Lecture 12:
Simple Linear Regression
STAT 310, Spring 2021

Scatterplots

- A scatterplot a graphical display used to study the relationship between two numerical variables x and y.
- Data displayed on a scatterplot are collected in pairs:

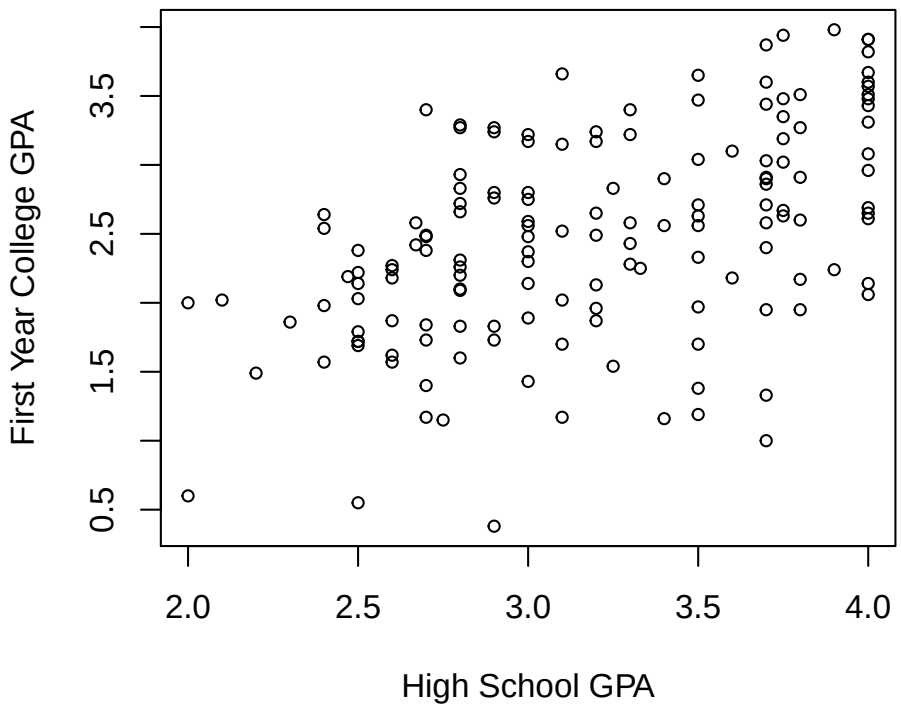
$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

where n denotes the total number of cases or pairs.

- A scatterplot provides insight into how two variables are related.

