Simple Linear Regression

CSU Chico, Math 314

2018-11-26

outline

- Recap
- Correlation
  - definition
  - examples
- Simple Linear Regression
  - lite example
  - assumptions
  - parameter
  - residuals
  - estimation
- Example
- References

outline

Recap

Correlation
- definition
- examples

Simple Linear Regression
- lite example
- assumptions
- parameter
- residuals
- estimation

Example

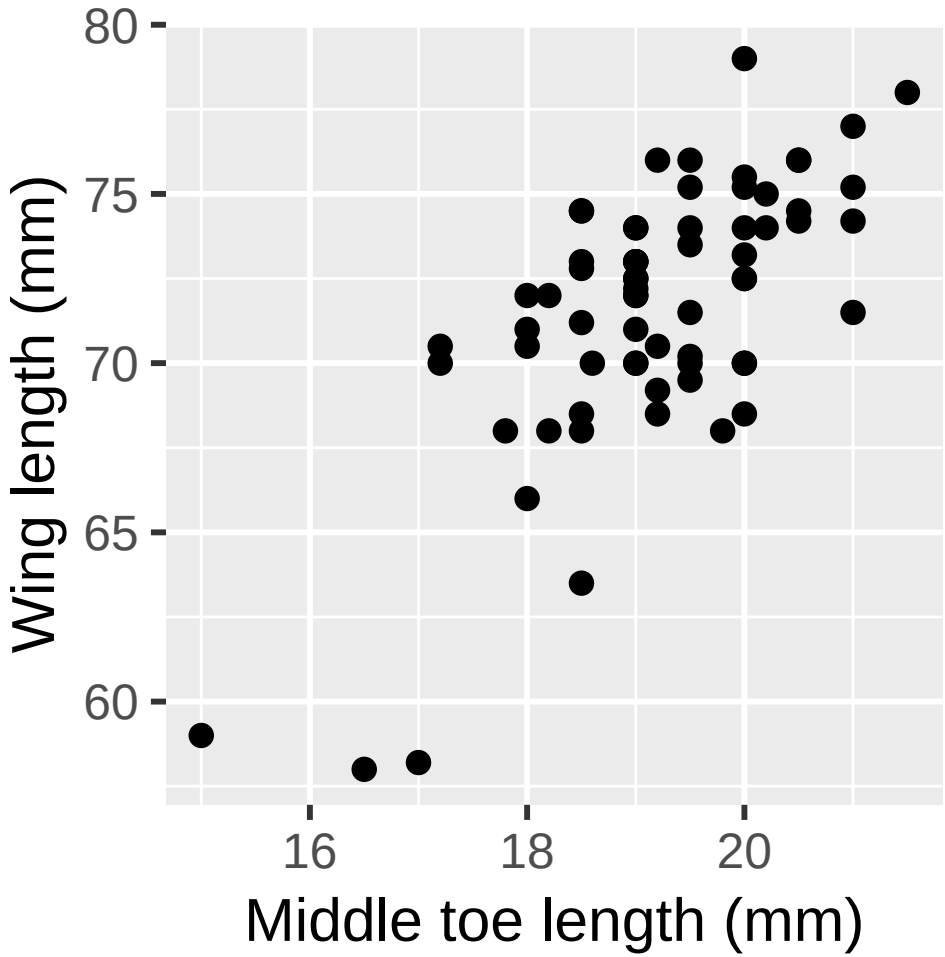
References

Recap, Take 1: ANOVA

ANOVA breaks up means of response variable Y by levels of one categorical variable.

Recap, Take 2: Scatterplots

Scatterplots are a graphical description of two numerical variables; consider Darwin's finch data



Recap, Take 2: Scatterplots

We used some keywords to describe scatterplots.

- associated or not.
- direction: possitive or negative association.
- structure: linear or nonlinear.

outline

Recap

Correlation
- definition
- examples

Simple Linear Regression
- lite example
- assumptions
- parameter
- residuals
- estimation

Example

References

Correlation, definition

Correlation is denoted by R and is the numeric analogue of the words above.

correlation

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables.

Correlation, notes

Notes on correlation

- More accurate name is Pearson correlation coefficient.
- Describes linear relationships only.
- Bounded by -1 and 1.
- The value 0 denotes no association.
- The sign dictates directionality.

Correlation, math

Mathematically, (Pearson) correlation is defined as

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

In R we should just use the function `cor`.

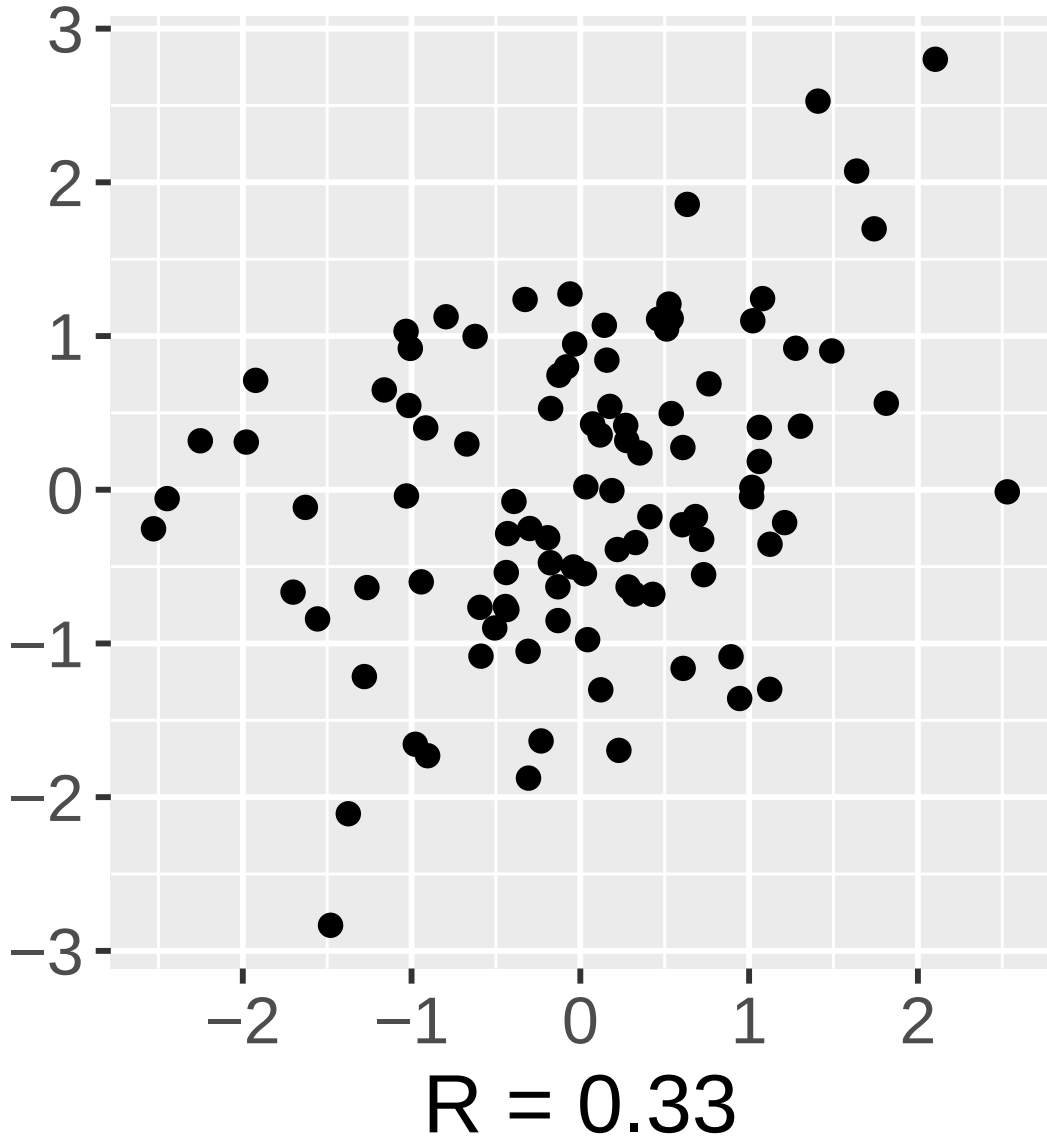
```
with(finch,
     cor(middletoelength, winglength))

## [1] 0.7034241
```

