REGRESSION/CORRELATION II (DA_2022)

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

Lecture n. 24, Rome mon 30th of May 2021

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

Outline L 24

MATHEMATICAL IDEAS REVIEWED LAST TIME

- Linear spaces (in n dimensions)
- Vectors (geometry/algebra)
- n-ples (coordinates)
- linear combinations of vector
- bases
- Linear transformations
- change of bases
- matrices
- .2

- Linear regression analysis independent/dependent variables)
- Least squares method
- W&S 17.1,17.2
- Rosner 11.1, 11.3, 11,11.4

CASE STUDY FOR LINEAR REGRESSION
(Greene & Touchstone 1963)
Estriol levels of the mother close to delivery and birthweight of the baby

Sample data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	17	17	32
2	9	25	18	25	32
3	9	25	19	27	34
4	12	27	20	15	34
5	14	27	21	15	34
6	16	27	22	15	35
7	16	24	23	16	35
8	14	30	24	19	34
9	16	30	25	18	35
10	16	31	26	17	36
11	17	30	27	18	37
12	19	31	28	20	38
13	21	30	29	22	40
14	24	28	30	25	39
15	15	32	31	24	43
16	16	32			

Source: Based on the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.

WHAT DOES LINEAR MEAN?

E.G. A LINEAR MODEL

If x = estriol level and y = birthweight, then we can postulate a linear relationship between y and x that is of the following form:

$$E(y|x) = \alpha + \beta x$$

If x = estriol level and y = birthweight, then we can postulate a linear relationship between y and x that is of the following form:

