Linear Regression
and
Correlation

Linear Regression
and
Correlation

- What is correlation?
- What is the correlation coefficient?
- What information does the correlation coefficient provide regarding linear regression?
- What is meant by the line of best fit?

Linear Trend

- For a strong linear trend
  - Points in scatter plot are tightly packed around a possible line
- For a weak linear trend
  - Points in scatter plot are loosely scattered around a possible line

Linear Trend

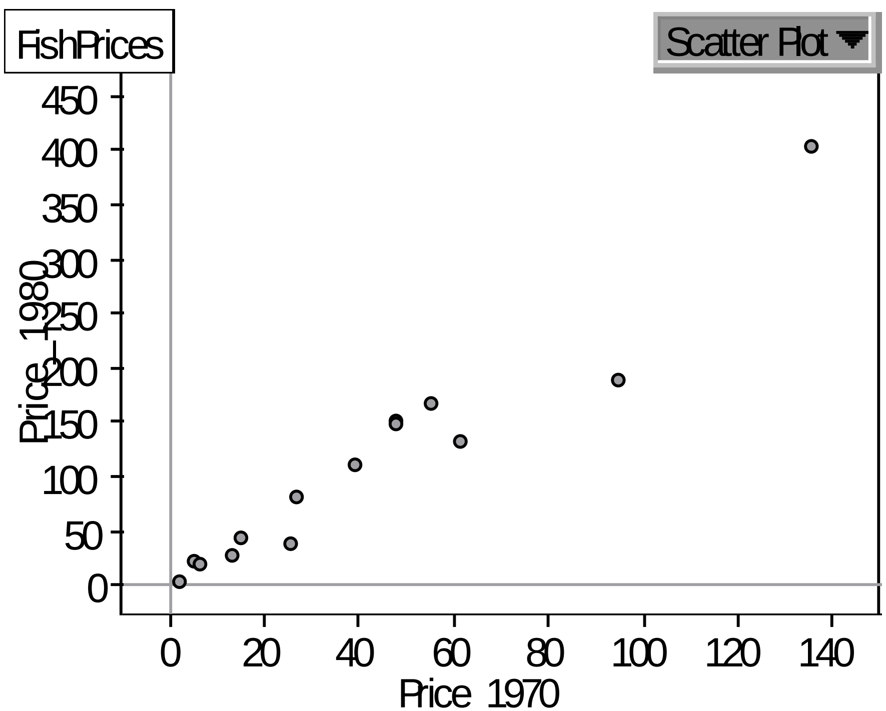
- For a strong linear trend
  - Points in scatter plot are tightly packed around the regression line
- For a weak linear trend
  - Points in scatter plot are loosely scattered around the regression line

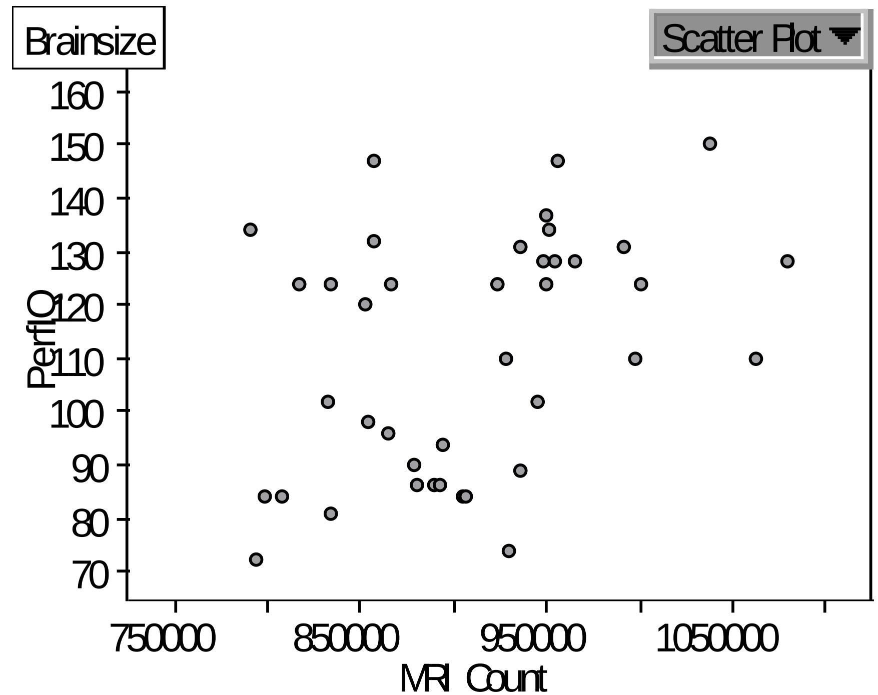
Linear Trend

- For a strong linear trend
  - Points in scatter plot are tightly packed around the line that best-fits the data

- For a weak linear trend
  - Points in scatter plot are loosely scattered around the line that best-fits the data

