# Correlation

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

December 6, 2011



## The Big Picture

- We have just completed a lengthy series of lectures on ANOVA where we considered models with a *normally distributed response variable* where the mean was determined by *one or more categorical explanatory variables*.
- We next consider *simple linear regression models* where there is again a normally distributed response variable, but where the means are determined by *a linear function of a single quantitative explanatory variable*.
- Before developing ideas about regression, we need to explore *scatter plots* to display two quantitative variables and *correlation* to numerically quantify the linear relationship between two quantitative variables.

## Data

- We will consider data sets of two quantitative random variables such as in this example.
- age is the age of a male lion in years;
- `proportion.black` is the proportion of a lion's nose that is black.

```
age  proportion.black
1.1  0.21
1.5  0.14
1.9  0.11
2.2  0.13
2.6  0.12
3.2  0.13
3.2  0.12
...
```



## Scatter Plots

- A *scatter plot* is a graph that displays two quantitative variables.
- Each observation is a single point.
- One variable we will call $X$ is plotted on the horizontal axis.
- A second variable we call $Y$ is plotted on the vertical axis.
- If we plan to use a model where one variable is a response variable that depends on an explanatory variable, $X$ should be the explanatory variable and $Y$ should be the response.

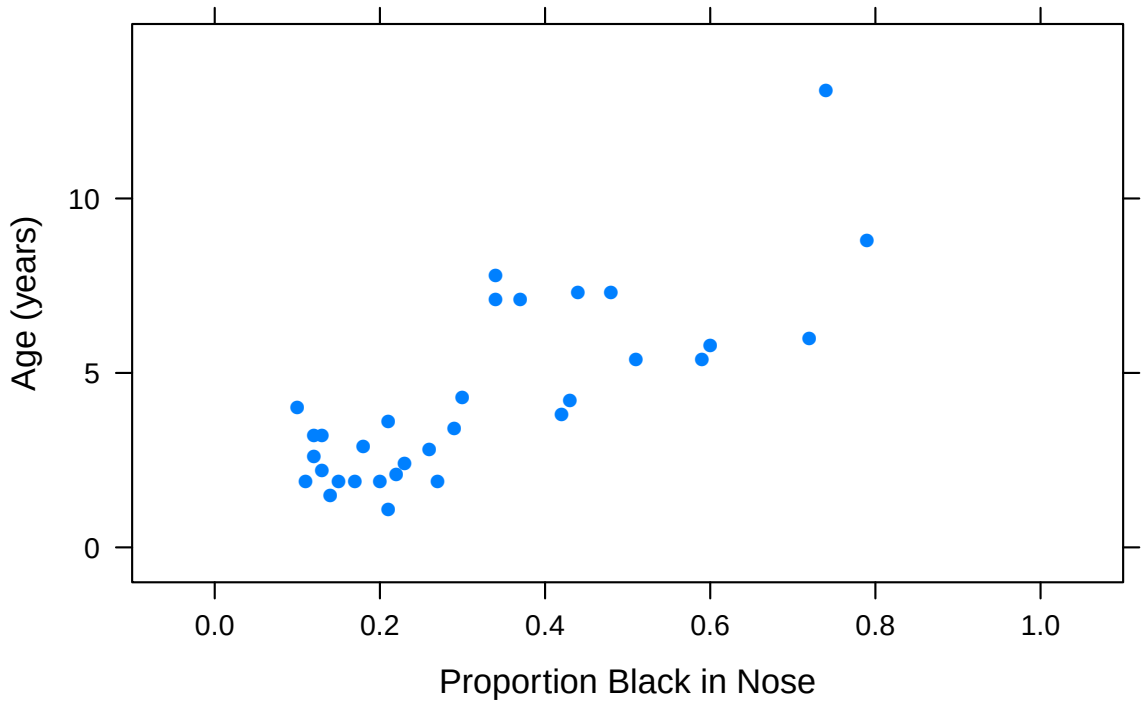
# Lion Example

## Case Study

- The noses of male lions get blacker as they age.
- This is potentially useful as a means to estimate the age of a lion with unknown age.
- (How one measures the blackness of a lion's nose in the wild without getting eaten is an excellent question!)
- We display a scatter plot of age versus blackness of 32 lions of known age (Example 17.1 from the textbook).
- The choice of $X$ and $Y$ is for the desired purpose of estimating age from observable blackness.
- It is also reasonable to consider blackness as a response to age.



