STA221

Neil Montgomery

Last edited: 2017-07-11 19:01

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

χ^2 tests—when are the p-values accurate?

The test statistic has an approximate χ^2_ν distribution:

- when the observations in the dataset are independent.
  - Not related to the "test of independence"
  - Cannot usually be verified—only assumed based on the way the data were collected.
- **and when expected cell counts are all over 5.**
  - Other suggestions are common—all are fine.
  - Similar in spirit to the $np \geq 5$ suggestion in the "normal approximation to the binomial"

The second condition is the main thing to check.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.
  - apparently some people try to apply the χ^2 procedures to other kinds of summary tables, which is why the readings emphasize this point as a warning.

χ^2 tests - other matters to ponder (or not)

The readings mention a few other things to "check" when it comes to goodness-of-fit tests. Here is some commentary:

- "counted data condition," which even the readings admit is silly.
  - the summary tables with counts are not datasets - they are summaries.
  - apparently some people try to apply the χ^2 procedures to other kinds of summary tables, which is why the readings emphasize this point as a warning.
- "randomization condition," which has more to do with the possibility of *inferring something about a larger population, or not* than anything to do with χ^2 tests per se.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have affected the results.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have affected the results.
  - **the expected cell counts were all much larger than 5.**

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have affected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have affected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.
  - The analysis was not on any sample at all - I used *all* numbers ever drawn!

examples of verifying the assumptions and conditions

- In the Lotto 6/49 example:
  - there was actually a very mild lack of independence among the rows in the dataset, but it wouldn't have affected the results.
  - **the expected cell counts were all much larger than 5.**
  - I did indeed analyse counts.
  - The analysis was not on any sample at all - I used *all* numbers ever drawn!
- All other examples satisfied the $E_{ij} \geq 5$ condition, which is the main thing that should always be verified and commented on.

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

These can be called *standardized residuals*, where the $O_i - E_i$ are just the *residuals*.

post-hoc investigations of χ^2 tests using residuals

The χ^2 tests are based on the following standardized deviation of *observed* from *expected*:

$$\frac{O_i - E_i}{\sqrt{E_i}}$$

(similar with *ij* subscripts.)

These can be called *standardized residuals*, where the $O_i - E_i$ are just the *residuals*.

These are approximately $N(0, 1)$, so one can glance at the cell-by-cell residuals to get information about which cells had the largest deviation from expected.

standardized residuals example - I (pipeline)

```
##      Pressure
## Leak  High  Low  Med  Sum
##   No   277  278  247  802
##   Yes   71   66   61  198
##   Sum  348  344  308 1000

##
##  Pearson's Chi-squared test
##
## data:  table(Leak, Pressure)
## X-squared = 0.16116, df = 2, p-value = 0.9226

##      Pressure
## Leak    High    Low    Med
##   No  -0.125  0.127 -0.001
##   Yes  0.253 -0.256  0.002
```

standardized residuals example - II (births)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  childbirth
## X-squared = 14.388, df = 1, p-value = 0.0001487
```

```
##          Breastfeeding
## Epidural     No    Yes
##      No  -1.728  1.418
##      Yes  2.429 -1.993
```

regression

linear models

Basic model: Output $=$ Input $+$ Noise

You've seen a few of these already; perhaps not expressed this way.

linear models

Basic model: Output $=$ Input $+$ Noise

You've seen a few of these already; perhaps not expressed this way.

The "one sample t-test" can be thought of a way to analyze data that can be modeled as:

$$Y_i = \mu + \varepsilon_i$$

where ε_i are independent $N(0, \sigma)$ and n is the sample size.

linear models

Basic model: Output = Input + Noise

You've seen a few of these already; perhaps not expressed this way.

The "one sample t-test" can be thought of a way to analyze data that can be modeled as:

$$Y_i = \mu + \varepsilon_i$$

where ε_i are independent $N(0, \sigma)$ and n is the sample size.

The "two sample t-test" could be modeled as:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, 2$ and the μ_i are the two population means. (There are a few ways to treat the ε_{ij}.)

several numerical variables

Suppose your dataset has a numerical variable we'll call y and other variable (typically also numerical) called x. Most datasets will have several!

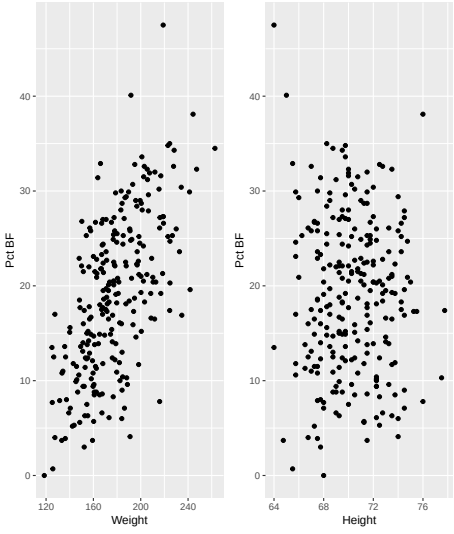
several numerical variables

Suppose your dataset has a numerical variable we'll call y and other variable (typically also numerical) called x. Most datasets will have several!

Let's consider the male body fat dataset that is discussed in the readings (Chapter 24).

