer	0.2907	0.093	3.110	0.002	0.107	0.474
Coarse Aggregate	0.0180	0.009	1.919	0.055	-0.000	0.036
Fine Aggregate	0.0202	0.011	1.883	0.060	-0.001	0.041
Age	0.1142	0.005	21.046	0.000	0.104	0.125

Omnibus:	5.379	Durbin-Watson:	1.281
Prob(Omnibus):	0.068	Jarque-Bera (JB):	5.305
Skew:	-0.174	Prob(JB):	0.0705
Kurtosis:	3.045	Cond. No.	1.06e+05

Figure: Degrees of freedom in the statsmodel output

Standard error uses $\widehat{\sigma^2}$

- The standard deviation of each coefficient (called Std Error) is calculated using $\widehat{\sigma^2}$, and not the unknown $\sigma^2 = \mathbb{V}(\xi_i)$:

$$DP(\hat{\beta}_j) = \sqrt{\mathbb{V}(\hat{\beta}_j)} = \hat{\sigma}\sqrt{(X'X)^{-1}[jj]}$$

OLS Regression Results

Dep. Variable:	Compressive Strength	R-squared:	0.615
Model:	OLS	Adj. R-squared:	0.612
Method:	Least Squares	F-statistic:	204.3
Date:	Mon, 21 Apr 2025	Prob (F-statistic):	6.76e-206
Time:	12:12:55	Log-Likelihood:	-3869.0
No. Observations:	1030	AIC:	7756.
Df Residuals:	1021	BIC:	7800.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.1638	26.588	-0.871	0.384	-75.338	29.010
Cement	0.1198	0.008	14.110	0.000	0.103	0.136
Blast Furnace Slag	0.1038	0.010	10.245	0.000	0.084	0.124
Fly Ash	0.0879	0.013	6.988	0.000	0.063	0.113
Water	-0.1503	0.040	-3.741	0.000	-0.229	-0.071
Superplasticizer	0.2907	0.093	3.110	0.002	0.107	0.474
Coarse Aggregate	0.0180	0.009	1.919	0.055	-0.000	0.036
Fine Aggregate	0.0202	0.011	1.883	0.060	-0.001	0.041
Age	0.1142	0.005	21.046	0.000	0.104	0.125

Omnibus:	5.379	Durbin-Watson:	1.281
Prob(Omnibus):	0.068	Jarque-Bera (JB):	5.305
Skew:	-0.174	Prob(JB):	0.0705
Kurtosis:	3.045	Cond. No.	1.06e+05

Figure: $DP(\hat{\beta}_j)$

Gauss–Markov Theorem

Introduction to the Gauss–Markov Theorem

Recap:

- **Aim:** to predict Y using p features in the vector $\boldsymbol{X}$.
- **Approach:** To use a linear predictor $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$.
- **Seen as a purely minimimzation problem:** Solution is

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y} = \arg\min_{\beta} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

- **Linear algebra view:** $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ is the orthogonal projection of $\boldsymbol{Y}$ onto $\mathcal{C}(\boldsymbol{X})$.

Introduction to the Gauss–Markov Theorem

- **We can also see the problem with a stochastic model:**
- Assume that $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.
- Then we can derive stochastic properties for the least squares estimator $\hat{\beta}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$:
  - It is unbiased: $\mathbb{E}(\hat{\beta}_{LS}) = \beta$. This tells us that the estimator is accurate to estimate β.
  - We also could find its variability: $Var(\hat{\beta}_{LS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. This allows us to know how precise is the estimator.
  - We used $\widehat{\sigma^2} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-p-1}$ to estimate σ^2.
- **The Gauss–Markov theorem** gives a remarkable optimality result: among a broad class of estimators, no other estimator can be more precise than $\hat{\beta}_{LS}$.

Informal description of the Gauss–Markov Theorem

- The Gauss–Markov theorem states that

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

  is the "best way" to estimate the unknown coefficient vector $\boldsymbol{\beta}$ of the linear model.
- We already found one "best way": the numerical approach that gave us that $\hat{\boldsymbol{\beta}}_{LS}$ minimizes the distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$
- Another "best way" is the route followed by the Gauss–Markov Theorem.
- It states that each coordinate of $\hat{\boldsymbol{\beta}}_{LS}$ is the best estimator of the corresponding unknown coefficient among all linear unbiased estimators, because it has the smallest squared estimation error on average.