# Lion Data Scatter Plot



## Observations

- We see that *age and blackness in the nose are positively associated*.
- As one variable increases, the other also tends to increase.
- However, there is not a perfect relationship between the two variables, as the points do not fall exactly on a straight line or simple curve.
- We can imagine other scatter plots where points may be more or less tightly clustered around a line (or other curve).
- Statisticians have invented a statistic called the *correlation coefficient* to quantify the strength of a linear relationship.
- Before developing simple linear regression models, we will develop our understanding of correlation.



## Correlation

### Definition

The *correlation coefficient* $r$ is measure of the strength of the linear relationship between two variables.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

$$= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

- Notice that the correlation is not affected by linear transformations of the data (such as changing the scale of measurement) as each variable is standardized by substracting the mean and dividing by the standard deviation.

## Observations about the Formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right)$$

- Notice that observations where $X_i$ and $Y_i$ are either both greater or both less than their respective means will contribute positive values to the sum, but observations where one is positive and the other is negative will contribute negative values to the sum.
- Hence, the correlation coeficient can be positive or negative.
- Its value will be positive if there is a stronger trend for $X$ and $Y$ to vary together in the same direction (high with high, low with low) than vice versa.



## Additional Observations about the Formula

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \Big/ (n-1)}{s_X \times s_Y}$$

- If $X = Y$, the numerator would be *the sample variance*.
- When $X$ and $Y$ are different variables, the expression in the numerator is called the *sample covariance*.
- The covariance has units of the product of the units of $X$ and $Y$.
- Dividing the covariance by the two standard deviations makes the correlation coeficient *unitless*.
- Also note that if $X = Y$, then the numerator and the denominator would both be equal to the variance of $X$, and $r$ would be 1.
- This suggests that $r = 1$ is the largest possible value (which is true, but requires more advanced mathematics than we assume to prove).
- If $Y = -X$, then $r = -1$, and this is the smallest that $r$ can be.

## Scolding the Textbook Authors

- Also note that throughout Chapters 16 and 17, the textbook authors fail to include subscripts with their summation equations, which is inexcusable. For example,

$$\sum (X - \bar{X})^2$$

should be

$$\sum_{i=1}^{n} (X_i - \bar{X})^2$$

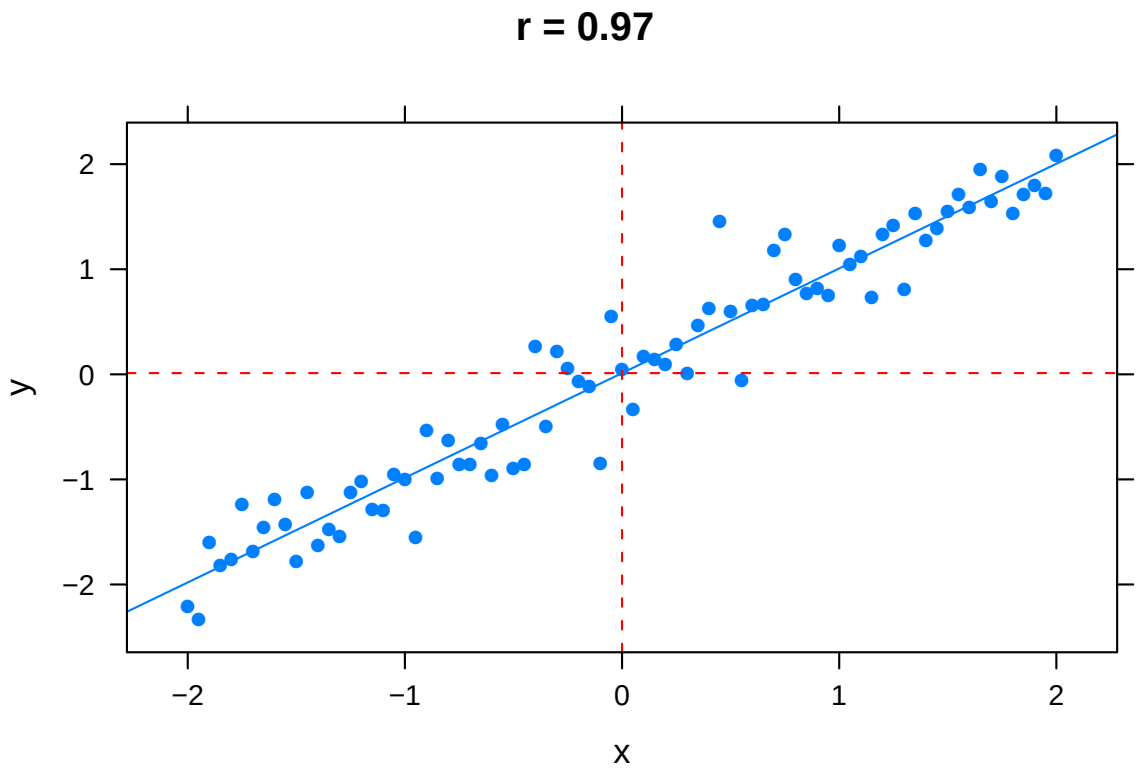
so that you, the reader, knows that values of $X_i$ change (potentially) from term to term, but $\bar{X}$ is constant.