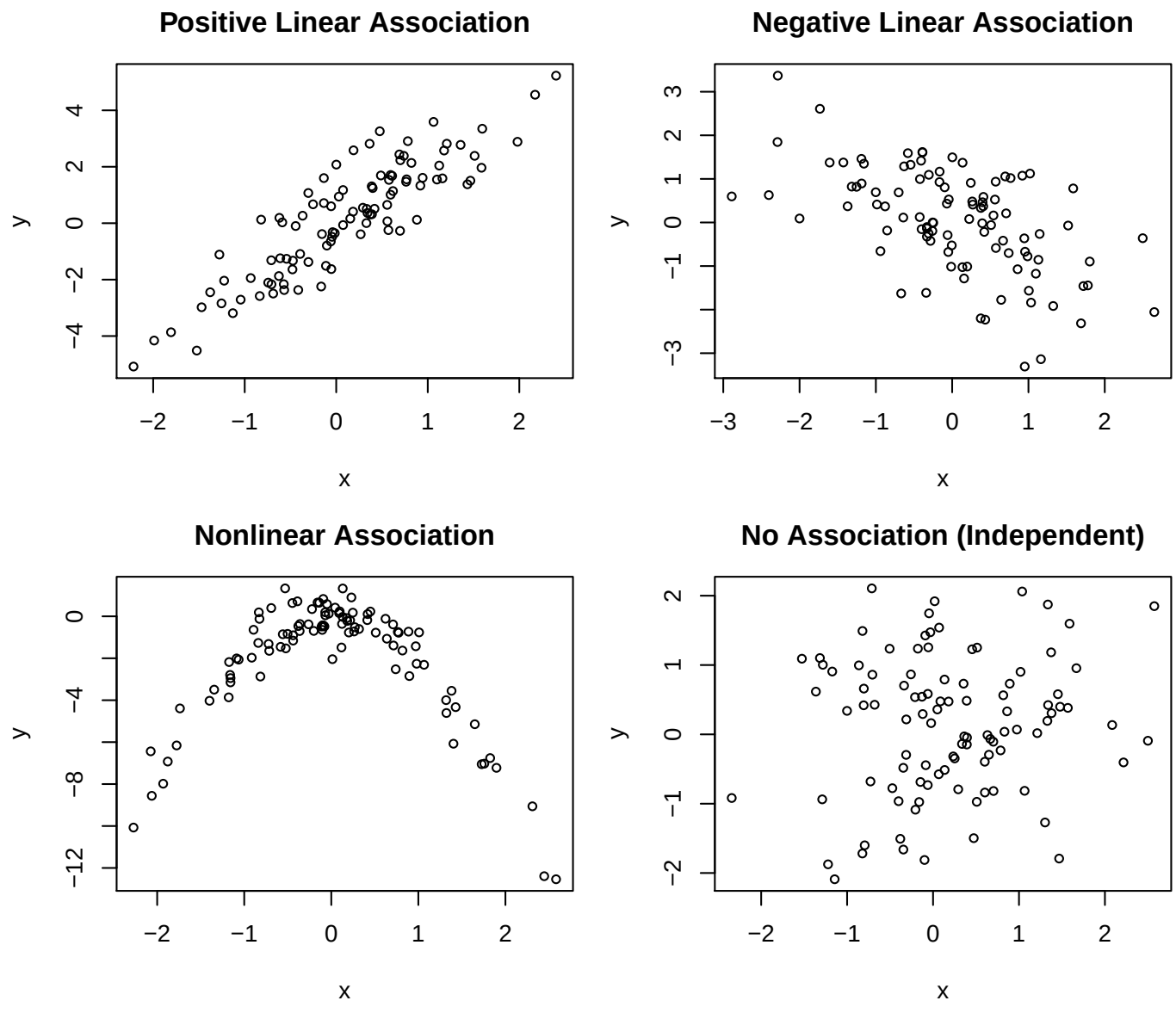
Example

A scatterplot showing the association between first year college GPA and high school GPA for a random sample of 150 students.



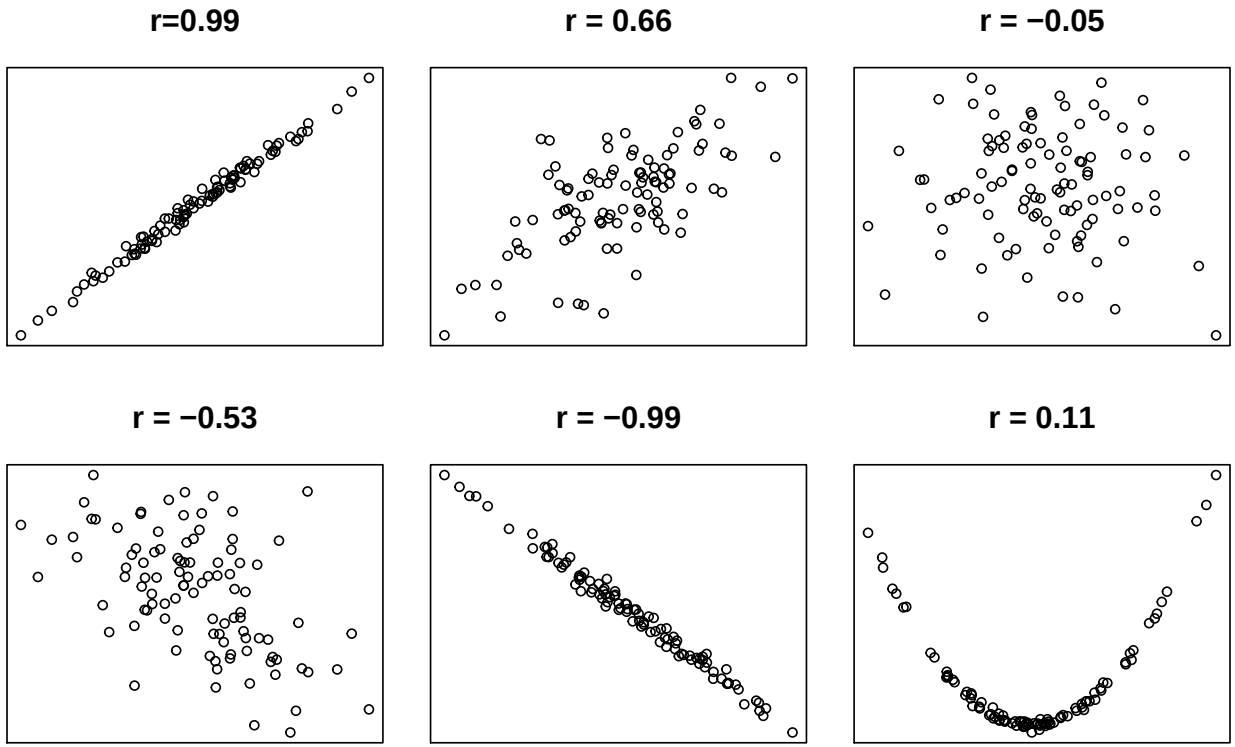
Types of Relationships Between Variables