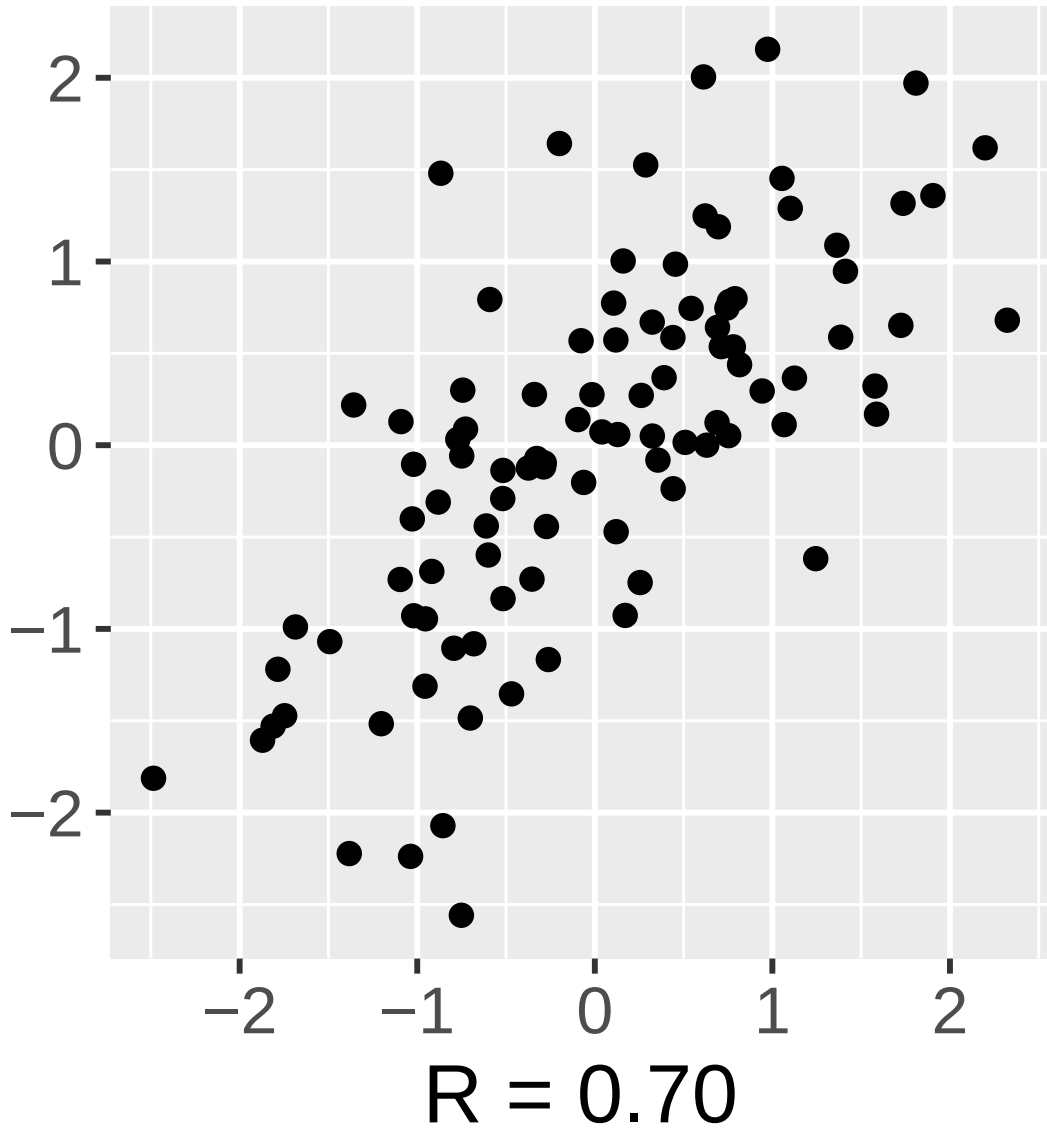
Correlation, there are others

Pearson correlation is the most common. That is to say our default assumption is linear relationships. If there is convincing evidence otherwise, then use Spearman's rho.

