SF 2930 Regression Analysis

Lecture 5

Multiple Linear Regression, Part 3

Timo Koski

KTH Royal Institute of Technology

2023

Learning outcomes

- Repetition of Distributions Related to the Normal Distribution
  - χ^2 distribution
  - Student's t-distribution
  - F-distribution
- Independence of Quadratic Forms
- Marginal Confidence Intervals for β in Ordinary Normal Multiple Regression.
- non-central F-distribution
- F-statistic, ANOVA Table for Significance of Multiple Regression
- distribution

Part I: Refreshment

Selected Topics from Preceding Lectures Required in this Lecture.

The (Ordinary) Multiple Linear Regression Model

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

$$\mathbf{Y} = X\beta + \varepsilon. \tag{1}$$

The following assumptions hold:

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

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

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

Hat Matrix

$$H := X(X^T X)^{-1} X^T. \tag{2}$$

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

SUMMARY: ORDINARY MULTIPLE REGRESSION

$$\mathbf{Y} = X\beta_* + \varepsilon. \quad \text{True model} \tag{3}$$

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

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

$$C_{\widehat{\beta}} = \sigma^2 (X^T X)^{-1} \tag{4}$$

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

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

SUMMARY: NORMAL (GAUSSIAN) MULTIPLE REGRESSION

$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbb{I}_n)$ and β_* such that

$$\mathbf{Y} = X\beta_* + \varepsilon \quad \text{True model} \tag{5}$$

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

$$\widehat{\beta} \sim N_{k+1}\left(\beta_*, \sigma^2 (X^T X)^{-1}\right) \tag{6}$$

and as shown in Lecture 3

$$\widehat{\beta} = \beta_* + (X^T X)^{-1} X^T \varepsilon \tag{7}$$

$$\widehat{\varepsilon} = \mathbf{Y} - H\mathbf{Y} = \varepsilon - H\varepsilon. \tag{8}$$

$$\widehat{\varepsilon} \sim N_n\left(\mathbf{0}_n, \sigma^2\left(\mathbb{I}_n - H\right)\right). \tag{9}$$

$$\widehat{\beta} \sim N_{k+1}\left(\beta_*, \sigma^2(X^T X)^{-1}\right) \tag{10}$$

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

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)}\mathbf{e}_{LSE}^T\mathbf{e}_{LSE} \tag{11}$$

$$(n-k-1)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(n-k-1) \tag{12}$$

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

Probability Distributions Related to the Normal Distribution:

- χ^2-distribution
- Student's t-distribution
- F-distribution

Probability Distributions Related to the Normal Distribution: chi-square

Definition

$\mathbf{X} \sim N_n(\mathbf{0}, \mathbb{I}_n)$ *(i.e.* $X_1, \ldots, X_n$ *are i.i.d.,* $X_i \sim N(0,1)$*).*

$$W := \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} X_i^2.$$

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

The pdf of W is

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

Student's distribution

Definition

$X \sim N(\mu, \sigma^2)$, $Z \sim \chi^2(n)$ are independent r.v.'s. Set

$$U = \frac{\frac{(X-\mu)}{\sigma}}{\sqrt{\frac{Z}{n}}} \tag{13}$$

Gut (2009) Chapter 1, Section 3, Problem 9: U has the (Student's) t-distribution with n degrees of freedom, symbolically $U \sim t(n)$

Student's t-distribution $t(n)$, pdf

The pdf of $t(n)$ is

$$f(t;\, n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < +\infty \tag{14}$$

When $n = 1$ we get, as $\Gamma(\frac{1}{2}) = \sqrt{\pi}$,$\Gamma(1) = 0! = 1$,

$$f(t;\, 1) = \frac{1}{\pi\left(1+t^2\right)},$$

which is the Cauchy distribution $C(0, 1)$.

The pdf of F-distribution with (n_1, n_2) degrees of freedom

$$f(x; n_1, n_2) = \frac{1}{\mathrm{B}\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{n_1/2} x^{n_1/2-1} \left(1 + \frac{n_1}{n_2} x\right)^{-(n_1+n_2)/2}, \tag{15}$$

where we have the Beta function

$$\mathrm{B}\left(\frac{n_1}{2}, \frac{n_2}{2}\right) = \frac{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2} + \frac{n_2}{2}\right)}$$

Probability Distributions Related to the Normal Distribution: F-distribution

Proposition

If $X_1 \sim \chi^2(n_1)$ *and* $X_2 \sim \chi^2(n_2)$ *are independent. Let*

$$V := \frac{X_1/n_1}{X_2/n_2} \tag{16}$$

V has the F-distribution with (n_1, n_2) *degrees of freedom, coded as* $V \sim F(n_1, n_2)$

This proposition is Problem 10. of Chapter 1 Section 3, in Gut, Allan: An Intermediate Course in Probability. Second Edition, Springer, 2009.

A Digression

How to find the pdf of $V = \frac{X_1/n_1}{X_2/n_2}$? A general trick: We have the ratio $Z = \frac{X}{Y}$ of two independent random variables with pdfs $f_X(x)$ and $f_Y(y)$, respectively. First, transform $(X, Y) \mapsto (Z, U)$ by

$$Z = \frac{Y}{X}, U = Y.$$

Find the Jacobian of the inverse transformation $X = X(Z, U) = ZU, Y = Y(Z, U) = U$ as $U = Y$. Then the rule of transformation of variables in a pdf (Gut Thm. 2.1 p. 21) gives $f_{Z,Y}(z, y) = f_{X,Y}(zy, y)|y| = f_X(zy)f_Y(y)|y|$ (by independence) and thus

$$f_Z(z) = \int_{-\infty}^{+\infty} |y| f_Y(y) f_X(zy) dy$$

Try the handwork at home to obtain (15) with (16). (This is NOT EXQ).

Step 1.: $\hat{\beta}$ and $\hat{\sigma}^2$ are independent

Independence of Linear Forms

Proposition

$\mathbf{X} \sim N(\mu, C)$. *Let A and B be conforming matrices. Then $A\mathbf{X}$ and $B\mathbf{X}$ are independent random variables if and only*

$$ACB^T = \mathbf{O}. \tag{17}$$

If A and B are symmetric and $C = \sigma^2\mathbb{I}$, then $A\mathbf{X}$ and $B\mathbf{X}$ are independent random variables if and only

$$AB = \mathbf{O}. \tag{18}$$

The proof is short showing that the joint momentgenerating function of the pair $(A\mathbf{X}, B\mathbf{X})$ is a product of the momentgenerating function of $A\mathbf{X}$ and the momentgenerating function of $B\mathbf{X}$. The details are omitted here.

A Remarkable Case of Statistical Independence

Proposition

If $\mathbf{Y} \sim N_n(X\beta_*, \sigma^2\mathbb{I}_n)$*, then* $\hat{\beta}$ *and* $\hat{\sigma}^2$ *are independent.*

Proof: We have (c.f. (7) above)

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_* + (X^T X)^{-1} X^T \boldsymbol{\varepsilon} \tag{19}$$

As an object of probability calculus $\hat{\sigma}^2$ is the random variable

$$\widehat{\sigma^2} = \frac{1}{(n-k-1)} \widehat{\boldsymbol{\varepsilon}}^T \widehat{\boldsymbol{\varepsilon}} \tag{20}$$

Hence we conclude by (19) and (20) that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent random variables, soon as $(X^T X)^{-1} X^T \varepsilon$ and $\widehat{\varepsilon}$ are independent random variables.