- Two variables are said to be **associated** if the scatterplot shows a discernible pattern or trend.
- An association is **positive** if y increases as x increases.
- An association is **negative** if y decreases as x increases.
- An association is **linear** if the scatterplot between x and y has a linear trend; otherwise, the association is called **nonlinear**.


Positive Linear Association
Negative Linear Association
Nonlinear Association
No Association (Independent)
x
y

Correlation Coefficient

The **correlation coefficient**, denoted by r, is a number between -1 and 1 that describes the strength of the linear association between two variables.



Correlation Coefficient

- $r \approx 1$ when there is a strong positive linear association between the variables.
- $r \approx -1$ when there is a strong negative linear association between the variables.
- $r \approx 0$ when there is no association between the variables (i.e., independent).
- The correlation coefficient is only useful for evaluating the linear association between two variables. It is not a useful measure for nonlinear relationships.

Correlation Coefficient

Formally, the correlation can be calculated using the following formula. The formula is rather complex, so we use software packages such as R to do the calculations for us.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- $\bar{x}$ and $\bar{y}$ are the sample means
- s_x and s_y are the sample standard deviations

Example

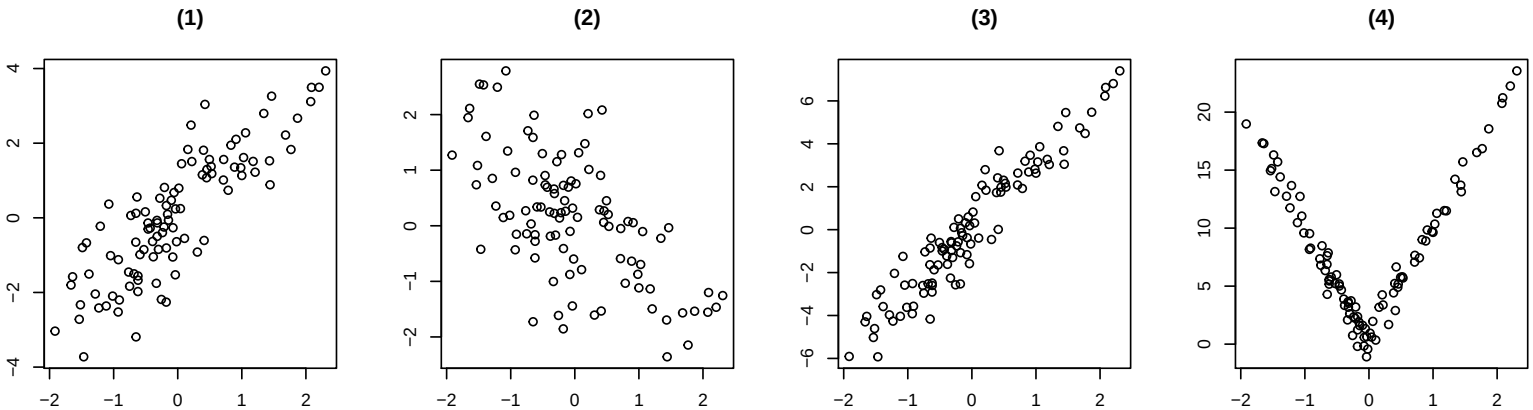
Match each correlation to the corresponding scatterplot.