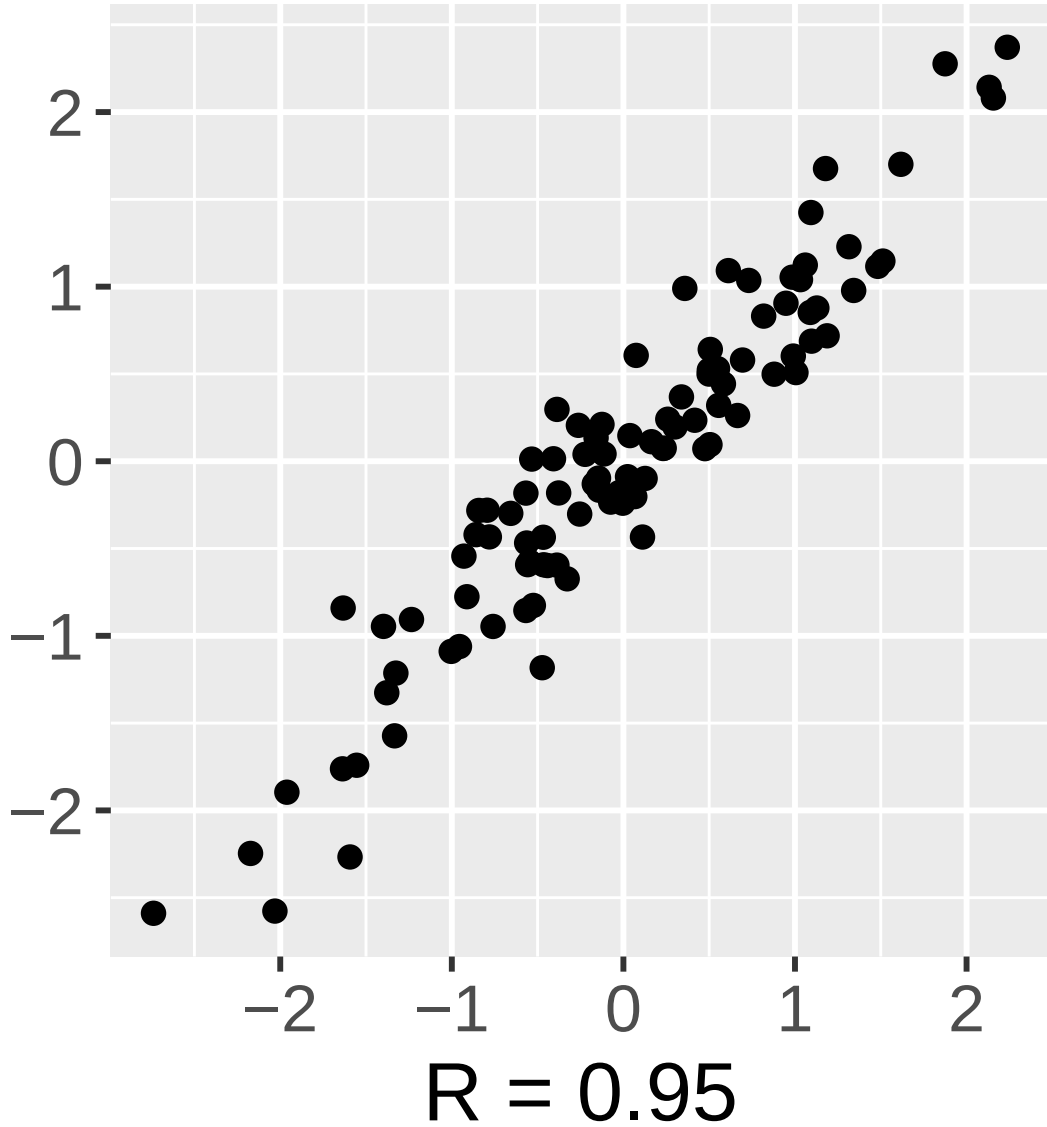
Correlation, example



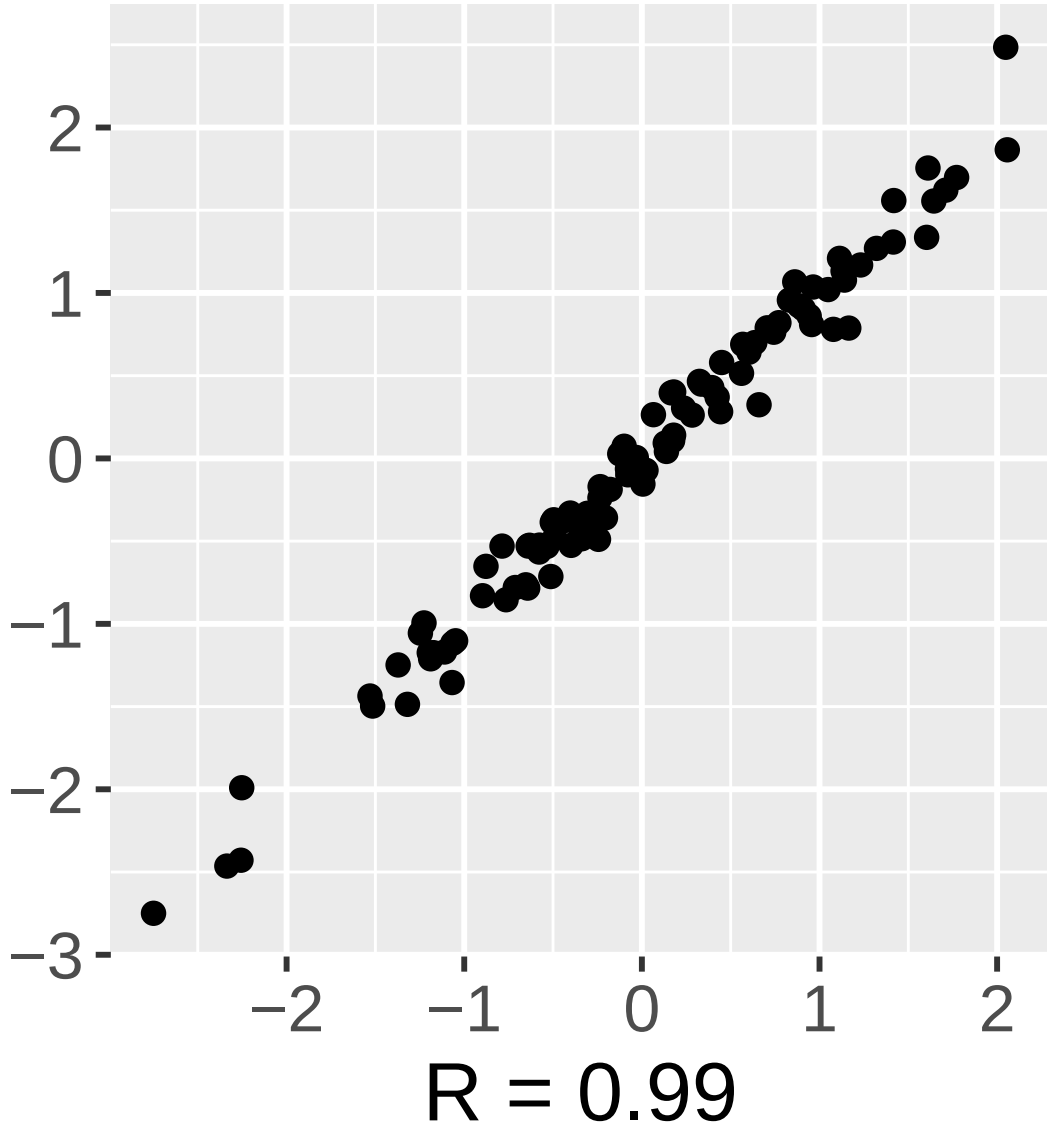
Correlation, example



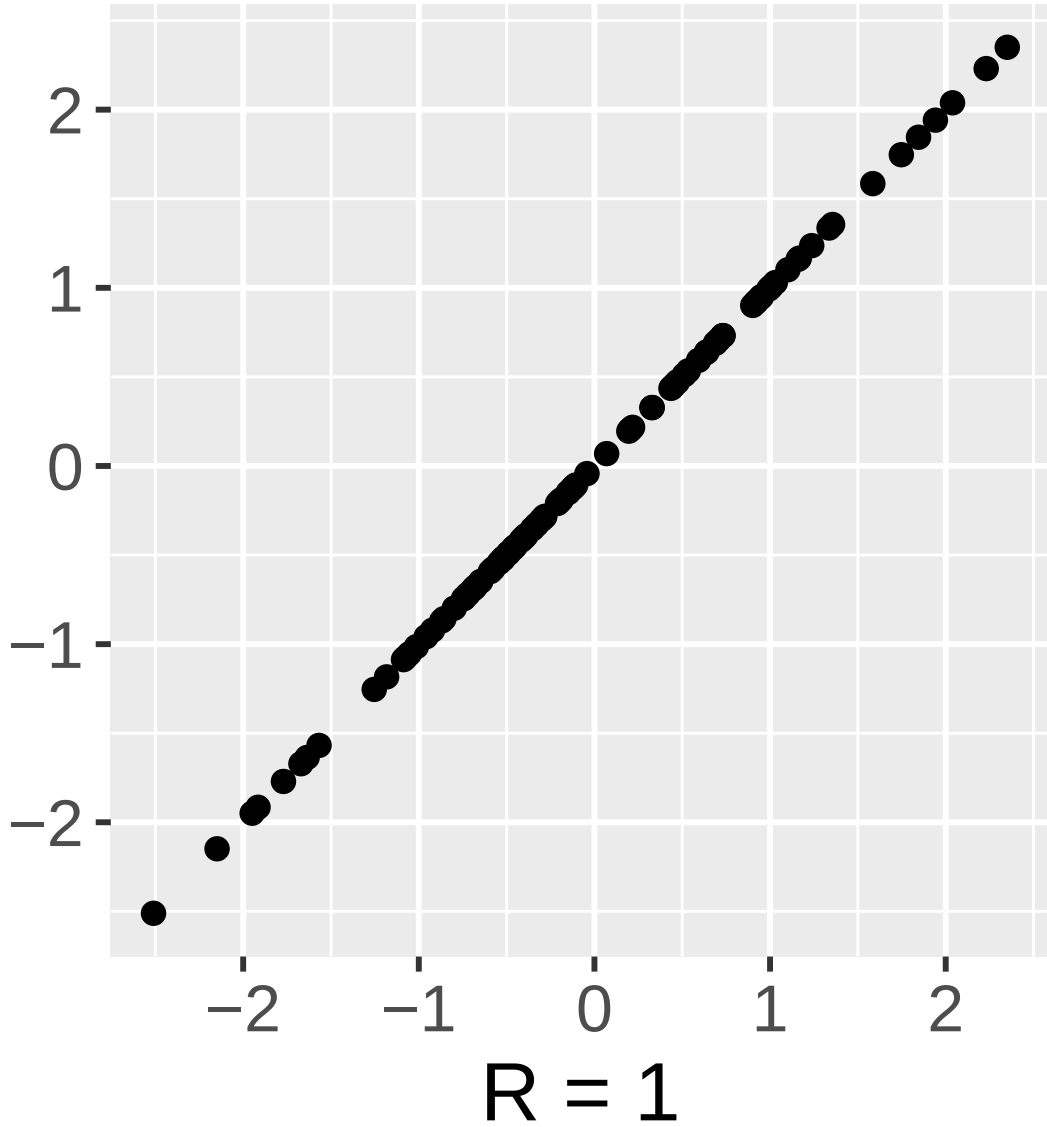
