SF 2930 Regression Analysis
Lecture 4
Multiple Linear Regression, Part 2

Timo Koski

KTH Royal Institute of Technology

2023

Learning outcomes

- Gauss Markov Theorem
- Prediction
- Statistical Properties of the Residuals
- Distributions of Quadratic Forms
    - χ^2-distribution

Part I: Refreshment

Selected Findings from Preceding Lectures Required in this Lecture.

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 $\hat{\varepsilon} \leftrightarrow \mathbf{e}_{LSE}$

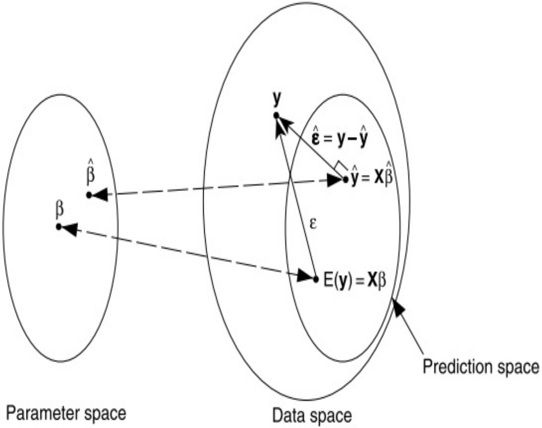


Figure 7.4 Geometric relationships of vectors associated with the multiple linear regression model.

Projection Geometrically for Simple Linear Regression

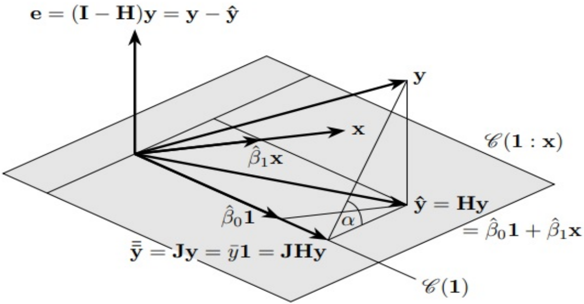


Figure 8.3 Projecting $\mathbf{y}$ onto $\mathscr{C}(\mathbf{1} : \mathbf{x})$.

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

Linear Transformations

Proposition

$\mathbf{Y}$ and $\mathbf{X}$ are random vectors, $\mu_{\mathbf{Y}} = E[\mathbf{Y}]$, $\mu_{\mathbf{X}} = E[\mathbf{X}]$, $\mathbf{X}$ has covariance matrix $C_{\mathbf{X}}$, A and B are $m \times n$ matrices. $\mathbf{a}$ and $\mathbf{b}$ are vectors of suitable dimension. Then we have

- $E[\mathbf{X} + \mathbf{Y}] = \mu_{\mathbf{X}} + \mu_{\mathbf{Y}}$
- $\mathbf{Z} = A\mathbf{X} + \mathbf{b}$,
$$E[\mathbf{Z}] = A\mu_{\mathbf{X}} + \mathbf{b}, \tag{1}$$
$$C_{\mathbf{Z}} = AC_{\mathbf{X}}A^T. \tag{2}$$
- $C_{\mathbf{X}} = E\left[\mathbf{X}\mathbf{X}^T\right] - \mu_{\mathbf{X}}\mu_{\mathbf{X}}^T$
- $$\mathrm{Var}\left[\mathbf{a}^T\mathbf{X}\right] = \mathbf{a}^T C_{\mathbf{X}}\mathbf{a} \tag{3}$$

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

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

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

The (Ordinary) Multiple Linear Regression Model. $k > 1$ Covariates/Predictors

$\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, $n \geq k+1$.

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{4}$$

The following assumptions hold:

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

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

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

Least Squares Estimation

$$Q(\beta) := \| \mathbf{y} - X\beta \|^2 \tag{5}$$

and the LSE is the minimizer

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

Proposition

If $X^T X$ is a positive definite matrix, then

$$\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{y}. \tag{6}$$

Hat Matrix

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

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

SUMMARY: Ordinary Multiple Regression

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \varepsilon. \quad \text{True model} \tag{8}$$

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

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

$$C_{\widehat{\boldsymbol{\beta}}} = \sigma^2 (X^T X)^{-1} \tag{9}$$

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

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

SUMMARY: Normal (Gaussian) Multiple Regression

$\varepsilon \in N\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$ and $\boldsymbol{\beta}_*$ such that

$$\mathbf{Y} = X\boldsymbol{\beta}_* + \varepsilon \quad \text{True model} \tag{10}$$

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

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}_*, \sigma^2 (X^T X)^{-1}\right) \tag{11}$$

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

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)} \mathbf{e}_{LSE}^T \mathbf{e}_{LSE} \tag{12}$$

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)} \mathbf{e}_{LSE}^T \mathbf{e}_{LSE} \tag{13}$$

The χ^2 distribution of the quadratic form in the LSE residuals determing $\widehat{\sigma^2}$, $\mathbf{e}_{LSE}^T \mathbf{e}_{LSE}$, will be eventually derived for the normal multiple regression in this Lecture.

PART II

Gauss-Markov Theorem

Gauss–Markov theorem states that $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$ has uniquely the lowest variance within the class of linear unbiased estimators

Gauss-Markov Theorem

Proposition

Let $\tilde{\beta}$ be any unbiased linear estimator of β_*. Let $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$, i.e. the unbiased ordinary LSE. Then it holds for an arbitrary $(k+1) \times 1$ vector $\mathbf{a}$ that

$$\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] \geq \mathrm{Var}\left[\mathbf{a}^T \widehat{\beta}\right]. \tag{14}$$

If $\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] = \mathrm{Var}\left[\mathbf{a}^T \widehat{\beta}\right]$, then $\tilde{\beta} = \widehat{\beta}$.