(a) $r = -0.63$

(b) $r = 0.85$

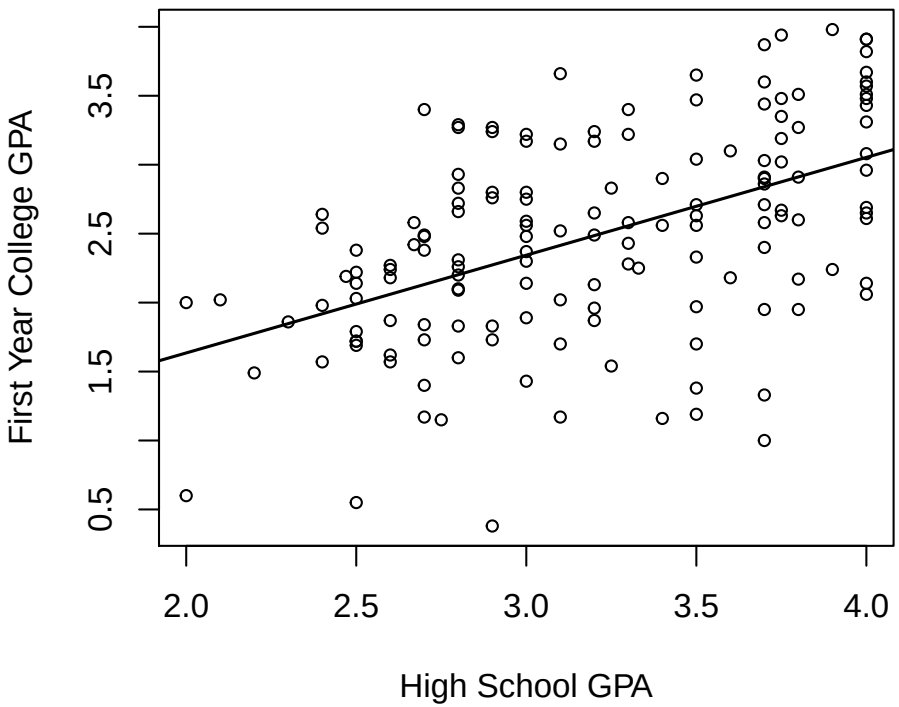
(c) $r = 0.19$

(d) $r = 0.95$



Simple Linear Regression

- **Simple linear regression** is a method for fitting a straight line to data that show a linear trend when displayed on a scatterplot.
- The method is useful for making predictions and explaining the relationship between two numerical variables.



Simple Linear Regression

A **simple linear regression model** expresses the relationship between two variables, x and y, as a straight line with some error:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y is called the **response** variable
- x is called the **explanatory** or **predictor** variable.
- β_0 is the **intercept** parameter
- β_1 is the **slope** parameter
- ϵ is called the **random error** term. It captures the variability in the points around the line.

Fitted Values and Residuals

- The line that we estimate, or fit to the data in the scatterplot, is written as

