Linear Regression

- Linear regression with one predictor
- Assess the fit of a regression model
  - Total sum of squares
  - Model sum of squares
  - Residual sum of squares
  - R^2
- Test for model significance – *F* test
- Interpret a regression model

What is Regression?

- A way of predicting the value of one variable from another.
  - It is a hypothetical model of the relationship between two variables.
  - The model used is a linear one.
  - Therefore, we describe the relationship using the equation of a straight line.

Assumptions of Simple Linear Regression

- For each value of x, Y are randomly sampled and independent.
- For any value of X in the pop'l there exists a normal distribution of Y values
- There is homogeneity of variances in the population. ie. the variance of the normal distribut. of Y values in pop'l are equal for all of values of x.
- The relationship of x and y is linear.
- X is measured without error

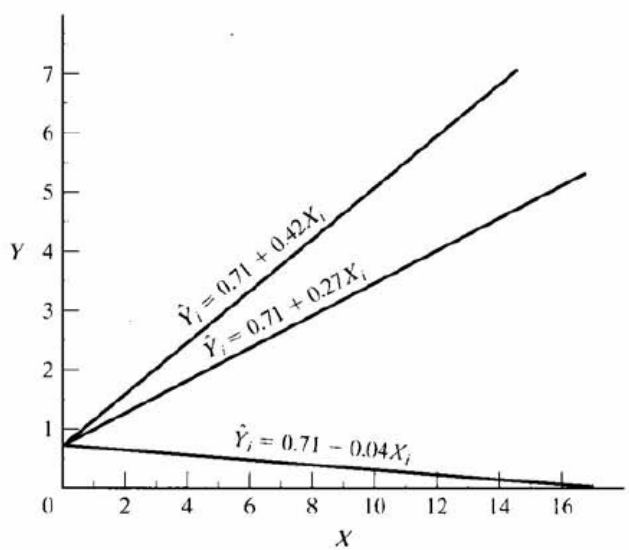
Describing a Straight Line

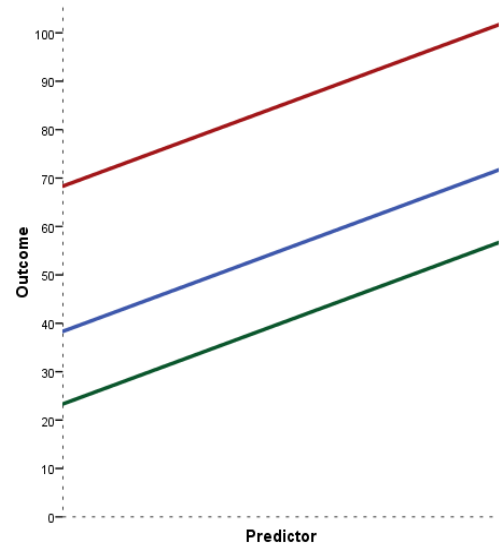
$$Y_i = b_0 + b_i X_i + \varepsilon_i$$

- b_i
  - Regression coefficient for the predictor
  - Gradient (slope) of the regression line
  - Direction/strength of relationship
- b_0
  - Intercept (value of *Y* when *X* = 0)
  - Point at which the regression line crosses the *Y*-axis (ordinate)