Proof: Let $\tilde{\boldsymbol{\beta}} = B\mathbf{Y} + g_o$ be another unbiased linear estimator, where g_o is a $(k+1) \times 1$ vector. Unbiasedness means that $E\left[\tilde{\boldsymbol{\beta}}\right] = \boldsymbol{\beta}_*$. On the other hand, by the rule (1) and the true model (8) above

$$E\left[\tilde{\boldsymbol{\beta}}\right] = E\left[B\mathbf{Y} + g_o\right] = BE\left[\mathbf{Y}\right] + g_o = BX\boldsymbol{\beta}_* + g_o.$$

For unbiasedness it must hold that

$$BX = \mathbb{I}_{k+1}, g_o = \mathbf{0}_{k+1} \tag{15}$$

Now we take without loss of generality that

$$B = (X^T X)^{-1} X^T + G. \tag{16}$$

Then

$$BX = (X^T X)^{-1} X^T X + GX = \mathbb{I}_{k+1} + GX. \tag{17}$$

In view of (15) it holds that

$$GX = \mathbf{0}_{k+1}. \tag{18}$$

Next we find the covariance matrix of $\tilde{\boldsymbol{\beta}}$. By the rule (2) and the true model (8)

$$C_{\tilde{\boldsymbol{\beta}}} = BC_{\mathbf{Y}}B^T = \sigma^2 BB^T.$$

Here

$$\begin{aligned} BB^T &= \left((X^TX)^{-1}X^T + G\right)\left((X^TX)^{-1}X^T + G\right)^T \\ &= \left((X^TX)^{-1}X^T + G\right)\left(X(X^TX)^{-1} + G^T\right) \\ &= (X^TX)^{-1}\underbrace{X^TX(X^TX)^{-1}}_{=\mathbb{I}_{k+1}} + (X^TX)^{-1}\underbrace{X^TG^T}_{=\mathbf{0}_{k+1}^T} \\ &\quad + \underbrace{GX}_{=\mathbf{0}_{k+1}}(X^TX)^{-1} + GG^T, \end{aligned}$$

where we used (18) twice, since $X^TG^T = (GX)^T = \mathbf{0}_{k+1}^T$.

Thus

$$C_{\tilde{\boldsymbol{\beta}}} = \sigma^2 \left((X^T X)^{-1} + GG^T \right) \tag{19}$$

Let now $\mathbf{a}$ be an arbitrary $(k+1) \times 1$ vector. By (3) and (19)

$$\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] = \mathbf{a}^T C_{\tilde{\boldsymbol{\beta}}} \mathbf{a} = \sigma^2 \mathbf{a}^T (X^T X)^{-1} \mathbf{a} + \sigma^2 \mathbf{a}^T GG^T \mathbf{a} \tag{20}$$

Put $\mathbf{z} = G^T \mathbf{a}$. Then $\mathbf{a}^T GG^T \mathbf{a} = \mathbf{z}^T \mathbf{z} = \| \mathbf{z} \|^2 \geq 0$. Hence

$$\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] \geq \sigma^2 \mathbf{a}^T (X^T X)^{-1} \mathbf{a}. \tag{21}$$

Due to (9) we have $\sigma^2 \mathbf{a}^T (X^T X)^{-1} \mathbf{a} = \mathbf{a}^T C_{\widehat{\boldsymbol{\beta}}} \mathbf{a}$. Again, by the rule (2) we have $\mathbf{a}^T C_{\widehat{\boldsymbol{\beta}}} \mathbf{a} = \mathrm{Var}\left[\mathbf{a}^T \widehat{\beta}\right]$. In other words, in (21)

$$\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] \geq \mathrm{Var}\left[\mathbf{a}^T \widehat{\beta}\right],$$

which is (14).

Next we prove uniqueness of $\widehat{\beta}$ as stated in the theorem. We have found in (20)

$$\mathrm{Var}\left[\mathbf{a}^T\tilde{\beta}\right] = \mathrm{Var}\left[\mathbf{a}^T\widehat{\beta}\right] + \sigma^2\mathbf{a}^T G G^T \mathbf{a}$$

Hence if $\mathrm{Var}\left[\mathbf{a}^T\tilde{\beta}\right] = \mathrm{Var}\left[\mathbf{a}^T\widehat{\beta}\right]$, we have for any $\mathbf{a}$ the equality

$$\mathbf{a}^T G G^T \mathbf{a} = 0.$$

As above we set $\mathbf{z} = G^T\mathbf{a}$, and then $\mathbf{a}^T G G^T \mathbf{a} = \mathbf{z}^T\mathbf{z} = \parallel \mathbf{z} \parallel^2 = 0$. But a norm is point-separating, i.e., $\parallel \mathbf{z} \parallel^2 = 0$ implies that $\mathbf{z} = \mathbf{0}_{k+1}$. Thus $G^T\mathbf{a} = \mathbf{0}_{k+1}$ for every $\mathbf{a}$. Hence, we are allowed to take $\mathbf{a}$ as the standard basis vector $\mathfrak{E}_j = (0, \ldots, 0, \underbrace{1}_{\text{position } j}, 0 \ldots, 1)^T$, so that $G^T\mathfrak{E}_j$ is the j:th row of G, which is $= \mathbf{0}^T_{k+1}$. In this way we recognize that every row in G is the zero vector $\mathbf{0}^T_{k+1}$.

Hence $G = \mathbf{0}$, = the $(k+1) \times n$ zero matrix, and by (16) it follows that

$$B = (X^T X)^{-1} X^T.$$

Hence we have shown that $\mathrm{Var}\left[\mathbf{a}^T \tilde{\beta}\right] = \mathrm{Var}\left[\mathbf{a}^T \widehat{\beta}\right]$, implies $\tilde{\beta} = \widehat{\beta}$, i.e., the asserted uniqueness property. $\square$

COROLLARY

For $j = 0, \ldots, k$

$$\mathrm{Var}\left[\tilde{\beta}_j\right] \geq \mathrm{Var}\left[\widehat{\beta}_j\right].$$

Proof: Use the standard basis vectors $\mathfrak{E}_j = (0, \ldots, 0, \underbrace{1}_{\text{position } j}, 0 \ldots, 1)^T$ in (14) for $j = 0, \ldots, k$. $\square$

Remark

The preceding proofs of the proposition and its corollary do not require the multivariate normal distribution, and the Gauss-Markov theorem is thus valid for any ordinary multiple LSE.

Prediction in Real Time

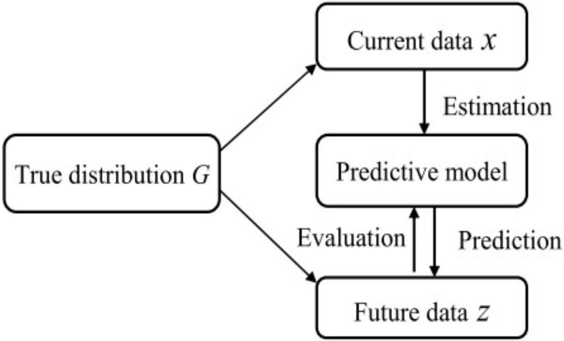


Fig. 1.2. Statistical modeling and the predictive point of view.