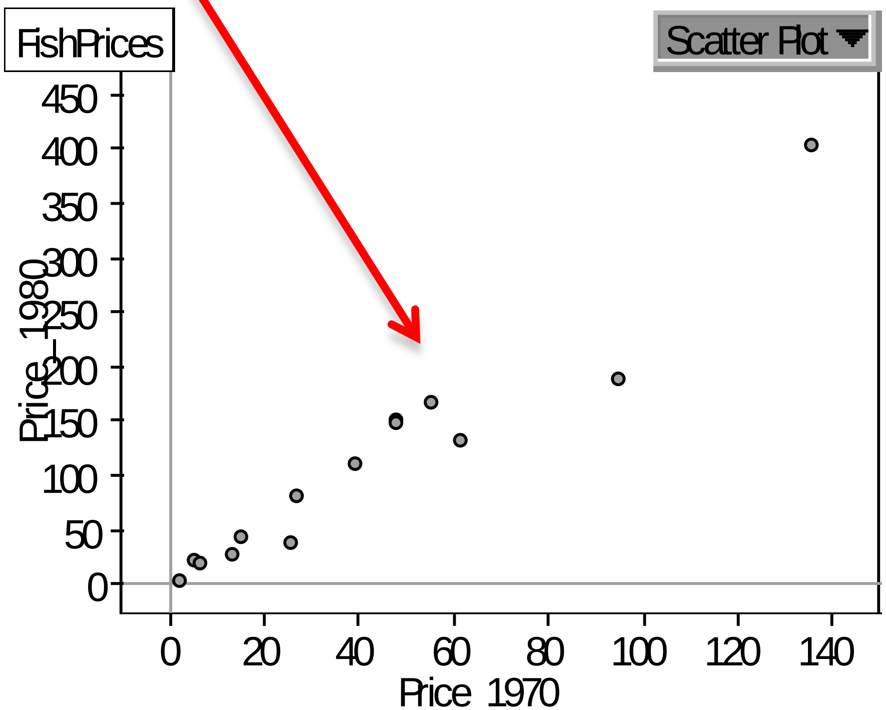
Linear Trend

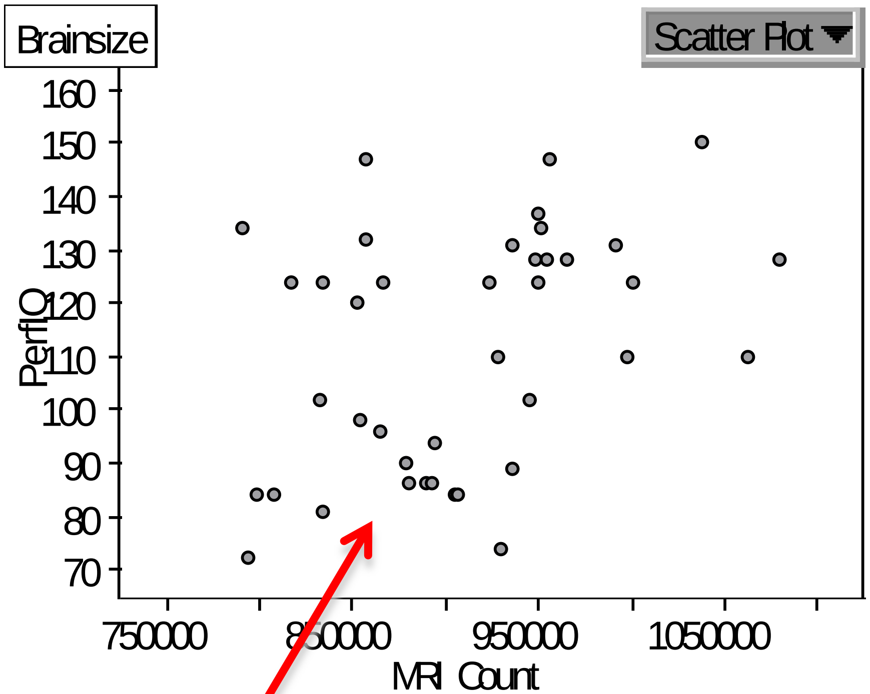




Linear Trend

Stronger linear trend

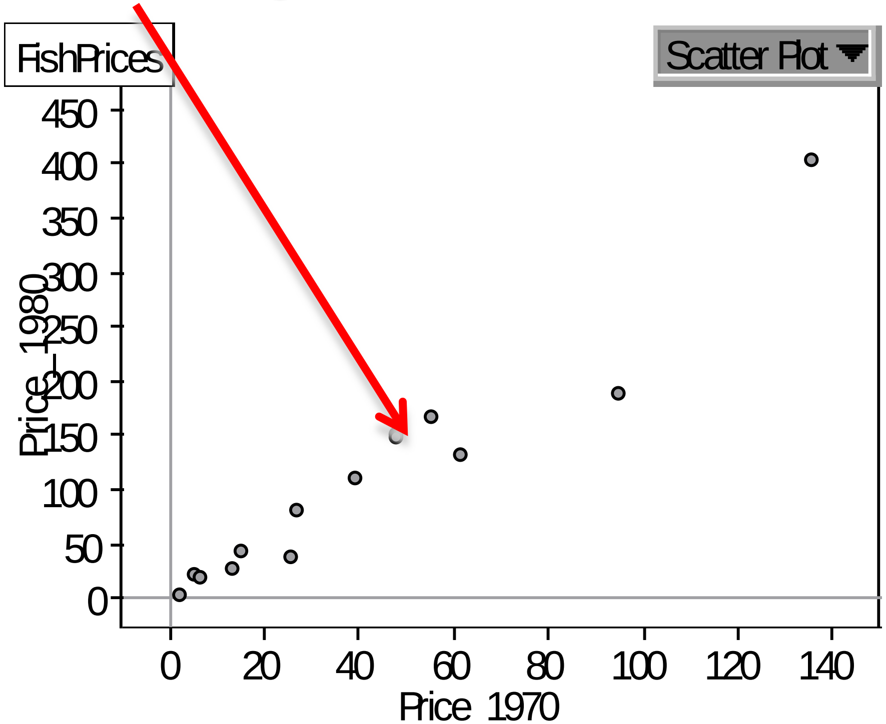


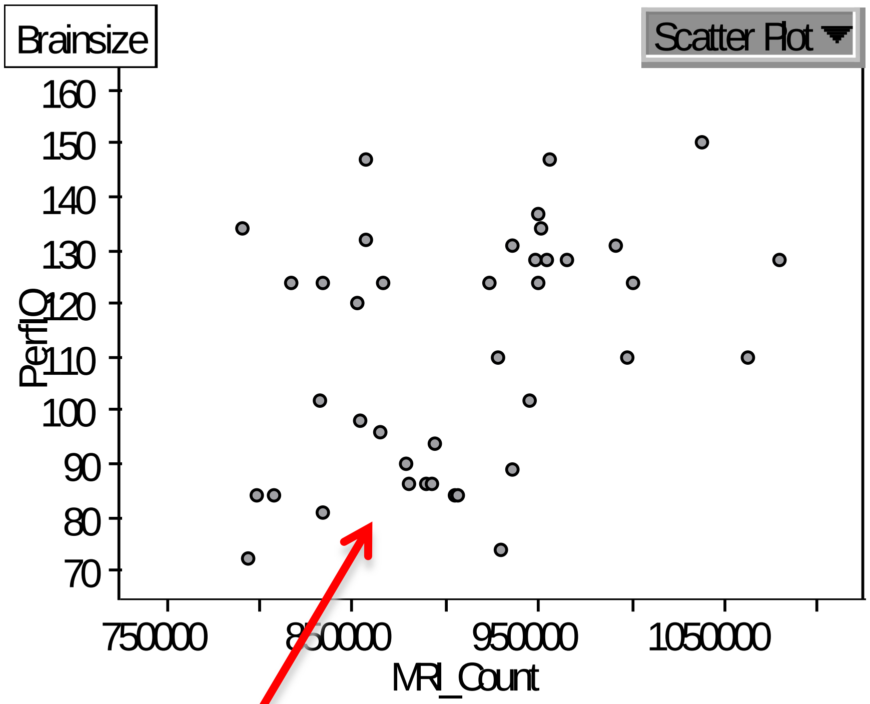