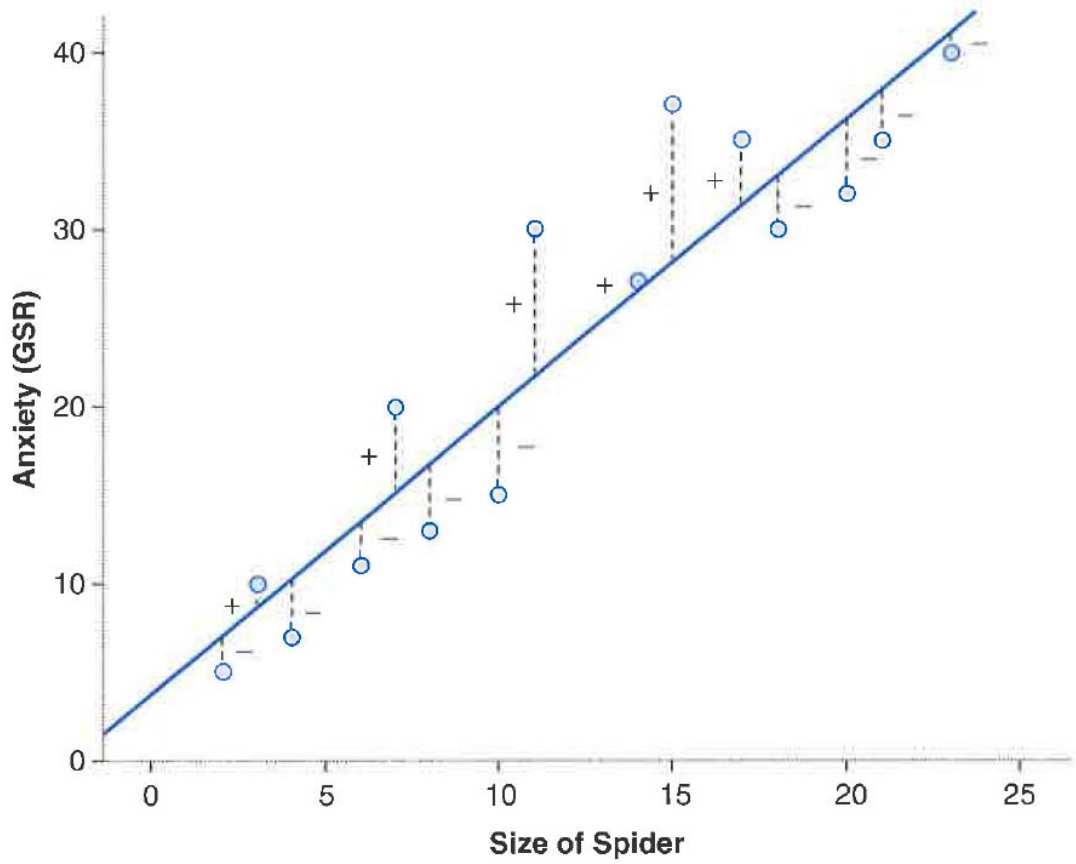
Intercepts and Gradients

$$Y_i = b_0 + b_i X_i + \varepsilon_i$$





The Method of Least Squares



This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

How Good Is the Model?

- The regression line is a model based on the data.
- This model might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
  - How?