Prediction in Real Time

A common situation is that we want to forecast a new value y_{n+1} based on the values of the $\mathbf{x}$- covariates. If we have the LSE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_*$, then an unbiased (to be shown) prediction is

$$\widehat{y}_{n+1} = \mathbf{x}_{n+1}^T \widehat{\boldsymbol{\beta}}, \quad \text{where } \mathbf{x}_{n+1}^T = (1, x_{n+11}, \ldots, x_{n+1k})$$

We can think of prediction in real time. We have observed the responses $y_1 \ldots, y_n$, up to time n and wish to predict the next value, y_{n+1}. We assume, of course, that the underlying "true" mechanism generating data is unchanged in the sense that y_{n+1} is an outcome of

$$Y_{n+1} = \mathbf{x}_{n+1}^T \boldsymbol{\beta}_* + \epsilon_{n+1}.$$

Note that ϵ_{n+1} is assumed to be independent of $\boldsymbol{\varepsilon} = (\epsilon_1 \ldots, \epsilon_n)^T$.

Prediction in Real Time

In order to simplify writing (and to accomodate to other possible cases of prediction) we set

$$\hat{y} = \widehat{y}_{n+1}, \mathbf{x} = \mathbf{x}_{n+1}, \quad e = \epsilon_{n+1}, Y_{n+1} = \mathbf{x}^T \boldsymbol{\beta}_* + e.$$

Prediction in Real Time: Error

Let us proceed to calculate the prediction error $Y_{n+1} - \widehat{y}$. We have

$$Y_{n+1} - \widehat{y} = \mathbf{x}^T \boldsymbol{\beta}_* + e - \mathbf{x}^T \widehat{\boldsymbol{\beta}} = \mathbf{x}^T \left(\boldsymbol{\beta}_* - \widehat{\boldsymbol{\beta}} \right) + e$$

We see that the expected prediction error equals zero

$$E\left[Y_{n+1} - \widehat{y}\right] = \mathbf{x}^T E\left[\left(\boldsymbol{\beta}_* - \widehat{\boldsymbol{\beta}}\right)\right] + E\left[e\right] = 0,$$

as $\widehat{\boldsymbol{\beta}}$ is unbiased and $e \sim N(0, \sigma^2)$. Next apply

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

to get

$$Y_{n+1} - \widehat{y} = -\mathbf{x}^T (X^T X)^{-1} X^T \boldsymbol{\varepsilon} + e.$$

Note that $e \sim N(0, \sigma^2)$ and $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$.

Mean Square Error (MSE) of Prediction in Real Time

We strive to compute $MSE := E\left[(Y_{n+1} - \widehat{y})^2\right]$. Here $\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon$ is a univariate r.v.. We square to get

$$(Y_{n+1} - \widehat{y})^2 = \left(\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon\right)^2 - 2\left(\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\right)\cdot e + e^2. \quad (22)$$

ϵ and e are independent, hence

$$E\left[\left(\mathbf{x}^T(X^TX)^{-1}X^T\epsilon\right)\cdot e\right] = \mathbf{x}^T(X^TX)^{-1}X^T E\left[\epsilon\right]\cdot E\left[e\right] = 0. \quad (23)$$

Mean Square Error (MSE) of Prediction in Real Time

Since $\mathbf{w}^T\mathbf{x} = \mathbf{x}^T\mathbf{w}$ holds for scalar products

$$E\left[\left(\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon\right)^2\right] = E\left[\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon \cdot \mathbf{x}^T(X^TX)^{-1}X^T\varepsilon\right]$$

$$= E\left[\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon \cdot \left(X(X^TX)^{-1}\mathbf{x}\right)^T\varepsilon\right]$$

$$= E\left[\mathbf{x}^T(X^TX)^{-1}X^T\varepsilon\varepsilon^T\left(X(X^TX)^{-1}\mathbf{x}\right)\right]$$

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

$$= \sigma^2\mathbf{x}^T(X^TX)^{-1}\mathbf{x} \tag{24}$$

Mean Square Error (MSE) of Prediction in Real Time

Since $E\left[e^2\right] = \sigma^2$, (22), (23) and (24) entail

$$MSE := E\left[(Y_{n+1} - \widehat{y})^2\right] = \sigma^2 \left(\mathbf{x}_{n+1}^T (X^T X)^{-1} \mathbf{x}_{n+1} + 1\right)$$

By (9) and (3)

$$MSE = \mathbf{x}_{n+1}^T C_{\widehat{\beta}} \mathbf{x}_{n+1} + \sigma^2 = \mathrm{Var}\left[\mathbf{x}_{n+1}^T \widehat{\beta}\right] + \sigma^2.$$

By the Gauss-Markov Theorem, there is no linear unbiased one-step predictor in real time with smaller MSE.

An Example of Prediction in Real Time: Quality of Wine Given Weather Predictor Data in the Current Year $(n+1)$ before Crushing, Extraction, Fermentation e.t.c..

We do ordinary multiple regression with three predicting variables x_1, x_2, x_3, and $Y =$ quality of wine, observed up to year n with

$$x_1 = \text{Precipitation during the winter months}$$

$$x_2 = \text{Average temperature during growing season}$$

$$x_3 = \text{Precipitation during harvesting season}$$

These are now the variables $\mathbf{x}_{n+1}^T = (1, x_{n+11}, x_{n+12}, x_{n+13})$.

Multiple Linear Regression: Quality of Wine

Statistical predictor (SPR) for prediction of the annual quality of Bordeaux wine (i.e., before anyone has tasted it) due to Orley Ashenfelter[1] is

$$\widehat{y}_{n+1} = \mathbf{x}_{n+1}^T \widehat{\boldsymbol{\beta}}, \quad \text{where } \mathbf{x}_{n+1}^T = (1, x_{n+11}, x_{n+12}, x_{n+13})$$

with

$$\widehat{\boldsymbol{\beta}} = (12.145, 0.00117, 0.0614, 0.00386)^T$$

$$\text{Wine Quality} = 12.145 + 0.00117 x_1 + 0.0614 x_2 - 0.00386 x_3$$

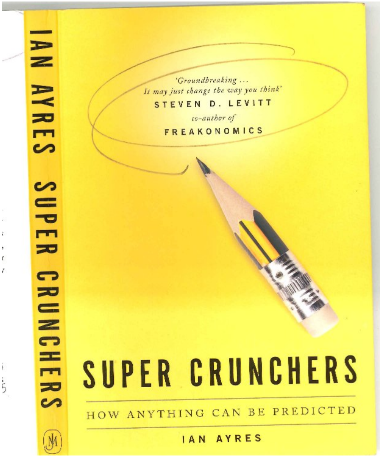
This is reported to be a succesful predictor, but is met with resent and embarrasment by many excellent experts on wine tasting.