Correlation, example



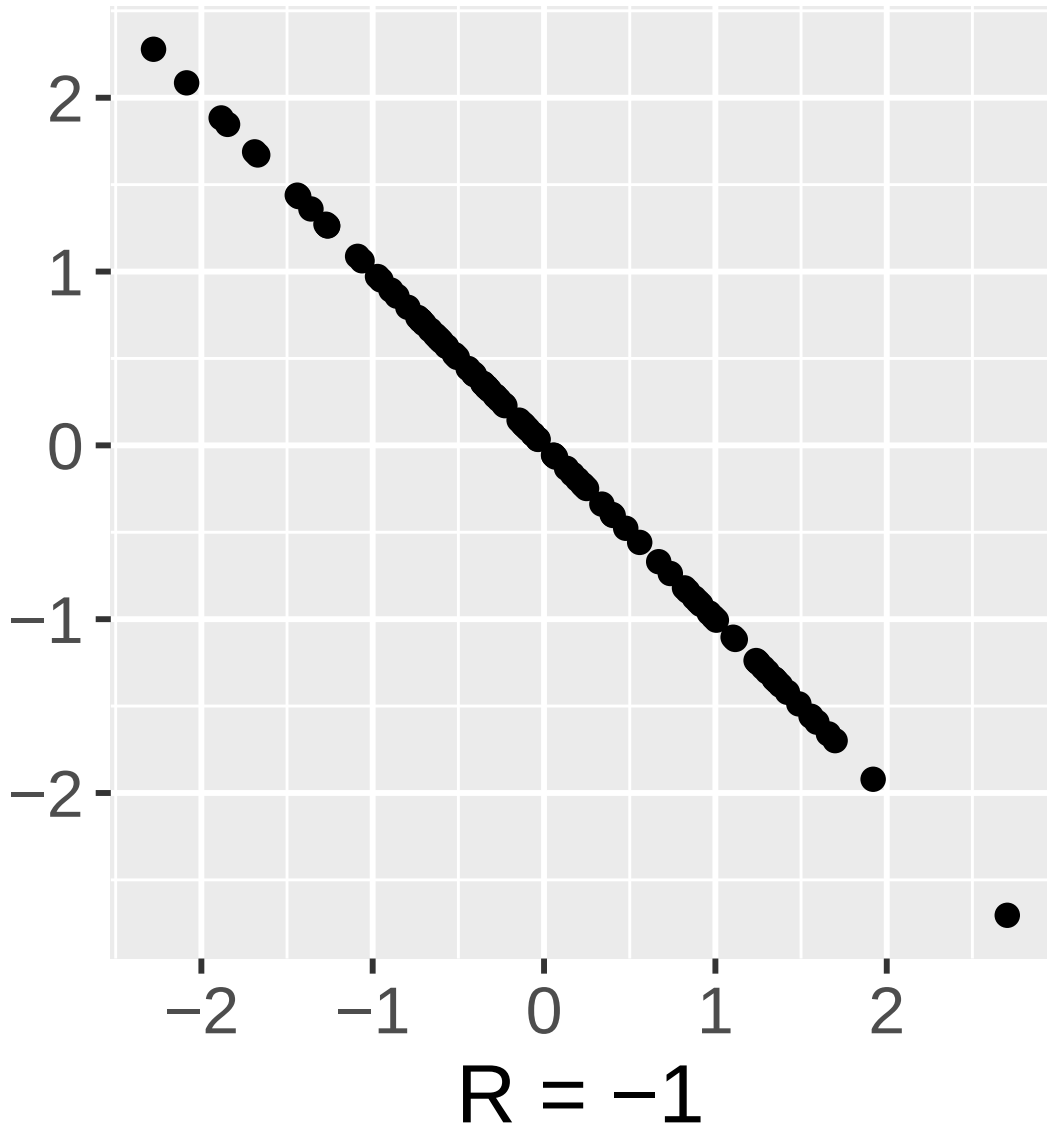
Correlation, example



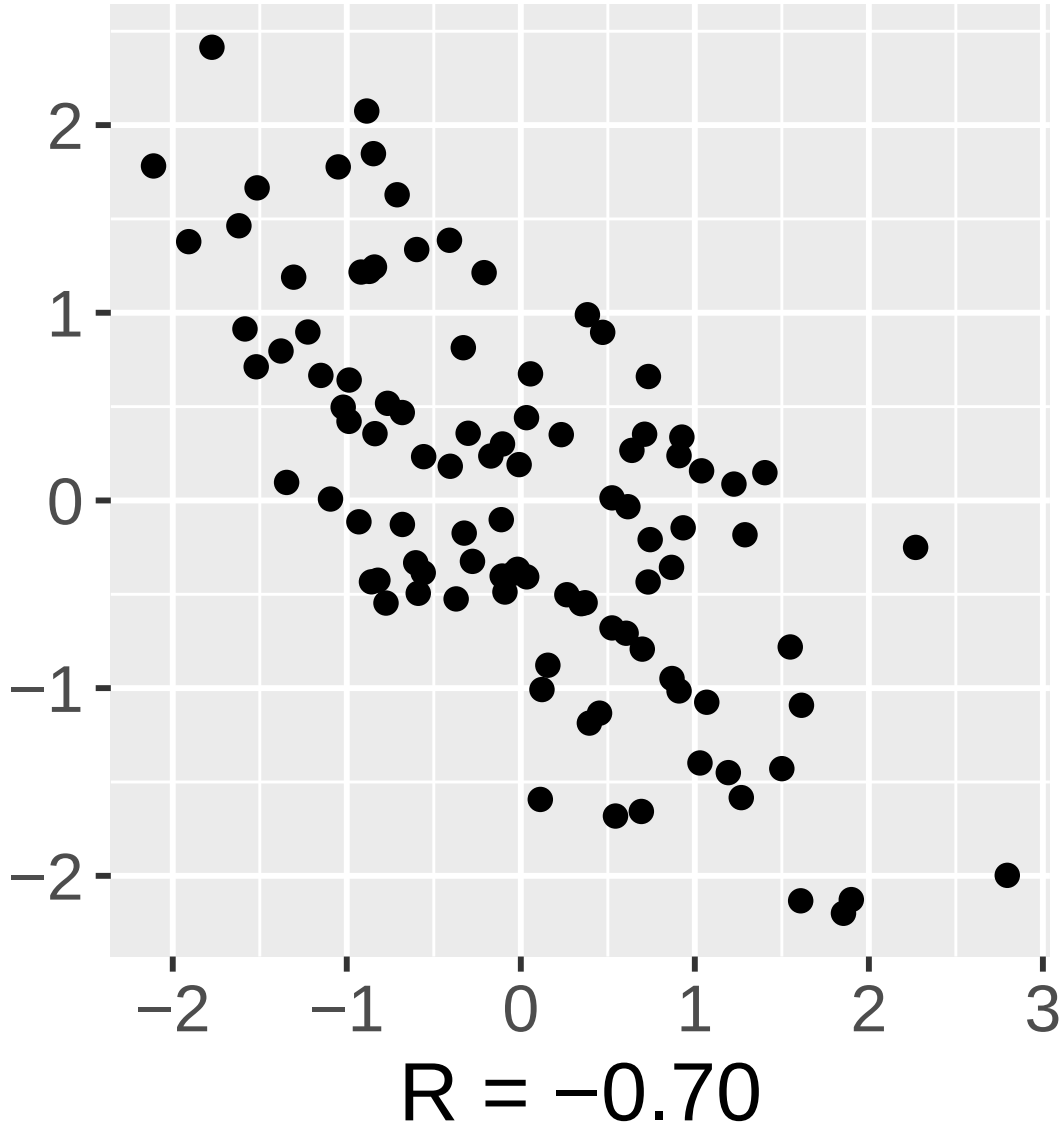
Correlation, example



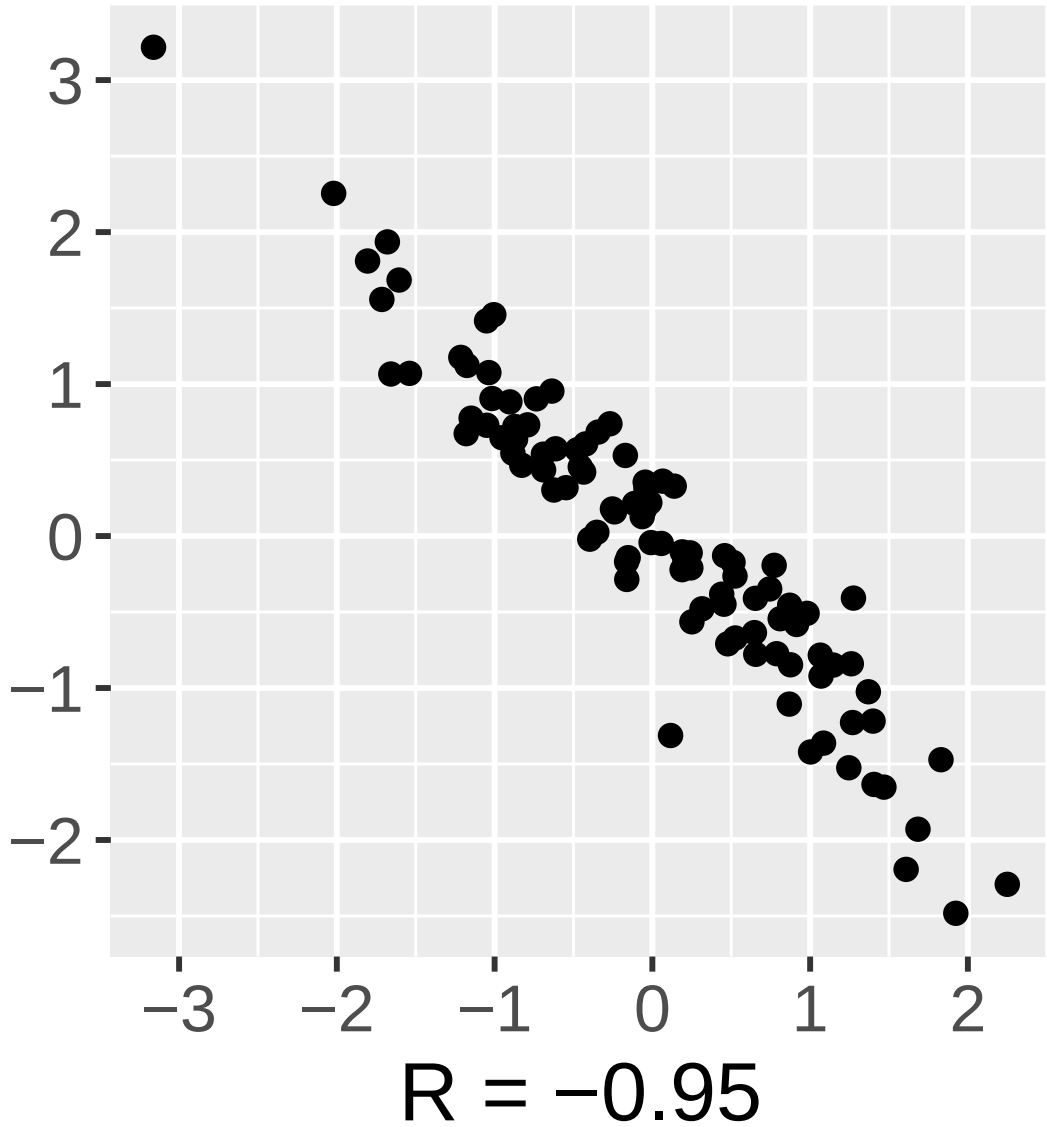