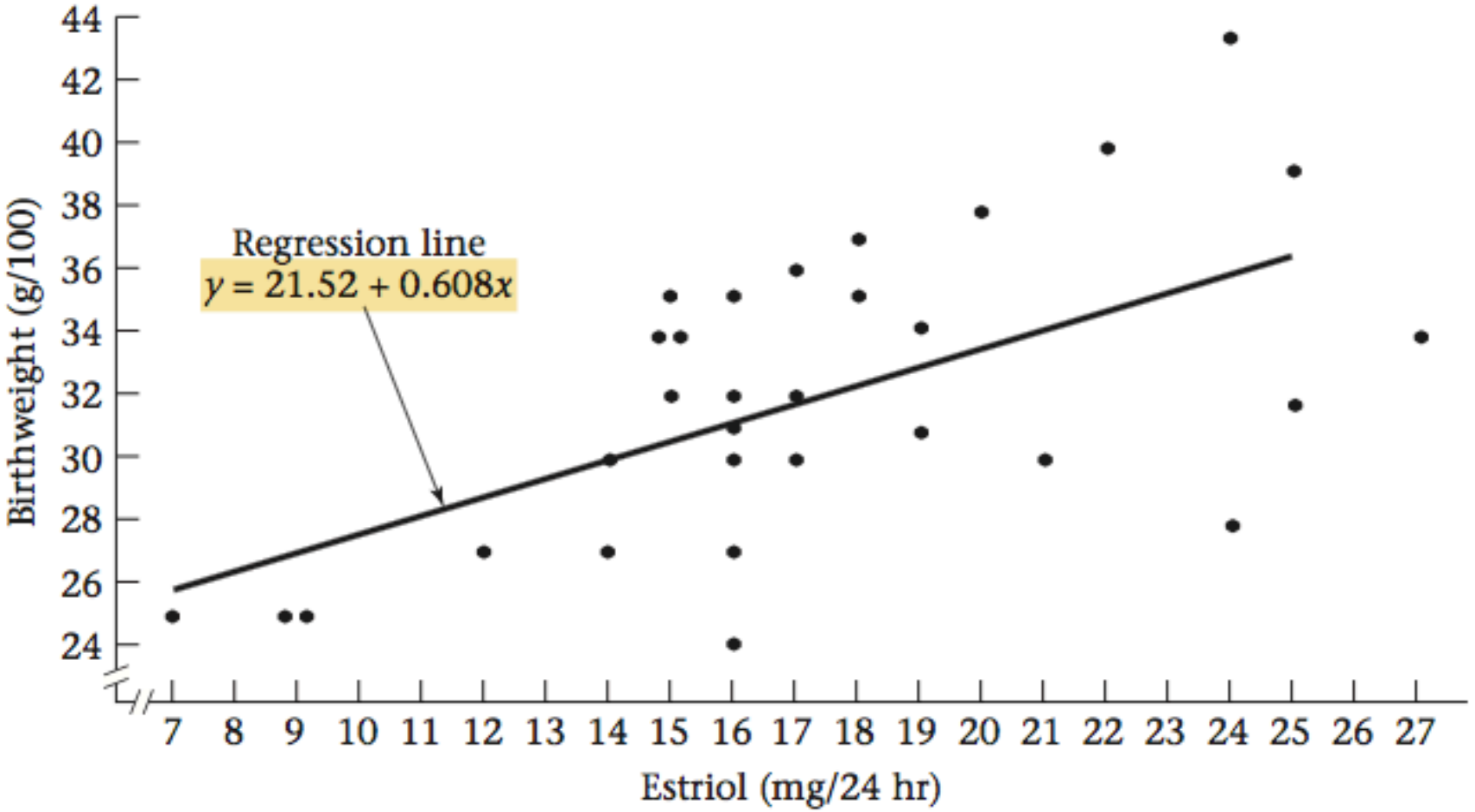
$$E(y|x) = \alpha + \beta x$$

where $E(y|x)$ = expected or average birthweight (y) among women with a given estriol level (x).

That is, for a given estriol-level x, the average birthweight $E(y|x) = \alpha + \beta x$.

The relationship $y = \alpha + \beta x$ is not expected to hold exactly for every woman. For example, not all women with a given estriol level have babies with identical birthweights. Thus, an error term e, which represents the variance of birthweight among all babies of women with a given estriol level x, is introduced into the model.

Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



Source: Based on the *American Journal of Obstetrics and Gynecology*, *85*(1), 1–9, 1963.

COMMENT:…DESCRIPTIVE / INFERENTIAL USE OF REGRESSION

Let's assume e follows a normal distribution, with mean 0 and variance σ^2. The full linear-regression model then takes the following form:

$$y = \alpha + \beta x + e$$

where e is normally distributed with mean 0 and variance σ^2.

For any linear-regression equation of the form $y = \alpha + \beta x + e$, y is called the **dependent variable** and x is called the **independent variable** because we are trying to predict y as a function of x.

Obstetrics Birthweight is the dependent variable and estriol is the independent variable for the problem posed in Example 11.3 because estriol levels are being used to try to predict birthweight.

One interpretation of the regression line is that for a woman with estriol level x, the corresponding birthweight will be normally distributed with mean $\alpha + \beta x$ and variance σ^2. If σ^2 were 0, then every point would fall exactly on the regression line, whereas the larger σ^2 is, the more scatter occurs about the regression line. This relationship is illustrated in Figure 11.2.

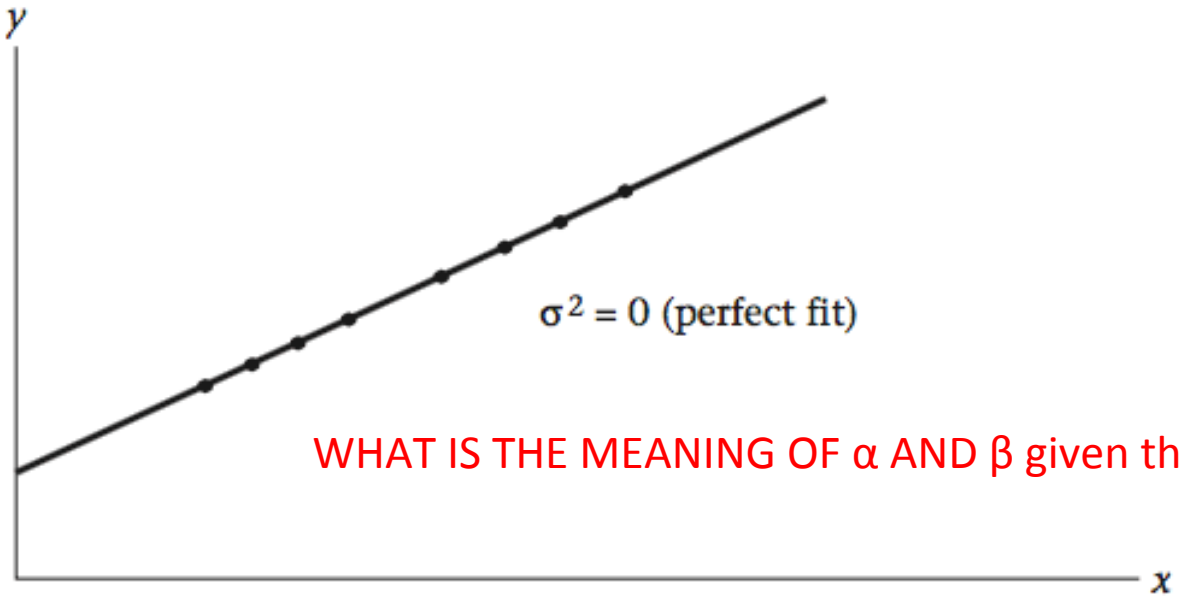
Assumptions Made in Linear-Regression Models