The Gauss–Markov Theorem (coordinate version)

Theorem

Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Consider the class of all estimators of $\boldsymbol{\beta}$ that are:

- linear in the observations. That is, they are written as $\hat{\boldsymbol{\beta}}_* = \boldsymbol{C}\boldsymbol{Y}$. for some matrix $\boldsymbol{C}$.
- They are unbiased to estimate $\boldsymbol{\beta}$. That is, $\mathbb{E}(\boldsymbol{C}\boldsymbol{Y}) = \boldsymbol{\beta}$.

Fix one coefficient β_j of the linear model. Among all linear unbiased estimators of β_j, the least squares estimator $\hat{\beta}_j$ has the smallest squared estimation error on average.

Proof of the Gauss–Markov Theorem

- We first prove the theorem for one coefficient β_j.
- This scalar version contains the main idea of the proof.
- The full matrix version follows from the same reasoning but requires some additional linear algebra. It is left as optional material.
- Write the least squares estimator as

$$\hat{\boldsymbol{\beta}}_{LS} = \boldsymbol{A}\boldsymbol{Y}, \qquad \boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

- Consider another linear candidate to estimate $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^* = \boldsymbol{C}\boldsymbol{Y}$$

- Write $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{D}$.
- For $\hat{\boldsymbol{\beta}}^*$ to be unbiased, we must have $\mathbb{E}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}$ for every possible $\boldsymbol{\beta}$.

The unbiasedness condition

- Compute the expectation:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{C}\mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{A} + \boldsymbol{D})\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{A}\boldsymbol{X} + \boldsymbol{D}\boldsymbol{X})\boldsymbol{\beta}.$$

- Since

$$\boldsymbol{A}\boldsymbol{X} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I},$$

  unbiasedness requires

$$\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}\,.$$

- This is the only restriction imposed by unbiasedness.

Minimizing the estimation error

- What does "better" mean?
- Since the candidate estimator is unbiased (and hence, $\mathbb{E}(\hat{\beta}_j^*) = \beta_j$)), the mean squared estimation error equals the variance:

$$\mathbb{E}(\hat{\beta}_j^* - \beta_j)^2 = \mathbb{E}(\hat{\beta}_j^* - \mathbb{E}(\hat{\beta}_j^*))^2 = \mathbb{V}(\hat{\beta}_j).$$

- Compute the covariance matrix of $\hat{\boldsymbol{\beta}}_*$:

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) = \boldsymbol{C}\mathbb{V}(\boldsymbol{Y})\boldsymbol{C}^T = \boldsymbol{C}(\sigma^2\boldsymbol{I})\boldsymbol{C}^T\sigma^2\boldsymbol{C}\boldsymbol{C}^T.$$

- Substituting $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{D}$ and expanding:

$$\begin{aligned}\mathbb{V}(\hat{\boldsymbol{\beta}}_*) &= \sigma^2(\boldsymbol{A} + \boldsymbol{D})(\boldsymbol{A} + \boldsymbol{D})^T \\ &= \sigma^2(\boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{D}^T + \boldsymbol{D}\boldsymbol{A}^T + \boldsymbol{D}\boldsymbol{D}^T).\end{aligned}$$

- The cross terms vanish because

$$\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{A}\boldsymbol{D}^T = \boldsymbol{0} = \boldsymbol{D}\boldsymbol{A}^T$$

Conclusion

- As a consequence,

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) = \sigma^2(\boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{D}\boldsymbol{D}^T).$$

- Using the expression for $\boldsymbol{A}$, we have

$$\sigma^2\boldsymbol{A}\boldsymbol{A}^T = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS})$$

- Therefore,

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) = \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) + \sigma^2\boldsymbol{D}\boldsymbol{D}^T.$$

- The diagonal entry $[\boldsymbol{D}\boldsymbol{D}^T]_{jj}$ is the inner product of row j of $\boldsymbol{D}$ with itself.
- Therefore, $[\boldsymbol{D}\boldsymbol{D}^T]_{jj} \geq 0$.
- Hence, for every coordinate j, we have $\mathbb{V}(\hat{\beta}_j^*) \geq \mathbb{V}(\hat{\beta}_j^{LS})$.
- Thus, least squares gives the smallest typical estimation errors for every coefficient.
- No linear unbiased estimator can estimate any coefficient with smaller variance.

Why do we need the full matrix version?