[1] Ashenfelter, Orley: Predicting the quality and prices of Bordeaux wine. *Journal of Wine Economics*, 5, 1, 40–52, 2010

A Pocket Book on SPR

For Orley Ashenfelter's SPR for the quality of wine see also

I.Ayres:*Super Crunchers. How anything can be predicted*. John Murray (Publishers), Paperback edition 2008, London.



A quote from I.Ayres: Super Crunchers

For a very wide range of prediction problems, statistical prediction rules (SPRs), often rules that are very easy to implement, make predictions that are as reliable as, and typically more reliable than, human experts. The success of SPRs forces us to reconsider our views about what is involved in understanding, explanation, and good reasoning.

(AI ?)

Climate Prediction Japan Meteorological Agency (JMA).

For the setting of JMA[2] we think first of

$$y(t) = \beta_0 + \beta_1 x_1(t) + \cdots + \beta_k x_k(t) + \varepsilon(t)$$

where t is continuous time. $y(t)$ is the temperature (^oC) in Tokyo at time t. The design: we sample these responses and covariates at n times $t_1, \ldots, t_n$ (winter) and set $y_i = y(t_i)$, $x_{ij} = x_j(t_i)$, $\varepsilon_i = \varepsilon(t_i)$ for $i = 1, \ldots, t_n$ $j = 1, \ldots, k$. Hence we obtain the ordinary multiple regression equations:

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

[2] Source: Statistical Methods for long-range forecast. By Syunji Takahashi / Climate Prediction Division/ JMA

Situation of Multiple Regression Model

	predictand	predictors								
year	Temp	NHPV	EEPV	NHZI	EEZI	OKHOTK	MIDH	OKINAW	OGASH	WPAH
1980	-13	-4.0	-9.2	6.1	-24.4	27.9	0.4	13.8	9.6	0.6
1981	8	-4.9	8.5	-13.6	-4.8	-1.2	0.3	-3.3	-5.0	-2.1
1982	-21	1.9	13.7	2.1	11.1	-8.2	-11.5	-21.4	-3.8	6.5
1983	-11	-25.4	-4.8	-12.7	-23.9	-6.1	-11.6	-3.9	2.9	14.9
1984	7	0.7	14.4	1.0	7.6	-19.9	-5.6	-6.6	-21.9	1.9
1985	5	16.4	16.6	10.2	5.5	-20.5	-6.8	-12.8	-11.6	-1.7
1986	-12	-14.8	-12.2	5.6	-28.3	11.1	-10.3	-8.7	-15.2	-2.4
1987	9	20.2	21.1	-0.4	3.6	-12.2	-2.7	5.9	8.5	0.0
1988	-19	8.7	14.8	3.2	11.2	33.8	2.6	-2.6	1.3	7.4
1989	-9	-18.4	-58.7	-17.3	0.8	26.1	2.2	-7.1	-1.7	-0.6
1990	6	25.1	19.1	10.6	3.5	-4.9	4.0	9.5	1.5	-2.3
1991	5	5.6	31.4	-7.2	-28.5	-0.8	-0.8	9.6	6.1	-0.9
1992	0	-20.8	-42.9	3.6	-0.8	-3.9	-4.1	-5.0	3.7	1.9
1993	-21	30.5	-12.0	-13.9	-24.8	7.1	-12.7	-3.0	2.8	9.3
1994	24	-33.2	-39.8	14.1	29.4	-14.2	13.8	8.0	-0.5	2.8
1995	5	-6.4	3.9	-12.8	-3.5	-15.5	-1.0	11.2	11.6	3.9
1996	6	-32.5	-27.9	4.0	6.2	1.4	5.2	10.6	-3.4	2.8
1997	4	-3.3	-8.7	-17.4	12.1	-5.3	1.5	-4.5	-2.3	0.4
1998	5	43.5	33.6	2.1	-27.1	50.0	14.2	14.5	16.0	-5.1
1999	4	3.2	19.5	14.9	43.6	-1.6	9.2	-6.6	-0.6	-14.3
2000	15	-3.1	-9.2	-0.1	-2.4	4.1	4.6	-7.4	-4.4	-15.1

Predictand Vector

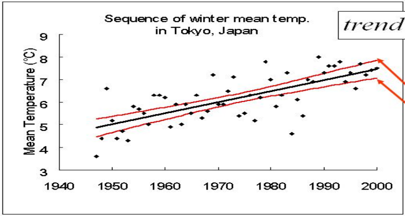
Independent Data

Predictor Matrix

The confidence intervals in the next figure are to be derived later.

SOURCE: STATISTICAL METHODS FOR LONG-RANGE FORECAST. BY SYUNJI TAKAHASHI CLIMATE PREDICTION DIVISION JMA

Property of Calculated Trend



$$trend = 4.90 \quad (°C/100 years)$$

Confidence Interval of the regression line

Confidence Interval of the estimated trend

$$3.6 < trend < 6.10 \quad (°C/100 years)$$

Warming trend in Tokyo is significant

Part III: Coefficient of Determination

$$R^2 = \frac{\widehat{\boldsymbol{\beta}}^T X^T \mathbf{y} - n\bar{y}^2}{\mathbf{y}^T \mathbf{y} - n\bar{y}^2}$$

R^2 is the fraction of response variance that is captured by the model.

Auxiliaries on LSE Residuals $\mathbf{e}_{LSE}$

$$\varepsilon = \mathbf{Y} - X\beta \quad \text{true residuals, unobservable r.v.}$$

$$\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\beta} \quad \text{observed LSE residuals as a random vector}$$

$\mathbf{e}_{LSE} = \mathbf{y} - X\widehat{\beta} = \mathbf{y} - H\mathbf{y} = \mathbf{y} - \widehat{\mathbf{y}}$ observed outcome of $\widehat{\varepsilon}$,
$\mathbf{e}_{LSE} = (\widehat{e}_1, \widehat{e}_2, \ldots, \widehat{e}_n)^T$.