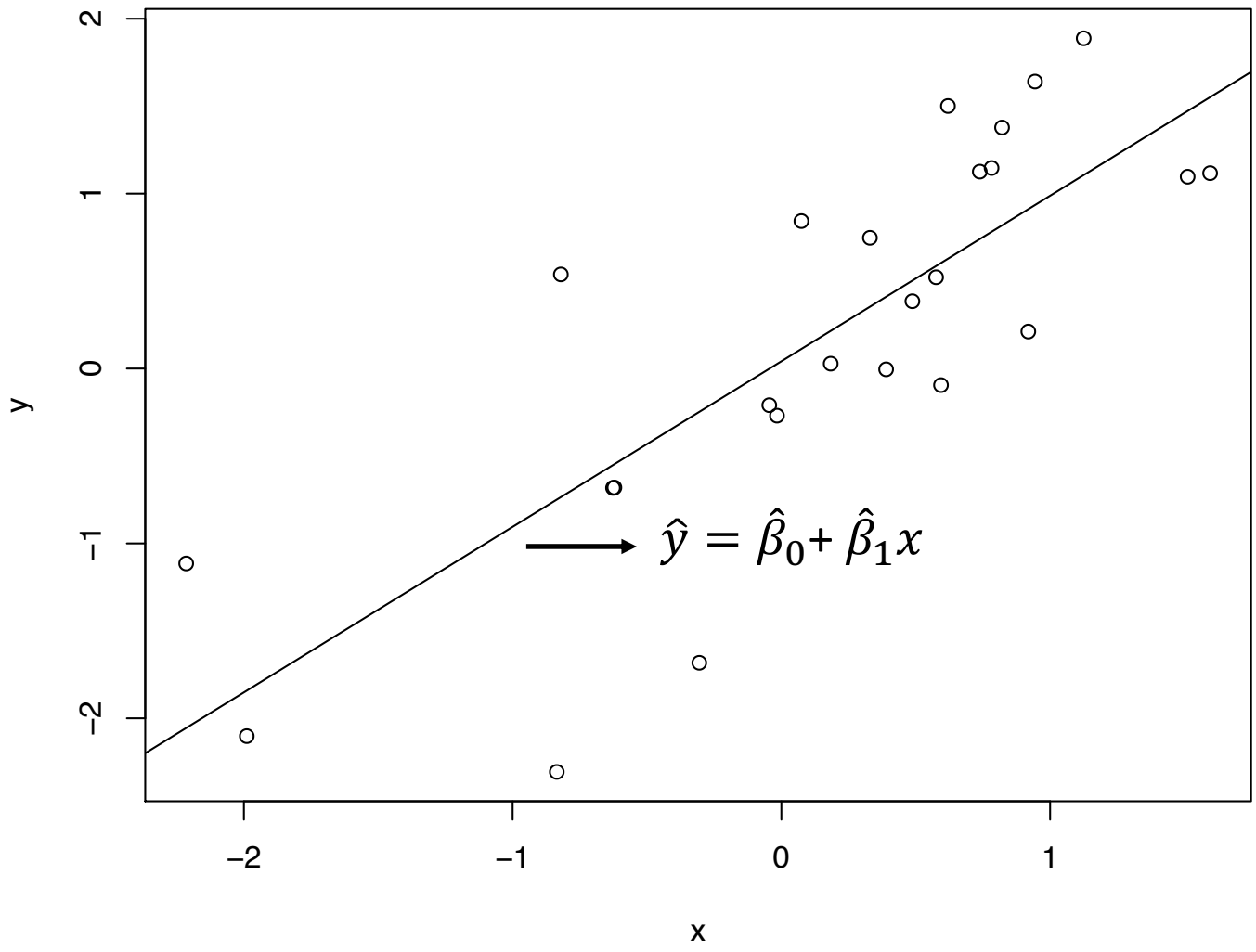
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