Sums of Squares

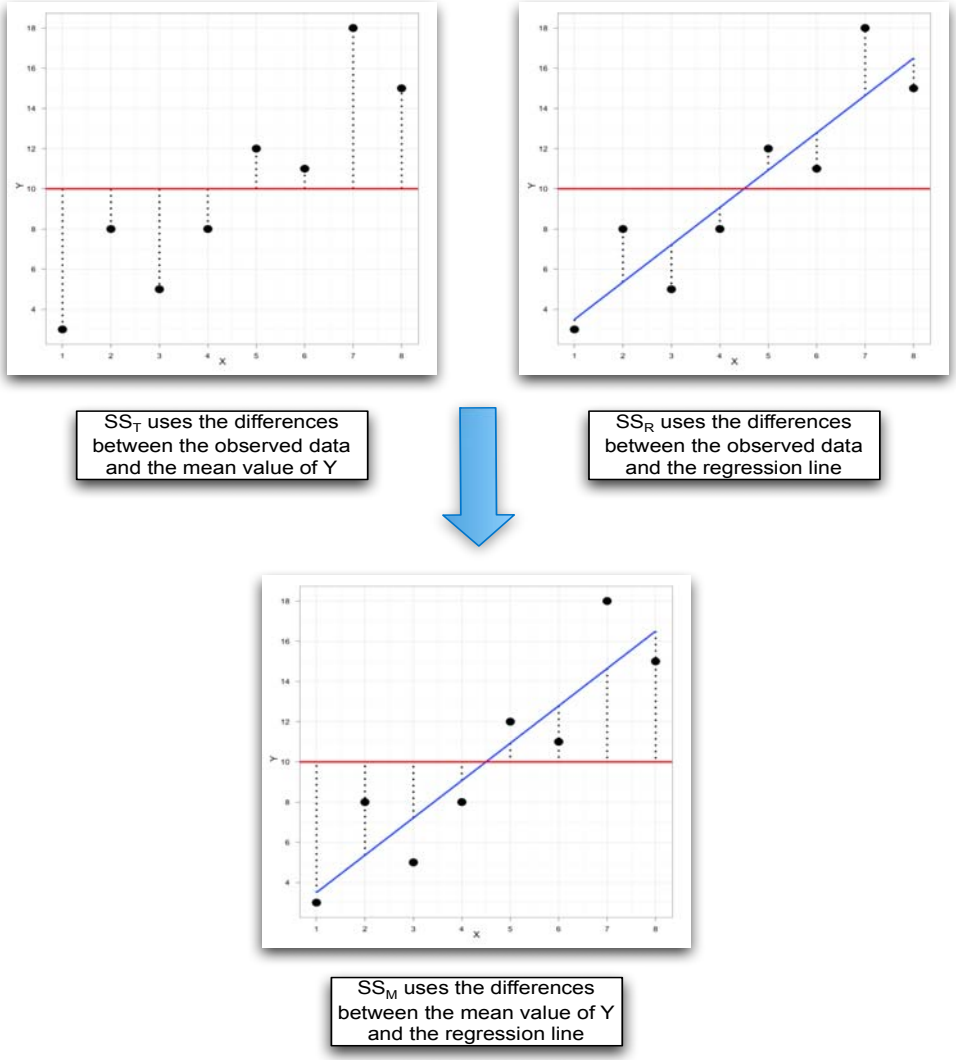
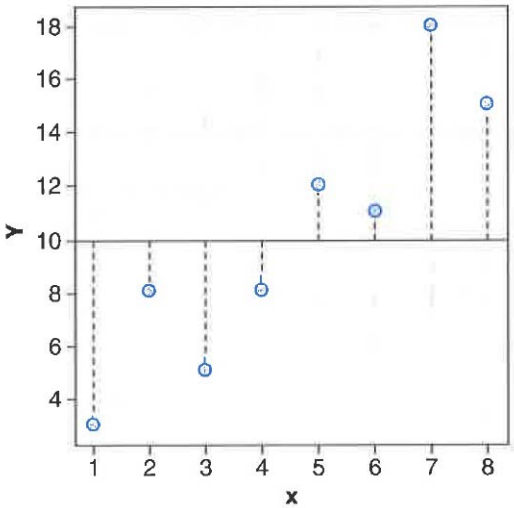


Diagram showing from where the regression sums of squares derive

Total SS (SS_T)

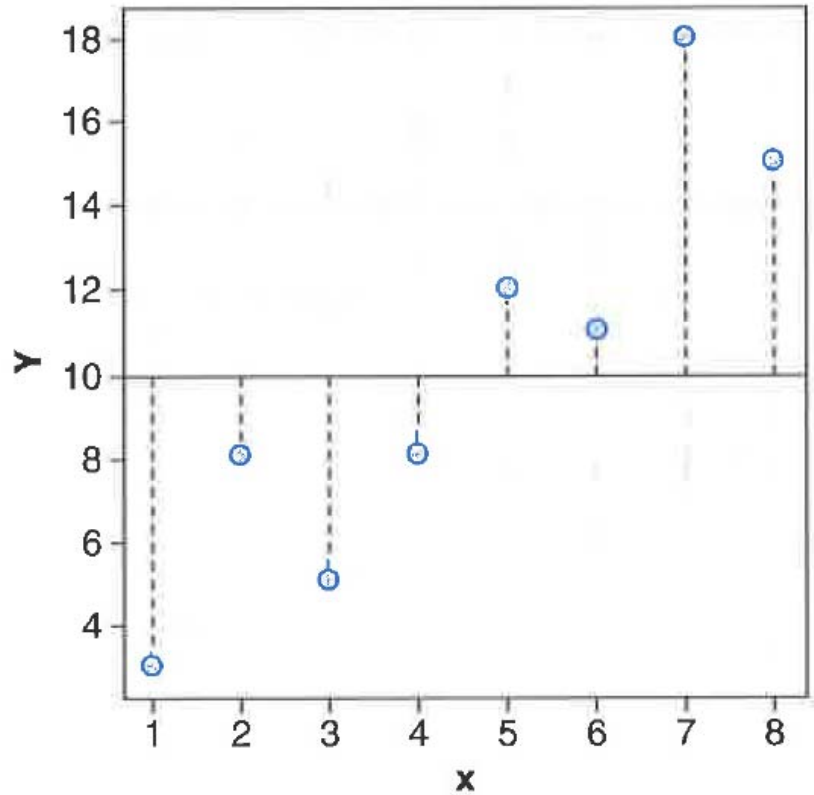
- SS_T
  - Total variability (variability between scores and the mean).
- TSS is the sum of the squared residuals when the most basic model is applied to the data.
- How good is the mean as a model to the observed data?