(1) For any given value of x, the corresponding value of y has an average value $\alpha + \beta x$, which is a linear function of x.

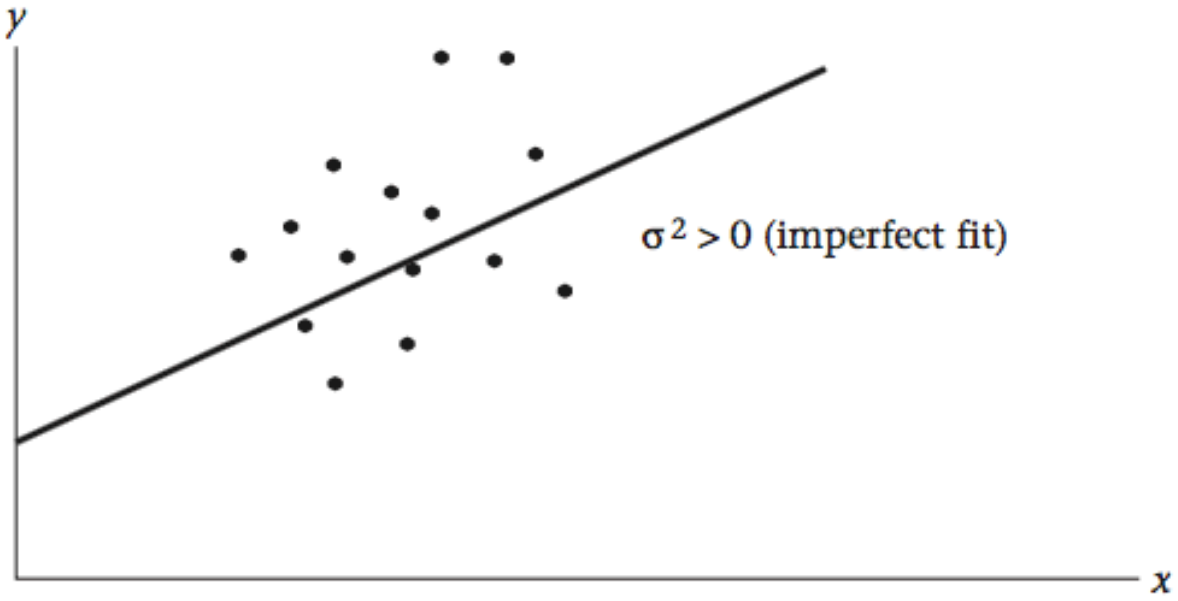
(2) For any given value of x, the corresponding value of y is normally distributed about $\alpha + \beta x$ with the same variance σ^2 for any x.

(3) For any two data points (x_1, y_1), (x_2, y_2), the error terms e_1, e_2 are independent of each other.