Weaker linear trend

Linear Trend

Notice that points are close together.

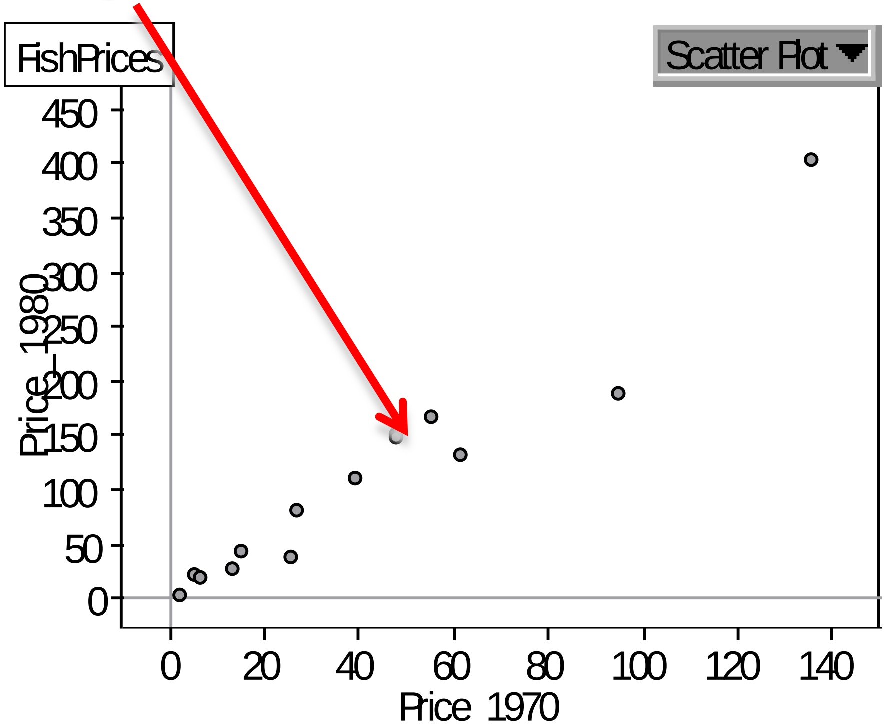


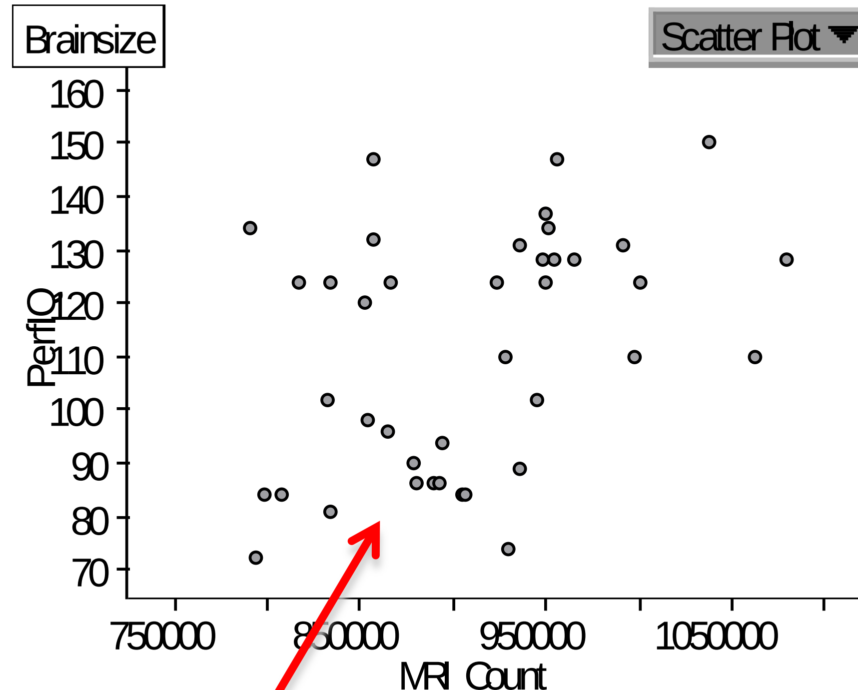


Notice that points are scattered.

Linear Trend

The points are somewhat aligned.