A Remarkable Case of Statistical Independence

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

Let us note from the previous (c.f. (8) above) that

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

By the true normal model assumptions $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right)$. Hence we consider the proposition above with $\mathbf{X} \leftrightarrow \boldsymbol{\varepsilon}$, $A = (X^T X)^{-1} X^T$ and $B = (\mathbb{I}_n - H)$ and $C = \sigma^2 \mathbb{I}_n$. Note that A is not symmetric, hence we try the case (17).
$A = (X^T X)^{-1} X^T$ is an $(k+1) \times n$ matrix and $B = (\mathbb{I}_n - H)$ is an $n \times n$ matrix. Hence A and B are conformable for the matrix product

$$ACB^T = \sigma^2 AB,$$

as $B = (\mathbb{I}_n - H)$ is symmetric.

A Remarkable Case of Statistical Independence

$$ACB^T = \sigma^2 AB = \sigma^2 (X^T X)^{-1} X^T (\mathbb{I}_n - H)$$
$$= \sigma^2 \left((X^T X)^{-1} X^T - (X^T X)^{-1} X^T H \right).$$

By definition of the hat matrix H we get

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

Hence

$$ACB^T = \sigma^2 \left((X^T X)^{-1} X^T - (X^T X)^{-1} X^T \right) = \mathbf{O}_{(k+1)\times n}$$

and the claim about independence holds as asserted. □

Applications of this Independence in Simple Linear Regression

We know that under the true normal model assumptions

$$\widehat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}_*, \sigma^2(X^T X)^{-1}\right) \tag{21}$$

We consider as our first example the simple linear regression $k = 1$. By (21) and the properties of joint normal distributions (here N_2), see Lecture 2, we know that in $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1\right)^T$, $\hat{\beta}_0$ and $\hat{\beta}_1$ have univariate marginal distrbutions that are normal distributions.

Applications of this Independence in Simple Linear Regression

We treat the case $k = 1$ and present confidence intervals with exact confidence degrees for β_0 and β_1^*, for the predictor as a staight line, and a prediction interval with an exact confidence degree.

It is found in eqn. (21) of Appendix B of slides of Lecture 3 that

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \tag{22}$$

Hence

$$\hat{\beta}_0 \sim N\left(\beta_0^*, \sigma^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1^*, \sigma^2 \frac{1}{S_{xx}}\right)$$

By Proposition (9) in Appendix of Lecture 1, slides, we recall

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

Simple Linear Regression

Thus we have

$$\hat{\beta}_0 \sim N\left(\beta_0^*, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1^*, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right).$$

which agrees with the formula derived in Lecture 1.

Simple Linear Regression

We established in Lecture 4 the distribution of the observed residuals as

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

Now we know in addition that $\widehat{\varepsilon}^T\widehat{\varepsilon}$ and $\hat{\beta}_1$ are independent. We recall a result/an exercise of probability calculus, see, e.g., Gut, Allan: An Intermediate Course in Probability. Second Edition, Springer, 2009, Chapter 1, Section 3, Problem 9.

We apply (13) with $X \mapsto \hat{\beta}_1$ and $Z \mapsto \frac{1}{\sigma^2}\widehat{\varepsilon}^T\widehat{\varepsilon}$. This gives

$$T_n := \frac{(\hat{\beta}_1 - \beta_1^*)}{\frac{\sigma}{\sqrt{S_{xx}}}} \frac{1}{\sqrt{\frac{\frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2}}{n-2}}} \sim t(n-2)$$

Let us recall that $\hat{\sigma}^2 = \frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{n-2}$ is the unbiased estimate of σ^2 . Then

$$T_n = \frac{(\hat{\beta}_1 - \beta_1^*)}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \sim t(n-2). \tag{23}$$

From this we derive a confidence interval (CI) for β_1^* by a well known technique.

Let us choose confidence level $1-\alpha$. By symmetry of $f(t;n-2)$, see (14), we can find (software or by table, section 17.8 in Råde, Lennart and Westergren, Bertil: *Mathematics handbook for science and engineering*) a positive number $t_{\alpha/2}(n-2)$ such that

$$P\left(-t_{\alpha/2}(n-2) \leq T_n \leq t_{\alpha/2}(n-2)\right) = 1-\alpha.$$

We insert T_n from (23). Some equivalent re-writing of the inequalities entails

$$P\left(\hat{\beta}_1 - t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}} \leq \beta_1^* \leq \hat{\beta}_1 + \frac{\hat{\sigma}}{\sqrt{S_{xx}}}\hat{\sigma}\right) = 1-\alpha.$$

This gives us the CI with the confidence degree $1-\alpha$:

$$I_{\beta_1^*} = \left[\hat{\beta}_1 - t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right].$$

CI for β_1^*

At Math.Dept/KTH this CI is also written down in the form

$$I_{\beta_1^*} = \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\,.$$

Test of hypothesis about β_1: CI method

There is no theoretic line of regression when $\beta_1 = 0$. We can make a significance test by the method of CI:

$$H_0 : \beta_1^* = 0$$

mot

$$H_1 : \beta_1^* \neq 0$$

Reject H_0 at the significance level α , if the observed interval does not inlude zero, i.e., reject H_0, if

$$0 \notin I_{\beta_1^*} = \hat{\beta}_1 \pm t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \, .if$$

If $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$ is "large', the length of this CI is small and the CI is thus informative.

EXQ

Show that the CI for β_0 at the confidence level $1-\alpha$ is

$$I_{\beta_0^*} = \hat{\beta}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}.$$

CI for the true line of regression

(3) $\underline{\theta = \beta_0 + \beta_1 x_0}$

Here we derive a CI for the theoretical line of regression at $x = x_0$. Let us first note that

$$E(\hat{\beta}_0 + \hat{\beta}_1 x_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0.$$

and with notations form Lecture 1

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{and} \quad \hat{\beta}_0 = \sum_{i=1}^{n} d_i y_i$$

where

$$c_i = (x_i - \bar{x})/S_{xx} \quad \text{and} \quad d_i = \frac{1}{n} - c_i \bar{x}. \tag{24}$$

$$\mathbf{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \mathbf{Var}\left(\sum_{i=1}^{n} d_i Y_i + \sum_{i=1}^{n} c_i Y_i x_0\right)$$

$$= \mathbf{Var}\left(\sum_{i=1}^{n} (d_i + c_i x_0) Y_i\right)$$

and due to independence

$$= \sum_{i=1}^{n} (d_i + c_i x_0)^2 \mathbf{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} (d_i + c_i x_0)^2 .$$

Confidence Interval

It is shown in Appendix A that

$$\mathbf{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \sum_{i=1}^{n} (d_i + c_i x_0)^2 = \sigma^2 \cdot \left(\frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx} \right).$$