```
## # A tibble: 250 × 15
##    `Pct BF`   Age Weight Height  Neck Chest Abdomen    waist   Hip
##       <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl>    <dbl> <dbl>
## 1      12.3    23 154.25  67.75  36.2  93.1    85.2 33.54331  94.5
## 2       6.1    22 173.25  72.25  38.5  93.6    83.0 32.67717  98.7
## 3      25.3    22 154.00  66.25  34.0  95.8    87.9 34.60630  99.2
## 4      10.4    26 184.75  72.25  37.4 101.8    86.4 34.01575 101.2
## 5      28.7    24 184.25  71.25  34.4  97.3   100.0 39.37008 101.9
## 6      20.9    24 210.25  74.75  39.0 104.5    94.4 37.16535 107.8
## 7      19.2    26 181.00  69.75  36.4 105.1    90.7 35.70866 100.3
## 8      12.4    25 176.00  72.50  37.8  99.6    88.5 34.84252  97.1
## 9       4.1    25 191.00  74.00  38.1 100.9    82.5 32.48031  99.9
## 10     11.7    23 198.25  73.50  42.1  99.6    88.6 34.88189 104.1
```

body fat EDA


Pct BF
40
30
20
10
0
120
160
200
240
Weight
Pct BF
40
30
20
10
0
64
68
72
76
Height

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

$\texttt{Pct BF} = \texttt{Weight} + \text{noise}$

$\texttt{Pct BF} = \texttt{Height} + \text{noise}$

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

`Pct BF` $=$ `Weight` $+$ noise

`Pct BF` $=$ `Height` $+$ noise

In general:

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

where y and x are the variables and ε is the random noise.

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

`Pct BF` $=$ `Weight` $+$ noise

`Pct BF` $=$ `Height` $+$ noise

In general:

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

where y and x are the variables and ε is the random noise.

When there are only two variables this is called a *simple regression model.*

linear model for two numerical variables

When there is a linear relationship between two variables, we might propose a linear model such as:

`Pct BF` $=$ `Weight` $+$ noise

`Pct BF` $=$ `Height` $+$ noise

In general:

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

where y and x are the variables and ε is the random noise.

When there are only two variables this is called a *simple regression model.*

The *parameter* β_1 is the slope of the line and is of primary interest. (The parameter β_0 is the y-intercept and not normally of any interest.)

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind calling them "response" and "predictor".

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind calling them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind calling them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind calling them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.
- It could be a pre-specified grid of values.

model details; terminology

y and x are *not* interchangeable; i.e. these are completely different:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad\qquad x = \beta_0 + \beta_1 x + \varepsilon$$

y can be called the "output" variable and x can be called the "input" variable. I think these are the best names.

I don't mind calling them "response" and "predictor".

I hate "dependent" and "independent" variables. These words are already being used by an important concept in probability.

Think of the model from the inside and move out. It starts with x, which can be anything.

- It doesn't have to be random.
- It could be a pre-specified grid of values.
- The "grid" could consist of as few as two values!

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

model details

Starting from the inside with x. Now consider the line $\mu_y = \beta_0 + \beta_1 x$.

μ_y is intended to suggest the (theoretical) mean of y at any x value. (I might have put $\mu_y(x)$ to emphasize the role of x.)

This line is the basis of the relationship between input and output.

Finally, to this line we add some random noise ε to get the final model:

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

For the moment we'll put these requirements on the random noise:

- the noise has constant variation
- the noise for each observation is independent

We'll add another requirement when the time comes.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the parameters.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the parameters.
- Use probability to make inferences using these estimates.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the parameters.
- Use probability to make inferences using these estimates.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the parameters.
- Use probability to make inferences using these estimates.

The classic method of regression parameter estimation given data is called *least squares regression*.

- The data come in pairs $(y_1, x_1), (y_2, x_2), \ldots (y_n, x_n)$.

estimating the slope (and intercept)

Now we proceed through the typical steps of a data analysis (the χ^2 procedures were an exception!)

- We have a model with unknown parameters.
- So we gather data, and use the data to estimate the parameters.
- Use probability to make inferences using these estimates.

The classic method of regression parameter estimation given data is called *least squares regression*.

- The data come in pairs $(y_1, x_1), (y_2, x_2), \ldots (y_n, x_n)$.
- For any "candidate" slope b_0^* and intercept b_1^* we could construct the set of "predictions" $\hat{y}_i = b_0^* + b_1^* x_i$ and their "residuals" $\varepsilon_i = y_i - \hat{y}_i$

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

The formula for the slope estimator b_1 turns out to be:

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

more least squares

Here's the actual "least squares" part...

It is possible to find the unique slope and intercept that makes this sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* y_i))^2$$

as small as possible.

We'll call the unique intercept and slope b_0 and b_1, respectively.

(More common to call them $\hat{\beta}_0$ and $\hat{\beta}_1$.)

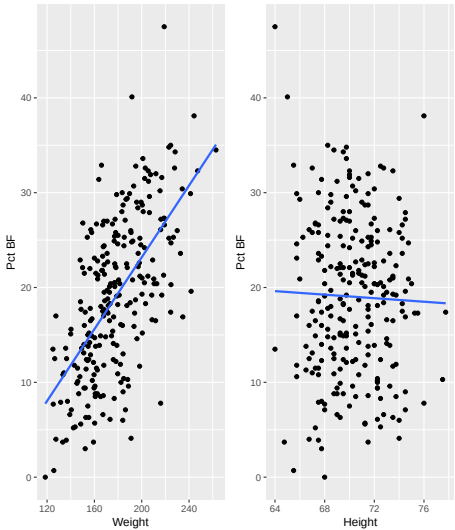
The formula for the slope estimator b_1 turns out to be:

$$b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

The formula for the intercept is $b_0 = \overline{y} - b_1 \overline{x}$

body fat examples

Here are the plots with the least squares regression lines added:



bodyfat calculation examples

Obviously don't do these by hand! Here is basic R regression output:

```
##
## Call:
## lm(formula = `Pct BF` ~ Weight, data = bodyfat)
##
## Coefficients:
## (Intercept)       Weight
##    -14.6931       0.1894
```

```
##
## Call:
## lm(formula = `Pct BF` ~ Height, data = bodyfat)
##
## Coefficients:
## (Intercept)       Height
##    25.58078     -0.09316
```

what is random and what is fixed?

The simple regression model:

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

What is random and what is fixed?

- β_0 and β_1 are fixed model parameters

what is random and what is fixed?

The simple regression model:

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

What is random and what is fixed?

- β_0 and β_1 are fixed model parameters
- the x values are treated as fixed (even if they aren't)

what is random and what is fixed?

The simple regression model:

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

What is random and what is fixed?

- β_0 and β_1 are fixed model parameters
- the x values are treated as fixed (even if they aren't)
- the error ε is random

what is random and what is fixed?

The simple regression model:

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

What is random and what is fixed?

- β_0 and β_1 are fixed model parameters
- the x values are treated as fixed (even if they aren't)
- the error ε is random
- therefore, y is random (as the sum of a fixed part and a random part)

the estimators b_0 and b_1 are also random

The intercept and slope parameter estimates are functions of the y and x.

the estimators b_0 and b_1 are also random

The intercept and slope parameter estimates are functions of the y and x.

Since the y are considered random, so are b_0 and b_1.

the estimators b_0 and b_1 are also random

The intercept and slope parameter estimates are functions of the y and x.

Since the y are considered random, so are b_0 and b_1.

What properties do they have? By simulation, and by examination of the formulae we will determine these properties. The properties of interest are:

- their distributions

the estimators b_0 and b_1 are also random

The intercept and slope parameter estimates are functions of the y and x.

Since the y are considered random, so are b_0 and b_1.

What properties do they have? By simulation, and by examination of the formulae we will determine these properties. The properties of interest are:

- their distributions
- their means and variances

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.
- independent observations in the dataset

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.
- independent observations in the dataset
  - hard to verify—usually assumed.

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.
- independent observations in the dataset
  - hard to verify—usually assumed.
  - one type of non-independence can sometimes be detected by plotting values versus time or the order in which they were observed.

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.
- independent observations in the dataset
  - hard to verify—usually assumed.
  - one type of non-independence can sometimes be detected by plotting values versus time or the order in which they were observed.
  - violation could be **fatal** but possibly not

review of model assumptions, plus a new one - I

In a simple regression analysis, we need:

- an actual linear relationship between y and x
  - check using scatterplot; *another more sensitive plot can be used TBA*
  - violation of this requirement is **fatal** to any analysis.
  - transforming one or both variables is a possible remedy.
- independent observations in the dataset
  - hard to verify—usually assumed.
  - one type of non-independence can sometimes be detected by plotting values versus time or the order in which they were observed.