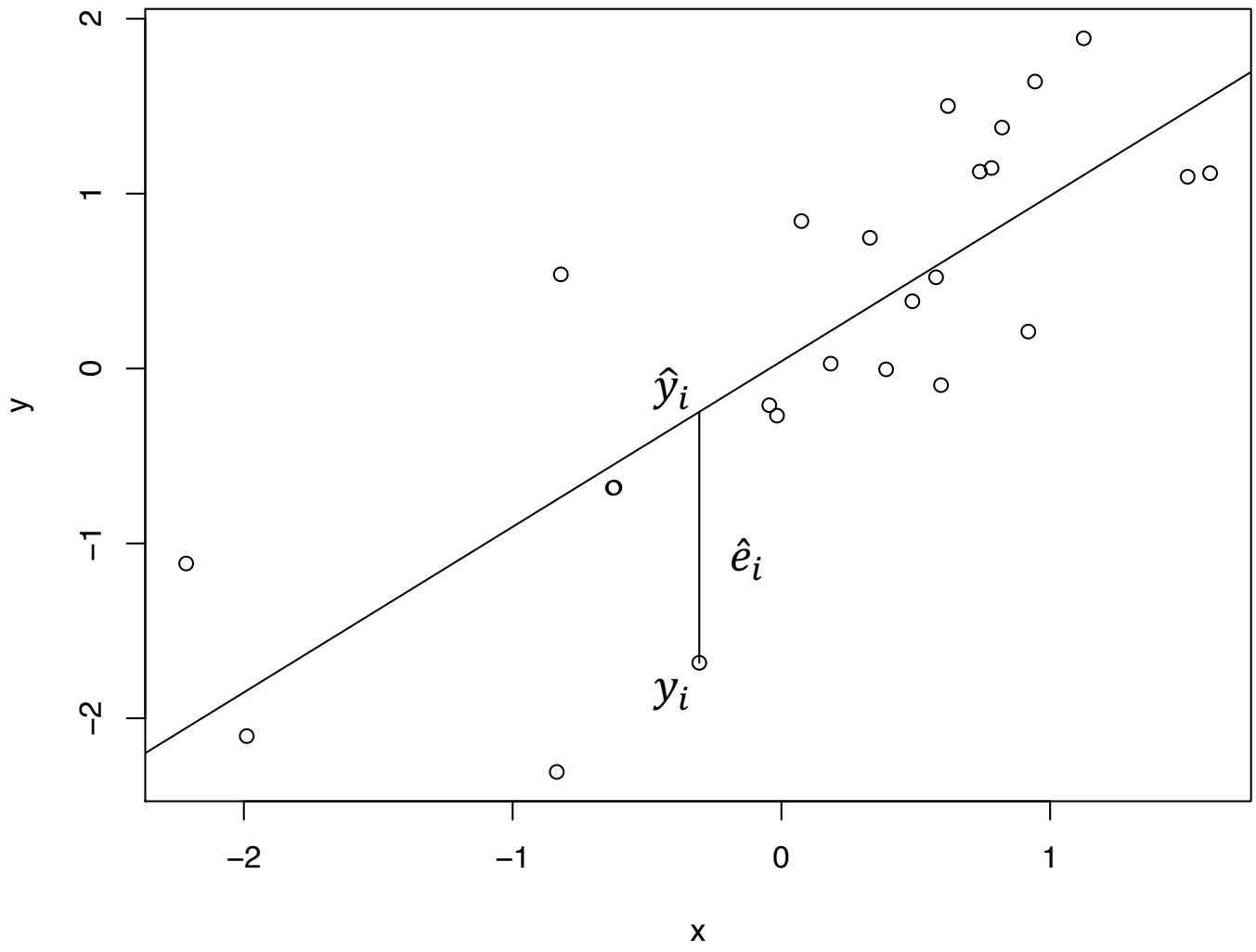
- The fitted (or predicted) value for the i^{th} observation (x_i, y_i):