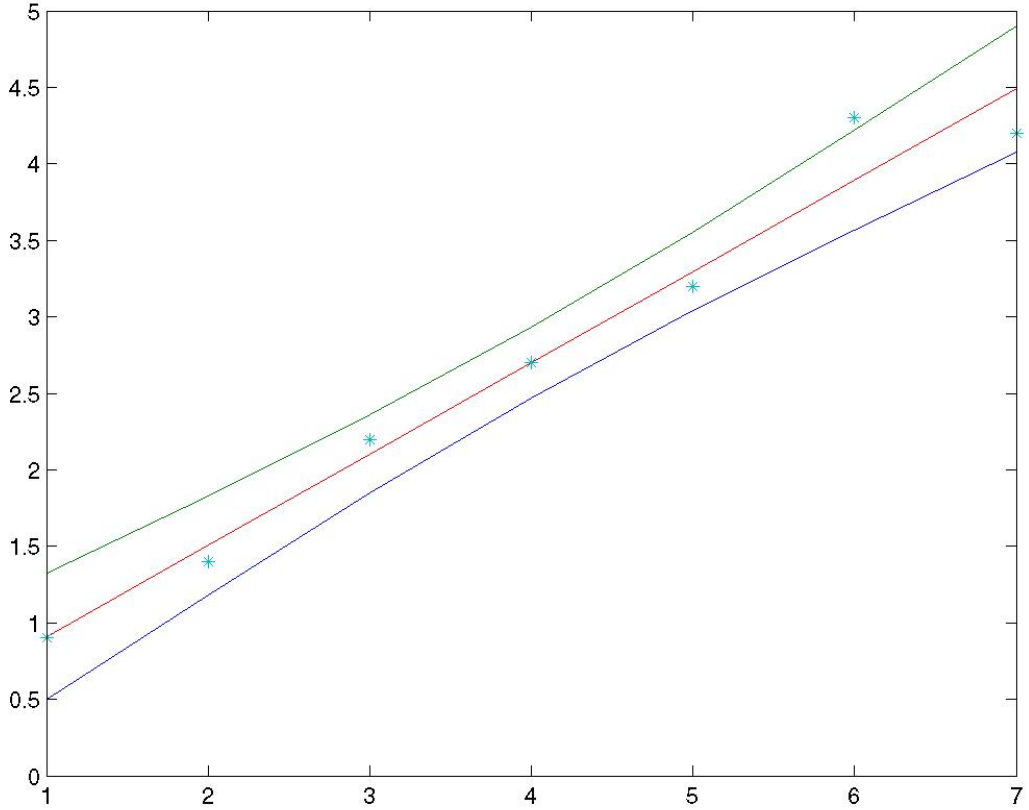
Hence the sought CI is

$$I_{\beta_0+\beta_1 x_0} = \hat{\beta}_0^* + \hat{\beta}_1^* x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}.$$

CI for the Toy Example of Lecture 1. above with $\alpha = 0.05$

Plot of the training set, the predictor and

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{0.05}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$



Prediction Interval

Let Y_0 be a new observed response to x_0, i.e., $Y_0 \in N(\beta_0^* + \beta_1^* x_0, \sigma^2)$. We predict Y_0 by the predictor

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Error of prediction

$$Y_0 - \hat{Y}_0$$

has a normal distribution with mean zero. Since Y_0 and $\hat{Y}_0$ are independent (Why?)

$$\mathbf{Var}\left(Y_0 - \hat{Y}_0\right) = V\left(Y_0\right) + \mathbf{Var}\left(\hat{Y}_0\right)$$

We have already observed above

$$\mathbf{Var}\left(\hat{Y}_0\right) = \mathbf{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \cdot \left(\frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}\right).$$

Prediction Interval

That is,

$$\mathbf{Var}\left(Y_0 - \hat{Y}_0\right) = \sigma^2\left(1 + \frac{1}{n} + (x_0 - \bar{x})^2/S_{xx}\right) \tag{25}$$

We estimate σ^2 with $\hat{\sigma}^2$. However, Since Y_0 and $\hat{Y}_0$ have zreo means,

$$\mathbf{Var}\left(Y_0 - \hat{Y}_0\right) = E\left[\left(Y_0 - \hat{Y}_0\right)^2\right] = MSE$$

The predictor $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$. is a specal case of the predictor $\mathbf{x}_{n+1}^T \widehat{\boldsymbol{\beta}}$ in Lecture 4. Does the MSE of prediction error above agree with the MSE of Lecture 4.?

A Check:

In Lecture 4.

$$MSE = \mathbf{x}_{n+1}^T C_{\widehat{\boldsymbol{\beta}}} \mathbf{x}_{n+1} + \sigma^2.$$

Now insert $C_{\widehat{\boldsymbol{\beta}}}$ from (22) and write $\mathbf{x}_{n+1}^T)(1, x_o)$. Expand the quadratic form and simplify using tricks of finite sums, then this yields (25), as it should.

Prediction Interval

EXQ: Explain why

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}}.$$

has a t-distribution with $n-2$ degrees of freedom. Then show that

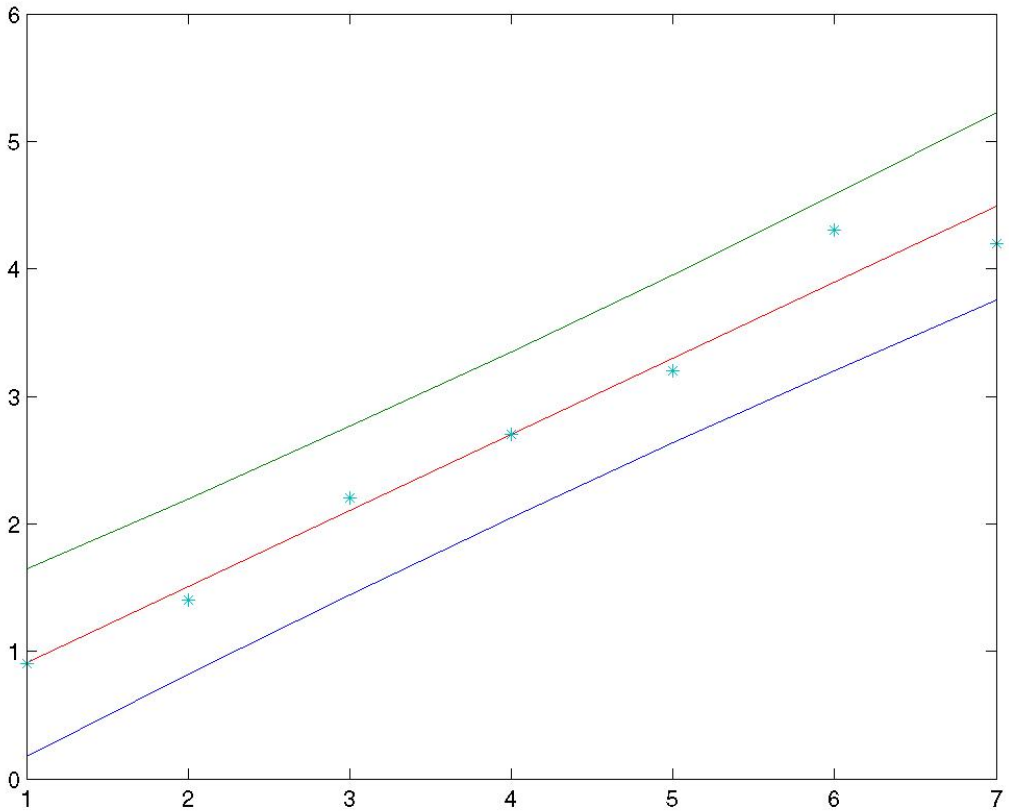
$$\hat{Y}_0 \pm t_{\alpha/2}(n-2)\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}$$

covers Y_0 with the probability $1 - \alpha$. The interval is called a $(1 - \alpha) \cdot 100\%$ prediction interval for Y_0

