STA221

Neil Montgomery

Last edited: 2017-02-01 15:06

erratum; and a comment

erratum; and a comment

The formula for the slope estimator b_1 turns out to be (corrected from class!):

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

erratum; and a comment

The formula for the slope estimator b_1 turns out to be (corrected from class!):

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

Recall 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

erratum; and a comment

The formula for the slope estimator b_1 turns out to be (corrected from class!):

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

Recall 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)

erratum; and a comment

The formula for the slope estimator b_1 turns out to be (corrected from class!):

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

Recall 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

erratum; and a comment

The formula for the slope estimator b_1 turns out to be (corrected from class!):

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

Recall 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.

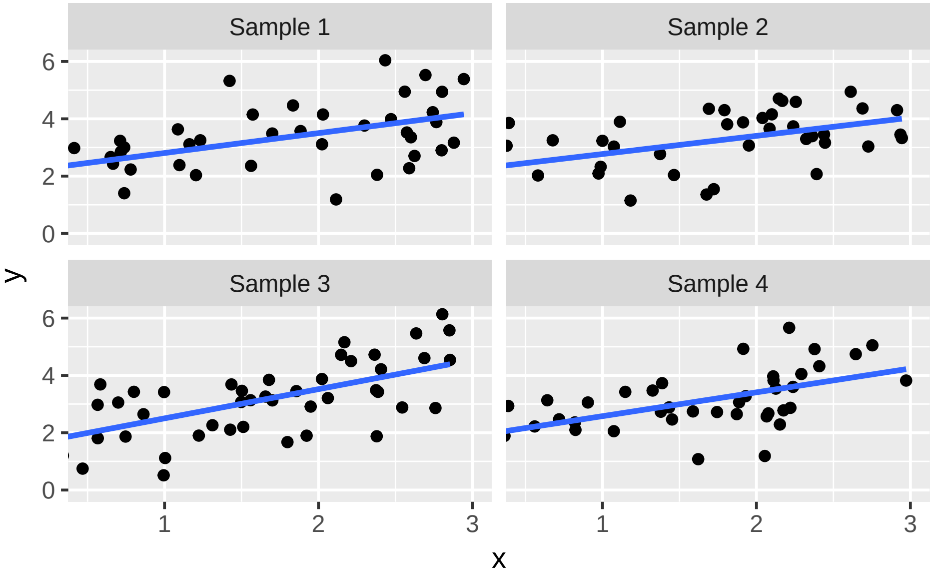
First, we'll look at the average value of b_0, using simulation. To do this I will start with a *fully known theoretical linear model*:

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

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.

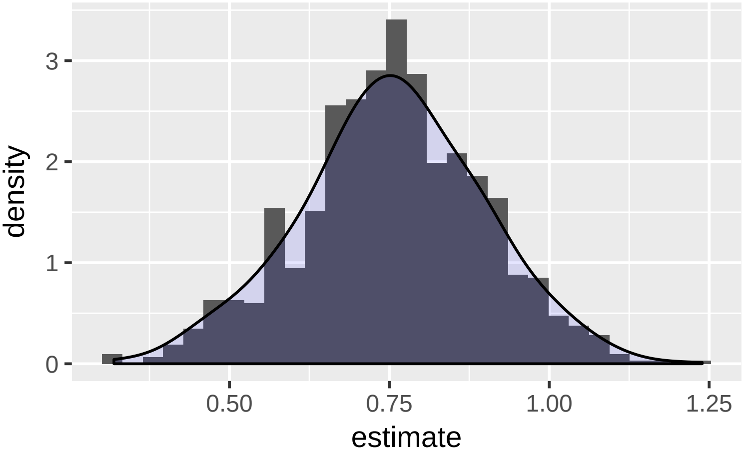
Here is a numerical summary of the 1000 simulated b_1 (and b_0 as well, since I have them):

term	Average	SD
(Intercept)	1.99499	0.22788
x	0.75465	0.14225

(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).

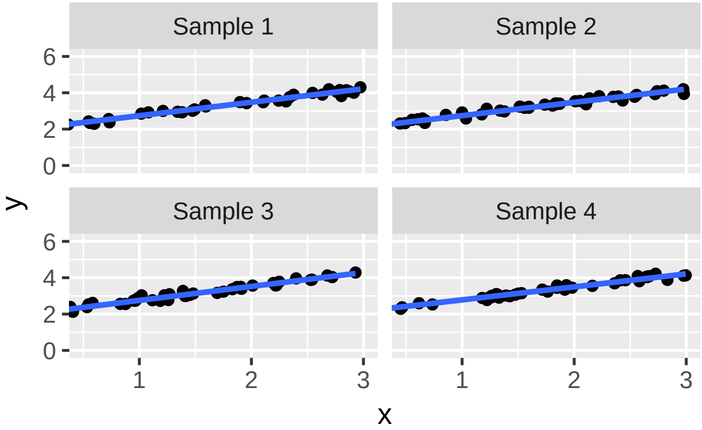
histogram of the simulated b_1



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:

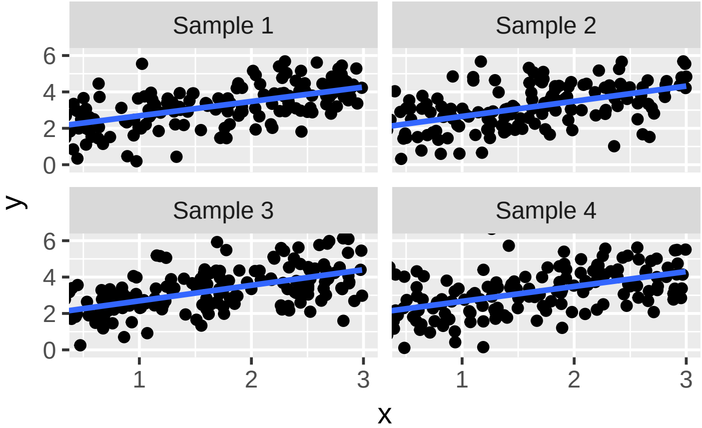
term	Average	SD
(Intercept)	1.99950	0.02333
x	0.75039	0.01432

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:

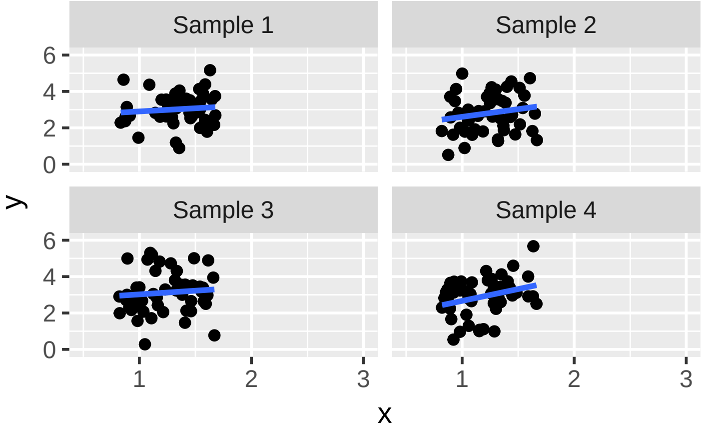
term	Average	SD
(Intercept)	2.00363	0.11522
x	0.74960	0.06986

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)	2.01478	0.73753
x	0.73827	0.57449

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 distribtion 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:

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

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 ??? distribution.

example - body fat versus weight