1 Partition of the Variation in Y and the Model F Test

1.1 Sums of Squares

- Variation in the observed responses $Y_1, Y_2, \ldots, Y_n$ emanates from two sources:

  1. Variation in $Y_1, Y_2, \ldots, Y_n$ **due to X, i.e. due to the linear relationship between Y and X.**
  2. Variation in $Y_1, Y_2, \ldots, Y_n$ **due to random error, i.e. due to all other factors *besides* X.**

- Define the ***total sum of squares***, denoted **SSTO**, to be:

Total Sum of Squares:

$$\text{SSTO} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

SSTO measures the **total variation** in the observed responses $Y_1, Y_2, \ldots, Y_n$.

- Recall that the ***error sum of squares***, denoted **SSE**, is:

Error Sum of Squares:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

SSE measures variation in $Y_1, Y_2, \ldots, Y_n$ due to **random error**.

- Also, we define the ***regression sum of squares***, denoted **SSR**, to be:

Regression Sum of Squares:

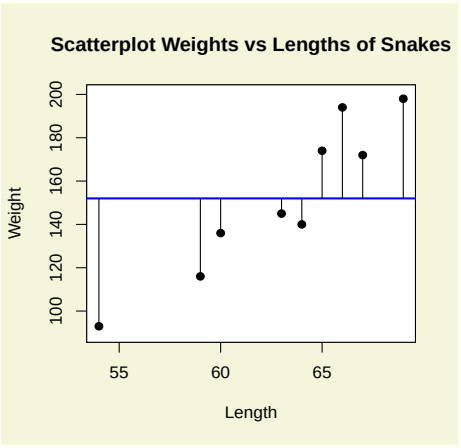
$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

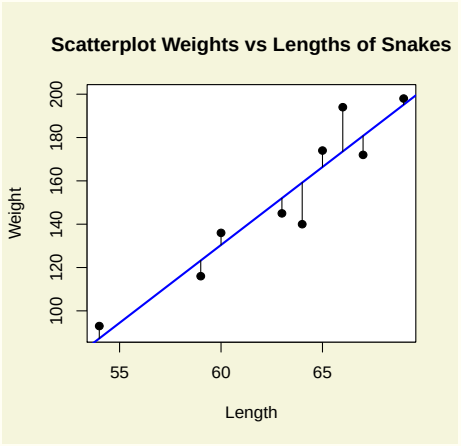
SSR measures variation in $Y_1, Y_2, \ldots, Y_n$ that's **due to X, i.e. due to the linear relationship between Y and X.**

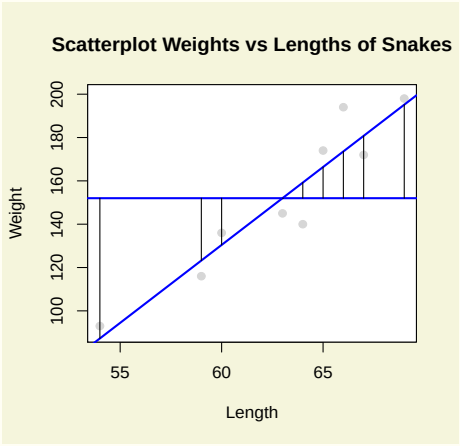
Example 1.1 The data and scatterplot below show the lengths (X) and weights (Y) of nine female snakes.

Lengths and Weights of Female Snakes

Snake	Length (cm)	Weight (g)
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145







SSTO measures the variation depicted in the top right plot, **SSE** the variation in the bottom left plot, and **SSR** that the bottom right plot.

- We can decompose a total deviation $Y_i - \bar{Y}$ into two parts:

$$Y_i - \bar{Y} \;=\; \hat{Y}_i - \bar{Y} \;+\; Y_i - \hat{Y}_i \tag{1}$$

If we square both sides and sum over i, it turns out that the cross-product terms sum to zero, i.e.

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \;=\; 0\,,$$

and we get

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \;=\; \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \;+\; \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2}$$

which is known as the ***partition of the total variation*** in the data.

Partition of the Total Variation in the Responses:

$$\text{SSTO} \;=\; \text{SSR} \;+\; \text{SSE} \tag{3}$$

This splits the total variation in $Y_1, Y_2, \ldots, Y_n$ (SSTO) into two parts, one due to X (SSR) and another due to all other factors, or random error (SSE).

1.2 Degrees of Freedom

- Associated with the partition (3) is a corresponding breakdown of the ***total degrees of freedom*** into ***degrees of freedom for regression*** and ***degrees of freedom for error***:

Degrees of Freedom:

$$\begin{aligned} \text{Total df (for SSTO)} &= n-1 \\ \text{df for Regression (SSR)} &= 1 \\ \text{df for Error (SSE)} &= n-2 \end{aligned}$$

Notice that **the degrees of freedom for SSE is the sample size n minus the number of parameters in the model, 2**, and that

$$\textbf{Total df} \;=\; \textbf{df for Regression} \;+\; \textbf{df for Error}$$

1.3 Mean Squares

- Dividing each sum of squares on the right side of (3) by its degrees of freedom gives the ***mean square***.

Mean Squares: The ***mean square for regression*** and ***mean squared error***, denoted **MSR** and **MSE**, respectively, are

$$\text{MSR} \;=\; \frac{\text{SSR}}{1}$$

and

$$\text{MSE} \;=\; \frac{\text{SSE}}{n-2}$$

1.4 Regression Model *F* Test

- It can be shown that

$$\text{E(MSE)} \;=\; \sigma^2$$

and

$$\text{E(MSR)} \;=\; \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Thus if

$$H_0 : \beta_1 = 0$$

is true,

$$\text{E(MSR)} \;=\; \text{E(MSE)} \;=\; \sigma^2 .$$

On the other hand, if

$$H_a : \beta_1 \neq 0$$

is true,

$$\text{E(MSR)} \;>\; \text{E(MSE)} .$$

- As an alternative to the t test discussed earlier, we could test

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned} \tag{4}$$

using the ***regression model F test***, with **test statistic**

Test Statistic: The test statistic for the *regression model F test* is

$$F \;=\; \frac{\text{MSR}}{\text{MSE}} . \tag{5}$$

Fact 1.1 Under the simple linear regression model, with the ϵ_i's independent $N(0, \sigma^2)$, if $H_0 : \beta_1 = 0$ is true,

$$F \;\sim\; F(1, n-2),$$

i.e. the F test statistic (5) follows an F distribution with numerator and denominator degrees of freedom 1 and $n-2$, respectively.

- **Large** values of F (i.e. values substantially greater than one) provide evidence against H_0. The **p-value** is the tail area to the **right** of the observed F value under the $F(1, n-2)$ distribution.
- The regression model F test and the t test for the slope of the regression line are equivalent:

Fact 1.2 The F statistic (5) is the square of the t statistic for testing (4) (Class Notes 2), i.e.

$$F = t^2$$

and the p-values for the two tests will be the same.

1.5 The Regression ANOVA Table

- The sums of squares, degrees of freedom, mean squares, F test statistic, and p-value are usually organized in a ***regression ANOVA table***:

Source of Variation	**df**	**Sum of Squares**	**Mean Square**	**F**	**P-value**
Regression Model	1	SSR	MSR = SSR/1	MSR/MSE	p
Error	$n-2$	SSE	MSE = SSE/$(n-2)$		
Total	$n-1$	SSTO			