MEDB 5501, Module05

2024-09-17

Topics to be covered

- What you will learn
  - Interpretation of linear regression coefficients
  - Computing linear regression in R
  - The least squares principle
  - The analysis of variance table
  - Computing the analysis of variance table in R
  - Confidence interval for the slope parameter
  - Computing confidence intervals in R
  - Your homework

Bad joke, 1 of 4

Speaker notes

I borrowed this image from a movie poster. Does anyone know what movie this is?

https://en.wikipedia.org/wiki/Airplane!

So I am using this as a setting for a bad Statistics joke.

Two statisticians are on an airplane, flying from Miami to Seattle. Fiften minutes into the flight, they hear a loud …

Bad joke, 2 of 4

Speaker notes

… BANG! The pilot comes on the PA system and says “Excuse me, Ladies and Gentlemen. We’ve just had an engine explode. We’ll be just fine with three engines, but instead of a three hour flight, this will now be a four hour flight.”

The statisticians go back to talking, and fifteen minutes later, they hear another loud …

Bad joke, 3 of 4

Speaker notes

… BANG! The pilot comes back on and says, “I’m sorry to report that we’ve had another engine explode. We can still make it to Seattle, but it will now be a six hour flight. I apologize for the additional delay.”

The statisticians shrug and start talking again when fifteen minutes later, you guessed it, they hear a third loud …

Bad joke, 4 of 4

Speaker notes

… BANG! The pilot comes on again and says “I’m sorry to report that a third engine has exploded. Each engine on this jet is very powerful and we can still make it to Seattle, but it is now going to be a nine hour flight.”

At this point, one statistician says to the other, “Boy, I hope this last engine doesn’t fail …”

“… or we’ll be up here forever!”

This is an example of a dangerous extrapolation. The experience with three, two, and then only one engines may be consistent, but don’t expect that trend to continue with zero engines.

Quote from “Peggy Sue Got Married”

MORE ON THIS QUOTE ››

“- Mr. Snelgrove: What's the meaning of this, Peggy Sue?
- Peggy Sue: Well, Mr Snelgrove, I happen to know that in the future I will not have the slightest use for algebra, and I speak from experience.”

[Peggy Sue hands in her algebra test]

KEN GRANTHAM - *Mr. Snelgrove*
KATHLEEN TURNER - *Peggy Sue*
[Tag:ability, foresight, mathematics]

Speaker notes

I want to get a quick feel for your background and interests. Here's a quote from a romantic comedy starring Kathleen Turner from 1986. A forty year old woman, played by Kathleen Turner, travels back in time to her high school senior year, 1960. She has an amusing interchange with her high school math teacher.

"I happen to know that in the future, I will not have the slightest use for algebra, and I speak from experience."

Think back to your high school algebra class.

1. Do you remember any important formulas from that class?
2. Did you hate, hate, hate high school algebra?
3. Did you love high school algebra?

Big question: Will you use high school algebra in your future?

Source: https://www.moviequotes.com/s-movie/peggy-sue-got-married/

Algebra formula for a straight line

- $Y = mx + b$
- $m = \Delta y / \Delta x$
- m = slope
- b = y-intercept

Speaker notes

One formula in algebra that most people can recall is the formula for a straight line. Actually, there are several different formulas, but the one that most people cite is

Y = m X + b

where m represents the slope, and b represents the y-intercept (we'll call it just the intercept here). They can also sometimes remember the formula for the slope:

$$m = \increment y / \increment x$$

In English, we would say that this is the change in y divided by the change in x.