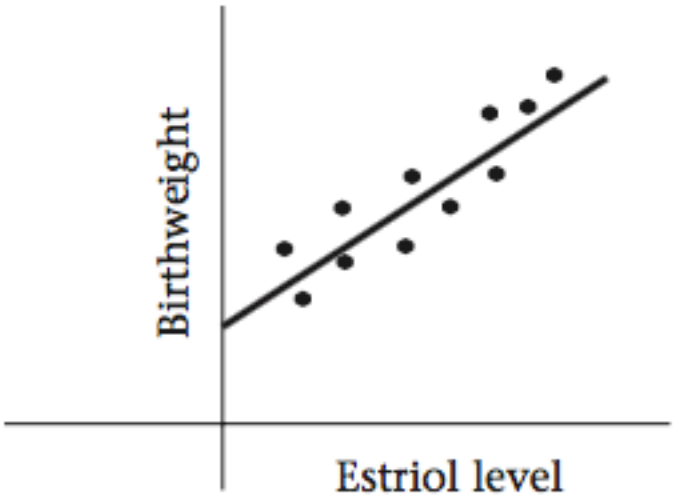
The effect of σ^2 on the goodness of fit of a regression line



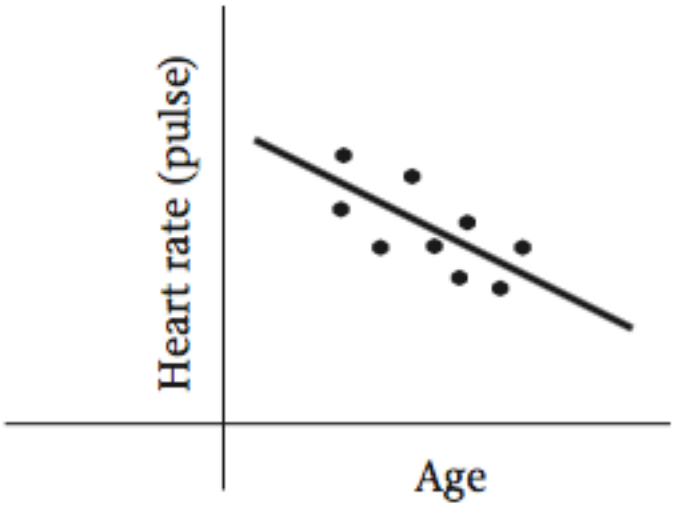
WHAT IS THE MEANING OF α AND β given the noise?



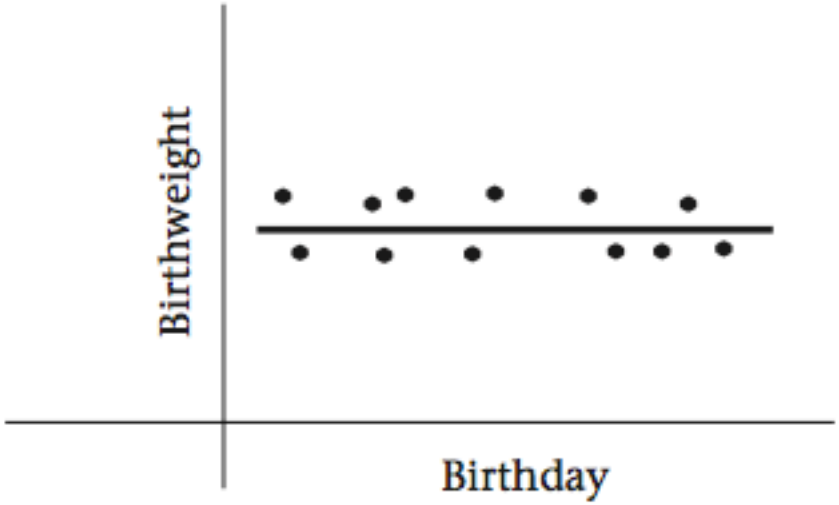
Interpretation of the regression line for different values of β



(a) $\beta > 0$



(b) $\beta < 0$



(c) $\beta = 0$

FITTING REGRESSION LINES TO DATA: THE METHOD OF LEAST SQUARE
(Inferential optimization method)

Find a, b such that the sum of deviations of the data from the regression line is minimum

S = sum of the squared distances of the points from the line

$$= \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Estimation of the Least-Squares Line