Correlation, example



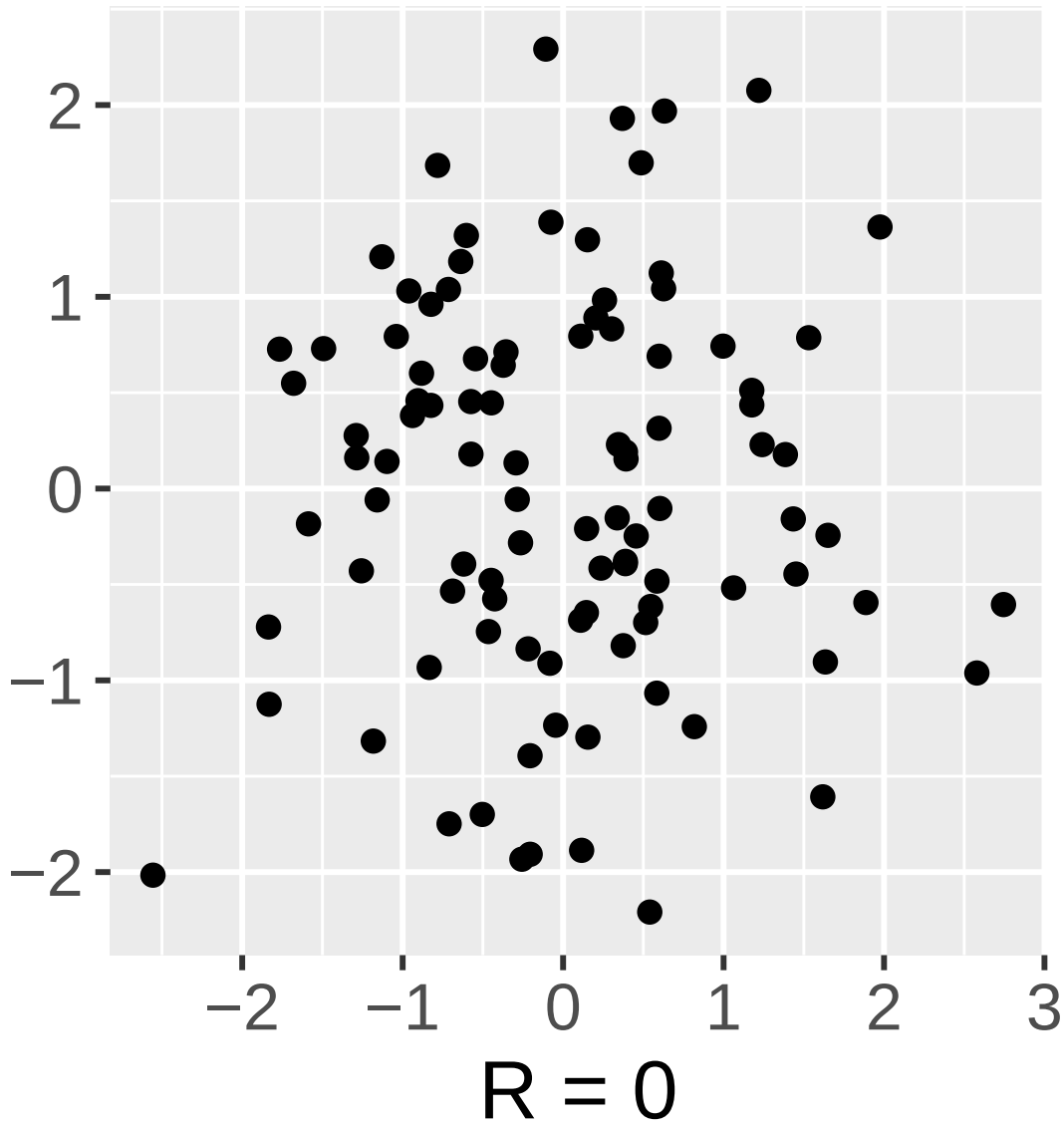
Correlation, example



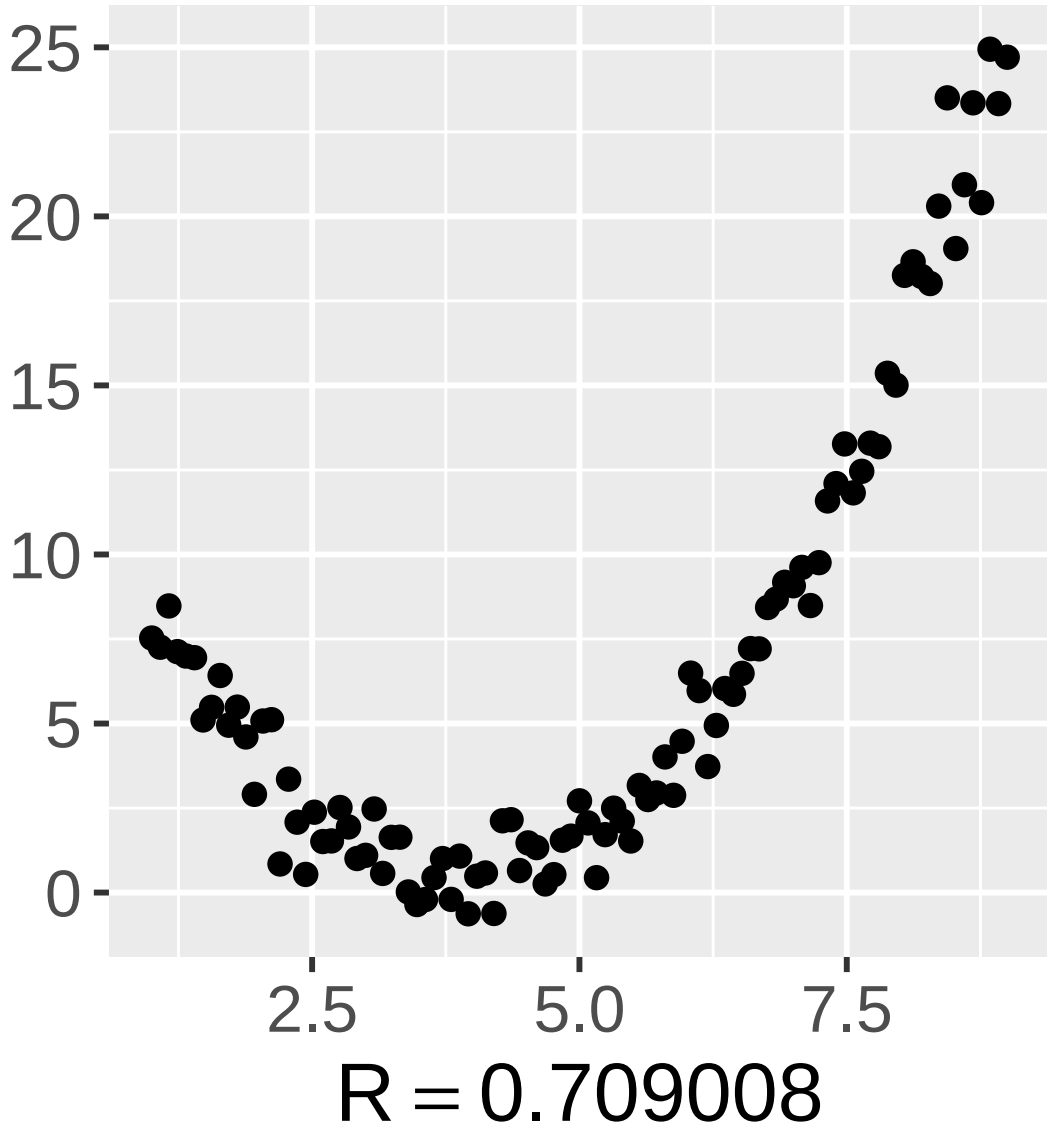
Correlation, example



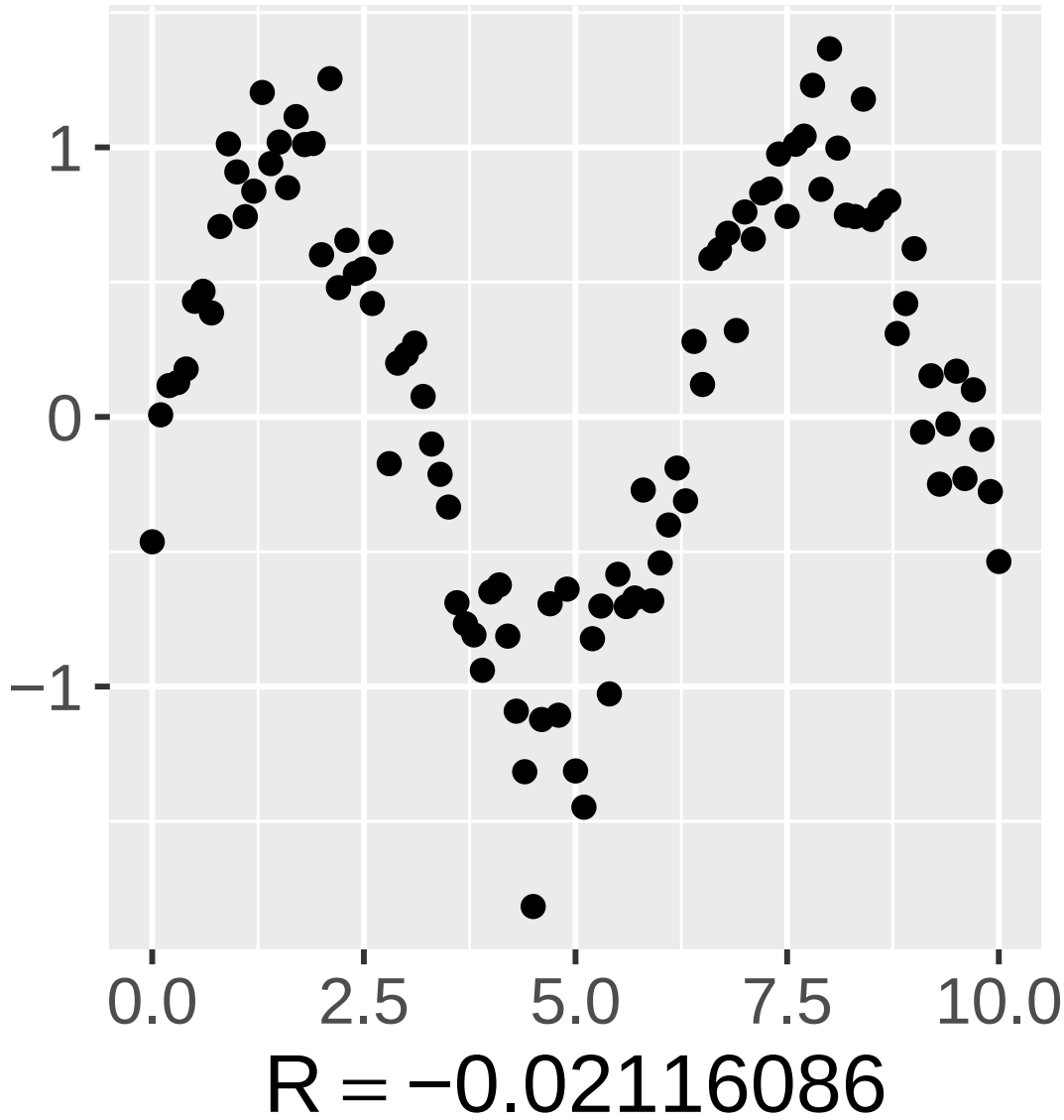
Correlation, example



Correlation, (watch out) example



Correlation, (watch out) example



Correlation, plots

This is another reason plots are so important.

