Example 1 (socio-economic study). The demand for a consumer product is affected by many factors. In one study, measurements on the relative urbanization (X_1), educational level (X_2), and relative income (X_3) of 9 randomly chosen geographic regions were obtained in an attempt to determine their effect on the product usage (Y). The data were:

X_1	X_2	X_3	Y
42.2	11.2	31.9	167.1
48.6	10.6	13.2	174.4
42.6	10.6	28.7	160.8
39.0	10.4	26.1	162.0
34.7	9.3	30.1	140.8
44.5	10.8	8.5	174.6
39.1	10.7	24.3	163.7
40.1	10.0	18.6	174.5
45.9	12.0	20.4	185.7

We fit the model: $Y = X\beta + \epsilon$, with $E(\epsilon) = 0$ and $Cov(\epsilon) = \sigma^2 I_n$. In this case, $n = 9$, $p = 4$. $\bar{y} = 167.07$ and the model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$.

We get $\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 60.0 \\ 0.24 \\ 10.72 \\ -0.75 \end{pmatrix}$. The detailed ANOVA (with mean) is

source	d.f.	SS	MS	F-ratio
mean	1	SSM $= n\bar{y}^2 =$ 251201.44	MSM = 251201.44	
regression (X_1, X_2, X_3)	3	$SS_{reg} =$ 1081.35	$MS_{reg} =$ 360.45	$F_{reg} = \frac{360.45}{39.57} = 9.11$
residual error	5	SSE = RSS = 197.85	MSE = 39.57	
Total (corrected)	8	1279.20		
Total	9	252480.64		

From this note that $s^2 = \text{RSS}/(n-r) = \text{MSE} = 39.57$, so $s = 6.29 = \hat{\sigma}$, and $R^2 = 1081.35/1279.20 = 84.5\%$. Abridged ANOVA is

source	d.f.	SS	MS	F-ratio
regression (X_1, X_2, X_3)	3	$\text{SS}_{reg} =$ 1081.35	$\text{MS}_{reg} =$ 360.45	$F_{reg} = \frac{360.45}{39.57} = 9.11$
residual error	5	SSE = RSS = 197.85	MSE = 39.57	
Total (corrected)	8	1279.20		

$R^2 = 84.5\%$ is substantial. What about $F = 9.11$? $F_{3,5}(.95) = 5.41$ and $F_{3,5}(.99) = 12.06$, so there is some evidence against the null and justifying the linear fit.

Example 2. X = height (cm) and Y = weight (kg) for a sample of $n = 10$ eighteen-year-old American girls:

X	Y
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

Upon fitting the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we get $\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} -36.9 \\ 0.582 \end{pmatrix}$, $s^2 = MSE = 71.50$, $s = 8.456$, $R^2 = 21.9\%$, $\bar{y} = 59.47$. ANOVA is

source	d.f.	SS	MS	F	R^2
X	1	159.95	159.95	2.24	21.9%
error	8	512.01	71.50		
Total (C)	9	731.96			

Note the following. (i) X is expected to be a useful predictor of Y, but the relationship may not be simple. (ii) $F_{1,8}(.90) = 3.46 = (1.86)^2 = t_8^2(.95)$, so is there a connection between the ANOVA F-test and a t-test?

Consider simple linear regression again: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, ϵ_i i.i.d. $N(0, \sigma^2)$. Then the F-ratio is the F statistic for testing the goodness of fit of the linear model, or for testing $H_0 : \beta_1 = 0$. Writing the linear model in the standard form, we have

$$X_{n\times 2} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \text{ and}$$

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

Therefore

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X'X)^{-1}X'Y = \frac{1}{n\sum_{i=1}^n (x_i-\bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Letting $S_{XX} = \sum_{i=1}^n (x_i-\bar{x})^2$, $S_{XY} = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})$, and extracting the least squares equations, we get,

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{S_{XX}}\left\{-n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i\right\} = \frac{S_{XY}}{S_{XX}}, \\ \hat{\beta}_0 &= \frac{1}{S_{XX}}\left\{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i\right\} = \frac{1}{S_{XX}}\left\{\bar{y}S_{XX} + n\bar{y}\bar{x}^2 - \bar{x}\sum_{i=1}^n x_i y_i\right\} \\ &= \frac{1}{S_{XX}}\left\{\bar{y}S_{XX} - \bar{x}\left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}\right)\right\} = \bar{y} - \bar{x}\hat{\beta}_1. \end{aligned}$$

Now, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{XX})$, so that, to test $H_0 : \beta_1 = 0$, use the test statistic,

$$\frac{\sqrt{S_{XX}}\hat{\beta}_1}{\sqrt{\text{RSS}/(n-2)}} \sim t_{n-2}, \qquad \text{or} \qquad \frac{\hat{\beta}_1^2 S_{XX}}{\text{MSE}} \sim F_{1,n-2},$$

if H_0 is true. The ANOVA table shows that

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= n\bar{y}^2 + \sum_{i=1}^n (y_i-\bar{y})^2 = n\bar{y}^2 + \text{RSS} + \text{SS}_{reg}, \text{ so} \\ \text{SS}_{reg} &= \sum_{i=1}^n (y_i-\bar{y})^2 - \text{RSS}. \end{aligned}$$

However,

$$
\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n}\Big(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})\Big)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 - 2\hat{\beta}_1\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2.
\end{aligned}
$$

Therefore, $\text{SS}_{reg} = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$, so that

$$
t^2 = \frac{\hat{\beta}_1^2 S_{XX}}{\text{RSS}/(n-2)} = \text{ F-ratio of ANOVA.}
$$

In Example 1, F-ratio tests $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. What if we want to test only $\beta_1 = \beta_3 = 0$? Then we have $H_0 : A\beta = 0$, where $A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}_{2\times 4}$ if of rank 2. Then apply the theorem: $\text{RSS}_{H_0} = (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0})$ where $\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-}A'(A(X'X)^{-}A')^{-1}(c - A\hat{\beta})$ and the test statistic is

$$
F = \frac{(\text{RSS}_{H_0} - \text{RSS})/q}{\text{RSS}/(n-r)} \sim F_{q,n-r} \text{ under } H_0.
$$