The coefficients of the least-squares line $y = a + bx$ are given by

$$b = L_{xy}/L_{xx} \quad \text{and} \quad a = \bar{y} - b\bar{x} = \left(\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \right) \Big/ n$$

Sometimes, the line $y = a + bx$ is called the *estimated or fitted regression line* or, more briefly, the *regression line.*

The **raw sum of squares for x** is defined by

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

The **corrected sum of squares for x** is denoted by L_{xx} and defined by

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

It represents the sum of squares of the deviations of the x_i from the mean. Similarly, the **raw sum of squares for y** is defined by

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

The **corrected sum of squares for y** is denoted by L_{yy} and defined by

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n$$

Notice that L_{xx} and L_{yy} are simply the numerators of the expressions for the sample variances of x (i.e., s_x^2) and y (i.e., s_y^2), respectively, because

$$s_x^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 \Big/ (n-1) \text{ and } s_y^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 \Big/ (n-1)$$

The **raw sum of cross products** is defined by

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

The **corrected sum of cross products** is defined by

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

which is denoted by L_{xy}.

It can be shown that a short form for the corrected sum of cross products is given by

$$\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n$$

NOTE: THIS L_{xy} TERM HAS TO DO WITH CORRELATION
(degree of equivariance of y and x)

The **predicted**, or **average, value of** y for a given value of x, as estimated from the fitted regression line, is denoted by $\hat{y} = a + bx$. Thus the point $(x, a + bx)$ is always on the regression line.

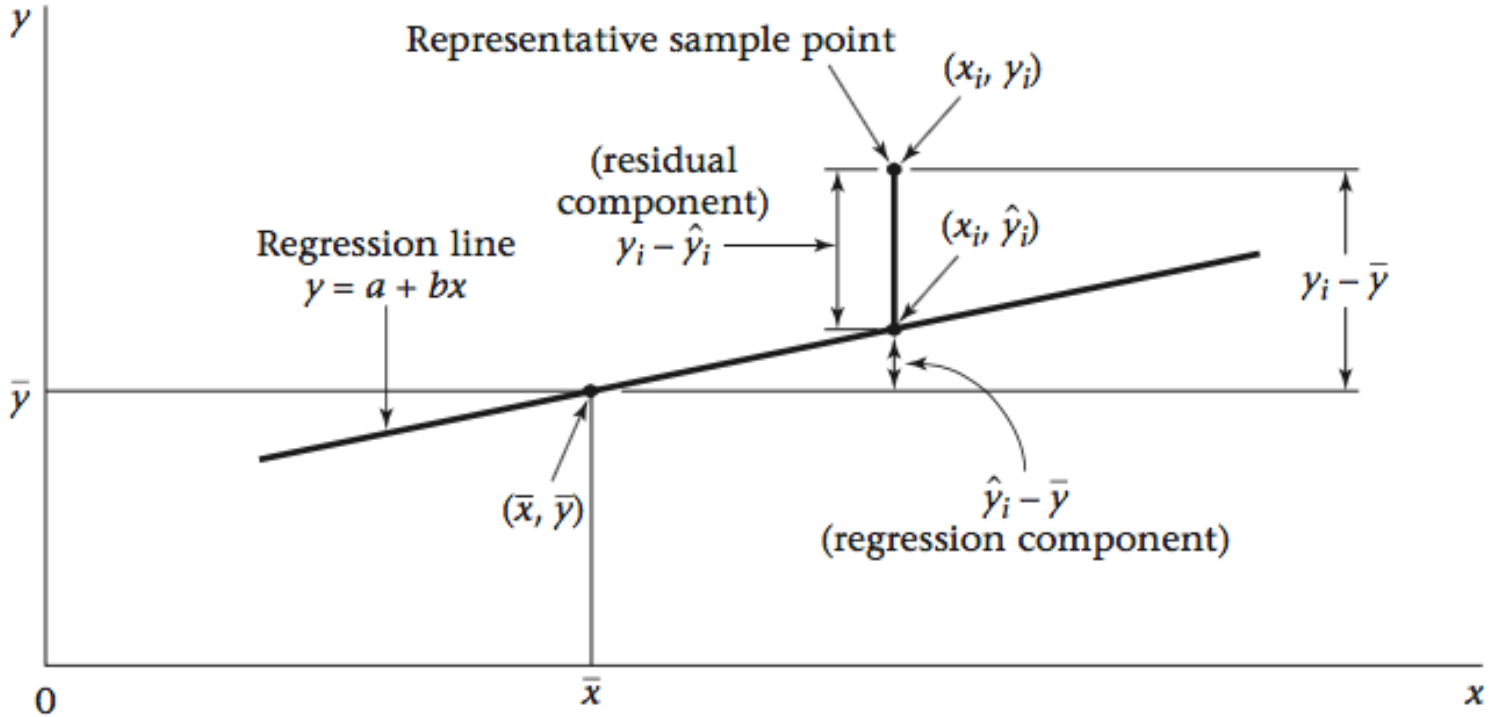
Obstetrics What is the estimated average birthweight if a pregnant woman has an estriol level of 15 mg/24 hr?