Total SS (SS_T)

- SS_T
  - Total variability (variability between scores and the mean).

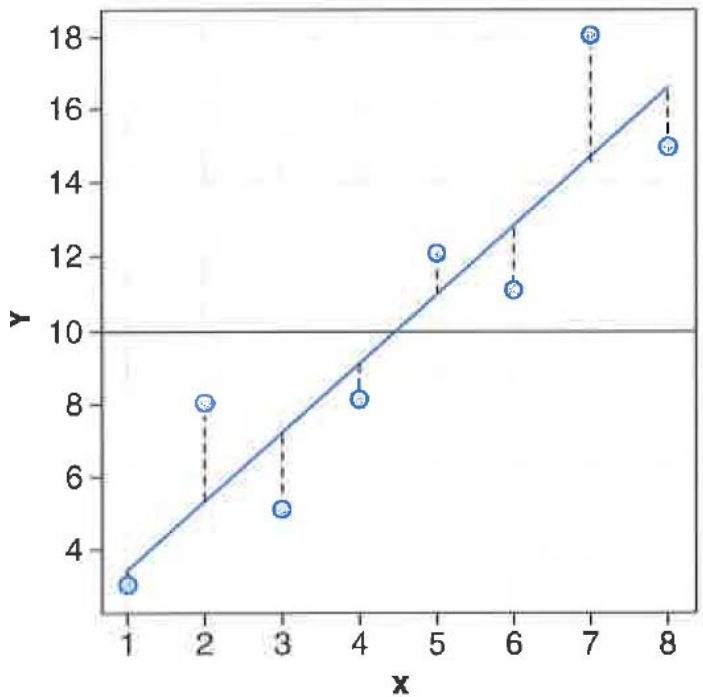
$$\text{total SS} = \sum (Y_i - \overline{Y})^2$$



SS_T uses the differences between the observed data and the mean value of Y

Residual SS or Error SS (SS_R)

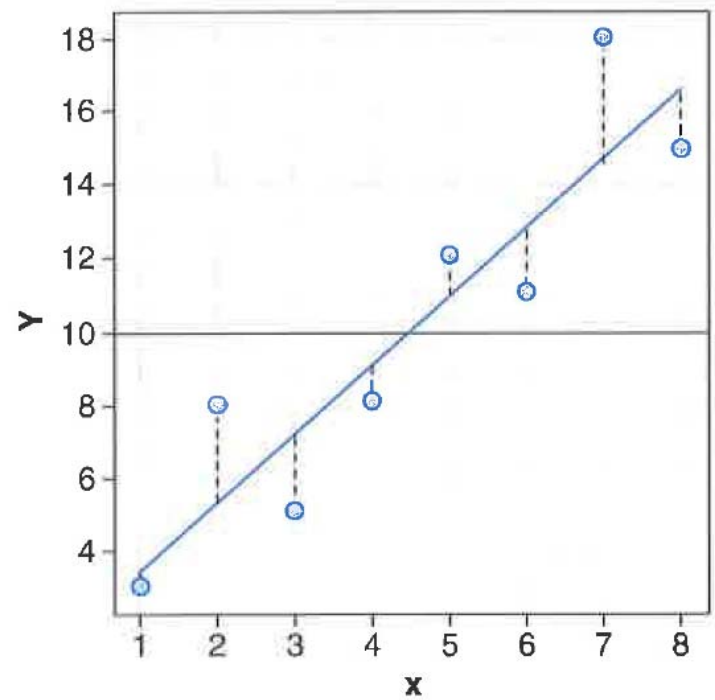
- SS_R
  - Residual/error variability (variability between the regression model and the actual data).
- Difference between the observed data and the model
- This represents the degree of inaccuracy when fitting the best fit model to the data.