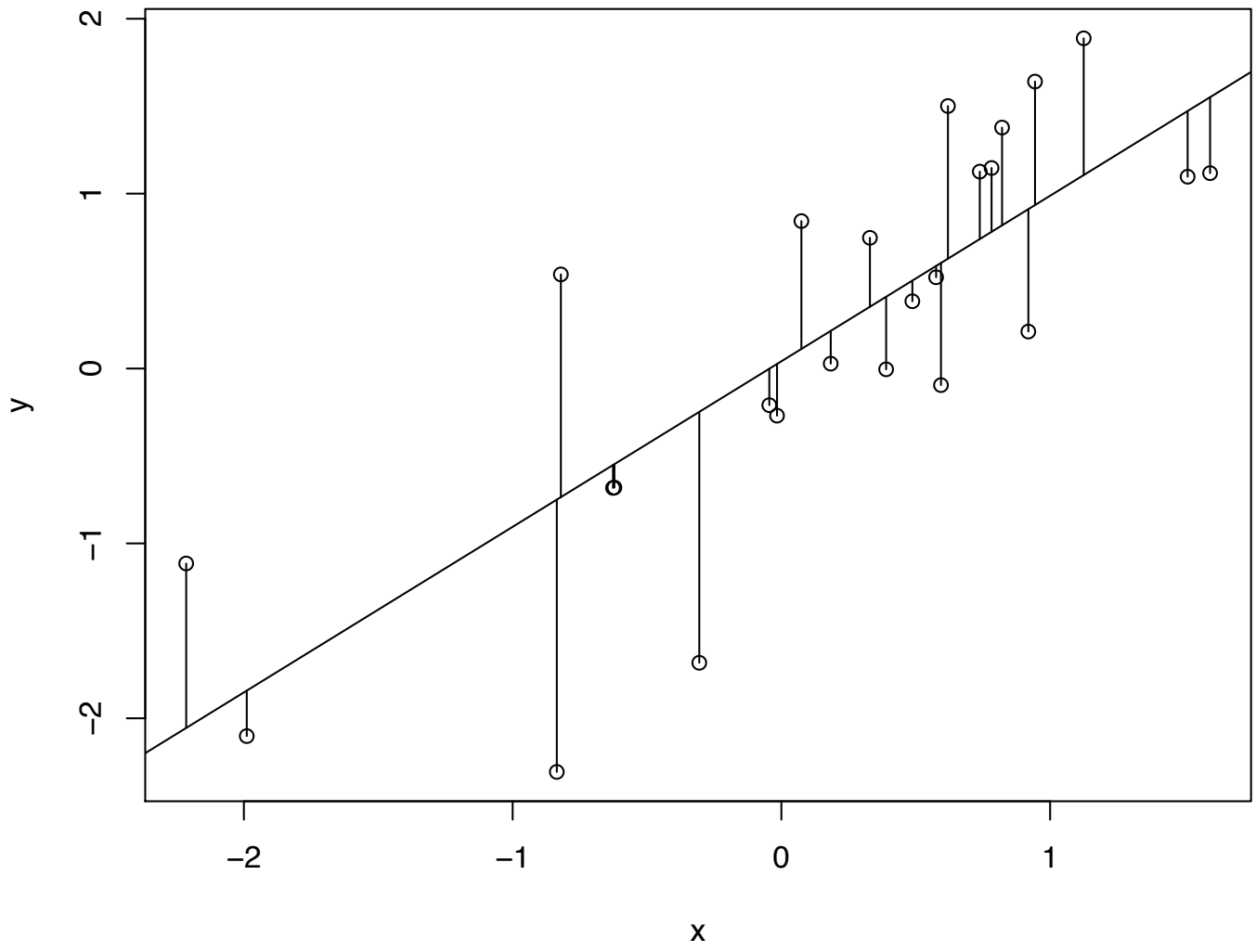
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The **residual** for the i^{th} observation is the difference between the observed value (y_i) and the predicted value $(\hat{y}_i)$:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$


2
1
0
−1
−2
y
−2
−1
0
1
x
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$


2
1
0
−1
−2
y
−2
−1
0
1
x
$\hat{y}_i$
$\hat{e}_i$
y_i


y
2
1
0
−1
−2
−2
−1
0
1
x

Sum of Squared Residuals

- Intuitively, a line that fits the data well has small residuals.
- The **least squares line** minimizes the **sum of squared residuals**:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- That is, out of all possible lines we could draw on the scatterplot, the least squares line is the "best fit" since it has the smallest sum of squared residuals.

Least Squares Estimates

It can be shown (using calculus) that the estimates of the intercept and slope that minimize the sum of squared residuals are given by the following formulas:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$