Solution: If the estriol level were 15 mg/24 hr, then the best prediction of the average value of y would be

$$\hat{y} = 21.52 + 0.608(15) = 30.65$$

Because y is in the units of birthweight (g)/100, the estimated average birthweight = $30.65 \times 100 = 3065$ g.

Goodness of fit of a regression line



A hypothetical regression line and a representative sample point have been drawn. First, notice that the point $(\bar{x}, \bar{y})$ falls on the regression line. This feature is common to all estimated regression lines because a regression line can be represented as

$$y = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x})$$

or, equivalently,

$$y - \bar{y} = b(x - \bar{x})$$

Decomposition of the Total Sum of Squares into Regression and Residual Components

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

or Total SS = Reg SS + Res SS

F Test for Simple Linear Regression

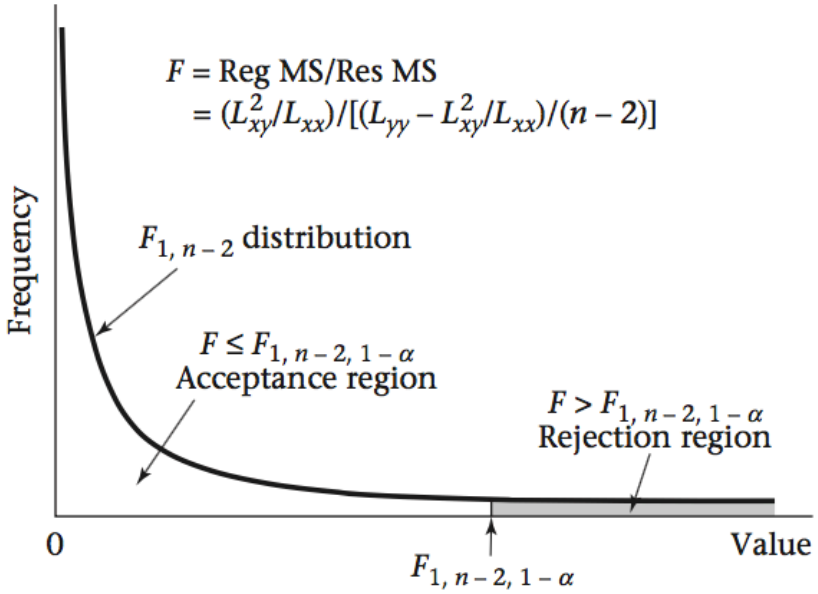
The criterion for goodness of fit used in this book is the ratio of the regression sum of squares to the residual sum of squares. A large ratio indicates a good fit, whereas a small ratio indicates a poor fit. In hypothesis-testing terms we want to test the hypothesis H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$, where β is the underlying slope of the regression line in Equation 11.2.

The following terms are introduced for ease of notation in describing the hypothesis test.