## Scatter plots

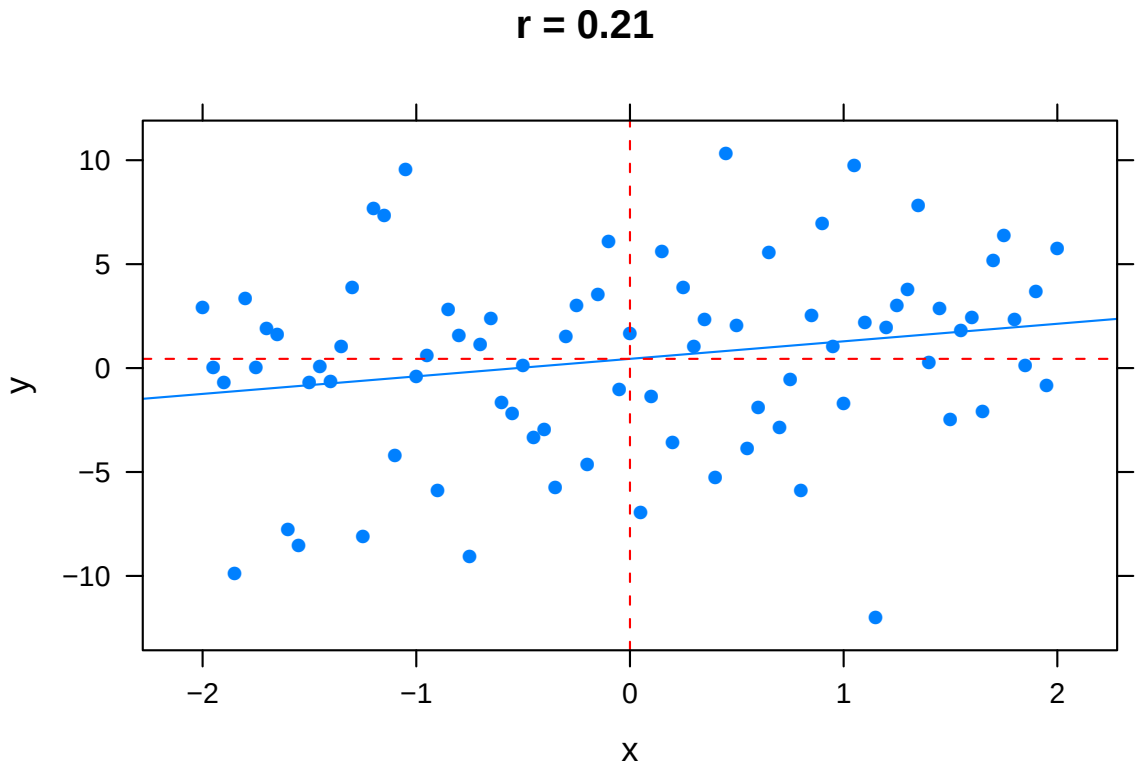
- The next several graphs will show scatter plots of sample data with different correlation coefficients so that you can begin to develop an intuition for the meaning of the numerical values.
- In addition, note that very different graphs can have the same numerical correlation.
- It is important to look at graphs and not only the value of $r$!