outline

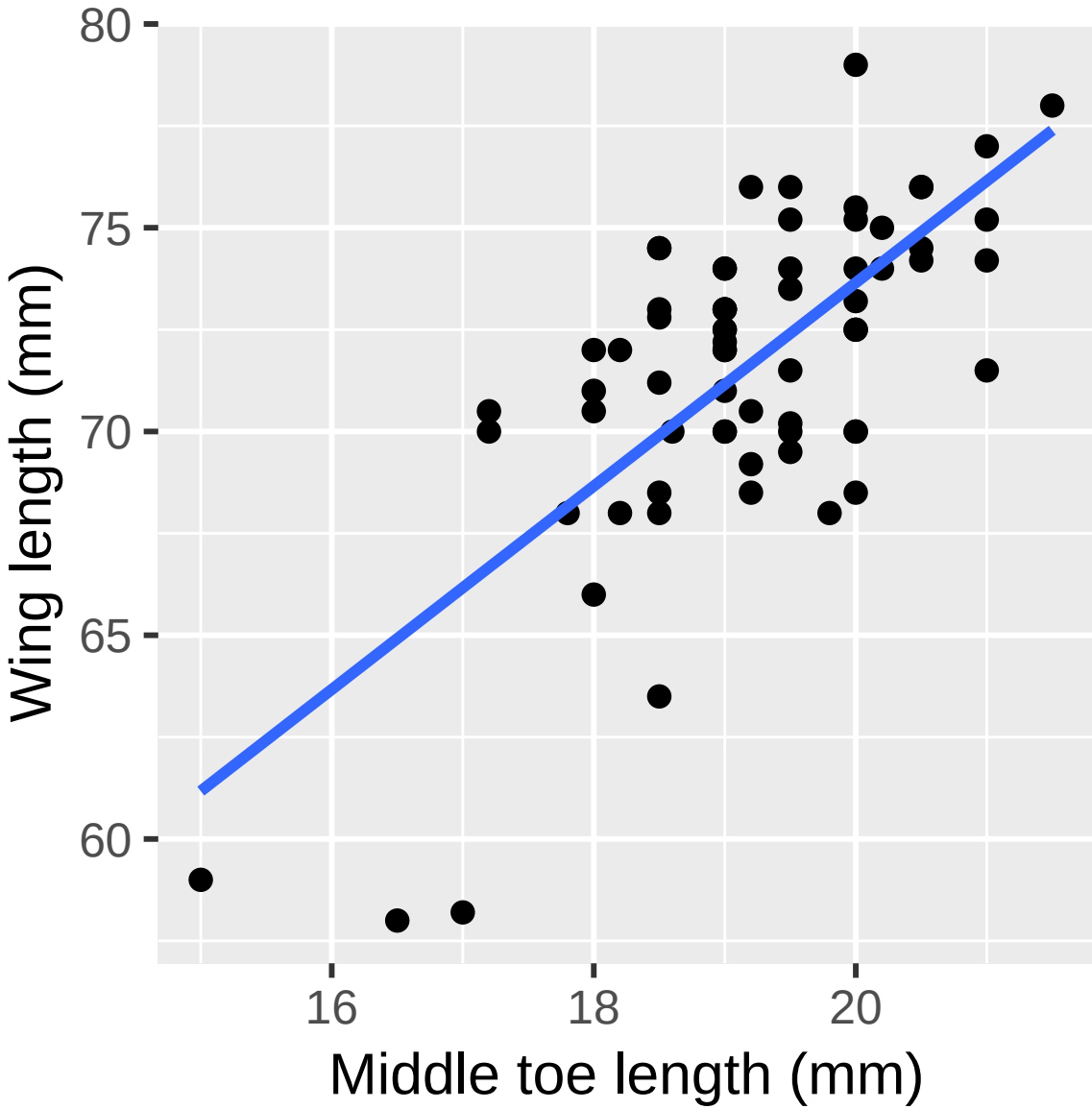
- Recap
- Correlation
  - definition
  - examples
- Simple Linear Regression
  - lite example
  - assumptions
  - parameter
  - residuals
  - estimation
- Example
- References

Simple Linear Regression, introduction

For all of the appropriately linear correlation examples above, it was easy to think about a line through the data and then ask, "How closely do the data fall onto that line?" That line through the data however, has a name and a mathematical definition.

Simple Linear Regression, plotted

The **least squares line** for Darwin's finch data is plot below in blue.



Simple Linear Regression, plotted

The coefficients of the **least squares line** for Darwin's finch data are

```
##      (Intercept) middletoelength
##        23.751547         2.494932
```

Simple Linear Regression, idea

Simple linear regression decomposes the response variable Y into three components:

- the **intercept**
  - the value Y takes on when X is equal to 0;
  - above, the length of a wing when the middle toe length is 0
- the **slope**
  - on the explanatory variable X
  - represents the increase in Y for a unit increase in X;
  - above, some increase in wing length for every mm increase in the middle toe length
- **errors/residuals**
  - some left over bits

Linear Regression, model

Given a response variable Y and an explanatory variable X, the simple linear regression model is

$$Y_i = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} X_i + \underbrace{\epsilon_i}_{\text{errors}}, \quad \epsilon_i \sim N(0, \sigma^2),$$

where $i = 1, \ldots, n$.

Simple Linear Regression, assumptions

Simple linear regression assmptions

- Linearity – the data should show a linear trend.
- Independent observations – no two points are dependent on each other.
- Constant Variability – variation of points around least squares line remains roughly constant.
- Normality – the residuals should be nearly normal.

Simple Linear Regression, parameters

The population parameters β_0 and β_1 are estimated with $\hat{\beta}_0$ and $\hat{\beta}_1$. These estimates within the linear regression equation are written

$$E(Y|X) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

The expected value, or fitted value, of Y is a function of the estimates of the intercept and slope, dependent on some value of X.

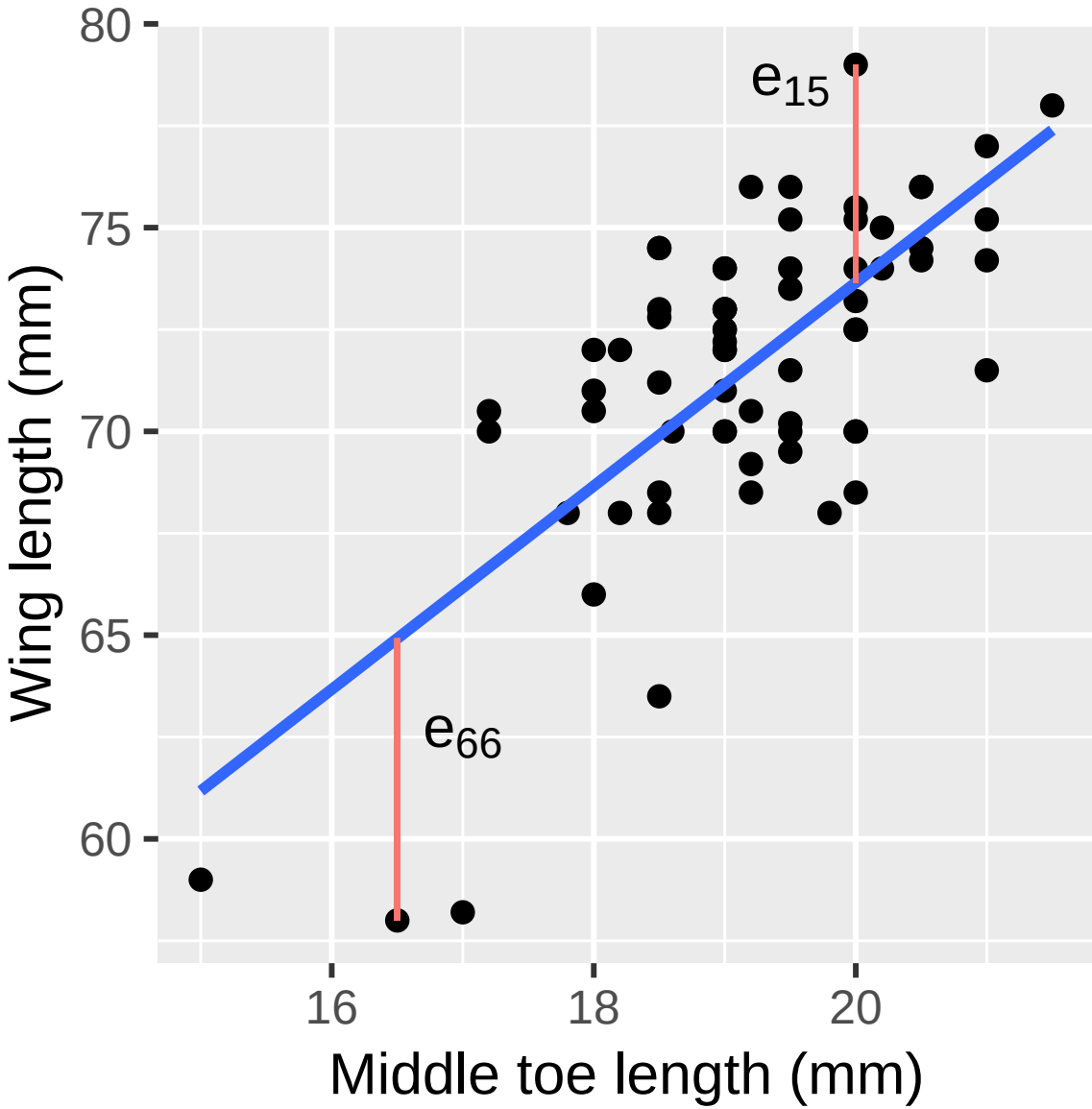
Linear Regression, residuals

Not every observation will fall of the least squares line. The difference between the true observation Y_i and the predicted value of $\hat{Y}_i$ at the X_i, is the ith residual

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

Simple Linear Regression, residuals by picture

Some residuals will be positive and some negative.