Linear regression interpretation of a straight line

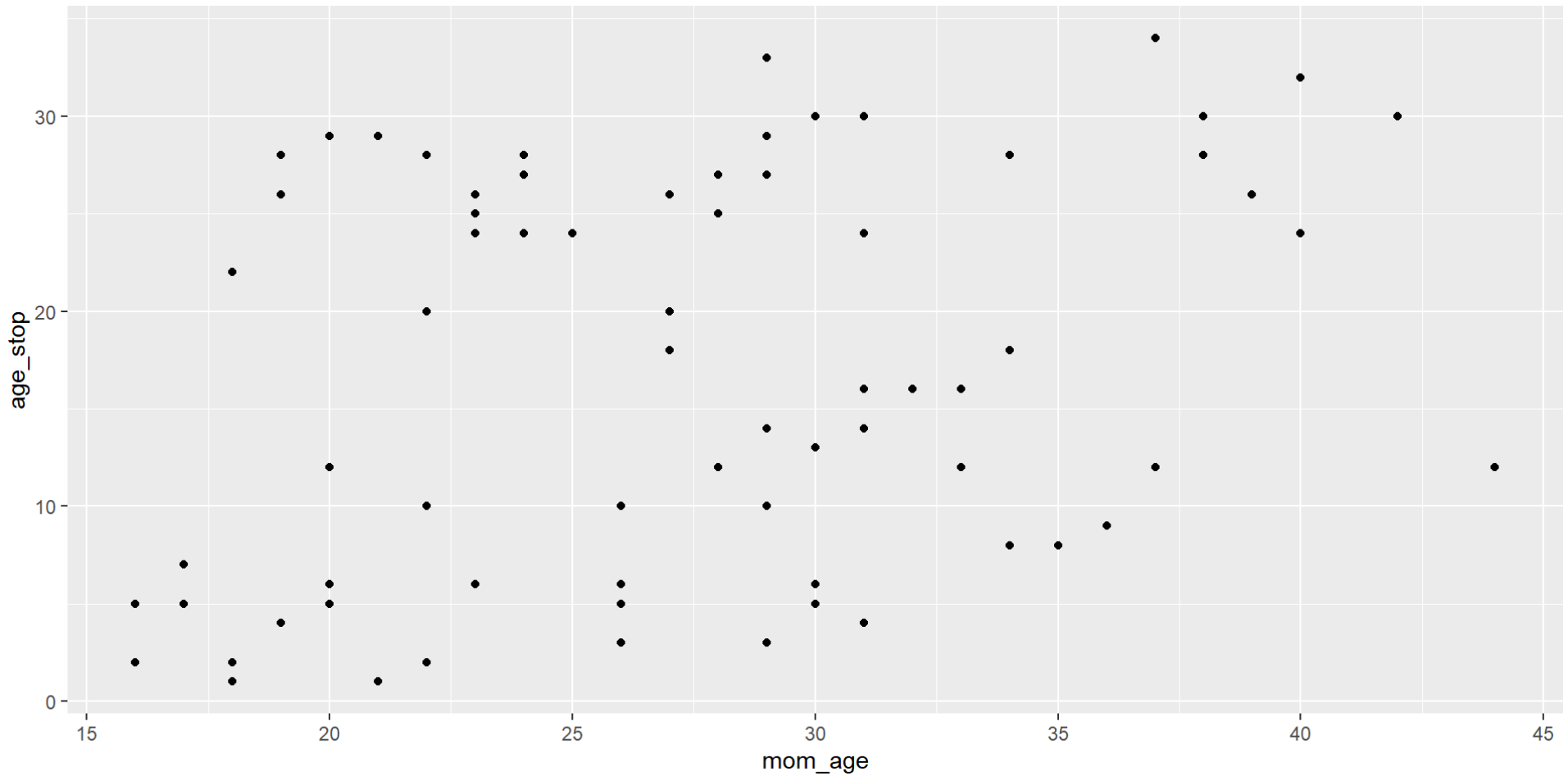
- The slope represents the estimated average change in Y when X increases by one unit.
- The intercept represents the estimated average value of Y when X equals zero.
- Terminology
  - X is the independent or predictor variable
  - Y is the dependent or outcome variable

Speaker notes

In linear regression, we use a straight linear to estimate a trend in data. We can’t always draw a straight line that passes through every data point, but we can find a line that “comes close” to most of the data. This line is an estimate, and we interpret the slope and the intercept of this line as follows:

Be cautious with your interpretation of the intercept. Sometimes the value X=0 is impossible, implausible, or represents a dangerous extrapolation outside the range of the data.