  - violation could be **fatal** but possibly not
  - "time series" methods are one way to deal with one type of non-independence.

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution
  - check using a normal quantile plot of the residuals

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution
  - check using a normal quantile plot of the residuals
  - violation is **not fatal*** as long as the sample size is "large enough"

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution
  - check using a normal quantile plot of the residuals
  - violation is **not fatal*** as long as the sample size is "large enough"

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution
  - check using a normal quantile plot of the residuals
  - violation is **not fatal*** as long as the sample size is "large enough"

These two assumptions can be rolled into one statement:

$$\varepsilon \sim N(0, \sigma)$$

review of model assumptions, plus a new one - II

- the amount of variation (up and down) around the line needs to be constant
  - check using a special scatterplot involving the residuals, TBA
  - violation is **fatal**
  - transformation of variables and more sophisticated models are possible remedies
- NEW the error should follow a normal distribution
  - check using a normal quantile plot of the residuals
  - violation is **not fatal*** as long as the sample size is "large enough"

These two assumptions can be rolled into one statement:

$$\varepsilon \sim N(0, \sigma)$$

* with one exception TBA

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?
- What factors affect the accuracy of the estimator?

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?
- What factors affect the accuracy of the estimator?

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?
- What factors affect the accuracy of the estimator?

Compare these issues with the simpler situation in which $\overline{X}$ is used to estimate μ, etc.

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?
- What factors affect the accuracy of the estimator?

Compare these issues with the simpler situation in which $\overline{X}$ is used to estimate μ, etc.

First, we'll look at the average value of b_0, using simulation. To do this I will start with a *fully known theoretical linear model*:

$$y = 2 + 0.75x + \varepsilon$$

with $\varepsilon \sim N(0, 1)$.

simulation for investigating b_0 and b_1

The properties of the slope parameter estimator b_1 is of most interest.

- What is its average value, variation, distribution?
- What factors affect the accuracy of the estimator?

Compare these issues with the simpler situation in which $\overline{X}$ is used to estimate μ, etc.

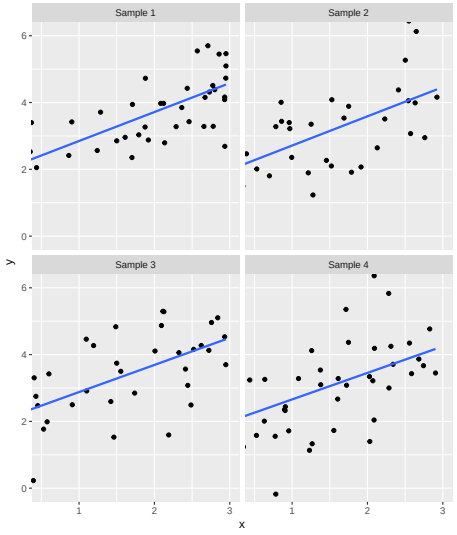
First, we'll look at the average value of b_0, using simulation. To do this I will start with a *fully known theoretical linear model*:

$$y = 2 + 0.75x + \varepsilon$$

with $\varepsilon \sim N(0, 1)$.

I will simulate fake datasets of size $n = 50$ from this model, compute the regression line for each dataset, and see what happens.

e.g. plots of four samples



properties of b_1 from 1000 samples

I would like to investigate the distribution of b_1 using simulation. So I will simulate 1000 replications, and see what happens.

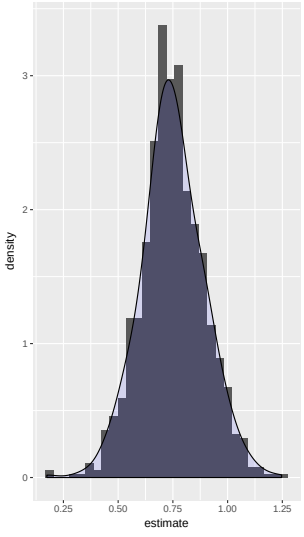
Here is a numerical summary of the 1000 simulated b_1 (and b_0 as well, since I have them):

	term	Average	SD
1	(Intercept)	2.00165	0.22706
2	x	0.74808	0.14264

(Note: these numbers *change* every time I render the lecture notes - the simulation is embedded right in them.)

Conclusion: the average values of b_1 (and b_0) are the true values β_1 (and β_0).

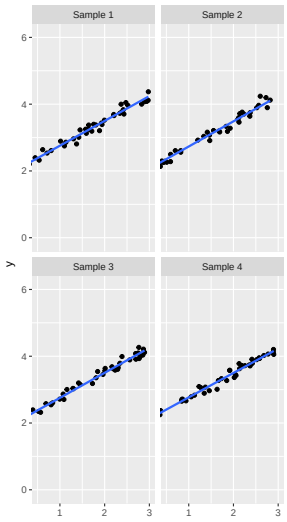
histogram of the simulated b_1



Looks symmetric and bell-shaped. Perhaps they have a normal distribution?