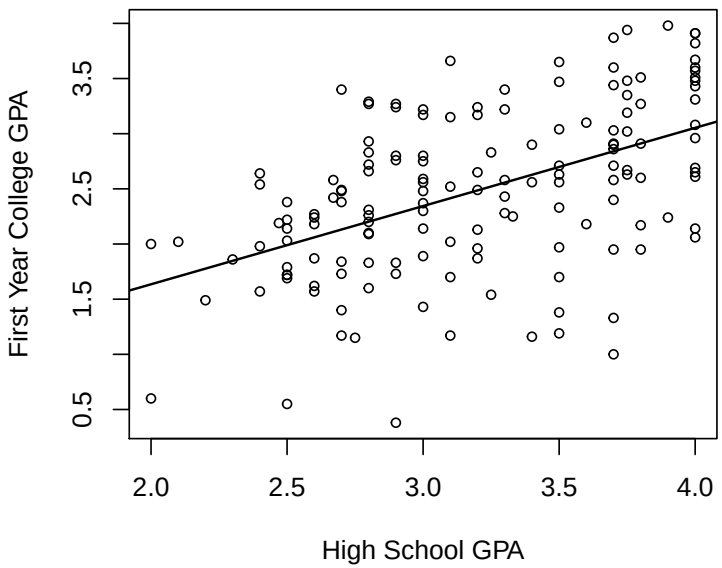
where r is the correlation coefficient, previously discussed. Note that the equation for the intercept guarantees the least squares line passes through the point $(\bar{x}, \bar{y})$.

Example

The least squares regression line in the scatterplot is given by:

$$\hat{y} = 0.217 + 0.709\,x$$

Suppose a student graduates high school with a 3.0 GPA. What is the predicted first year college GPA for this student?

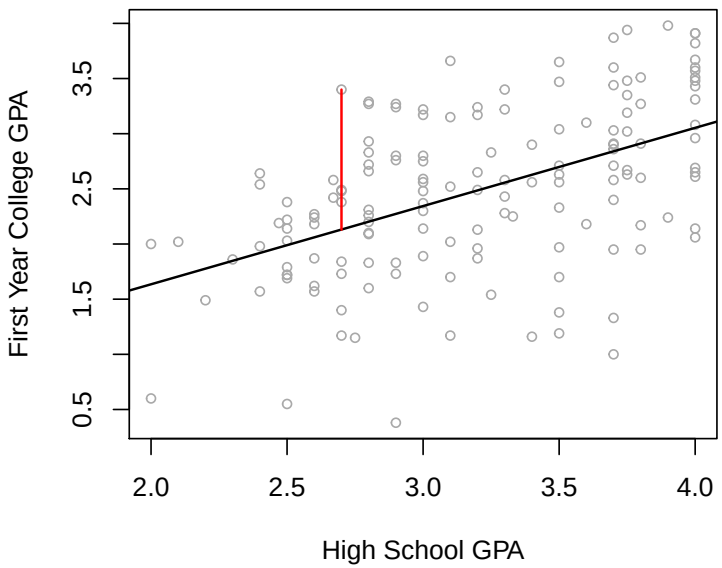


Example

The least squares regression line in the scatterplot is given by:

$$\hat{y} = 0.217 + 0.709\,x$$

Calculate the residual (show in red) for a student, in this data set, that had a 2.7 high school GPA and a 3.4 college GPA.



Example

Use the summary statistics in the table below to manually calculate the slope and intercept of the least squares line.

	HS GPA (x)	FY College GPA (y)
Mean	$\bar{x} = 3.196$	$\bar{y} = 2.48$
SD	$s_x = 0.534$	$s_y = 0.753$
	correlation	$r = 0.502$

Interpreting Coefficients

- **Slope**: an increase in the explanatory variable (x) by one unit is associated with a change of $\hat{\beta}_1$ in the predicted response ($\hat{y}$).

- **Intercept**: the prediction for the response variable ($\hat{y}$) when the value for the explanatory variable is zero ($x = 0$). It may not make sense to try to interpret the intercept depending on the application.

Interpreting Coefficients

The least squares regression line for predicting first year college GPA (y) from high school GPA (x) is given by:

$$\hat{y} = 0.217 + 0.709\,x$$

Interpret the slope and intercept of this model.

Coefficient of Determination (R^2)

- The coefficient of determination (R^2) is a measure of how well the linear regression model fits the data.
- The R^2 can be computed as the correlation coefficient r squared.
- R^2 can be interpreted as the proportion of variability in the response variable y that is explained by x.
- R^2 is always between 0 and 1; the closer R^2 is to 1, the better the linear regression model fits the data.

Coefficient of Determination (R^2)

- Going back to the example, the correlation between high school GPA and first year college GPA is 0.502.
- So $R^2 = (0.502)^2 = 0.252$.
- This tells us that about 25% of the variability in first year college GPA (y) can be explained by high school GPA (x).