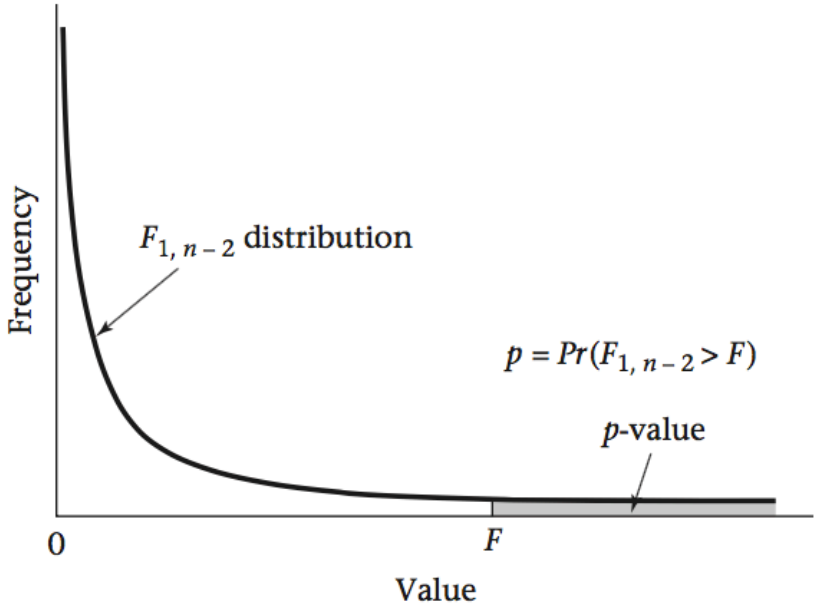
The **regression mean square**, or **Reg MS**, is the Reg SS divided by the number of predictor variables (k) in the model (not including the constant). Thus, Reg MS = Reg SS/k. For simple linear regression, which we have been discussing, $k = 1$ and thus Reg MS = Reg SS. For multiple regression in Section 11.9, k is >1. We will refer to k as the degrees of freedom for the regression sum of squares, or Reg *df*.

The **residual mean square**, or **Res MS**, is the ratio of the Res SS divided by $(n - k - 1)$, or Res MS = Res SS/$(n - k - 1)$. For simple linear regression, $k = 1$ and Res MS = Res SS/$(n - 2)$. We refer to $n - k - 1$ as the degrees of freedom for the residual sum of squares, or Res *df*. Res MS is also sometimes denoted by $s^2_{y \cdot x}$ in the literature. The Res MS is an estimate of $s^2_{y \cdot x}$ the variation of y for a given value of x.

Acceptance and rejection regions for the simple linear-regression *F* test



Computation of the *p*-value for the simple linear-regression *F* test



A summary measure of goodness of fit frequently referred to in the literature is R^2.

R^2 is defined as Reg SS/Total SS.

R^2 can be thought of as the proportion of the variance of y that is explained by x. If $R^2 = 1$, then all variation in y can be explained by variation in x, and all data points fall on the regression line. In other words, once x is known y can be predicted exactly, with no error or variability in the prediction. If $R^2 = 0$, then x gives no information about y, and the variance of y is the same with or without knowing x. If R^2 is between 0 and 1, then for a given value of x, the variance of y is lower than it would be if x were unknown but is still greater than 0. In particular, the best estimate of the variance of y given x (or σ^2 in the regression model in Equation 11.2) is given by Res MS (or $s^2_{y\cdot x}$). For large n, $s^2_{y\cdot x} \approx s^2_y(1-R^2)$. Thus, R^2 represents the proportion of the variance of y that is explained by x.

Obstetrics Compute and interpret R^2 and $s^2_{y.x}$ for the birthweight–estriol data in Example 11.12.

Solution: From Table 11.3, the R^2 for the birthweight–estriol regression line is given by 250.57/674 = .372. Thus, about 37% of the variance of birthweight can be explained by estriol level. Furthermore, $s^2_{y.x}$= 14.60, as compared with

$$s_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1) = 674 / 30 = 22.47$$

Thus, for the subgroup of women with a specific estriol level, such as 10 mg/24 hr, the variance of birthweight is 14.60, whereas for *all* women with any estriol level, the variance of birthweight is 22.47. Note that

$$s^2_{y\cdot x} / s^2_y = 14.60 / 22.47 = .650 \approx 1 - R^2 = 1 - .372 = .628$$

THAT’S ALL FOLKS FOR THIS YEAR!