# Correlation Plots

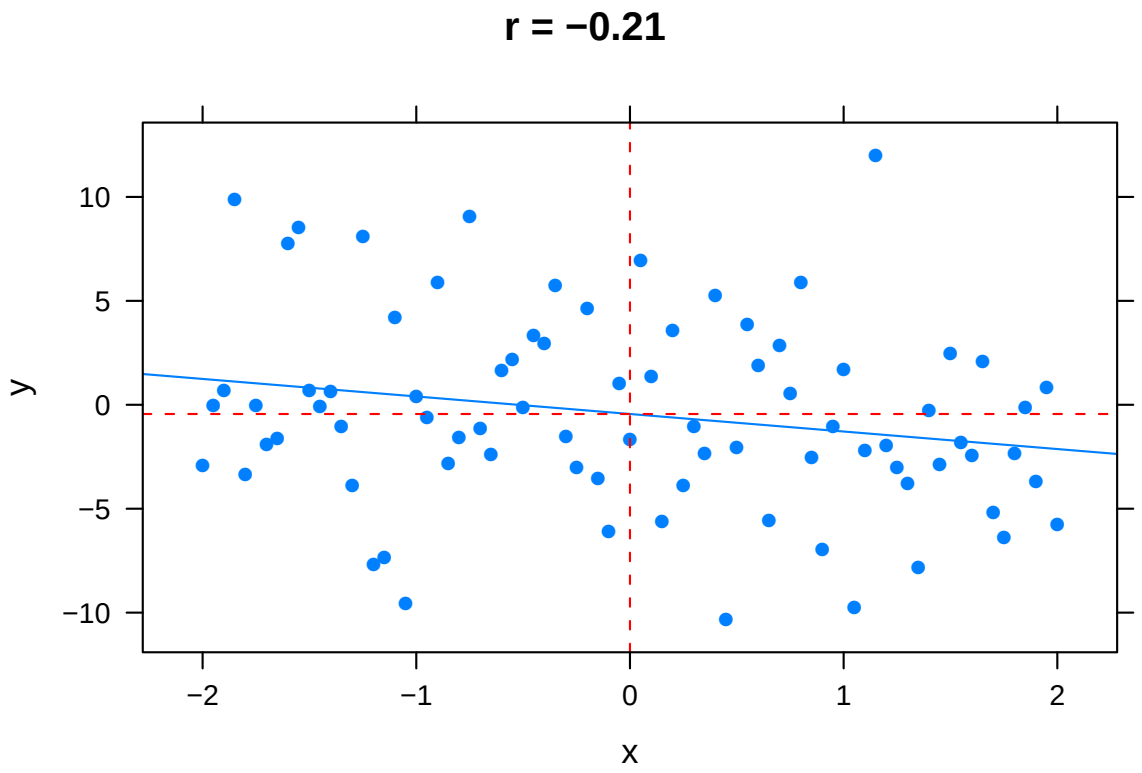





# Correlation Plots

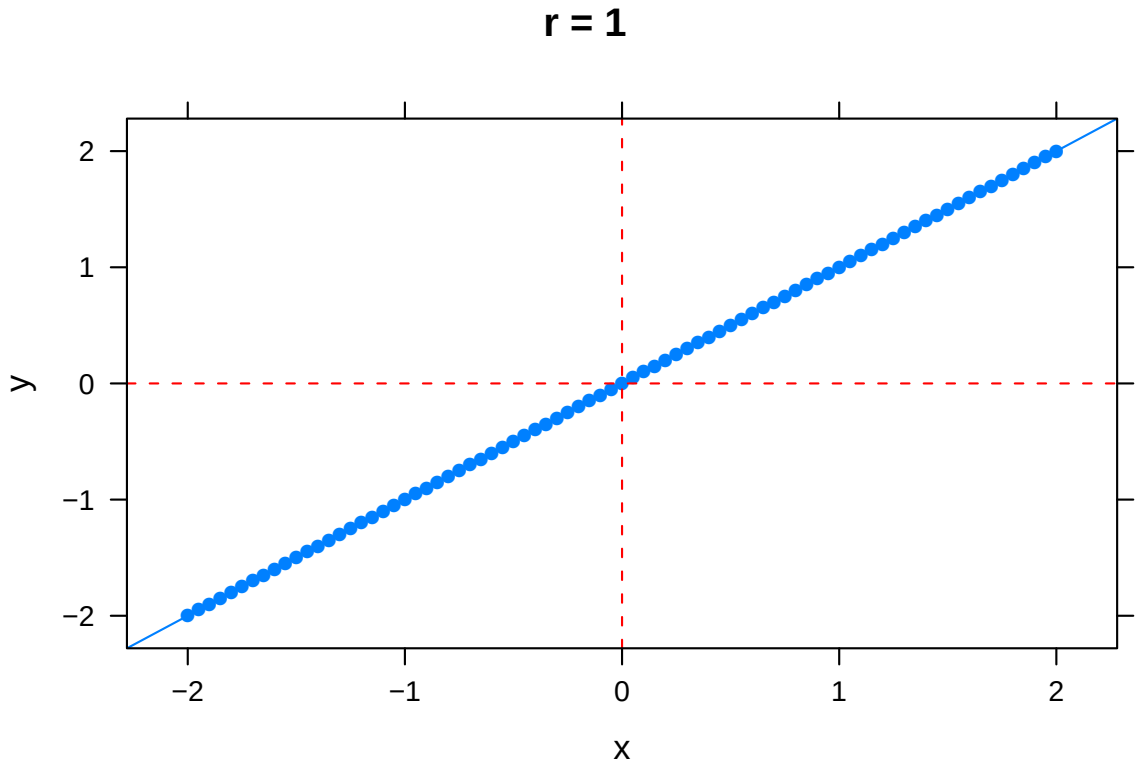


# Correlation Plots

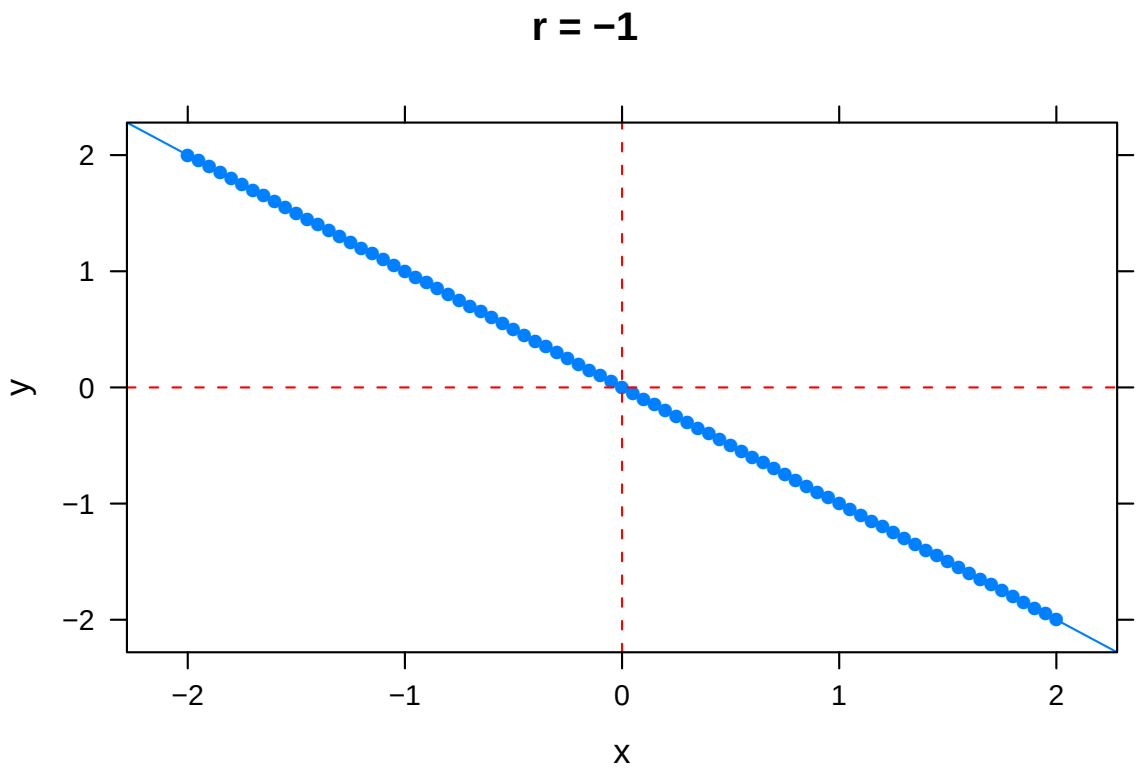