Prediction interval in the Toy Example: $\alpha = 0.05$

Plot of the training set, predictor and

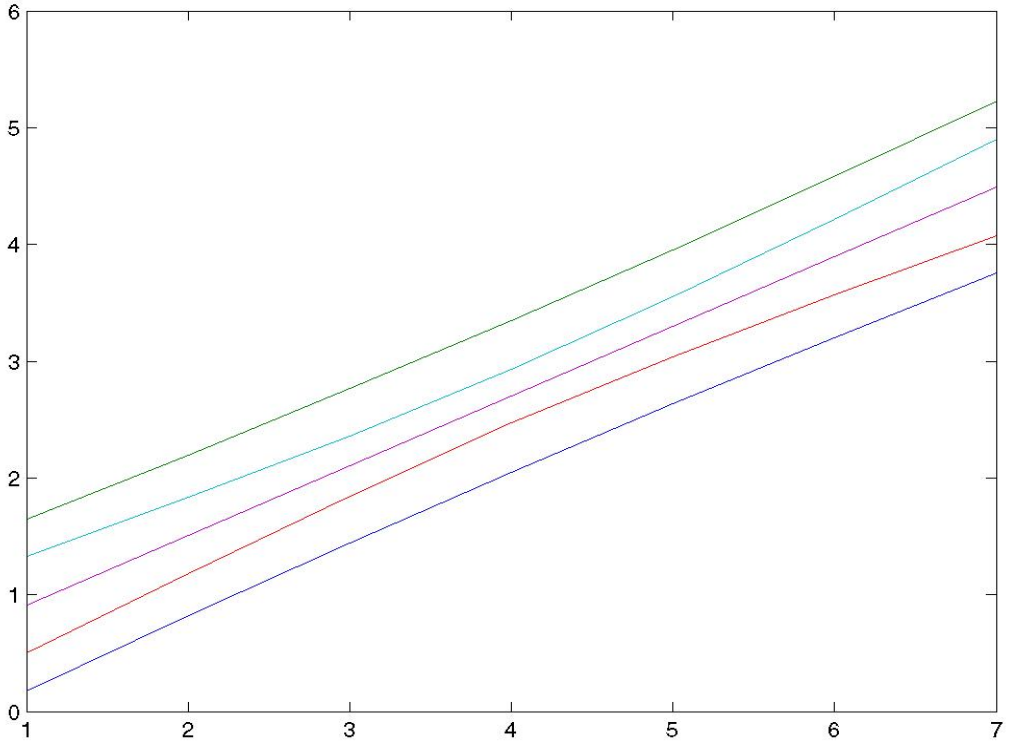
$$\hat{Y}_0 \pm t_{0.025}(n-2)\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 / S_{xx}}$$



Prediction interval and Confidence interval in the Toy Example plotted together

Plot of the predictor, $\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{0.025}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ and

$\hat{Y}_0 \pm t_{0.025}(n-2)\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2/S_{xx}}$



CI FOR β_i IN MULTIPLE NORMAL REGRESSION

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

By the properties of multivariate normal distribution it holds for every componet $\hat{\beta}_j$, $j = 0, \ldots, k+1$ that

$$\hat{\beta}_j \sim N\left(\boldsymbol{\beta}_j^*, \sigma^2 c_{jj}\right)$$

where c_{jj} is the array on the main diagonal of $(X^TX)^{-1}$. Hence we can apply the preceding study to find confidence intervals for a single β_i.

End of treatment of the case $k = 1$

Bootstrap and Multiple Linear Regression

The Annals of Statistics
1981, vol. 9, No. 6, 1218–1228

BOOTSTRAPPING REGRESSION MODELS

By D. A. Freedman[1]

University of California, Berkeley

The regression and correlation models are considered. It is shown that the bootstrap approximation to the distribution of the least squares estimates is valid, and some error bounds are given.

Bootstrap and Multiple Linear Regression

as classical methods. In the regression model, it is appropriate to resample the centered residuals. More specifically, the observable column n-vector $\hat{\varepsilon}(n)$ of residuals is given by $\hat{\varepsilon}(n) = Y(n) - X(n)\hat{\beta}$. However, $\hat{\mu}_n = (1/n) \sum_{i=1}^{n} \hat{\varepsilon}_i(n)$ need not vanish, for the column space of X need not include the constant vectors. Let $\hat{F}_n$ be the empirical distribution of $\hat{\varepsilon}(n)$, centered at the mean, so $\hat{F}_n$ puts mass 1/n at $\hat{\varepsilon}_i(n) - \hat{\mu}_n$ and $\int x \, d\hat{F}_n{}^x = 0$. Given $Y(n)$, let $\varepsilon_1^*, \cdots, \varepsilon_n^*$ be conditionally independent, with common distribution $\hat{F}_n$; let $\varepsilon^*(n)$ be the

sampled with replacement

Bootstrap and Multiple Linear Regression

let $\varepsilon_1^*, \cdots, \varepsilon_n^*$ be conditionally independent, with common distribution $\ddot{F}_n$; let $\varepsilon^*(n)$ be the n-vector whose ith component is ε_i^*; and let

$$Y^*(n) = X(n)\hat{\beta}(n) + \varepsilon^*(n).$$

Informally, ε^* is obtained by resampling the centered residuals. And Y^* is generated from the data, using the regression model with $\hat{\beta}$ as the vector of parameters and $\hat{F}_n$ as the distribution of the disturbance terms ε. Now imagine giving the starred data (X, Y^*) to another statistician, and asking for an estimate of the parameter vector. The least squares estimate is $\hat{\beta}^* = (X^T X)^{-1} X^T Y^*$. The bootstrap principle is that the distribution of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$, which can be computed directly from the data, approximates the distribution of $\sqrt{n}(\hat{\beta} - \beta)$. As will be shown in Section 2 below, this approximation is likely to be very

Bootstrap and Multiple Linear Regression

One repeats the computation of $\widehat{\beta}^*$ B times, finds their empirical distribution $\widehat{F}_{\widehat{\beta}^*}$, and from this one can find confidence intervals for β. Here no assumptions of normal distrbution are required.

Bootstrap and Multiple Linear Regression

Statistics and Probability Letters
134, 2018, pp.
141–149
journal homepage: www.elsevier.com/locate/stapro

Bootstrapping for multivariate linear regression models

Daniel J. Eck

Department of Biostatistics, Yale School of Public Health, 60 College St. LEPH PO Box 208034, New Haven, CT, 06510, USA

Step 2: Fundamental Analysis of Variance Identity (Once More)

Fundamental Analysis of Variance Identity (Once More)

From Lecture 4. the Fundamental Analysis of Variance Identity is written as

$$\underbrace{\mathbf{y}^T C_{ce} \mathbf{y}}_{=SS_{\mathrm{T}}} = \underbrace{\sum_{i=1}^{n} \left(\widehat{y}_i - \bar{\widehat{y}} \right)^2}_{SS_{\mathrm{R}}} + \underbrace{\mathbf{e}_{LSE}^T \mathbf{e}_{LSE}}_{=SS_{\mathrm{Res}}}. \tag{26}$$

where SS_{R} is the regression or model sum of squares, SS_{Res} is the Residual Sum of Squares and where we recall the centering matrix

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

SS_{R} AND SS_{Res} ARE INDEPENDENT

By the definition of the predictor

$$\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}}, \quad \widehat{\mathbf{Y}} = \left(\widehat{Y}_1 \dots \widehat{Y}_n\right)^T, \quad \bar{\widehat{\mathbf{Y}}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{Y}_i$$

and by (26) we have the random variable $SS_{\mathrm{R}} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{\widehat{\mathbf{Y}}}\right)^2$.

PROPOSITION

SS_{R} and SS_{Res} are independent random variables.

Proof: We know by the above that $\widehat{\beta}$ and $\widehat{\varepsilon}$ are independent r.v.s. SS_{R} and SS_{Res} are functions[1] of $\widehat{\beta}$ and $\widehat{\varepsilon}$, respectively, and thus independent.