change σ from 1 to 0.1

I will simulate again, but this time with $\varepsilon \sim N(0, 0.1)$. Four example plots:



properties of b_1 from 1000 samples ($\sigma = 0.1$ version)

The averages and SDs of the 1000 estimators:

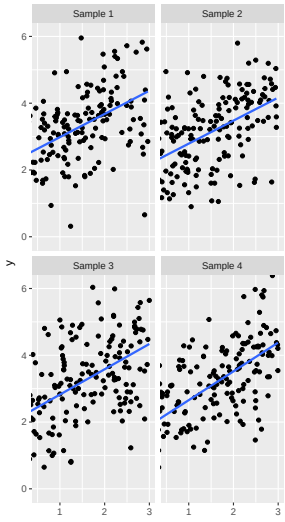
term	Average	SD
(Intercept)	2.00012	0.02251
x	0.74945	0.01356

The histogram looks the same.

Conclusion: when the *inherent underlying noise is smaller* the parameter estimators are *more accurate*.

put σ back to 1; increase the sample size to $n = 200$

Four sample plots:



properties of b_1 from 1000 samples ($n = 200$ version)

The averages and SDs of the 1000 estimators:

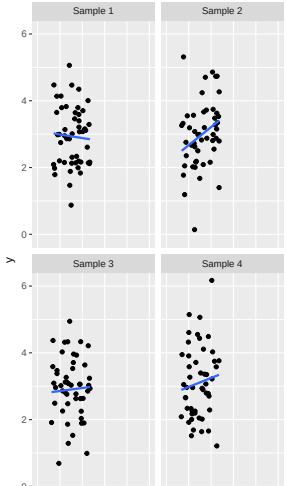
term	Average	SD
(Intercept)	2.00461	0.11211
x	0.74626	0.06999

The histogram looks the same.

Conclusion: when the *sample size is larger* the parameter estimators are *more accurate*.

back to $n = 50$; properties of b_1 when the x values are less spread out

This one is a little more subtle. It turns out the x values affect the accuracy of the parameter estimates. I re-simulate with less spread in the x values. Four sample plots with x values 4 times "less spread out":



properties of b_1 (x less spread version)

The averages and SDs of the 1000 estimators:

term	Average	SD
(Intercept)	1.98094	0.73876
x	0.76641	0.57547

The histogram looks the same.

Conclusion: when the x *values are less spread out* the parameter estimators are *less accurate*.

statistical properties of b_1

Start with the basic simple linear regression model:

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

in which the error follows a $N(0, \sigma)$ distribution.

The slope estimator b_1 turns out to follow a normal distribution with mean β_1 and standard deviation:

$$\frac{\sigma}{\sqrt{S_{xx}}}$$

(Recall $S_{xx} = \sum (x_i - \overline{x})^2$)

(Note: there is a typo on the first formula in section 24.2 - the s_x should not be under the $\sqrt{\ }$.)

statistical properties of b_1

Therefore we have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

statistical properties of b_1

Therefore we have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

Except we would never know the true value of σ. This is the third simple regression parameter—a nuisance we'll have to deal with.

statistical properties of b_1

Therefore we have:

$$\frac{b_1 - \beta_1}{\sigma / \sqrt{S_{xx}}} \sim N(0, 1)$$

and p-values and confidence intervals come from this—BAM we're done.

Except we would never know the true value of σ. This is the third simple regression parameter—a nuisance we'll have to deal with.

We can estimate σ using the "average" of the squared residuals:

$$s_e = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

statistical properties of b_1

Who wants to guess what distribution this will have:

$$\frac{b_1 - \beta_1}{s_e/\sqrt{S_{xx}}}$$

statistical properties of b_1

Who wants to guess what distribution this will have:

$$\frac{b_1 - \beta_1}{s_e/\sqrt{S_{xx}}} \sim t_{n-2}$$

hypothesis testing for β_1

The principal hypothesis test concerns whether there is any linear relationship at all between x and y. The null hypothesis immediately presents itself:

$$H_0 : \beta_1 = 0$$

hypothesis testing for β_1

The principal hypothesis test concerns whether there is any linear relationship at all between x and y. The null hypothesis immediately presents itself:

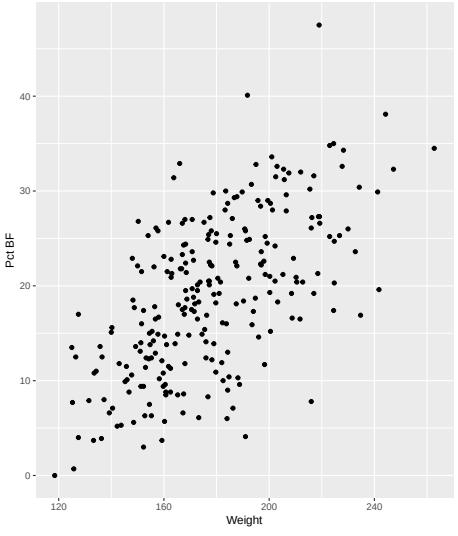
$$H_0 : \beta_1 = 0$$

And it works the same way any other hypothesis test works. Use the data to compute:

$$\frac{b_1 - 0}{s_e/\sqrt{S_{xx}}}$$

and get the probability of being "further away" from H_0, according to the t_{n-2} distribution.

example - body fat versus weight



example - body fat versus weight