- $X^T \mathbf{e}_{LSE} = \mathbf{0}_k$ (Check this!). When you look at the scalar product of the first row in X^T and $\mathbf{e}_{LSE}$ this means

$$\sum_{i=1}^{n} \widehat{e}_i = 0. \tag{25}$$

- Since $\widehat{y}_i = y_i + \widehat{e}_i$, (25) gives

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{y}_i = \bar{y}. \tag{26}$$

Fundamental Analysis of Variance Identity

LEMMA

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\widehat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n} \widehat{e}_i^2 \tag{27}$$

Proof: Let $\mathbf{1}_n$ be the $n \times 1$ vector with all entries equal to 1. We need first to study the $n \times n$ centering matrix C_{ce} defined by

$$C_{ce} := \mathbb{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T.$$

It has been discussed earlier that C_{ce} is idempotent and symmetric.

Fundamental Analysis of Variance Identity

Take next an $n \times 1$ vector $\mathbf{a}$. Then

$$C_{ce}\mathbf{a} = \mathbb{I}_n \mathbf{a} - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^T \mathbf{a} = \mathbf{a} - \frac{\sum_{i=1}^{n} a_i}{n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 - \bar{a} \\ a_2 - \bar{a} \\ \vdots \\ a_n - \bar{a} \end{pmatrix}. \tag{28}$$

Fundamental Analysis of Variance Identity

We compute the quadratic form $\mathbf{a}^T C_{ce} \mathbf{a}$. Idempotency, symmetry and (28) entail

$$\mathbf{a}^T C_{ce} \mathbf{a} = \mathbf{a}^T C_{ce} C_{ce} \mathbf{a} = \left(C_{ce}^T \mathbf{a} \right)^T C_{ce} \mathbf{a}$$

$$= \left(C_{ce} \mathbf{a} \right)^T C_{ce} \mathbf{a} = \sum_{i=1}^{n} (a_i - \bar{a})^2 \,,$$

that is

$$\mathbf{a}^T C_{ce} \mathbf{a} = \sum_{i=1}^{n} (a_i - \bar{a})^2 \,. \tag{29}$$

Fundamental Analysis of Variance Identity

We have $\mathbf{y} = H\mathbf{y} + \mathbf{e}_{LSE}$, where $\widehat{\mathbf{y}} = H\mathbf{y}$. Thus

$$C_{ce}\mathbf{y} = C_{ce}\widehat{\mathbf{y}} + C_{ce}\mathbf{e}_{LSE}.$$

Since $\sum_{i=1}^{n} \widehat{e}_i = 0$, see (25) above, we have

$$C_{ce}\mathbf{e}_{LSE} = \mathbf{e}_{LSE} - \frac{1}{n}\mathbf{1}_n \underbrace{\mathbf{1}_n^T\mathbf{e}_{LSE}}_{=\sum_{i=1}^{n}\widehat{e}_i=0} = \mathbf{e}_{LSE}.$$

i.e.,

$$C_{ce}\mathbf{e}_{LSE} = \mathbf{e}_{LSE}. \tag{30}$$

Hence

$$\mathbf{y}^T C_{ce} = \widehat{\mathbf{y}}^T C_{ce} + \mathbf{e}_{LSE}^T.$$

Fundamental Analysis of Variance Identity

$\mathbf{y}^T C_{ce} = \widehat{\mathbf{y}}^T C_{ce} + \mathbf{e}_{LSE}^T$. Then we multiply and use symmetry, idempotency and (30)

$$
\begin{aligned}
\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} &= \left(\widehat{\mathbf{y}}^T C_{ce} + \mathbf{e}_{LSE}^T\right)\left(C_{ce}\widehat{\mathbf{y}} + \mathbf{e}_{LSE}\right) \\
&= \widehat{\mathbf{y}}^T C_{ce} C_{ce} \widehat{\mathbf{y}} + \widehat{\mathbf{y}}^T \underbrace{C_{ce}\mathbf{e}_{LSE}}_{=\mathbf{e}_{LSE}} \\
&\quad + \underbrace{\underbrace{\mathbf{e}^T C_{ce} \widehat{\mathbf{y}}}}_{= (\mathbf{C}_{ce}^T \mathbf{e}_{LSE})^T \widehat{\mathbf{y}} = (C_{ce}\mathbf{e}_{LSE})^T \widehat{\mathbf{y}} = \mathbf{e}_{LSE}^T \widehat{\mathbf{y}}} + \mathbf{e}_{LSE}^T \mathbf{e}_{LSE} \\
&= \widehat{\mathbf{y}}^T C_{ce} \widehat{\mathbf{y}} + \widehat{\mathbf{y}}^T \mathbf{e}_{LSE} + \mathbf{e}_{LSE}^T \widehat{\mathbf{y}} + \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}. \qquad (31)
\end{aligned}
$$

Fundamental Analysis of Variance Identity : The Final Result

We have found:

$$\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} = \widehat{\mathbf{y}}^T C_{ce} \widehat{\mathbf{y}} + \widehat{\mathbf{y}}^T \mathbf{e}_{LSE} + \mathbf{e}_{LSE}^T \widehat{\mathbf{y}} + \mathbf{e}_{LSE}^T \mathbf{e}_{LSE}. \tag{32}$$

- $\mathbf{y}^T C_{ce} C_{ce} \mathbf{y} = \mathbf{y}^T C_{ce} \mathbf{y} = \sum_{i=1}^n (y_i - \bar{y})^2$ by idempotency and (29).
- $\widehat{\mathbf{y}}^T C_{ce} \widehat{\mathbf{y}} = \sum_{i=1}^n \left(\widehat{y}_i - \bar{\widehat{y}} \right)^2$, again by (29) and by (26).
- $\mathbf{e}_{LSE}^T \widehat{\mathbf{y}} = \widehat{\mathbf{y}}^T \mathbf{e}_{LSE} = 0$, since the LSE residuals are orthogonal to $\widehat{\mathbf{y}} = H\mathbf{y}$, as found in Lecture 3.
- $\mathbf{e}_{LSE}^T \mathbf{e}_{LSE} = \sum_{i=1}^n \widehat{e}_i^2$ by definition of the scalar product.

Hence (27) holds, as claimed. □

Fundamental Analysis of Variance Identity : The Quadratic Forms

The Fundamental Analysis of Variance Identity in (32) is thus also written as

$$\underbrace{\mathbf{y}^T C_{ce} \mathbf{y}}_{=SS_{\mathrm{T}}} = \underbrace{\sum_{i=1}^{n} \left(\widehat{y}_i - \bar{\widehat{y}} \right)^2}_{SS_{\mathrm{R}}} + \underbrace{\mathbf{e}_{LSE}^T \mathbf{e}_{LSE}}_{=SS_{\mathrm{Res}}} . \tag{33}$$

We have found that

$$\sum_{i=1}^{n} \left(\widehat{y}_i - \bar{y} \right)^2 = \sum_{i=1}^{n} \left(\widehat{y}_i - \bar{\widehat{y}} \right)^2$$

The decomposition (33) will turn out to be important in Lecture 5, once we can represent SS_{R} as a quadratic form, too.