Simple regression example with interpretation, 1

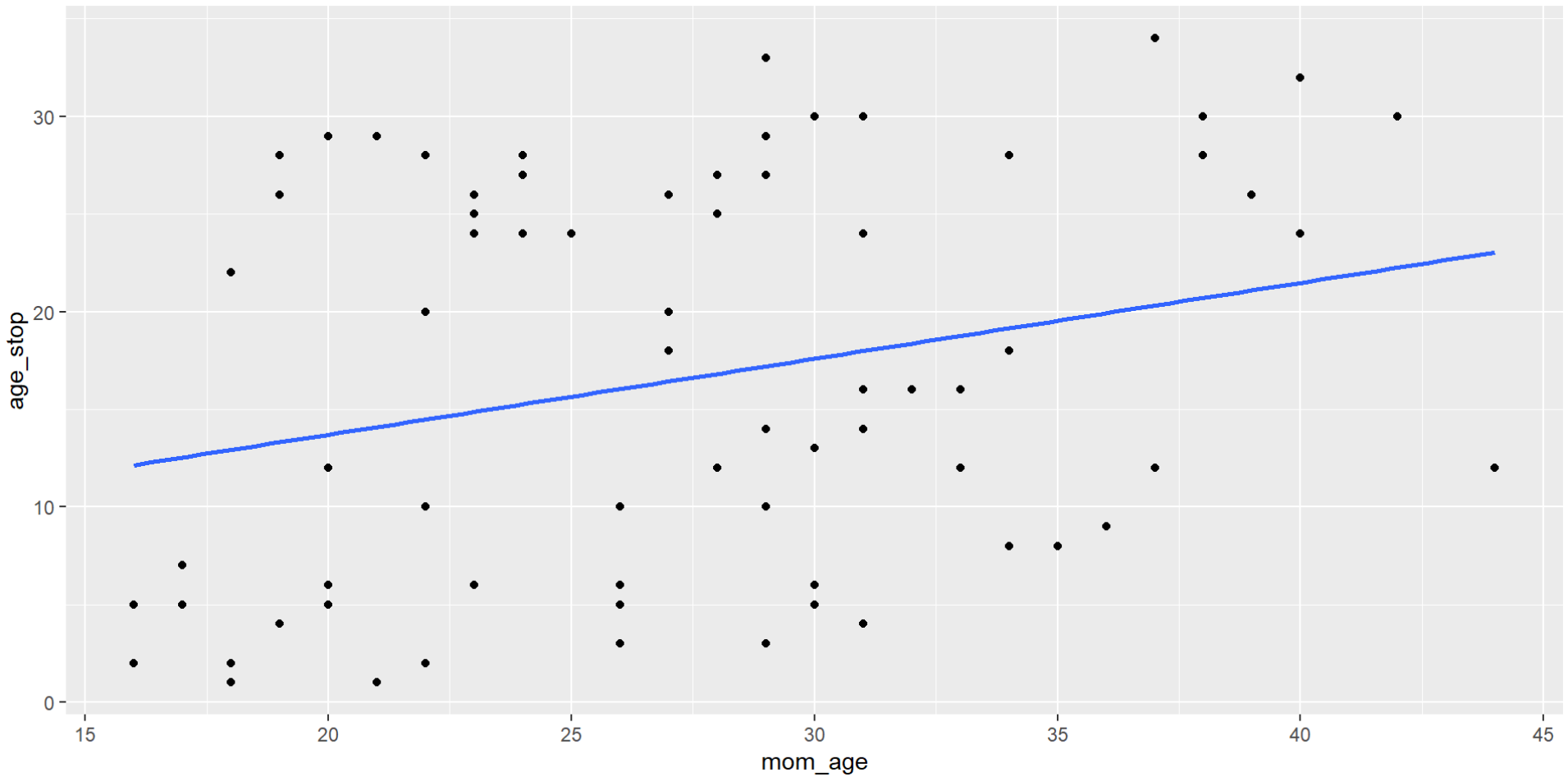


Speaker notes

Here is an illustration of a linear regression model. This data is from a study of breast feeding in pre-term infants. Successful breast feeding is more difficult for a pre-term infant because the mother goes home from the birth hospital before the infant. The independent variable, X, is the mother’s age. The dependent variable is the age at which the infant stopped breast feeding. The goal is to reach at least six months of breast feeding.