```
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -14.69314    2.76045  -5.323 0.000000229 ***
## Weight        0.18938    0.01533  12.357     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.538 on 248 degrees of freedom
## Multiple R-squared:  0.3811, Adjusted R-squared:  0.3786
## F-statistic: 152.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e/\sqrt{S_{xx}}}$	the p-value

---

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e/\sqrt{S_{xx}}}$	the p-value

---

A line of questionable utility.

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e/\sqrt{S_{xx}}}$	the p-value

---

A line of questionable utility.

Residual standard error: s_e on $n-2$ degrees of freedom

this table translated into formulae

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
b_0	(Not	often	very	relevant)
b_1	$\frac{S_{xy}}{S_{xx}}$	$\frac{s_e}{\sqrt{S_{xx}}}$	$\frac{b_1 - 0}{s_e/\sqrt{S_{xx}}}$	the p-value

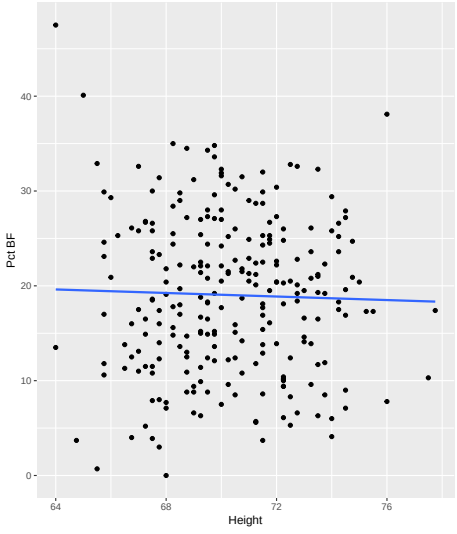
---

A line of questionable utility.

Residual standard error: s_e on $n-2$ degrees of freedom

Other stuff at the bottom not yet explained...

example - body fat versus height



example - body fat versus height

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719 .
## Height      -0.09316    0.20119  -0.463   0.6438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.307 on 248 degrees of freedom
## Multiple R-squared:  0.0008637,  Adjusted R-squared:  -0.003165
## F-statistic: 0.2144 on 1 and 248 DF,  p-value: 0.6438
```

confidence interval for the true slope β_1

95% confidence intervals are all pretty much the same, based on:

$$\frac{\text{estimator} - \text{parameter}}{SE(\text{estimator})} \sim \text{ something symmetric and bell shaped}$$

resulting in a formula:

$$\text{estimator} \pm \text{“2”} SE(\text{estimator})$$

confidence interval for the true slope β_1

95% confidence intervals are all pretty much the same, based on:

$$\frac{\text{estimator} - \text{parameter}}{SE(\text{estimator})} \sim \text{ something symmetric and bell shaped}$$

resulting in a formula:

$$\text{estimator} \pm \text{“2”} SE(\text{estimator})$$

In the case of β_1 we have:

$$\frac{b_1 - \beta_1}{s_e/\sqrt{S_{xx}}} \sim t_{n-2}$$

result in a 95% C.I. of:

$$b_1 \pm t^*_{n-2} \frac{s_e}{\sqrt{S_{xx}}}$$

example C.I.'s for β_1 - body fat versus weight and height

Since $n = 250$, our value of "2" is in this case: 1.9695757