Fundamental Analysis of Variance Identity : Illustration

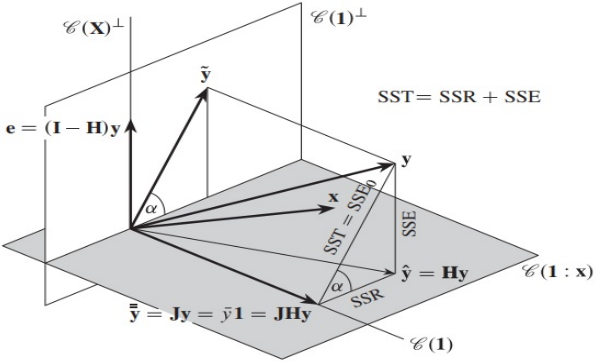


Figure 2 Illustration of SST = SSR + SSE.

By Courtesy of: Puntanen, S. and Isotalo, J. and Styan, GPH
Formulas Useful for Linear Regression Analysis and Related Matrix

Coefficient of Determination, R^2

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\widehat{e}_i^2$$

The **coefficient of determination** R^2 is defined by

$$R^2 \stackrel{\text{def}}{=} \frac{\sum_{i=1}^{n}\left(\widehat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

This was defined for simple linear regression in Lecture 1, but we have now seen that this makes sense for $\widehat{y}$ computed in multiple regression, roo. We shall now show that

$$R^2=\frac{\widehat{\beta}^T X^T\mathbf{y}-n\bar{y}^2}{\mathbf{y}^T\mathbf{y}-n\bar{y}^2} \tag{34}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}.$$

The equality $\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \mathbf{y}^T\mathbf{y} - n\bar{y}^2$ is rule (9) in Appendix C of the slides for Lecture 1. By the same rule we get

$$\sum_{i=1}^{n}\left(\widehat{y}_i - \bar{y}\right)^2 = \widehat{\mathbf{y}}^T\widehat{\mathbf{y}} - n\bar{y}^2.$$

We have $\widehat{\mathbf{y}} = H\mathbf{y} = X\widehat{\boldsymbol{\beta}}$. Hence

$$\begin{aligned}\sum_{i=1}^{n}\left(\widehat{y}_i - \bar{y}\right)^2 &= \widehat{\boldsymbol{\beta}}^T X^T X\widehat{\boldsymbol{\beta}} - n\bar{y}^2 \\ &= \widehat{\boldsymbol{\beta}}^T X^T X(X^T X)^{-1}X^T\mathbf{y} - n\bar{y}^2 \\ &= \widehat{\boldsymbol{\beta}}^T X^T\mathbf{y} - n\bar{y}^2.\end{aligned}$$

Hence we have (34).

We can hence write the regression or model sum of squares as

$$\boxed{SS_R = \widehat{\boldsymbol{\beta}}^T X^T \mathbf{y} - n\bar{y}^2. \qquad (35)}$$

Part IV: Statistical Properties of the LSE Residuals

$$\varepsilon = \mathbf{Y} - X\boldsymbol{\beta}_* \quad \text{true residuals, unobservable r.v.}$$

$$\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\boldsymbol{\beta}} \quad \text{observed LSE residuals as a random vector}$$

The statistical properties of $\widehat{\varepsilon}$ are studied in this part of the lecture.

Part IV: Statistical Properties of the LSE Residuals: Expectation

Proposition

$$E\left[\widehat{\varepsilon}\right] = \mathbf{0}_n. \tag{36}$$

Proof: We use the hat matrix to write $\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\boldsymbol{\beta}} = \mathbf{Y} - H\mathbf{Y} = (\mathbb{I}_n - H)\,Y$, that is

$$\widehat{\varepsilon} = (\mathbb{I}_n - H)\,Y. \tag{37}$$

Then

$$E\left[\widehat{\varepsilon}\right] = (\mathbb{I}_n - H)\,E\left[Y\right] =$$

and the rule (1) with the true model

$$= (\mathbb{I}_n - H)\left(X\boldsymbol{\beta}_* + E\left[\varepsilon\right]\right)$$

$$= (\mathbb{I}_n - H)\,X\boldsymbol{\beta}_* = X\boldsymbol{\beta}_* - HX\boldsymbol{\beta}_*$$

But $HX = X(X^TX)^{-1}X^TX = X$. Hence (39) follows.

Part IV: Statistical Properties of the LSE Residuals: Covariance Matrix

Proposition

$$C_{\widehat{\varepsilon}} = \sigma^2 \left(\mathbb{I}_n - H \right). \tag{38}$$

Proof: By (37) and the rule (2) for covariance matrices of linearly mapped random vectors

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

But in the true model, $C_Y = \sigma^2 \mathbb{I}_n$. Hence

$$C_{\widehat{\varepsilon}} = \sigma^2 \left(\mathbb{I}_n - H \right) \left(\mathbb{I}_n - H \right)^T = \sigma^2 \left(\mathbb{I}_n - H \right) \left(\mathbb{I}_n - H \right) = \sigma^2 \left(\mathbb{I}_n - H \right),$$

where we used the symmetry and idempotency of $\mathbb{I}_n - H$ □

Part IV: Statistical Properties of the LSE Residuals: Multivariate Normal Distribution

Proposition

In the normal true model

$$\widehat{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\left(\mathbb{I}_n - H\right)\right). \tag{39}$$

Proof: The two preceding propositions give the mean vector and covariance matrix as stated. In the true normal model $Y \sim N_n\left(X\boldsymbol{\beta}_*, \sigma^2\mathbb{I}_n\right)$. Since $\widehat{\varepsilon} = \left(\mathbb{I}_n - H\right)Y$, the random vector $\widehat{\varepsilon}$ has a normal distribution. □

Remark

Since

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

we find on the main diagonal that

$$\mathrm{Var}\left(\widehat{\varepsilon}_i\right) = \sigma^2 \left(1 - h_{ii}\right).$$

Because a variance is nonnegative, it must hold that $h_{ii} \leq 1$*. It will be shown later that* $0 < h_{ii} < 1$*, since* H *is idempotent.*

Distributions of Quadratic Forms of Normal Vectors

We shall next study the statistical properties of the quadratic form

$$\widehat{\varepsilon}^T \widehat{\varepsilon}$$

It is shown below that $\widehat{\varepsilon} = (\mathbb{I}_n - H)\, Y = (\mathbb{I}_n - H)\, \varepsilon$. We choose now to continue with

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

since ε is a normal vector with n independent components. First we collect some facts about $\mathbb{I}_n - H$.