Notice a weak trend here. Older mother’s seem to do a bit better than younger mothers, but there are some 20 year old moms who breast feed for quite a long time and some 40 year old moms who stop breast feeding early. Still, there is a tendency for older moms to breast feed longer than younger moms.

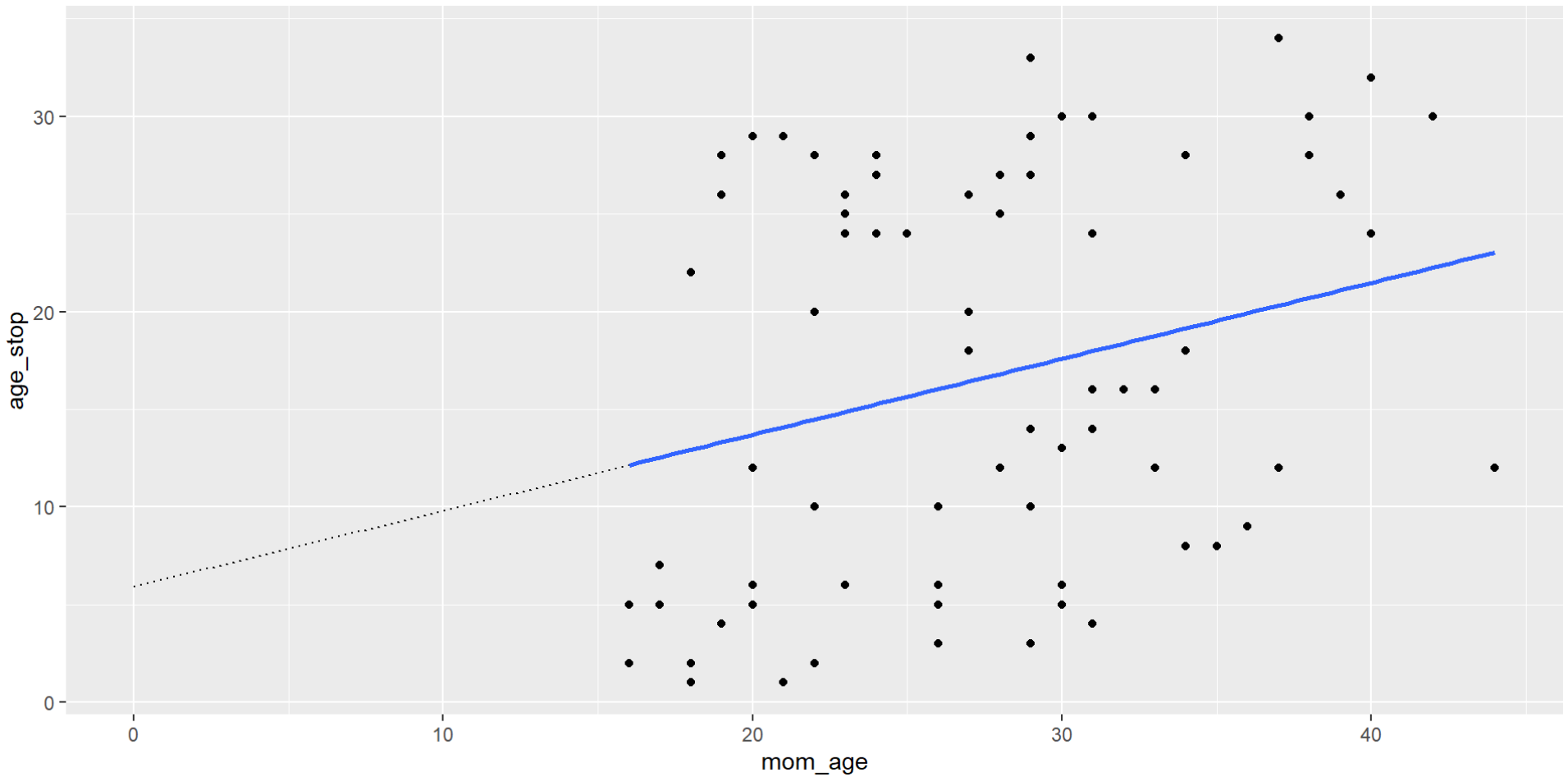
Simple regression example with interpretation, 2