[1] Borelmeasurable functions

SS_R as a Quadratic Form

$\sum_{i=1}^{n} \left(\widehat{y}_i - \bar{\widehat{y}}\right)^2$ will written as a quadratic form. Let use the centering matrix with

$$C_{ce}\widehat{\mathbf{y}} = \begin{pmatrix} \widehat{y}_1 - \bar{\widehat{y}} \\ \widehat{y}_2 - \bar{\widehat{y}} \\ \vdots \\ \widehat{y}_n - \bar{\widehat{y}} \end{pmatrix}. \tag{27}$$

Hence

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

We have $\left(C_{ce}H\right)^T = H^T C_{ce}^T = HC_{ce}$, since we have at earlier stage checked that H and C_{ce} are symmetric.

SS_R as a Quadratic Form

Hence

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

where we used the fact that C_{ce} is idempotent, as checked earlier. By definition of the centering matrix,

$$C_{ce} H = \left(\mathbb{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T\right) H = H - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T H = H - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T$$

where we used $\mathbf{1}_n^T H = \left(H^T \mathbf{1}_n\right)^T = \left(H \mathbf{1}_n\right)^T = \mathbf{1}_n^T$, see Lecture 3.

SS_R as a Quadratic Form

Hence

$$HC_{ce}H = H\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) = HH - \frac{1}{n}H\mathbf{1}_n\mathbf{1}_n^T = H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$$

since H is idempotent, and $H\mathbf{1}_n = \mathbf{1}_n$, see Lecture 3. In summary, we have found the desired quadratic form for the regression component:

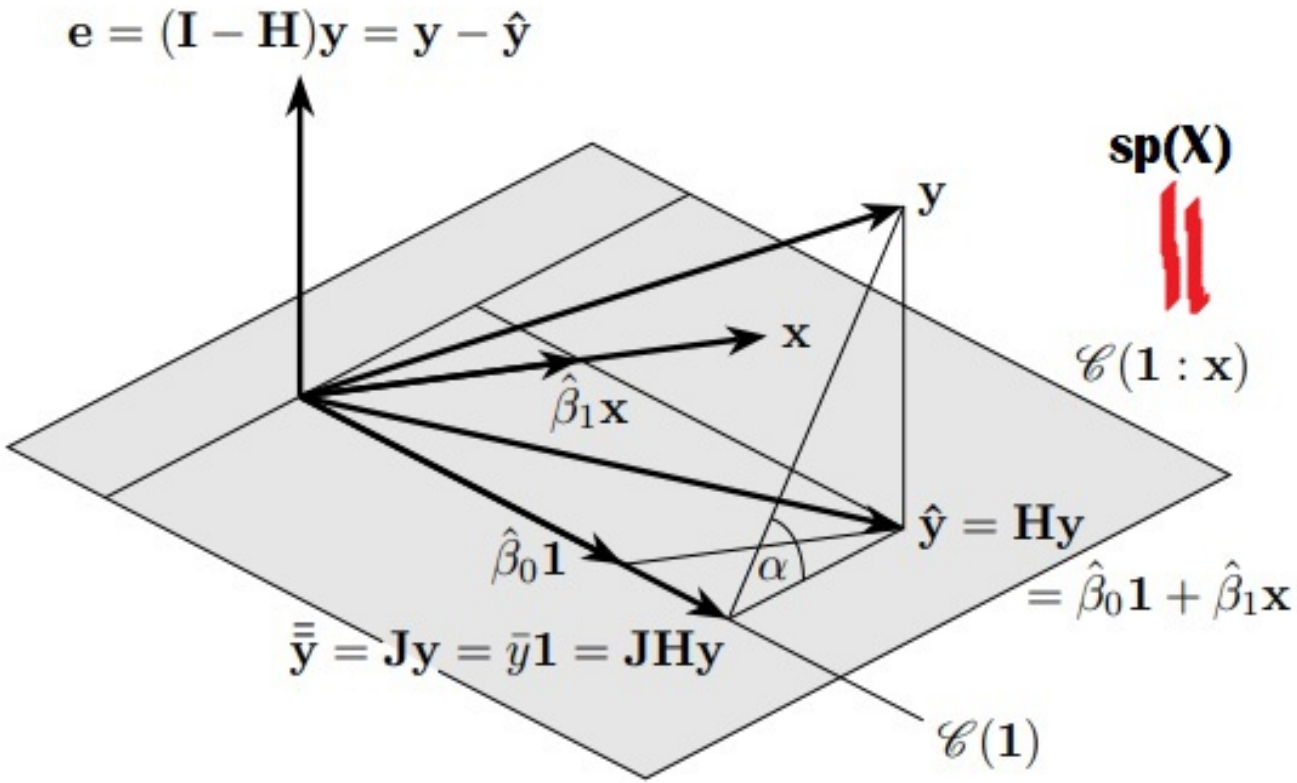
$$SS_R = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 = \mathbf{y}^T\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{y} \tag{28}$$

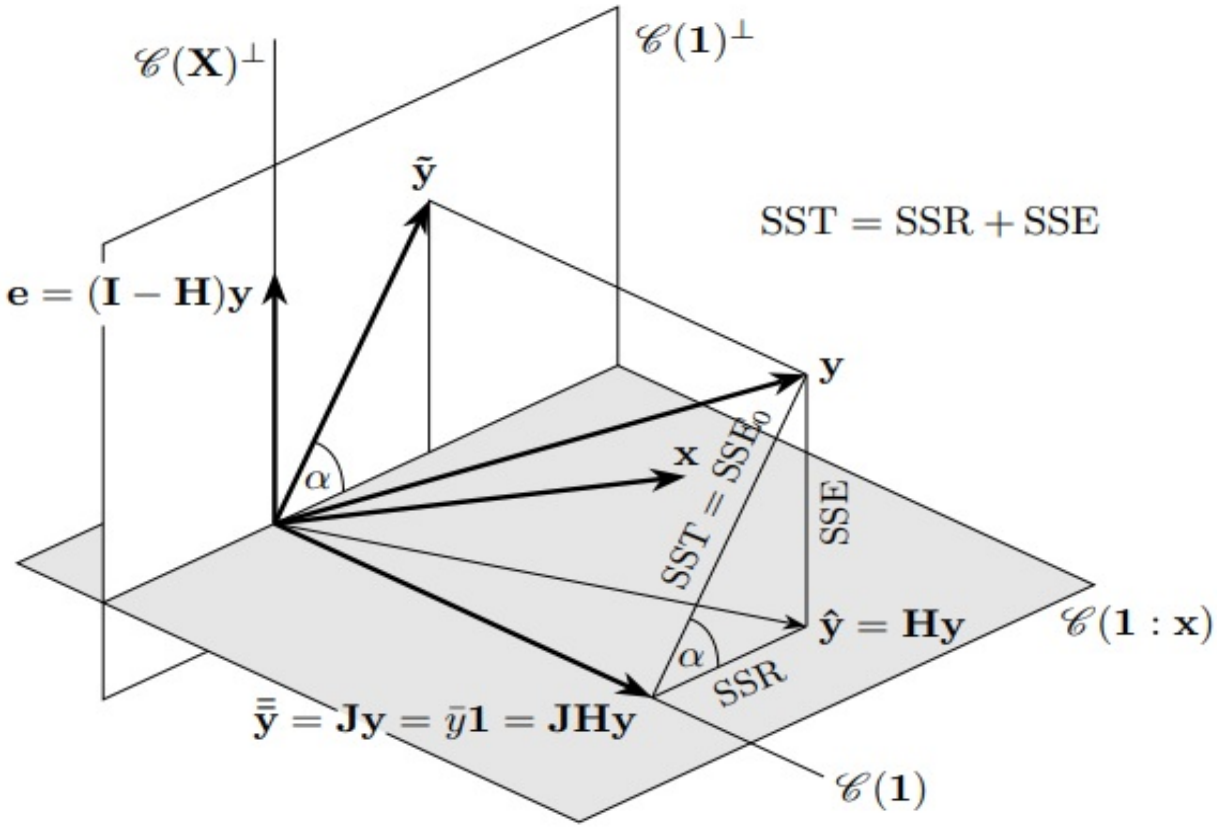
$H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is symmetric and we saw in Lecture 3 that it is idempotent.