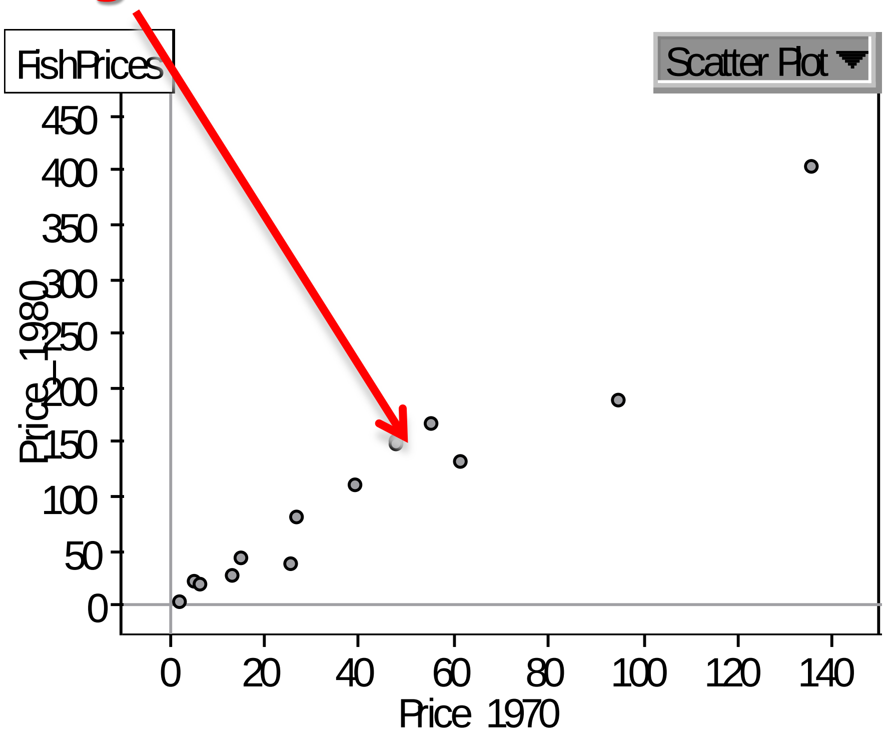


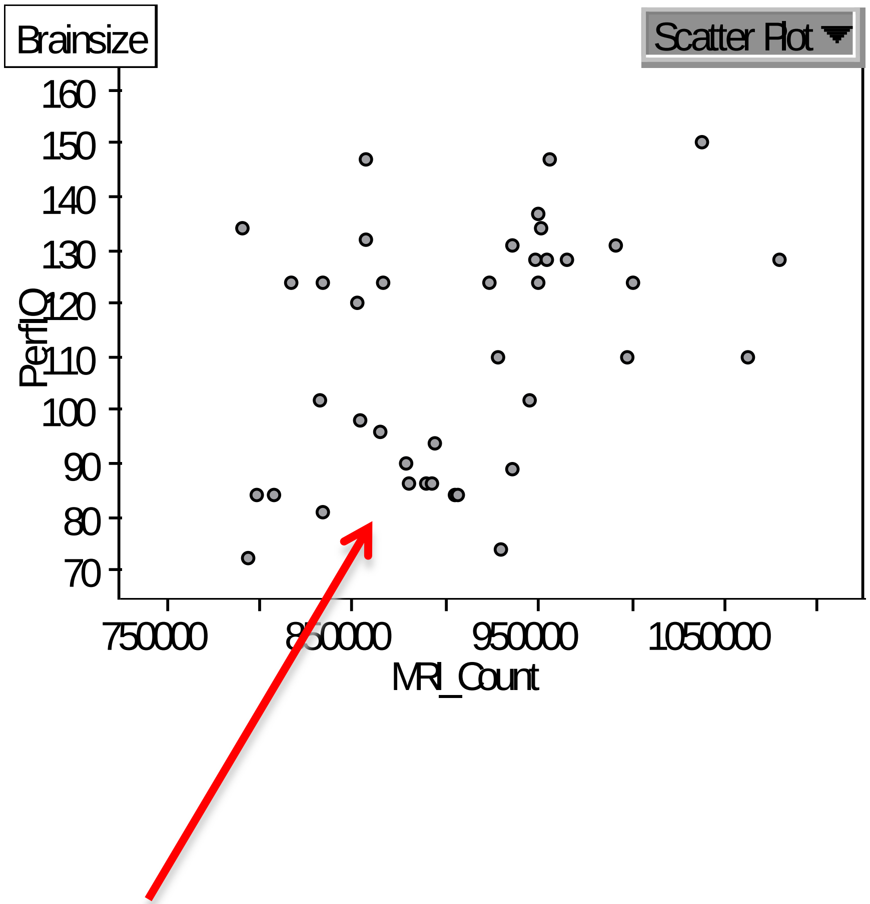


The points are scattered and cloud-like.

Linear Trend

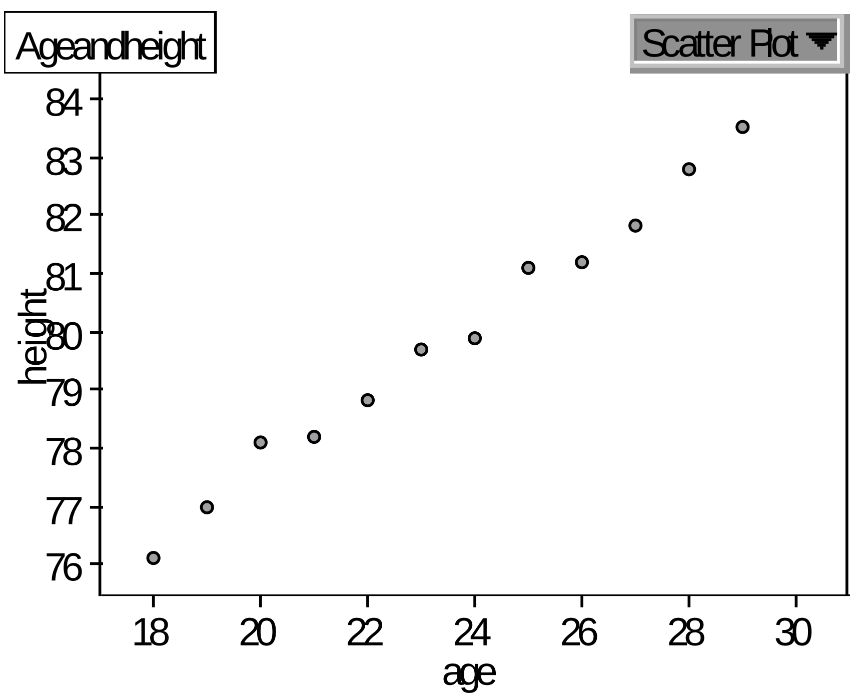
The points are somewhat aligned.