Simple Linear Regression, best

The word "best" is cleverly defined and not without debate. The most common definition of best means the line that minimizes the sum of the squared residuals. This idea is intuitive. We are to find the values of β_0 and β_1 that

- take $e_i = Y_i - \hat{Y}_i$, for all i,
- square each residual, e_i^2, and
- minimize $\sum_{i=1}^{n} e_i^2$.

Simple Linear Regression

Our model implies $Y \sim N(\beta_0 + X\beta_1, \sigma^2)$. Parameter estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma^2}$, are found via likelihood.

Simple Linear Regression in R

In R we use the function lm to fit linear regression. The format is much the same as that for ANOVA.

```
fit <- lm(winglength~middletoelength, data=finch)
## summary(fit) # RStudio
```

Simple Linear Regression, interpretation

The coefficients from the model can be extracted with the function `coefficients`.

```
(beta <- coefficients(fit))

##     (Intercept) middletoelength
##       23.751547        2.494932
```

Thus, our fitted linear model is written as

$$\hat{Y} = E(Y|X) = 23.75 + 2.49X.$$

outline

- Recap
- Correlation
  - definition
  - examples
- Simple Linear Regression
  - lite example
  - assumptions
  - parameter
  - residuals
  - estimation
- Example
- References

Elmhurst College Data

We'll consider a dataset named `elmhurst`. With these data, we might have the question, "How is family income related to the amount of gift aid a student receives from the college?"

```
url <- "https://roualdes.us/data/elmhurst.csv"
elmhurst <- read.csv(url)
head(elmhurst)
```

```
##   family_income gift_aid price_paid
## 1        92.922    21.72      14.28
## 2         0.250    27.47       8.53
## 3        53.092    27.75      14.25
## 4        50.200    27.22       8.78
## 5       137.613    18.00      24.00
## 6        47.957    18.52      23.48
```

Elmhurst College Data

Step 1?

Elmhurst College Data

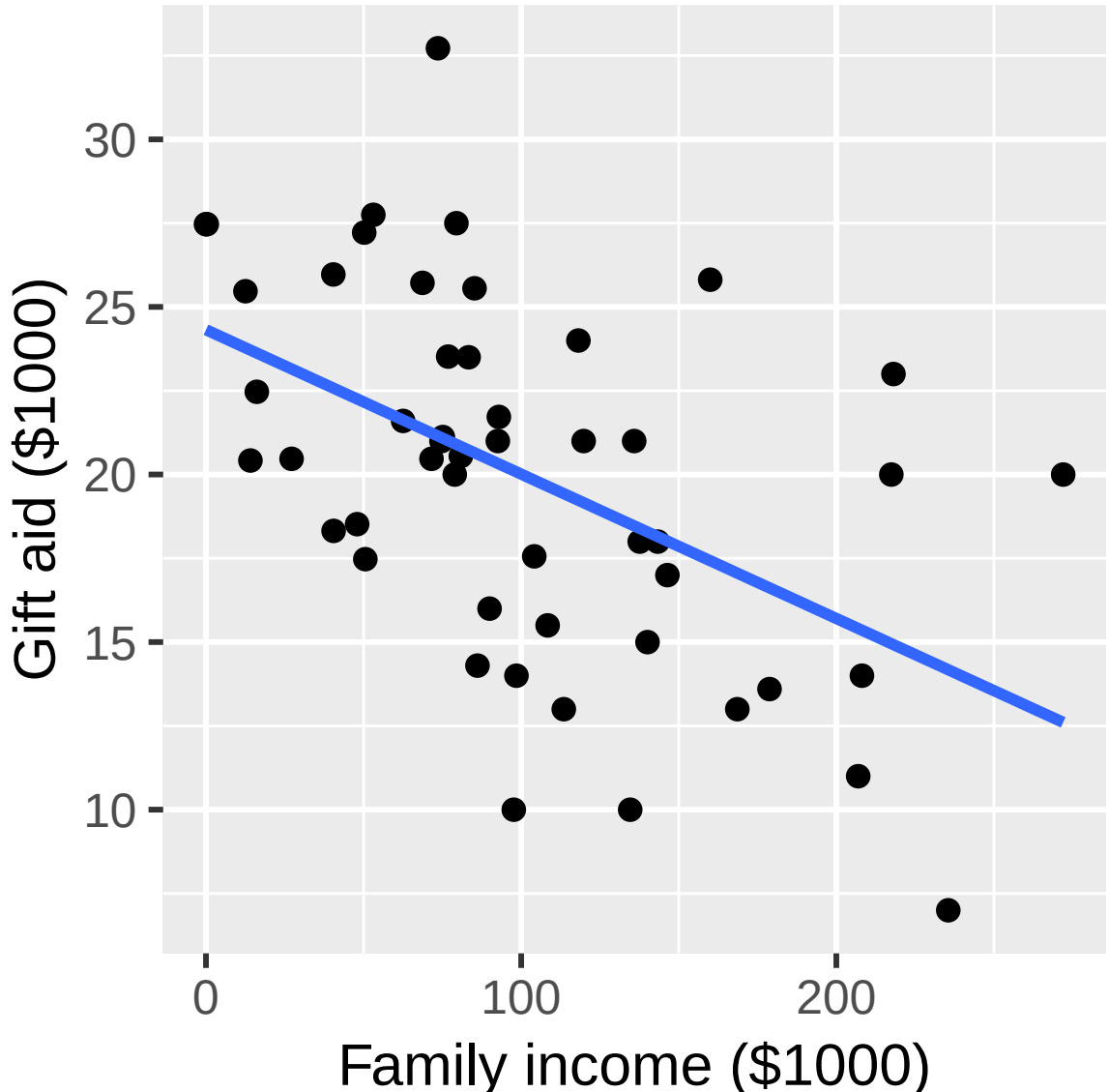
Plot the data!

```
## p <- qplot(family_income, gift_aid, data=elmhurst,
##            ylab="Gift aid ($1000)",
##            xlab="Family income ($1000)")
p <- ggplot(elmhurst, aes(family_income, gift_aid)) +
    geom_point() +
    labs(x="Family income ($1000)",
         y="Gift aid ($1000)")
with(elmhurst,
     cor(family_income, gift_aid))

## [1] -0.4985561
```

Elmhurst College Data

```
p + stat_smooth(method="lm", se=FALSE) # no standard errors
```



Elmhurst College Data

```
elmReg <- lm(gift_aid ~ family_income, data=elmhurst)
# summary(elmReg) ## RStudio
```

Elmhurst College Data

Our estimated linear model looks like

$$\widehat{aid} = 24.32 + -0.04 \times family\_income.$$

How do we interpret this?

- intercept – when family income is 0, average gift aid is expected to be 24.32 thousands of dollars.
- slope – for every \$1000 increase in family income, average gift aid is expected to decrease by -\$43.07.

Elmhurst College Data

Can we make causal connections from this model?

Take away

- Correlation is a helpful summary statistic between two numerical variables
- Linear regression fits "best" line through scatterplot
  - best means minimized squared residuals
  - expected value of response given some value of explanatory
  - the assumptions are important, we will return to them many times

outline

- Recap
- Correlation
  - definition
  - examples
- Simple Linear Regression
  - lite example
  - assumptions
  - parameter
  - residuals
  - estimation
- Example
- References

references I