The Geometry of the Fundamental Analysis of Variance Identity

The following two Figures[a] *illustrate the geometry of the decomposition. The notation in the Figures* $\mathcal{C}(1:n) = \mathbf{sp}(X)$ *and* $\mathbf{J} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$, $SS_{\mathrm{T}} = SST$ *and* $SS_{\mathrm{R}} = SSR$

[a] Copied from S. Puntanen & K. Vehkalahti: *Matriiseja tilastotieteilijöille*, Report 56/2017, Faculty of Information Sciences, TUNI


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


$\mathscr{C}(\mathbf{X})^\perp$
$\mathscr{C}(\mathbf{1})^\perp$
$\tilde{\mathbf{y}}$
$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$
$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$
$\mathbf{y}$
$\mathbf{x}$
α
$\mathrm{SST} = \mathrm{SSE}_0$
SSE
$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$
$\mathscr{C}(\mathbf{1} : \mathbf{x})$
α
SSR
$\bar{\tilde{\mathbf{y}}} = \mathbf{J}\mathbf{y} = \bar{y}\mathbf{1} = \mathbf{J}\mathbf{H}\mathbf{y}$
$\mathscr{C}(\mathbf{1})$

Non-central chi-square

Definition

$\mathbf{X} \sim N_n(\mu, \mathbb{I}_n)$ *(i.e.* $X_1, \ldots, X_n$ *are independent,* $X_i \sim N(\mu_i, 1)$*). Set*

$$W := \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} X_i^2, \quad \lambda := \sum_{i=1}^{n} \mu_i^2.$$

W has the **non-central chi-square distribution** *with n degrees of freedom and non-centrality parameter* λ*, coded as* $W \sim \chi^2(n, \lambda)$

Note that $\chi^2(n, 0) = \chi^2(n)$

The expression in this slide is NOT Required in the Examination

The pdf of $W \sim \chi^2(n, \lambda)$ is

$$f_X(x; n, \lambda) = \frac{1}{2} e^{-(x+\lambda)/2} \left(\frac{x}{\lambda}\right)^{n/4-1/2} I_{n/2-1}(\sqrt{\lambda x})$$

where $I_\nu(y)$ is a modified Bessel function of the first kind.

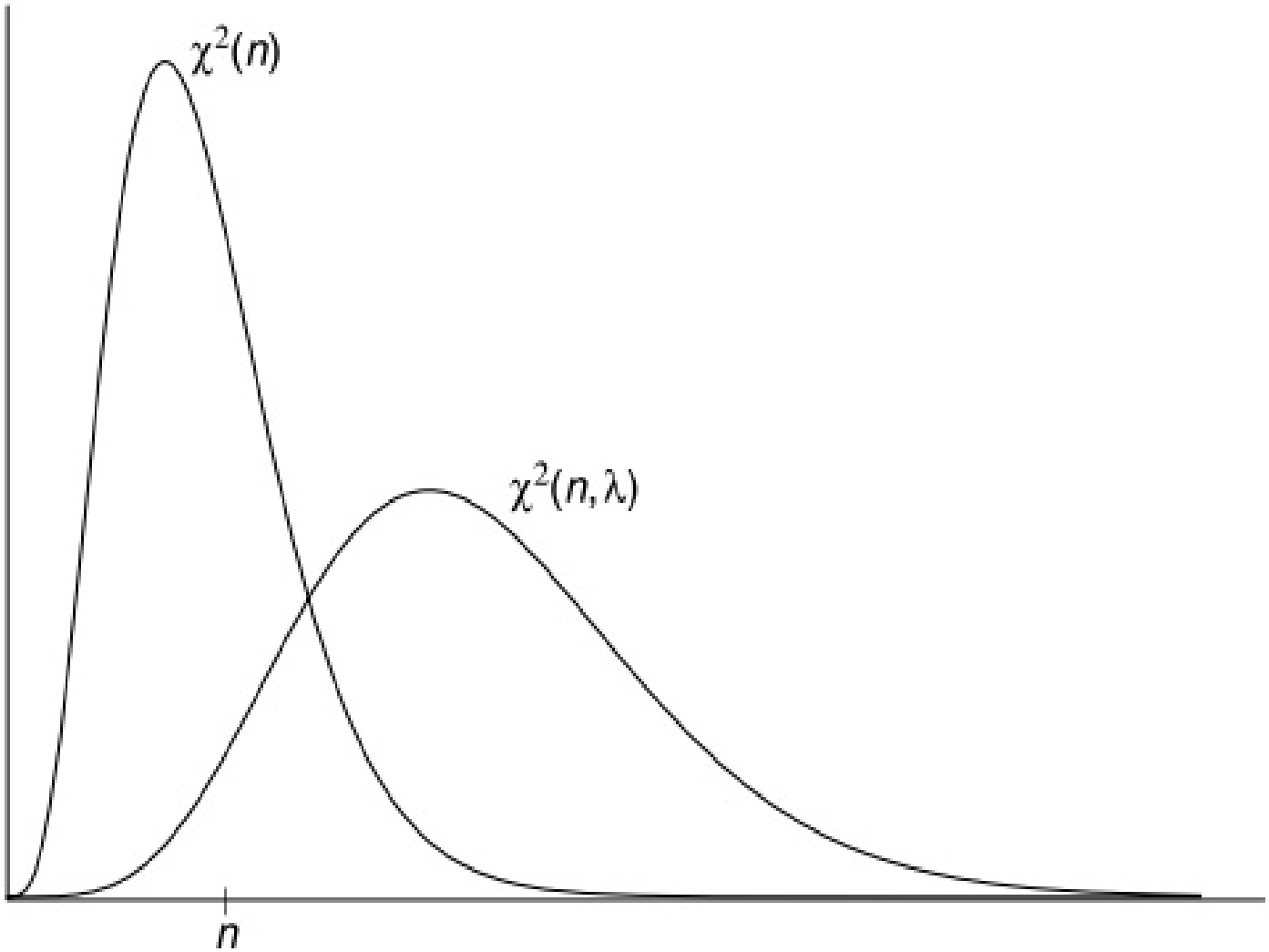


Figure 5.1 Central and noncentral chi-square densities.

Non-central chi-square & Quadratic Forms

Proposition

Let $\mathbf{X} \sim N_n(\mu, \Sigma)$, let A be a symmetric $n \times n$ matrix of constants of rank r, and let $\lambda := \frac{1}{2}\mu^T A\mu$. Then $\mathbf{X}^T A\mathbf{X} \sim \chi^2(r, \lambda)$ if and only if $A\Sigma$ is idempotent.

Theorem 5.5 on p. 117 in Rencher, Alvin C and Schaalje, G Bruce: *Linear Models in Statistics*, 2008. Proof by momentgenerating functions.
We shall now apply this to the quadratic forms SS_{T} and SS_{R}.

Non-central chi-square & Quadratic Forms

$\mathbf{X} \sim N_n(\mu, \Sigma)$, A symmetric $n \times n$ matrix of rank r, $\lambda := \frac{1}{2}\mu^T A \mu$. $\mathbf{X}^T A \mathbf{X} \sim \chi^2(r, \lambda) \Leftrightarrow$ $A\Sigma$ is idempotent.