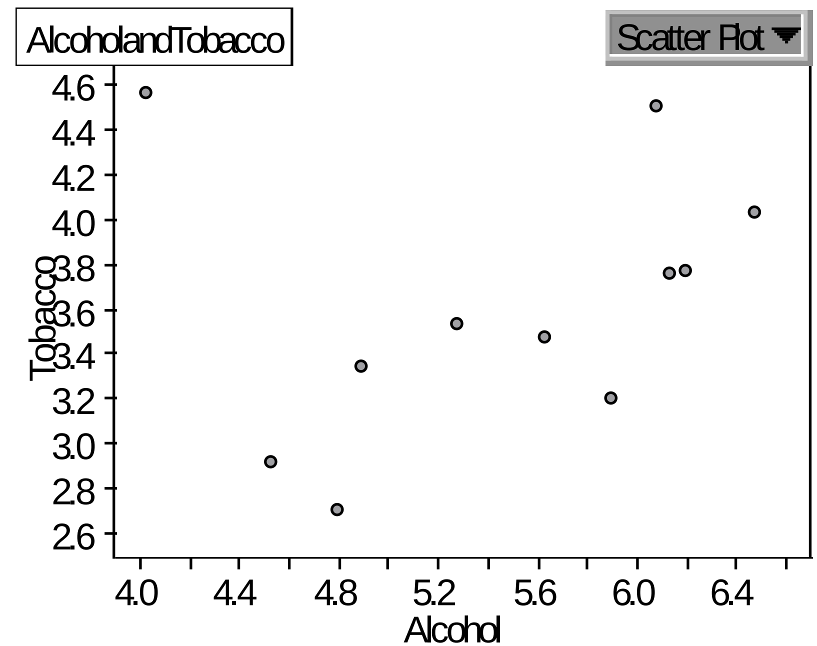




The points are not aligned.