Residual SS

- SS_R
  - Residual/error variability (variability between the regression model and the actual data).

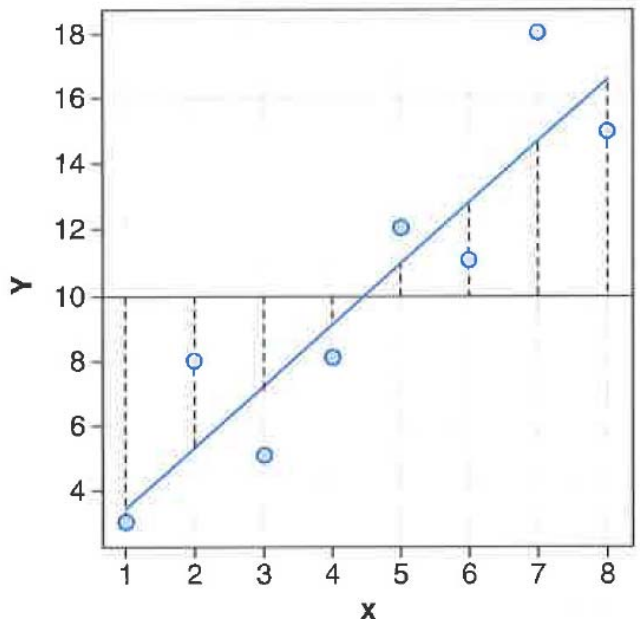
$$\text{residual SS} = \sum (Y_i - \hat{Y}_i)^2$$



SS_R uses the differences between the observed data and the regression line

Model SS or Regression SS (SS_M)

- SS_M
  - Model variability (difference in variability between the model and the mean).
- This is the improvement we get from fitting the model to the data relative to the null model.