# Correlation Plots

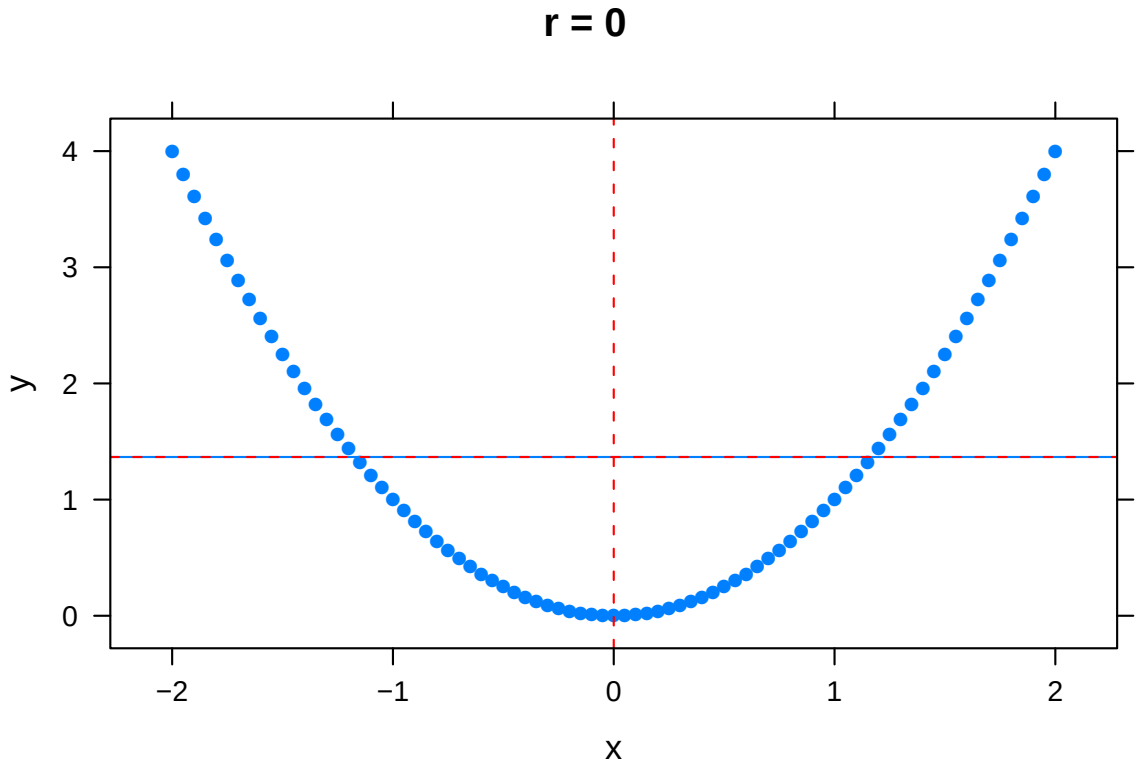


# Correlation Plots






# Correlation Plots

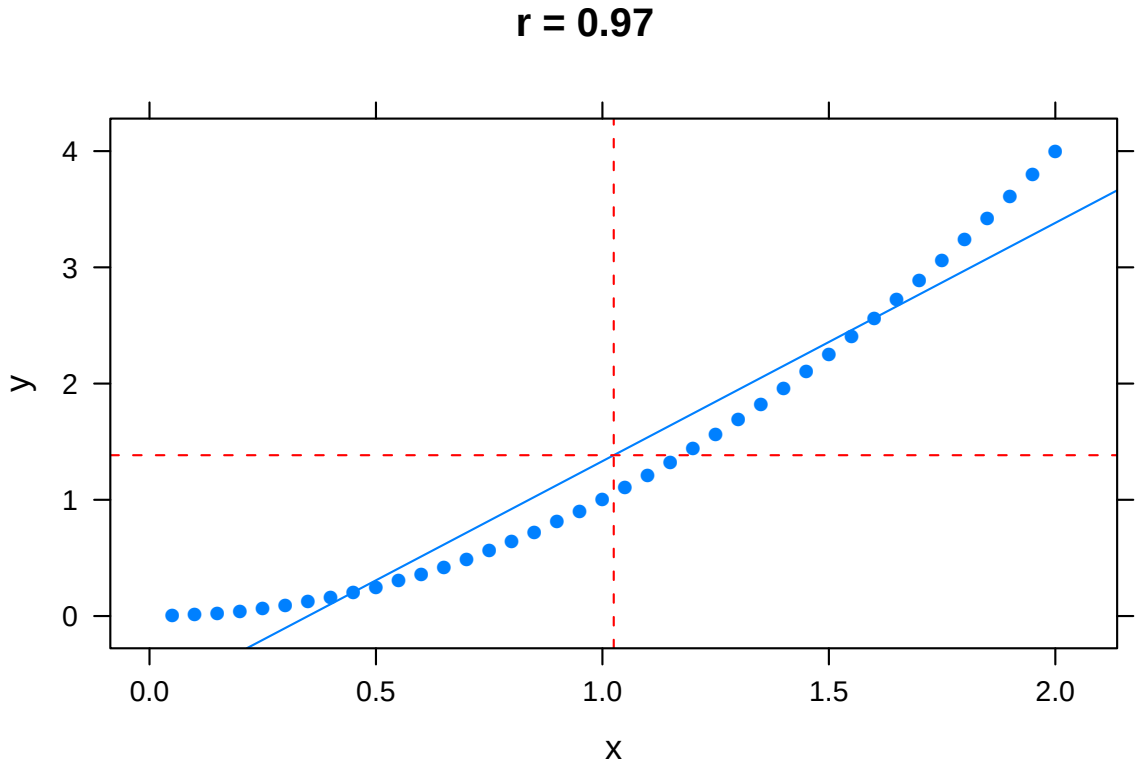


## Comment

- Notice that when $r = 0$, this means that there is no strong *linear* relationship between $X$ and $Y$, but there could be a strong *nonlinear relationship* between $X$ and $Y$.
- *You need to plot the data and not just calculate r!*
- $r$ will be close to zero whenever the *sum of squared residuals around the best fitting straight line* is about the same as the *sum of squared residuals around the best fitting horizontal line*.