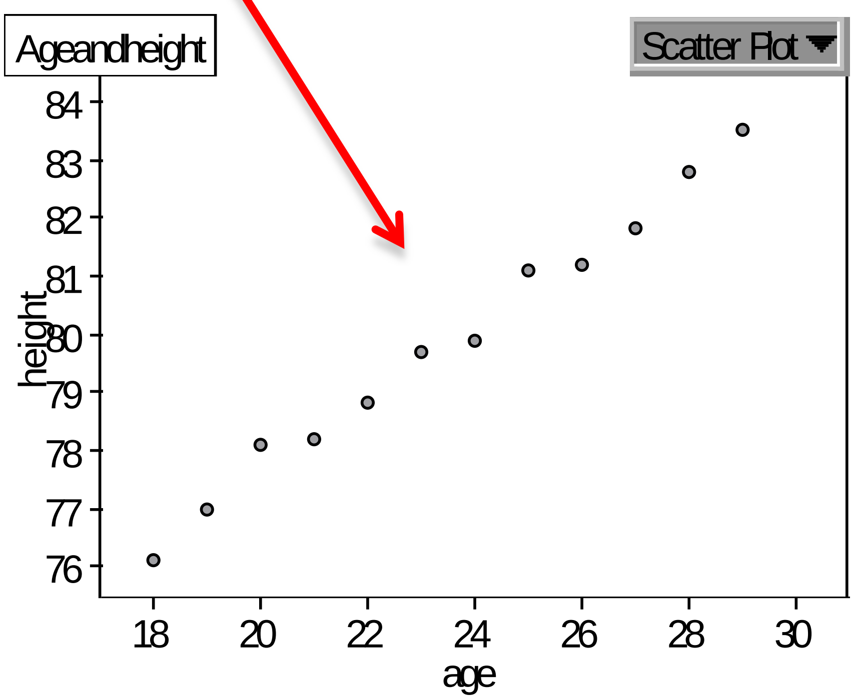
Linear Trend

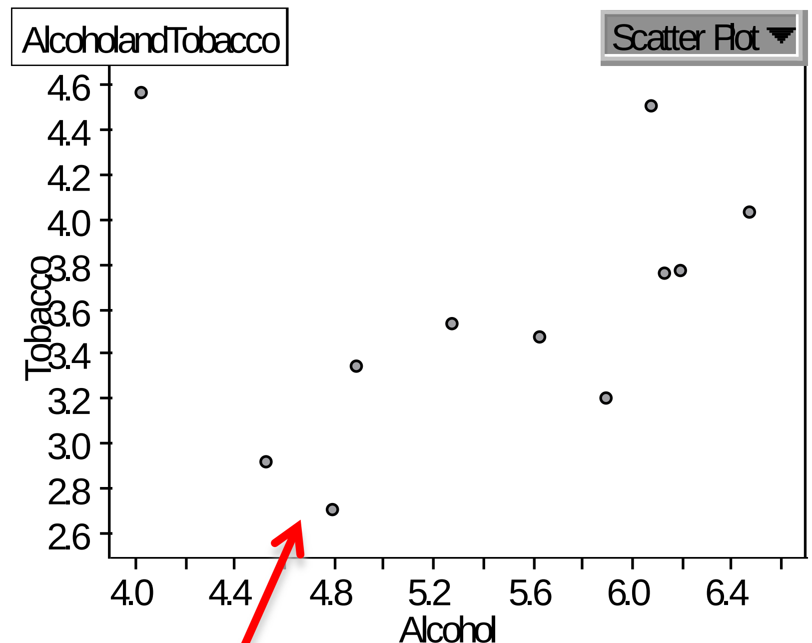




Linear Trend

Stronger linear trend

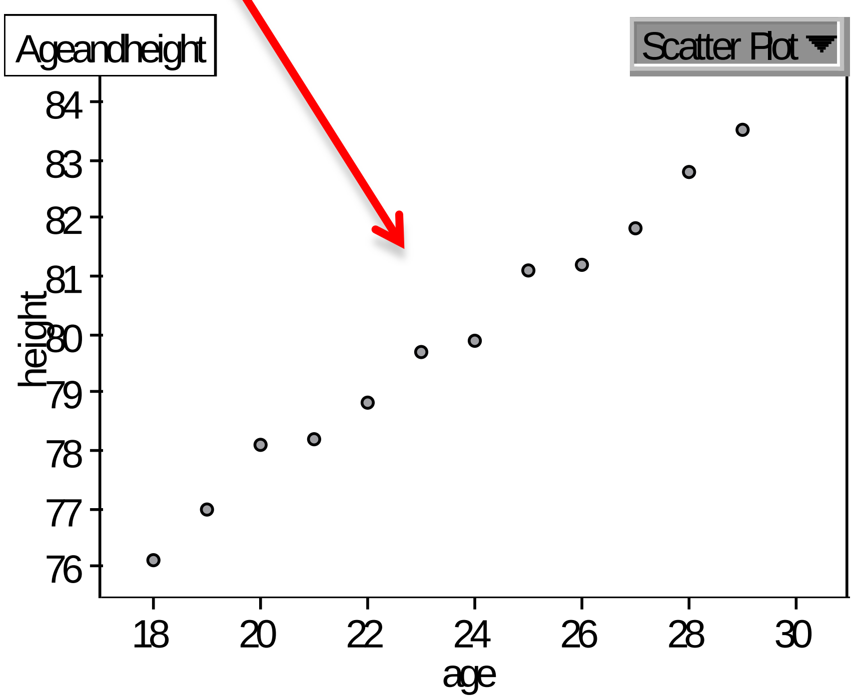


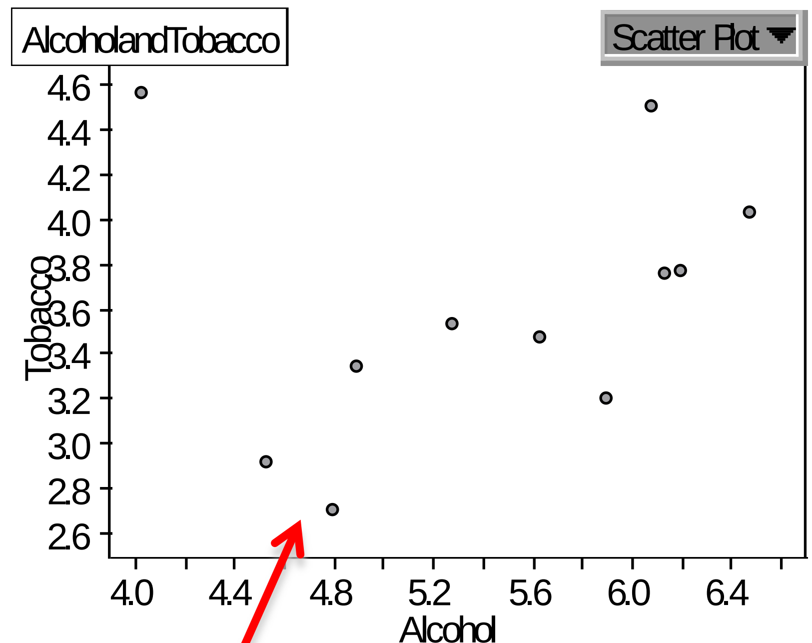


Weaker linear trend

Linear Trend

Very strong linear trend

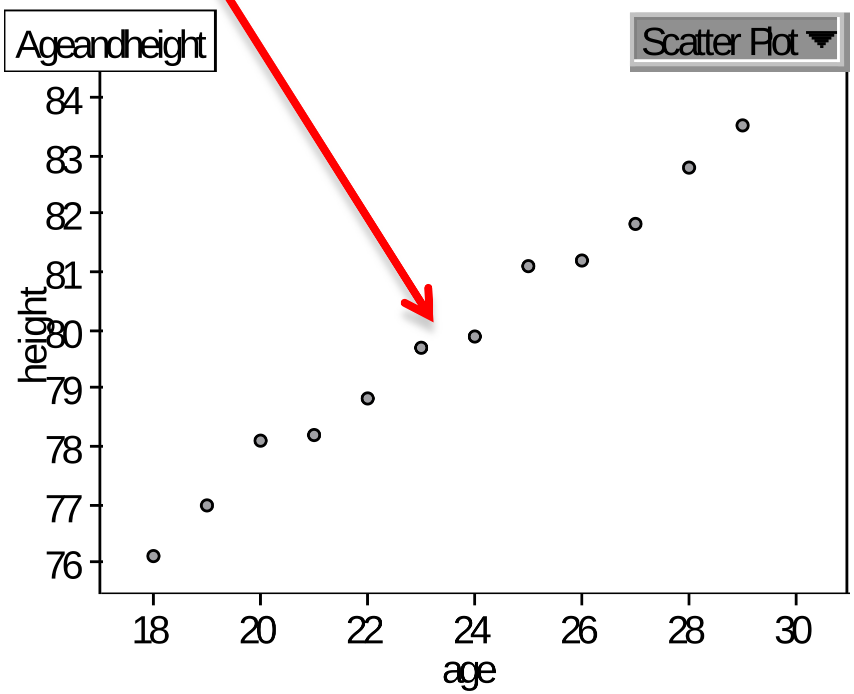


