Let's explain **why** one-way ANOVA and regression give **exactly the same F-statistic** — with a small, intuitive numeric example.

---

## 🧩 1️⃣ The core reason

Both **ANOVA** and **linear regression** are doing the *same decomposition of total variation*:

$$\text{Total Variation (SST)} = \text{Explained Variation (SSR)} + \text{Unexplained Variation (SSE)}$$

- **ANOVA** calls them:
  - *Between-group sum of squares* (SSB)
  - *Within-group sum of squares* (SSW)
- **Regression** calls them:
  - *Regression sum of squares* (SSR)
  - *Residual sum of squares* (SSE)

They're **numerically identical**, just using different names for the same quantities.

## 🧩 2️⃣ Let's use the same simple data

| Group | Y values |
|---|---|
| A | 10, 12, 8 |
| B | 20, 22, 18 |
| C | 15, 13, 17 |

---

## 🧩 3️⃣ Step 1: Compute means

| Group | Mean |
|---|---|
| A | 10 |
| B | 20 |
| C | 15 |

Overall mean:
$\bar{Y} = (10 + 12 + 8 + 20 + 22 + 18 + 15 + 13 + 17)/9 = 15$

---

### 🧩 4️⃣ Step 2: ANOVA view — partitioning total variation

#### (a) Total Sum of Squares (SST)

$$SST = \sum(Y_i - \bar{Y})^2$$

Compute it:

Compute it:

| Y | $Y_i - 15$ | $(Y_i - 15)^2$ |
|---|---|---|
| 10 | -5 | 25 |
| 12 | -3 | 9 |
| 8 | -7 | 49 |
| 20 | 5 | 25 |
| 22 | 7 | 49 |
| 18 | 3 | 9 |
| 15 | 0 | 0 |
| 13 | -2 | 4 |
| 17 | 2 | 4 |

$$SST = 25 + 9 + 49 + 25 + 49 + 9 + 0 + 4 + 4 = 174$$

**(b) Between-group Sum of Squares (SSB)**

$$SSB = \sum n_j(\bar{Y}_j - \bar{Y})^2$$

For each group (each has $n_j = 3$):

| Group | $\bar{Y}_j - 15$ | $(\bar{Y}_j - 15)^2$ | $3(\bar{Y}_j - 15)^2$ |
|---|---|---|---|
| A | -5 | 25 | 75 |
| B | 5 | 25 | 75 |
| C | 0 | 0 | 0 |

$$SSB = 75 + 75 + 0 = 150$$

### (c) Within-group Sum of Squares (SSW)

$$SSW = SST - SSB = 174 - 150 = 24$$

So:

- $SSB = 150$
- $SSW = 24$

---

### (d) F-statistic (ANOVA)

$$F = \frac{MSB}{MSW} = \frac{(SSB/(k-1))}{(SSW/(N-k))}$$

Here $k = 3$, $N = 9$:

$$F = \frac{(150/2)}{(24/6)} = \frac{75}{4} = 18.75$$

✅ ANOVA F-statistic = **18.75**

## 🧩 5️⃣ Step 3: Regression view

Regression with dummy variables:

| Group | $x_B$ | $x_C$ | Y |
|---|---|---|---|
| A | 0 | 0 | 10 |
| A | 0 | 0 | 12 |
| A | 0 | 0 | 8 |
| B | 1 | 0 | 20 |
| B | 1 | 0 | 22 |
| B | 1 | 0 | 18 |
| C | 0 | 1 | 15 |
| C | 0 | 1 | 13 |
| C | 0 | 1 | 17 |

Model:
$Y_i = \beta_0 + \beta_B x_{B,i} + \beta_C x_{C,i} + \varepsilon_i$

We already found:

- $\beta_0 = 10$
- $\beta_B = 10$
- $\beta_C = 5$

---

**Predicted values ($\hat{Y}_i$) from regression:**

| Group | Predicted $\hat{Y}$ | Residual = $Y - \hat{Y}$ |
|---|---|---|
| A | 10 | 0, 2, -2 |
| B | 20 | 0, 2, -2 |
| C | 15 | 0, -2, 2 |

For each group $j$:
$SSR_j = n_j \times (\hat{Y}_j - \bar{Y})^2$

| Group | $n_j$ | $\hat{Y}_j$ | $(\hat{Y}_j - 15)$ | $(\hat{Y}_j - 15)^2$ | $n_j(\hat{Y}_j - 15)^2$ |
|---|---|---|---|---|---|
| A | 3 | 10 | -5 | 25 | 75 |
| B | 3 | 20 | +5 | 25 | 75 |
| C | 3 | 15 | 0 | 0 | 0 |

Compute sums of squares:

- **Regression SS (SSR)** = variation in fitted means from overall mean = **150**
- **Residual SS (SSE)** = variation within groups = **24**

Exactly same as ANOVA's SSB and SSW!

---

### F-statistic (regression)

$$F = \frac{(SSR/2)}{(SSE/6)} = \frac{(150/2)}{(24/6)} = \frac{75}{4} = 18.75$$

✅ Regression F-statistic = **18.75**, same as ANOVA.

## 🧭 6️⃣ Why they match conceptually

Because **ANOVA** and **regression** are two languages describing the same model:

| ANOVA term | Regression term | Meaning |
|---|---|---|
| Between-group SS (SSB) | Regression SS (SSR) | Variation explained by group means |
| Within-group SS (SSW) | Residual SS (SSE) | Unexplained (random) variation |
| Total SS (SST) | Total SS (SST) | Overall variation in data |

Both partition variance in the exact same way and compute

$$F = \frac{\text{explained variation per df}}{\text{unexplained variation per df}}$$

---

| Concept | One-way ANOVA | Multiple Linear Regression | Meaning |
|---|---|---|---|
| **Explained variation** | $SS_{Between} = \sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2$ | $SS_{Reg} = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$ | Variation explained by the model (differences in fitted means) |
| **Unexplained variation** | $SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$ | $SS_{Res} = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$ | Variation not explained by the model (residuals) |
| **Total variation** | $SS_{Total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y})^2$ | $SS_{Tot} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$ | Total variation in the data around the overall mean |