```
##
## Coefficients:
##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -14.69314    2.76045  -5.323 0.000000229
## Weight        0.18938    0.01533  12.357     < 2e-16
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.58078   14.15400   1.807   0.0719
## Height      -0.09316    0.20119  -0.463   0.6438
```

R^2

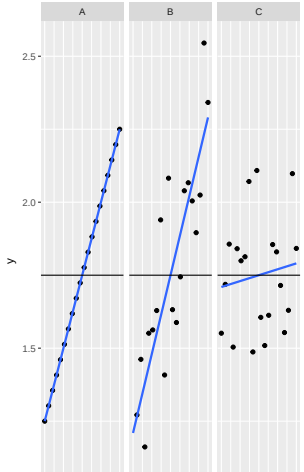
R^2

The y values are random. They aren't all the same. What "explains" the differences in the y values?

R^2

The y values are random. They aren't all the same. What "explains" the differences in the y values?

A = all "model" | B = "typical" | C = all "error":



R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y $=$ variation due to the model $+$ variation due to error

$$\sum (y_i - \overline{y})^2 = \qquad\qquad +$$

R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y $=$ variation due to the model $+$ variation due to error

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 +$$

R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2$$

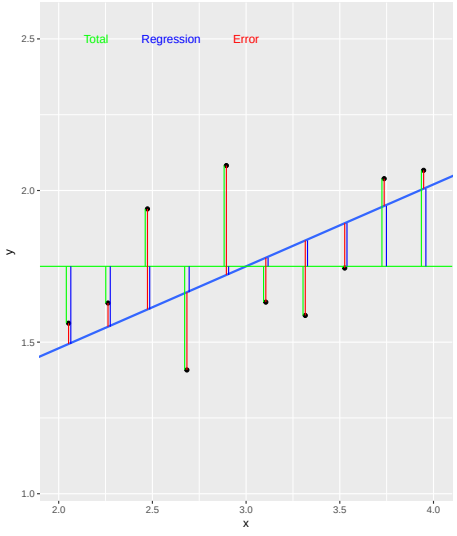
R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y = variation due to the model + variation due to error

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