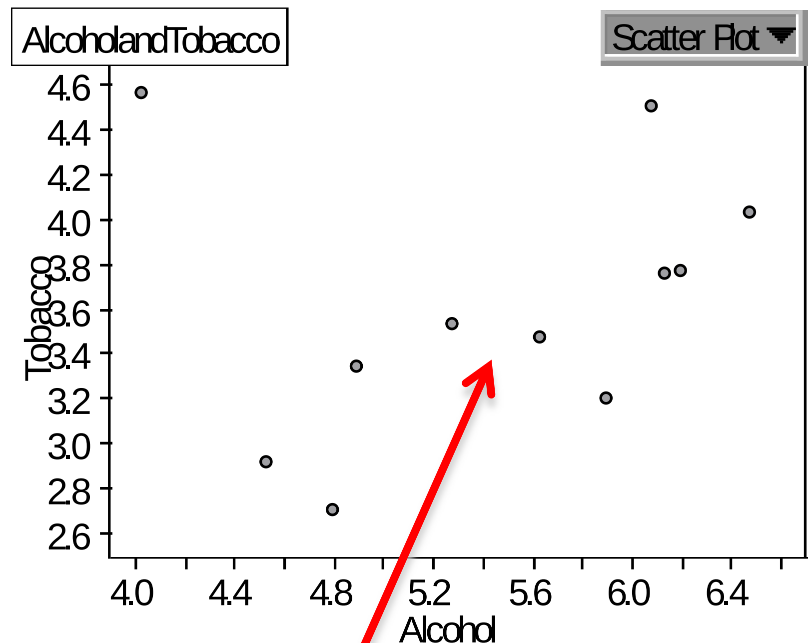


Weaker linear trend

Linear Trend

The points are somewhat aligned.

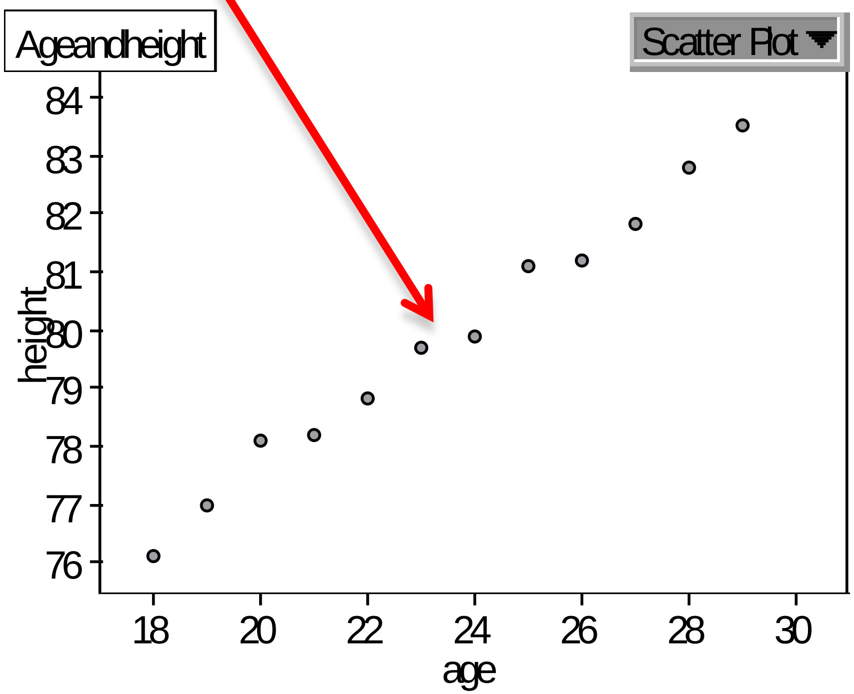


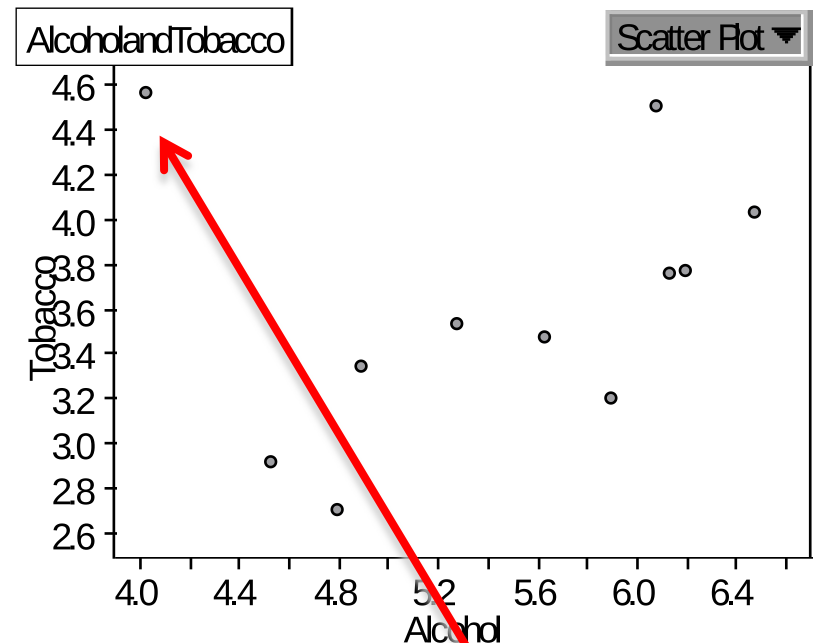


While the points are not aligned, they are grouped together …

Linear Trend

The points are somewhat aligned.