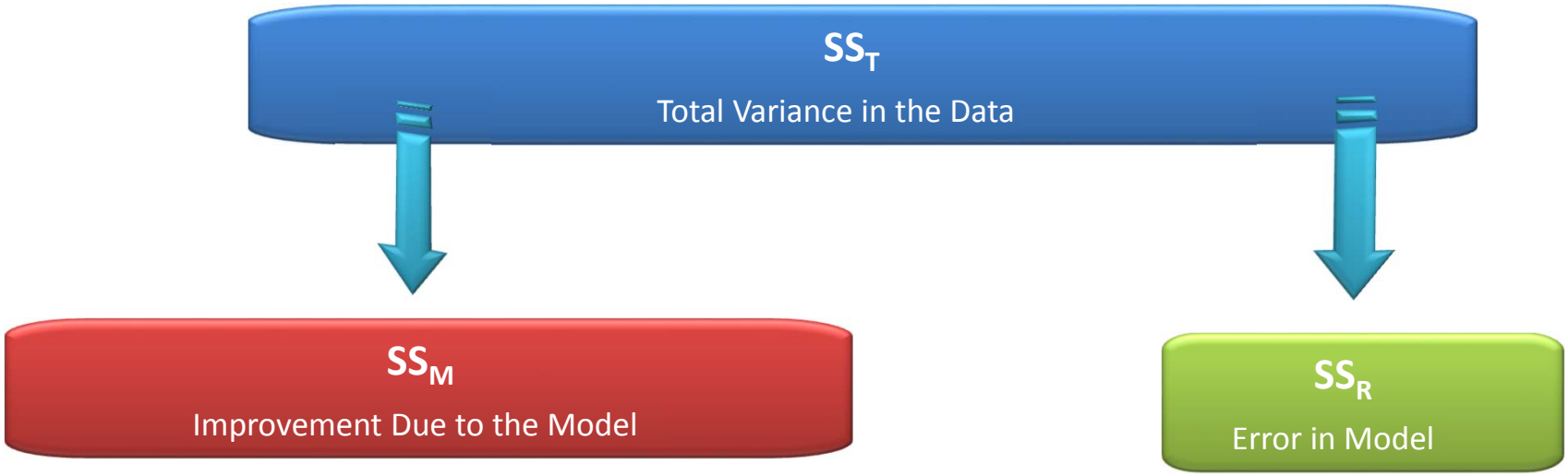


$$\text{regression SS} = \sum(\hat{Y}_i - \overline{Y})^2$$

SST = SSR + SSM

- How to we get large SSM?
- What happens if the SSM is large?
- Regression model is much different from using the mean as the outcome, therefore regression model improves the outcome.
- So, we can calculate the proportion of improvement due to the model.
- SSM/SST, percentage of variation explained by the model.

Testing the Model: ANOVA



- If the model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R
- SST = SSM + SSR

Evaluating the quality of the Model: R^2

- R^2
  - The proportion of variance accounted for by the regression model.
  - The Pearson Correlation Coefficient Squared

$$R^2 = \frac{SS_M}{SS_T}$$

SS for model testing

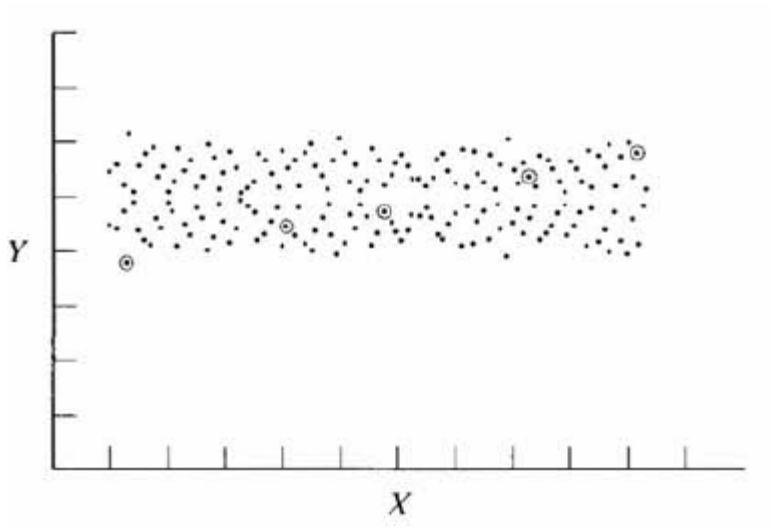
- A second use of the sum of squares values is to test the model.

Testing H_0: $\beta = 0$ against H_A: $\beta \neq 0$

- Evaluate the amount of systematic variance (regression) divided by the amount of unsystematic (residual) variance.
- The magnitude of the sum of squares is dependent on the number of observations

SS for model testing

Testing H_0: $\beta = 0$ against H_A: $\beta \neq 0$



SS for model testing

- F test – “termed variance ratio test”

1. We divide the SSM and SSR by their respective degrees of freedom (DF).
   - DF for SSM is the number of parameters in the model.
   - DF for SSR number of obs – number of parameters in the model.

Degrees of freedom

- Given a statistic (mean, var) and sample size of a population.
- DF are the number of terms that are independent, such that when any of the other terms are known, the value can be estimated.

Testing the Model: ANOVA

- Mean squared error
  - Sums of squares are total values.
  - They can be expressed as averages, divided by DF terms.
  - These are called mean squares, MS.

$$F = \frac{MS_M}{MS_R}$$

EXAMPLE 17.1 Wing Lengths of 13 Sparrows of Various Ages. The Data Are Plotted in Figure 17.1.

Age (days) (X)	*Wing length* (cm) (Y)
3.0	1.4
4.0	1.5
5.0	2.2
6.0	2.4
8.0	3.1
9.0	3.2
10.0	3.2
11.0	3.9
12.0	4.1
14.0	4.7
15.0	4.5
16.0	5.2
17.0	5.0