sum of squares decomposition, graphically



R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

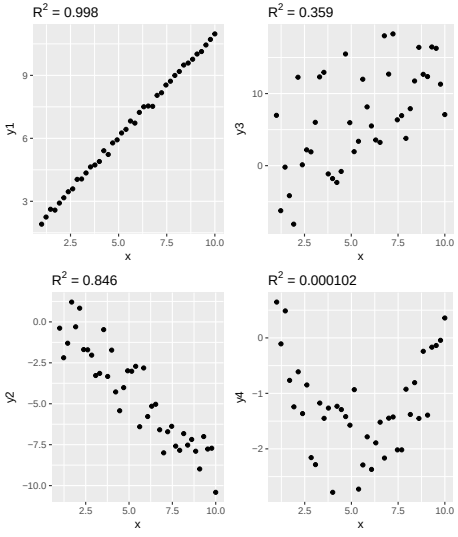
"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

Keep in mind it is *one number* that is being used to summarize an entire empirical bivariate relationship. And it isn't even the *best* number.

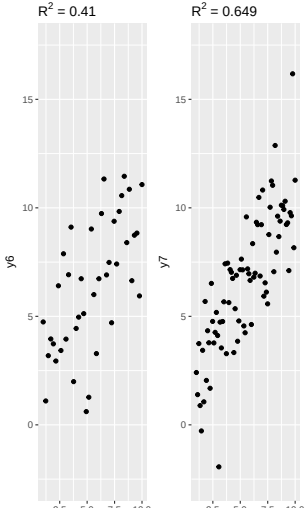
Some R^2 values



Another limitation: sample size effect

Both simulated datasets are from the ***same underlying model***

(happens to be $y = 1 + 1 \cdot x + \varepsilon$ with $\varepsilon \sim N(0, 2)$)



regression model assumption (etc.) verification

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

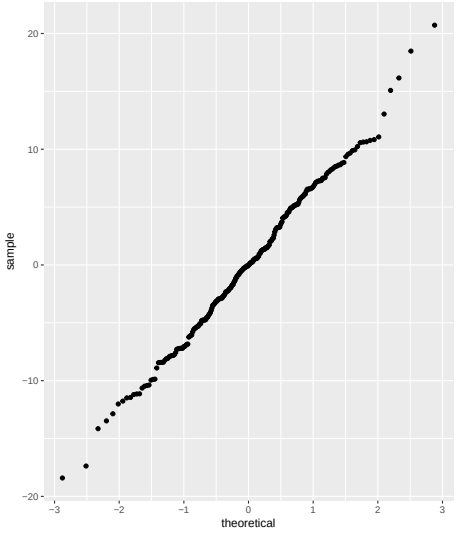
The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

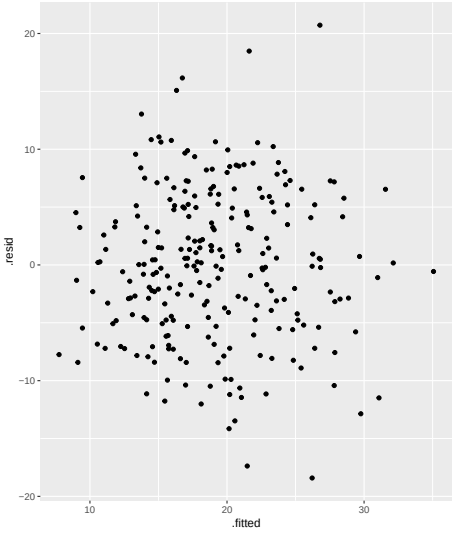
Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

We will verify graphically, using various plots of the *residuals* $\hat{\varepsilon}_i = y_i - \hat{y}_i$

verify normality with normal quantile (or normal probability) plot of $\hat{\varepsilon}_i$



verify linearity with plot of $\hat{\varepsilon}_i$ versus $\hat{y}_i$



verify equal variance with (same!) plot of $\hat{\varepsilon}_i$ versus $\hat{y}_i$