Speaker notes

If you ask R to add a regression line, it stops at the range of the data. No dangerous extrapolations here!

Simple regression example with interpretation, 3



Speaker notes

Here is a modification of the graph that expands the limits of the X axis to include X=0. This graph also extends the line beyond the range of the data all the way down to X=0. I used a dotted line and a different color to emphasize that this is an extrapolation beyond the range of the data.

The graph shown below represents the relationship between mother’s age and the duration of breast feeding in a research study on breast feeding in pre-term infants.

The regression coefficients are shown below. The intercept, 6, is represented the estimated average duration of breast feeding for a mother that is zero years old. This is an impossible value, so the interpretation is not useful. What is useful, is the interpretation of the slope, approximately 0.4. The estimated average duration of breast feeding increases by 0.4 weeks for every extra year in the mother’s age.

Simple regression example with interpretation

```
Call:
lm(formula = age_stop ~ mom_age, data = bf)

Coefficients:
(Intercept)      mom_age
      5.920        0.389
```

Speaker notes

The actual values are pretty close to the rough estimates that we got from the graph.

Predicted values, 1

- How long would you expect a 20 year old mom to breast feed?

```
# A tibble: 5 × 2
  mom_age age_stop
    <dbl>    <dbl>
1      20        5
2      20       29
3      20        6
4      20       NA
5      20       12
```

Speaker notes

Easy way out is to find a 20 year old mother in the data. That is actually not such a good idea. But there are five of them, four if you are only counting non-missing values.

In a different dataset maybe you have data on 19 and 21 year olds but no 20 year olds.

If you predict using a linear regression model, you are incorporating information from all mothers into your prediction, not just 20 year old mothers. Unless there are some major problems with the regression model, this is a much better choice.

Predicted values, 2

- For an existing value in the data, X_i
  - $\hat{Y}_i = b_0 + b_1 X_i$
- For a new value of X
  - $\hat{Y}_{new} = b_0 + b_1 X_{new}$
  - Do not predict outside the range of X values

Speaker notes

To make a prediction at an existing value in the dataset, use the first formula. If you want to make a prediction at a value that is not in your dataset, use the same formula, but notice that the subscript is “new” to emphasize that this is a new value not seen before in the data. Be careful here. It is okay to make predictions inside the range of the X values. In the breast feeding example that I have been talking about, it is okay to make predictions for a mother between the ages of 16 and 44, but making predictions for a 14 year old mother or a 50 year old mother is risky. You could be making a dangerous extrapolation.

Why predict for a value you already have seen?

- Future Y may differ from previous Y
- $\hat{Y}_i$ is more precise
- Comparison of $\hat{Y}_i$ to existing Y_i.

Predicted values, 3

Predicted age_stop = 5.92 + 0.389*20 = 13.7

```
# A tibble: 1 × 2
  mom_age .fitted
    <dbl>   <dbl>
1      20    13.7
```

Speaker notes

Here is the predicted value from R.