Properties of $\mathbb{I}_n - H$

- $\mathbb{I}_n - H$ is symmetric and idempotent. (Check !) Hence $\mathbb{I}_n - H$ is singular (has no inverse matrix) by an Appendix to Lecture 3.
  - Set $A = X(X^T X)^{-1}$, $B = X^T$. By the rule 2. in Appendix B,

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

    Hence by rule 3. in Appendix B,

    $$\mathrm{Tr}\,(\mathbb{I}_n - H) = \mathrm{Tr}\, \mathbb{I}_n - \mathrm{Tr}\, H = n - (k + 1). \tag{40}$$

  - $\mathbb{I}_n - H$ is positive semidefinite by Appendix XXX below
- Since $\mathbb{I}_n - H$ is positive semidefinite, it follows from (40) by a result in Appendix XXX below that

$$\mathrm{rank}\,(\mathbb{I}_n - H) = n - (k + 1). \tag{41}$$

Distributions of Quadratic Form of Normal Vectors

- $\mathbf{X} \sim N_n(\mu, \Sigma)$ is an $n \times 1$ Gaussian vector, where Σ is positive definite. The quadratic form is

$$(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu).$$

We use the factorization of Σ^{-1} in (46) of Appendix XXXX to get

$$(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu) = \left(\Sigma^{-1/2} (\mathbf{X} - \mu)\right)^T \Sigma^{-1/2} (\mathbf{X} - \mu).$$

Let $\mathbf{Z} := \Sigma^{-1/2} (\mathbf{X} - \mu)$. Then our rules of computation give that Z has the mean vector $E[Z] = \mathbf{0}_n$ and Z has the covariance matrix

$$C_{\mathbf{Z}} = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = \Sigma^{-1/2} \Sigma^{1/2} \Sigma^{1/2} \Sigma^{-1/2} = \mathbf{I}_n.$$

Distributions of Quadratic Form of Normal Vectors

I.e., $\mathbf{Z} \sim N_n(\mathbf{0}_n, \mathbb{I}_n)$. Hence

$$(\mathbf{X} - \mu)^T \Sigma^{-1} (\mathbf{X} - \mu) = \mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

This is Thm 9.1. in Gut, Allan:*An Intermediate Course in Probability. Second Edition*. Note that $z_i \sim N(0, 1)$ are indepedent and that a finite sum of sauares of independent standard normal r.v.'s has $\chi^2(n)$ (chi-squared distribution with n degrees of freedom) (as follows by moment generating functions, c.f., the textbook by Allan Gut).

Chi-square

Definition

$X_1, \ldots, X_n$ are i.i.d., $X_i \sim N(0,1)$.

$$W = \sum_{i=1}^{n} X_i^2.$$

W has the **chi-square distribution with** *n* **degrees of freedom***, symbolically $W \sim \chi^2(n)$*

The pdf of W is

$$f(x;n) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Here $\Gamma(.)$ is the Gamma function.

Distributions of a Quadratic Form of OLS Residuals in the Normal Multiple Regression

The purpose of all this is to determine the statistical distribution of the unbiased estimator of σ^2 found earlier as the random variable[3]

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)} \widehat{\varepsilon}^T \widehat{\varepsilon}.$$

[3] $\widehat{\sigma^2}$ in (12) is the outcome of the current quadratic form

Distribution of the Estimator of Variance in the Normal Multiple Regression

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-k-1) \tag{42}$$

$\chi^2(n-k-1)$ is the chi-squared distribution with $n-k-1$ degrees of freedom.

We shall now establish this important fact.

We have by definition of $\widehat{\sigma^2}$

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} = \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2}.$$

Degrees of Freedom ?

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} = \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2} \sim \chi^2(n-k-1) \tag{43}$$

is now our claim to be proved. We start by re-checking from Lecture 3 that $\widehat{\varepsilon} = (\mathbb{I}_n - H)\,\varepsilon$.

By the true model and the hat matrix $H = X(X^T X)^{-1} X^T$

$$\widehat{\varepsilon} = (\mathbb{I}_n - H)\,\mathbf{Y} = (\mathbb{I}_n - H)\,(X\boldsymbol{\beta}_* + \varepsilon)$$

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

$$= X\boldsymbol{\beta}_* + \varepsilon - X\boldsymbol{\beta}_* - H\varepsilon$$

$$= (\mathbb{I}_n - H)\,\varepsilon$$

Distributions of a Quadratic Form of OLS Residuals in the Normal Multiple Regression

Hence

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} = \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2} = \frac{\varepsilon^T\left(\mathbb{I}_n - H\right)^T\left(\mathbb{I}_n - H\right)\varepsilon}{\sigma^2}.$$

The preceding argument about χ-square distribution of quadratic forms has to be revised, when dealing with

$$\frac{\varepsilon^T\left(\mathbb{I}_n - H\right)^T\left(\mathbb{I}_n - H\right)\varepsilon}{\sigma^2}$$

for the obvious reason that $\mathbb{I}_n - H$ is not invertible.

We quote

PROPOSITION

If $\mathbf{Z} \sim N_n(\mathbf{0}_n, \mathbb{I}_n)$*, then*

$$\mathbf{Z}^T A\mathbf{Z} \sim \chi^2(r) \tag{44}$$

if and only if A *is an idempotent matrix with* $\operatorname{rank} A = r$.

This is Corollary 1 to Theorem 5.5 on pp. 117−118 in Rencher, Alvin C and Schaalje, G Bruce: *Linear Models in Statistics*, 2008. The proof in loc.cit. is based on the moment generating function of the quadratic form. Details are omitted here.