While the points are not aligned, they are grouped together

except for this point.

Trend/Association

Trend/Association

- Positive Trend/Association
  - Two variables are *positively* associated if the values of the output increase as the values of the input increase
- Negative Trend/Association
  - Two variables are *negatively* associated if the values of the output decrease as the values of the input increase

Since graphs are read from left to right, you examine the trend as the value of the input variable increases.

Trend/Association

- Positive Trend/Association
  - Two variables are *positively* associated if the values of the output increase as the values of the input increase
- Negative Trend/Association
  - Two variables are *negatively* associated if the values of the output decrease as the values of the input increase

Since graphs are read from left to right, you examine the trend as the value of the explanatory variable increases.

Trend/Association

- Positive Trend/Association
  - Two variables are *positively* associated if the values of the output increase as the values of the input increase
- Negative Trend/Association
  - Two variables are *negatively* associated if the values of the output decrease as the values of the input increase

Correlation

Correlation

- The degree to which two or more quantities are associated

Linear Correlation

- The degree to which two or more quantities are *linearly* associated

Linear Correlation Coefficient

- Also known as the *Pearson product moment correlation coefficient*

Linear Correlation Coefficient

- Also known as the *Pearson product moment correlation coefficient*
- A measurement or quantification of a linear relation between two variables
- A measure of strength of a linear relation between two variables

Linear Correlation Coefficient

- Represented by r
  - Correlation coefficient, r, is a measure of the strength of the linear relation between an explanatory variable and the response variable.

Linear Correlation Coefficient

- Properties
  - The correlation coefficient is always between –1 and 1, inclusive, that is, $-1 \le r \le 1$

Linear Correlation Coefficient

- Properties
  - *The correlation coefficient is always between –1 and 1, inclusive, that is, $-1 \leq r \leq 1$*
  - $r = 1$ is the perfect positive linear relation between the explanatory variable and the response variable.

Linear Correlation Coefficient

- Properties
  - *The correlation coefficient is always between –1 and 1, inclusive, that is, $-1 \leq r \leq 1$*
  - r = -1 is the perfect negative linear relation between the explanatory variable and the response variable

Linear Correlation Coefficient

- Properties
  - The closer r is to 1 the stronger the evidence of *positive* association between two variables
  - The closer r is to -1 the stronger the evidence of *negative* association between two variables

Linear Correlation Coefficient

- Properties
  - If r is close to 0 then there is evidence of no linear relation between the two variables
  - ***CAUTION***: r close to zero provides evidence that there is no *linear* relation, but it does *not* imply that there is no relation between the variables

Linear Correlation Coefficient

$$r=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2-\left(\sum x\right)^2}\sqrt{n\sum y^2-\left(\sum y\right)^2}}$$

Linear Correlation Coefficient

- Properties
  - The linear correlation coefficient is a unit-less measure of association

Linear Correlation Coefficient

- Properties
  - The linear correlation coefficient is a unit-less measure of association
  - The units of measure of x and y do not affect the interpretation of r

Linear Correlation Coefficient

- Properties
  - The linear correlation coefficient is a unit-less measure of association
  - The units of measure of x and y do not affect the interpretation of r
  - The closer the value of |r| is to 1, the stronger the linear relation is between the explanatory variable and the response variable

Regression Line

Regression Line

- Also known as
  - Least Squares Line
  - Least Squares Regression Line
  - Line of Best-Fit

Regression Line

- Also known as
  - Least Squares Line
  - Least Squares Regression Line
  - Line of Best-Fit

  is the line that *best-fits* the data

Regression Line

- Line for which
  - The sum of the squared errors, SSE, is a small as possible

$$SSE = \sum \left(y_k - \hat{y}_k \right)^2$$

y_k - data value $\hat{y}_k$ - predicted value

Regression Line

- Line for which
  - The sum of the squared errors, SSE, is a small as possible

$$SSE = \sum \left(y_k - \hat{y}_k \right)^2$$

y_k values of response variable

$\hat{y}_k$ values of least squares line for x_k

Least Squares Regression Line

Least Squares Regression Line

- Sum of the squared errors (SSE) is as small as possible

Least Squares Regression Line

- Sum of the squared errors (SSE) is as small as possible
  - Sum and mean of the residuals $y_k - \hat{y}_k$ are zero

Least Squares Regression Line

- Sum of the squared errors (SSE) is as small as possible
  - Sum and mean of the residuals $y_k - \hat{y}_k$ are zero
  - Variation in the residuals is as small as possible

Least Squares Regression Line

- Sum of the squared errors (SSE) is as small as possible
  - Sum and mean of the residuals $y_k - \hat{y}_k$ are zero
  - Variation in the residuals is as small as possible
  - The line contains the point of averages $\left(\bar{x}, \bar{y}\right)$

Least Squares Regression Line

- For the least squares regression line $\hat{y} = mx + b$,
  - Slope: $$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$
  - Units on slope: $\dfrac{y - \text{units}}{x - \text{units}}$

Least Squares Regression Line

- For the least squares regression line $\hat{y} = mx + b$,
  - Slope: $$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$
  - Y-coordinate of y-intercept: $$b = \frac{1}{n}\left(\sum y - m\sum x\right)$$

Least Squares Regression Line

- For the least squares regression line $\hat{y} = mx + b$,
  - Slope: $$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$
  - Y-coordinate of y-intercept: $$b = \frac{\sum y}{n} - m\left(\frac{\sum x}{n}\right)$$

Least Squares Regression Line

- For the least squares regression line $\hat{y} = mx + b$,
  - Slope: $$m = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$
  - Y-coordinate of y-intercept: $$b = \overline{y} - m\overline{x}$$