## Correlation Plots

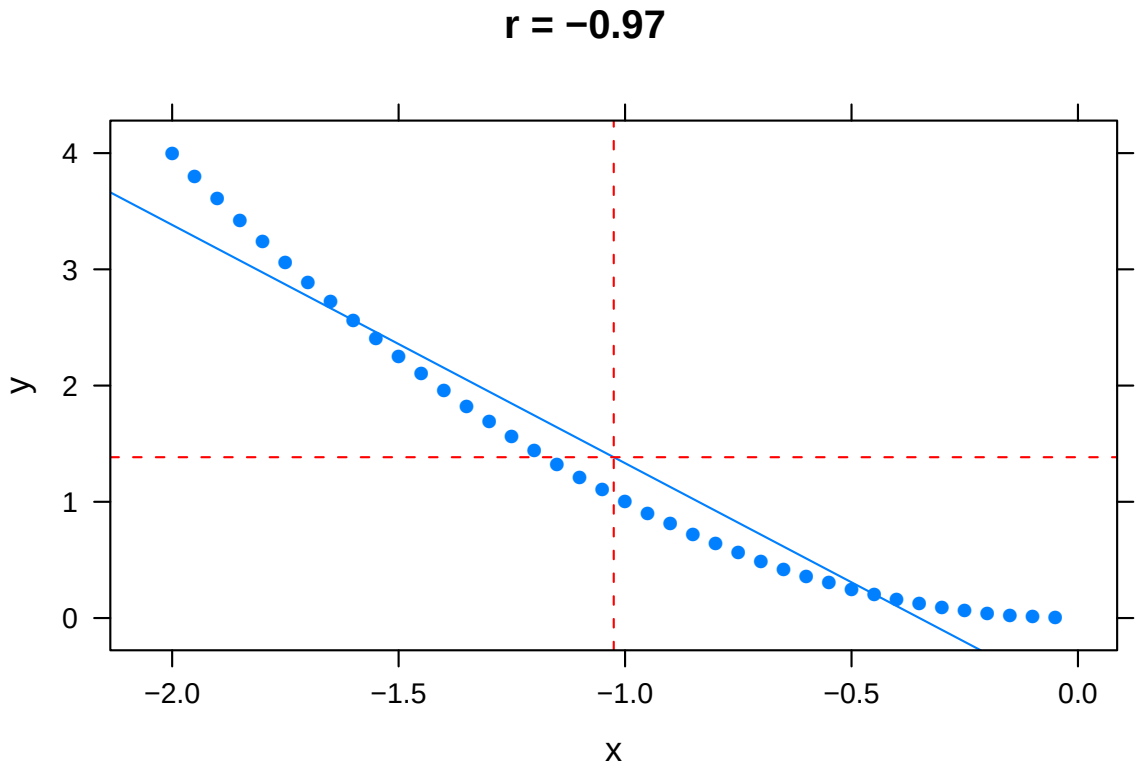


## Comment

- Notice that when $r$ is close to 1 (but not exactly one), there is a strong *linear* relationship between $X$ and $Y$ with a positive slope.
- However, there could be a *stronger nonlinear relationship* between $X$ and $Y$.
- *You need to plot the data and not just calculate r!*
- $r$ will be close to one whenever the *sum of squared residuals around the best fitting straight line with a positive slope* is much smaller than *sum of squared residuals around the best fitting horizontal line*.



## Correlation Plots



## Comment

- Notice that when $r$ is close to (but not exactly) $-1$, there is a strong *linear* relationship between $X$ and $Y$ with a negative slope.
- However, there could be a *stronger nonlinear relationship* between $X$ and $Y$.
- *You need to plot the data and not just calculate r!*
- $r$ will be close to $-1$ whenever the *sum of squared residuals around the best fitting straight line with a negtive slope* is much smaller than *sum of squared residuals around the best fitting horizontal line*.



## Summary of Correlation

- The correlation coefficient $r$ measures the strength of the linear relationship between two quantitative variables, on a scale from $-1$ to 1.
- The correlation coefficient is $-1$ or 1 only when the data lies perfectly on a line with negative or positive slope, respectively.
- If the correlation coefficient is near one, this means that *the data is tightly clustered around a line with a positive slope*.
- Correlation coefficients near 0 indicate weak linear relationships.
- However, $r$ does not measure the strength of nonlinear relationships.
- If $r = 0$, rather than $X$ and $Y$ being unrelated, it can be the case that they have a strong *nonlinear* relationshsip.
- If $|r|$ is close to 1, *it may still be the case that a nonlinear relationship is a better description of the data than a linear relationship*.

## A final thought

In case you missed a major theme of this lecture . . .

*Always plot your data!*



## A final thought

In case you missed a major theme of this lecture . . .

*Always plot your data!*