- The coordinate-wise theorem tells us that each coefficient $\hat{\beta}_j$ has the smallest variance among all linear unbiased estimators.
- But in practice, we often care about **linear combinations** of coefficients:

$$\theta = \boldsymbol{a}^T \boldsymbol{\beta} = a_0\beta_0 + \cdots + a_p\beta_p.$$

- Examples:
  - predictions,
  - treatment contrasts,
  - marginal effects,
  - sensitivity analyses.
- The variance of the estimated combination is

$$\mathbb{V}(\boldsymbol{a}^T \hat{\boldsymbol{\beta}}) = \boldsymbol{a}^T \mathbb{V}(\hat{\boldsymbol{\beta}}) \boldsymbol{a}.$$

- Therefore, to guarantee optimality for *every* linear combination, we need to compare the entire covariance matrices.

Optional: the full Gauss–Markov Theorem

Theorem (matrix version)

Let

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n).$$

Consider the class of all estimators of β that are:

- linear in the observations, that is, $\hat{\beta}_* = \boldsymbol{C}\boldsymbol{Y}$ for some matrix $\boldsymbol{C}$;
- unbiased, that is, $\mathbb{E}(\hat{\beta}_*) = \beta$.

Then the least squares estimator

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

has the smallest covariance matrix in this class. More precisely,

$$\mathbb{V}(\hat{\beta}_*) - \mathbb{V}(\hat{\beta}_{LS})$$

is positive semidefinite.

Why the full theorem is almost already proved

- In the previous proof, we considered another linear unbiased estimator

$$\hat{\boldsymbol{\beta}}_* = \boldsymbol{C}\boldsymbol{Y}, \qquad \boldsymbol{C} = \boldsymbol{A} + \boldsymbol{D},$$

  where

$$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

- Unbiasedness forced the condition

$$\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}.$$

- We also proved that

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) = \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) + \sigma^2\boldsymbol{D}\boldsymbol{D}^T.$$

- For the coordinate-wise theorem, we only used the diagonal entries of $\boldsymbol{D}\boldsymbol{D}^T$.
- For the full theorem, we use the stronger fact that

$$\boldsymbol{D}\boldsymbol{D}^T$$

  is positive semidefinite.

Proof of the full Gauss–Markov Theorem

- Recall that
$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) - \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2 \boldsymbol{D}\boldsymbol{D}^T.$$
- Therefore, it only remains to show that
$$\boldsymbol{D}\boldsymbol{D}^T$$
is positive semidefinite.
- Let $\boldsymbol{u} \in \mathbb{R}^{p+1}$ be any vector. Then
$$\boldsymbol{u}^T(\boldsymbol{D}\boldsymbol{D}^T)\boldsymbol{u} = (\boldsymbol{D}^T\boldsymbol{u})^T(\boldsymbol{D}^T\boldsymbol{u}) = \|\boldsymbol{D}^T\boldsymbol{u}\|^2.$$
- Since a squared norm is always nonnegative,
$$\|\boldsymbol{D}^T\boldsymbol{u}\|^2 \geq 0 \qquad \text{for every } \boldsymbol{u}.$$
- Hence,
$$\boldsymbol{D}\boldsymbol{D}^T$$
is positive semidefinite, and therefore
$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) - \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) \succeq 0.$$

Interpretation of the full theorem

- The coordinate-wise version says that, for every coefficient β_j,

$$\mathbb{V}(\hat{\beta}_j^*) \geq \mathbb{V}(\hat{\beta}_j^{LS}).$$

- The full matrix version is stronger: it compares the entire covariance matrices, not only their diagonal entries.
- Thus, least squares is optimal not only coefficient by coefficient, but also in the joint multivariate sense.
- This is why $\hat{\beta}_{LS}$ is called the **Best Linear Unbiased Estimator (BLUE)**.

What the full theorem really means

- The matrix statement says

$$\mathbb{V}(\hat{\boldsymbol{\beta}}_*) - \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) \succeq 0.$$

- By definition of positive semidefinite, for every vector $\boldsymbol{a} \in \mathbb{R}^{p+1}$,

$$\boldsymbol{a}^T \left(\mathbb{V}(\hat{\boldsymbol{\beta}}_*) - \mathbb{V}(\hat{\boldsymbol{\beta}}_{LS}) \right) \boldsymbol{a} \geq 0.$$

- But this is exactly

$$\mathbb{V}(\boldsymbol{a}^T \hat{\boldsymbol{\beta}}_*) - \mathbb{V}(\boldsymbol{a}^T \hat{\boldsymbol{\beta}}_{LS}) \geq 0.$$

- Therefore, every linear combination

$$\boldsymbol{a}^T \hat{\boldsymbol{\beta}}$$

  is estimated with the smallest possible variance by least squares.

- This is the strongest interpretation of the BLUE property.