Distribution of a Quadratic Form of OLS Residuals in the Normal Multiple Regression

We apply this with $A = \mathbb{I}_n - H$ and $\mathbf{Z} := \frac{\varepsilon}{\sigma}$. Then (check this) $\mathbf{Z} \sim N_n\left(\mathbf{0}_n, \mathbb{I}_n\right)$ and

$$\frac{\varepsilon^T \left(\mathbb{I}_n - H\right)^T \left(\mathbb{I}_n - H\right) \varepsilon}{\sigma^2} = \mathbf{Z}^T \left(\mathbb{I}_n - H\right)^T \left(\mathbb{I}_n - H\right) \mathbf{Z}.$$

By the proposition 7 above we have by idempotency and the previous computation of $\mathrm{rank}(\mathbb{I}_n - H)$

$$\frac{1}{\sigma^2}\widehat{\varepsilon}^T\widehat{\varepsilon} = \mathbf{Z}^T \left(\mathbb{I}_n - H\right)^T \left(\mathbb{I}_n - H\right) \mathbf{Z} = \mathbf{Z}^T \left(\mathbb{I}_n - H\right) \mathbf{Z} \sim \chi^2(n-k-1)$$

as was to be proved. $\square$

APPENDIX A

Appendix B : 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:

$$\mathrm{Tr} \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} = \sum_1^k a_{jj}$$

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 a is a column-vector, then $a^T a = \mathrm{Tr}\left(aa^T\right)$.
- 3. For any real numbers a and b, $\mathrm{Tr}(aC + bD) = a\,\mathrm{Tr}\, C + b\,\mathrm{Tr}\, D$

Appendix : Wonderings

A Wondering

We have also the identity $\mathbf{Y} = \widehat{\mathbf{Y}} + \widehat{\varepsilon}$, where $\widehat{\mathbf{Y}} = H\mathbf{Y}$. Then

$$\widehat{\varepsilon} = (\mathbb{I}_n - H)\,\mathbf{Y} =$$

$$= \widehat{\mathbf{Y}} + \widehat{\varepsilon} - H\widehat{\mathbf{Y}} - H\widehat{\varepsilon} = \widehat{\mathbf{Y}} + \widehat{\varepsilon} - HH\mathbf{Y} - H\widehat{\varepsilon}$$

$$= H\mathbf{Y} + \widehat{\varepsilon} - H\mathbf{Y} - H\widehat{\varepsilon}$$

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

as $H\widehat{\varepsilon} = X(X^TX)^{-1}X^T\widehat{\varepsilon} = \mathbf{0}_n$, since

$$X^T\widehat{\varepsilon} = X^T\left(\mathbf{Y} - \widehat{\mathbf{Y}}\right) = X^T\left(\mathbf{Y} - H\mathbf{Y}\right)$$

$$= X^T\mathbf{Y} - X^T H\mathbf{Y} = X^T\mathbf{Y} - X^T X(X^TX)^{-1}X^T\mathbf{Y} = X^T\mathbf{Y} - X^T\mathbf{Y} = \mathbf{0}_{k+1}.$$

Factorization and Square Root of Covariance Matrices

If Σ is an $n \times n$ symmetric matrix, then Σ can be written as

$$\Sigma = ADA^T,$$

where A is an orthogonal matrix ($A^T A = AA^T = I_n$) and D is an $n \times n$ diagonal matrix, with the eigenvalues on the main diagonal. If Σ is a covariance matrix, its eigenvalues λ_i are non-negative. Then

$$\Sigma^{1/2} = AD^{1/2}A^T,$$

where $D^{1/2}$ is an $n \times n$ diagonal matrix, with $\sqrt{\lambda_i}$ on the main diagonal. One checks now that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.

Factorization and Square Root of Covariance Matrices

When Σ is positive definite, its eigenvalues are positive and we define

$$\Sigma^{-1/2} = AD^{-1/2}A^T, \qquad (45)$$

where $D^{-1/2}$ is an $n \times n$ diagonal matrix, with $1/\sqrt{\lambda_i}$ on the main diagonal. $\Sigma^{-1/2}$ is symmetric, since $D^{-1/2}$ is symmetric. Clearly, $D^{-1/2}D^{-1/2} = D^{-1}$. Then

$$\Sigma^{-1} = \Sigma^{-1/2}\Sigma^{-1/2}. \qquad (46)$$

Appendix D: On Symmetric Idempotent Matrices

- For A positive semidefinite,

$$\operatorname{rank} A = \text{ the number of positive eigenvalues of } A. \tag{47}$$

- For any square A

$$\operatorname{Tr} A = \text{ the sum of eigenvalues of } A. \tag{48}$$

If A is a singular, symmetric and idempotent, then A is positive semidefinite.

- *Proof*: By symmetry $A = A^T$, and by idempotency $A^2 = A$. Then

$$A = A^2 = AA = AA^T.$$

But then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T AA^T\mathbf{x} = \left(A^T\mathbf{x}\right)^T A^T\mathbf{x} = \parallel A\mathbf{x} \parallel^2 \geq 0.$$

□

Eigenvalues of a symmetric and idempotent matrix

If A is an $n \times n$ symmetric and idempotent matrix with $\operatorname{rank} A = r$, then A has r eigenvalues equal to 1 and $n - r$ eigenvalues equal to 0.

- *Proof*: Let $\mathbf{x}$ satisfy $A\mathbf{x} = \lambda \mathbf{x}$. Then

$$A^2\mathbf{x} = A(A\mathbf{x}) = \lambda A\mathbf{x} = \lambda^2\mathbf{x}.$$

Also

$$A^2\mathbf{x} = A\mathbf{x} = \lambda\mathbf{x}.$$

That is, $\lambda^2\mathbf{x} = \lambda\mathbf{x} \Leftrightarrow (\lambda - \lambda^2)\mathbf{x} = \mathbf{0}$. But an eigenvector is not the zero vector. Hence $(\lambda - \lambda^2) = \lambda(1 - \lambda) = 0$, which holds for $\lambda = 0$ and $\lambda = 1$.
Since A is positive semidefinite, by (47) there are r eigenvalues equal to 1 and $n - r$ eigenvalues equal to 0. □

Eigenvalues of a symmetric and idempotent matrix

If A is an $n \times n$ symmetric and idempotent matrix with $\operatorname{rank} A = r$, then $\operatorname{Tr} A = r$.

- *Proof*: This follows by (48), as by the preceding statement the sum eigenvalues of A is r. $\square$

The following can be established without direct reference to momentgenerating functions, as shown next.

PROPOSITION

If $\mathbf{Z} \sim N_n(\mathbf{0}_n, \mathbb{I}_n)$ *and* A *is an idempotent matrix with* $\operatorname{rank} A = r$*, then*

$$\mathbf{Z}^T A \mathbf{Z} \sim \chi^2(r). \tag{49}$$

Proof: Söderström -Stoica: System Identification. Prentice Hall, 1986.