$n = 13$

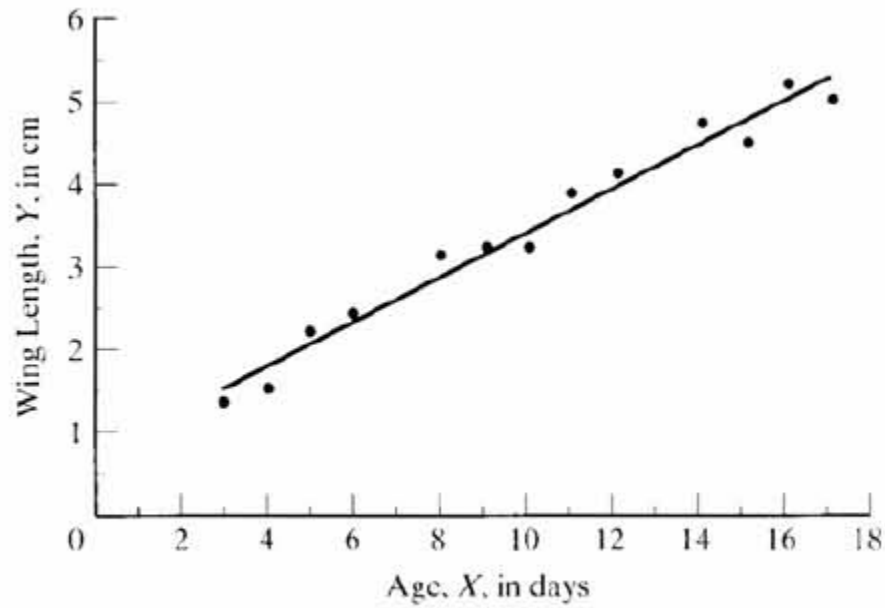


FIGURE 17.1: Sparrow wing length as a function of age. The data are from Example 17.1.

$$b = \frac{\sum xy}{\sum x^2} \qquad \alpha = \overline{Y} - \beta\overline{X}.$$

Worked example

- TSS

$$\text{total SS} = \sum y^2 = 171.30 - \frac{(44.4)^2}{13}$$
$$= 171.30 - 151.6431$$
$$= 19.656923$$

- Model SS

$$\text{regression SS} = \frac{\left(\sum xy\right)^2}{\sum x^2} = \frac{(70.80)^2}{262.00}$$
$$= \frac{5012.64}{262.00}$$
$$= 19.132214$$

Worked example

TABLE 17.1: Summary of the Calculations for Testing $H_0: \beta = 0$ against $H_A: \beta \neq 0$ by an Analysis of Variance

Source of variation	Sum of squares (SS)	DF	Mean square (MS)
Total $[Y_i - \overline{Y}]$	$\sum y^2$	$n - 1$	
Linear regression $[\hat{Y}_i - \overline{Y}]$	$\frac{(\sum xy)^2}{\sum x^2}$	1	$\frac{\text{regression SS}}{\text{regression DF}}$
Residual $[Y_i - \hat{Y}_i]$	total SS − regression SS	$n - 2$	$\frac{\text{residual SS}}{\text{residual DF}}$

DF for Regression (model DF) is 1 in simple linear regression

Residual DF (Error DF) is equal n - 2

Worked example

Source of variation	SS	DF	MS
Total	19.656923	12	
Linear regression	19.132214	1	19.132214
Residual	0.524709	11	0.047701

$$F = \frac{19.132214}{0.047701} = 401.1$$

$$F_{0.05(1),1,11} = 4.84$$

Therefore, reject H_0.

$$P \ll 0.0005 \quad [P = 0.00000000053]$$

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and downloads) in the week after release
- Predictor variable:
  - The amount (in units of £1000) spent promoting the record before release.

Output of a Simple Regression

- In R:

```
summary(albumSales.1)

>Coefficients:
              Estimate      Std. Error    t value     Pr(>|t|)
(Intercept) 1.341e+02   7.537e+00   17.799      <2e-16 ***
adverts      9.612e-02   9.632e-03    9.979      <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65.99 on 198 degrees of freedom
Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313
F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

Using the Model

$$\begin{aligned}\text{Record Sales}_i &= b_0 + b_1 \text{Advertising Budget}_i \\ &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right)\end{aligned}$$

$$\begin{aligned}\text{Record Sales}_i &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right) \\ &= 134.14 + \left(0.09612 \times 100\right) \\ &= 143.75\end{aligned}$$