SS_T: $\mathbf{Y} \sim N_n\left(X\beta_*, \sigma^2 \mathbb{I}_n\right)$, and the quadratic form is $\frac{1}{\sigma^2}\mathbf{Y}^T C_{ce}\mathbf{Y}$. C_{ce} is a symmetric and idempotent matrix. Hence its rank equals (see Slides for Lecture 4) is its trace, and by rules of trace,

$$\operatorname{rank} C_{ce} = \operatorname{Tr} C_{ce} = \operatorname{Tr} \mathbb{I}_n - \frac{1}{n} \operatorname{Tr} \mathbf{1}_n \mathbf{1}_n^T = n - 1$$

(Recall that $\mathbf{1}_n \mathbf{1}_n^T$ is an $n \times n$ matrix of ones). Thus $\operatorname{rank} A = \frac{1}{\sigma^2} C_{ce} = n - 1$. Compute $A\Sigma = \frac{1}{\sigma^2} C_{ce}\sigma^2 \mathbb{I}_n = C_{ce}$. As C_{ce} is idempotent, we have

$$\frac{1}{\sigma^2} SS_T \sim \chi^2(n-1, \lambda), \quad \lambda := \frac{1}{2} (X\beta_*)^T C_{ce} X\beta_*.$$

Non-central chi-square & Quadratic Forms

$\mathbf{X} \sim N_n(\mu, \Sigma)$, A symmetric $n \times n$ matrix of rank r, $\lambda := \frac{1}{2}\mu^T A \mu$. $\mathbf{X}^T A \mathbf{X} \sim \chi^2(r, \lambda) \Leftrightarrow$ $A\Sigma$ is idempotent.

SS_R: $\mathbf{Y} \sim N_n\left(X\beta_*, \sigma^2 \mathbb{I}_n\right)$, and the quadratic form is $\frac{1}{\sigma^2}\mathbf{Y}^T\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{Y}$. $A = \frac{1}{\sigma^2}\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)$. Since $H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is a symmetric and idempotent matrix. Its rank equals (see Slides for Lecture 4) its trace, and by rules of trace,

$$\operatorname{rank}\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) = \left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) = \operatorname{Tr} H - \frac{1}{n}\operatorname{Tr}\mathbf{1}_n\mathbf{1}_n^T = k.$$

$\operatorname{Tr} H = k + 1$ was found in Appendix C of Lecture 3. Thus $\operatorname{rank} A = k$. Next $A\Sigma = \frac{1}{\sigma^2}\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\sigma^2\mathbb{I}_n = \left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)$.

$$\frac{1}{\sigma^2}SS_R \sim \chi^2(k, \lambda), \quad \lambda := \frac{1}{2}(X\beta_*)^T\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)X\beta_*.$$

Degrees of Freedom

$$\frac{1}{\sigma^2} SS_T = \frac{1}{\sigma^2} SS_{\mathrm{R}} + \frac{1}{\sigma^2} SS_{\mathrm{Res}}$$

$$\updownarrow$$

$$\underbrace{\frac{1}{\sigma^2}\mathbf{Y}^T C_{ce}\mathbf{Y}}_{\text{degrees of freedom}=n-1} = \underbrace{\frac{1}{\sigma^2}\mathbf{Y}^T\left(H - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)\mathbf{Y}}_{\text{degrees of freedom}=k} + \underbrace{\frac{1}{\sigma^2}\widehat{\varepsilon}^T\widehat{\varepsilon}}_{\text{degrees of freedom}=n-k-1}$$

See (11) and (12) for the case $\frac{1}{\sigma^2} SS_{\mathrm{Res}}$. We note

$$n - 1 = k + (n - k - 1).$$

If $X \sim \chi^2(r, \lambda)$, $Y \sim \chi^2(s)$ and X and Y are independent, then

$$Q = \frac{X/r}{Y/s} \sim F(r, s, \lambda) \tag{29}$$

Here $F(r, s, \lambda)$ is the **non-central F-distribution** with non-centrality parameter λ. The pdf is is a noncentral F-distributed random variable. The probability density function (pdf) for the noncentral F-distribution is

$$f_Q(q) = \sum_{k=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^k}{B\left(\frac{s}{2}, \frac{r}{2} + k\right) k!} \left(\frac{r}{s}\right)^{\frac{r}{2}+k} \left(\frac{s}{s + rq}\right)^{\frac{r+s}{2}+k} q^{r/2-1+k}$$

The noncentral F-distribution is implemented in R.

Hence, by the preceding

$$\frac{SS_{\mathrm{R}}/k}{SS_{\mathrm{Res}}/(n-k-1)} \sim F(k, n-k-1, \lambda) \tag{30}$$

The ratio

$$F = \frac{SS_{\mathrm{R}}/k}{SS_{\mathrm{Res}}/(n-k-1)}$$

is thus called the F-statistic.

F-test of a Hypothesis on the Regression coefficients

$$H_o: \quad \boldsymbol{\beta} = \mathbf{0}_{k+1}$$

Under this null hypothesis the F-statistic has a central F-distribution:

$$F \sim F(k, n-k-1, 0) = F(k, n-k-1)$$

We choose a significance level α, find $F_\alpha(k, n-k-1)$, the upper a percentage point of the central $F(k, n-k-1)$- distribution, and refuse the null hypothesis, as soon as

$$F = \frac{SS_{\mathrm{R}}/k}{SS_{\mathrm{Res}}/(n-k-1)} \geq F_\alpha(k, n-k-1)$$

F-distribution, critical values $F_\alpha(r, s)$ of $F(r, s)$

E.g., $F_{0.05}(3, 8) = 4.07$, c.f., MVP Table A.4, p. 548.

s/r	1	2	3	4	5	6
1	161	200	216	225	230	234
2	18.5	19	19.2	19.2	19.3	19.3
3	10.1	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.1	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.2	3.09

F-Test

Source	df	Sum of Squares	MSS
Regression	k	SS_{R}	SS_{R}/k
Residual	$n-k-1$	SS_{Res}	$\widehat{\sigma}^2$=SS_{Res}/(n-k-1)
Total	$n-1$	SS_{T}	

Source = source of variation, df= degrees of freedom, SS= sum of squares, MSS= mean sum of squares.

F-test of a Hypothesis on the Regression coefficients

$$H_o: \quad \boldsymbol{\beta} = \mathbf{0}_{k+1}$$

The noncentral F distribution can often be used to evaluate the power of an F- test. The power of a test is the probability of rejecting H_o for a given value of λ.
Let $F_\alpha(p, q)$ be the upper a percentage point of the central $F(p, q)$ distribution. Let $Z \sim F(p, sq, \lambda)$. Then the **power** of the F-test , $P(p, q, \alpha, \lambda)$, is defined as

$$P(p, q, \alpha, \lambda) = P\left(Z \geq F_\alpha(p, q)\right).$$

Power of an F-test of a Hypothesis on the Regression coefficients:

$$P(p, q, \alpha, \lambda) = P\left(Z \geq F_\alpha(q, q)\right).$$

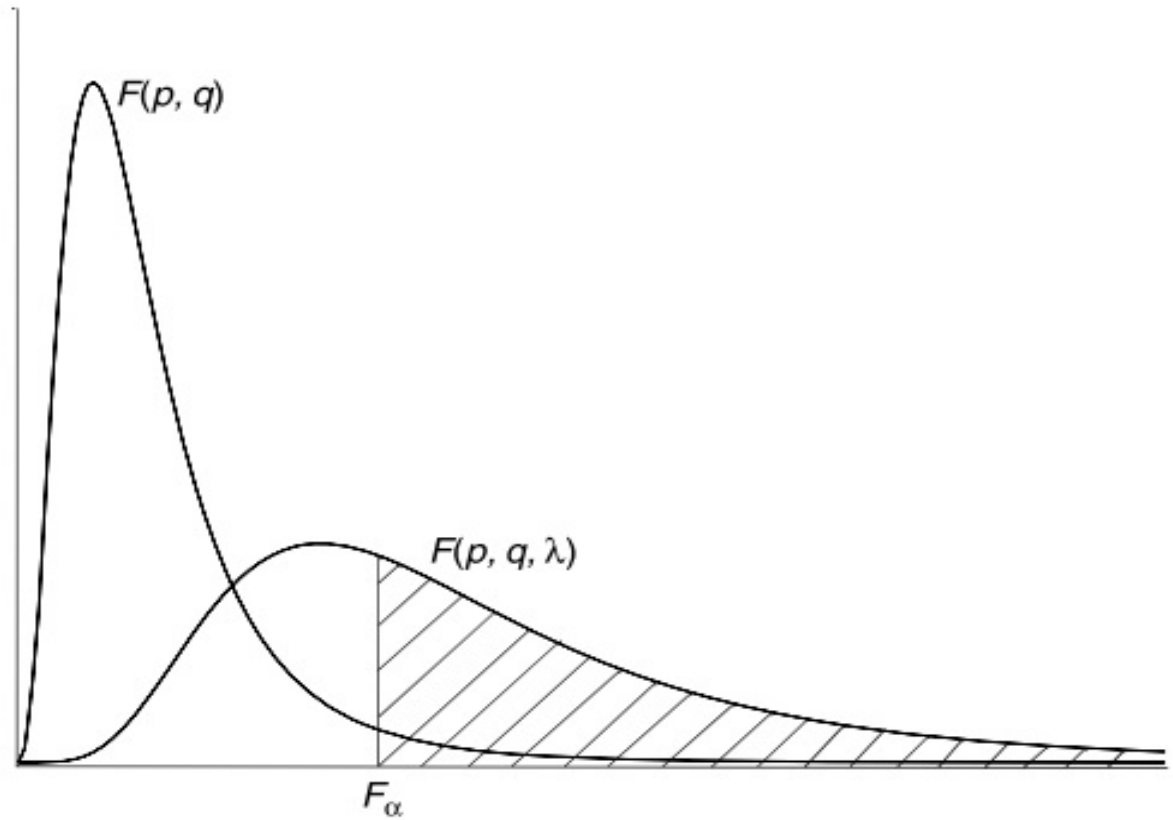


Figure 5.2 Central F, noncentral F, and power of the F test (shaded area).

Appendix A : Computations for Prediction Intervals

$\sum_{i=1}^{n}(d_i + c_i x_0)^2 = \frac{1}{n} + (x_0 - \bar{x})^2/S_{xx}$

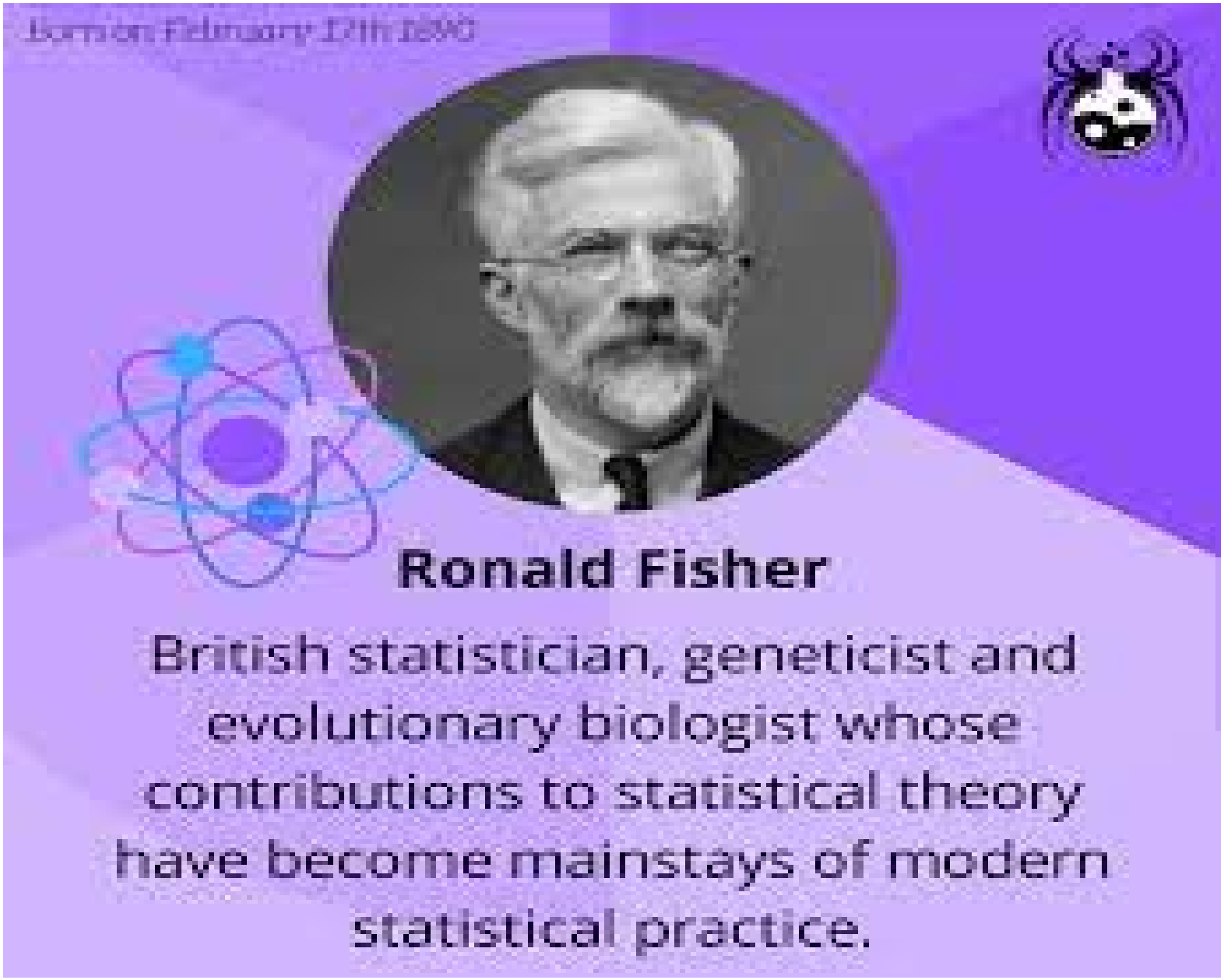
We insert $d_i = \frac{1}{n} - c_i\bar{x}$ to get

$$\sum_{i=1}^{n}(d_i + c_i x_0)^2 = \sum_{i=1}^{n}\left(\frac{1}{n} + c_i(x_0 - \bar{x})\right)^2 =$$

$$= \sum_{i=1}^{n}\frac{1}{n^2} + 2\frac{(x_0 - \bar{x})}{n}\sum_{i=1}^{n}c_i + (x_0 - \bar{x})^2\sum_{i=1}^{n}c_i^2 = \frac{1}{n} + (x_0 - \bar{x})^2/S_{xx}.$$

since the auxiliary (II) in Lecture 1. says $\sum_{i=1}^{n} c_i = 0$ and the auxiliary (IV) in Lecture 1. gives $\sum_{i=1}^{n} c_i^2 = 1/S_{xx}$.

Appendix B: The person behind F


Born on February 17th 1890
Ronald Fisher
British statistician, geneticist and evolutionary biologist whose contributions to statistical theory have become mainstays of modern statistical practice.