Residuals, 1

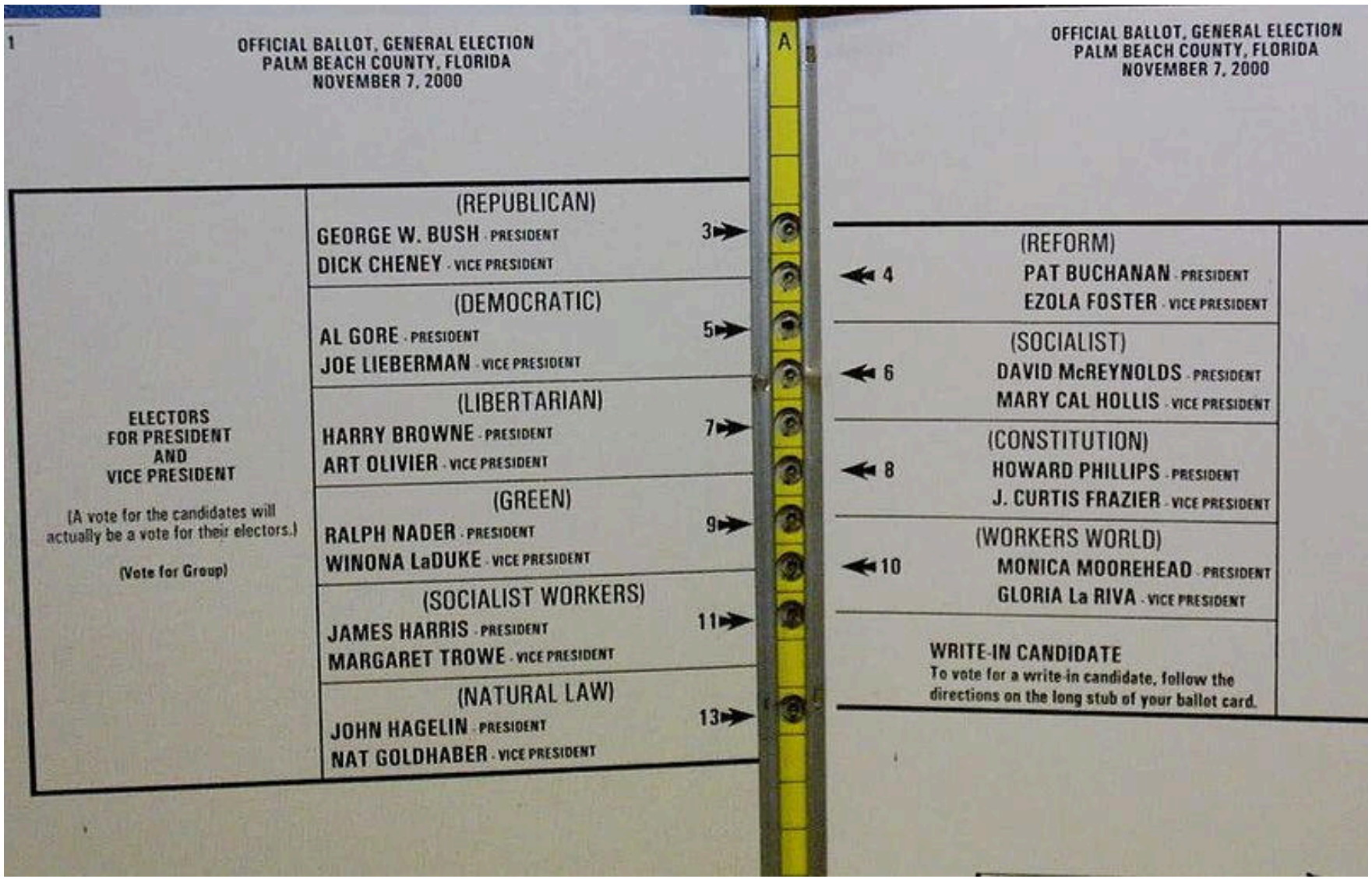
- $e_i = Y_i - \hat{Y}_i$
  - Residual = Observed - Predicted
- Very helpful in assessing assumptions

Speaker notes

The residual is also important. It represents the deviation between what was actually observed and what was predicted.

The residual is very helpful in assessing the assumptions needed for linear regression. This is a topic I will reserve for a later module.

Residuals, 2

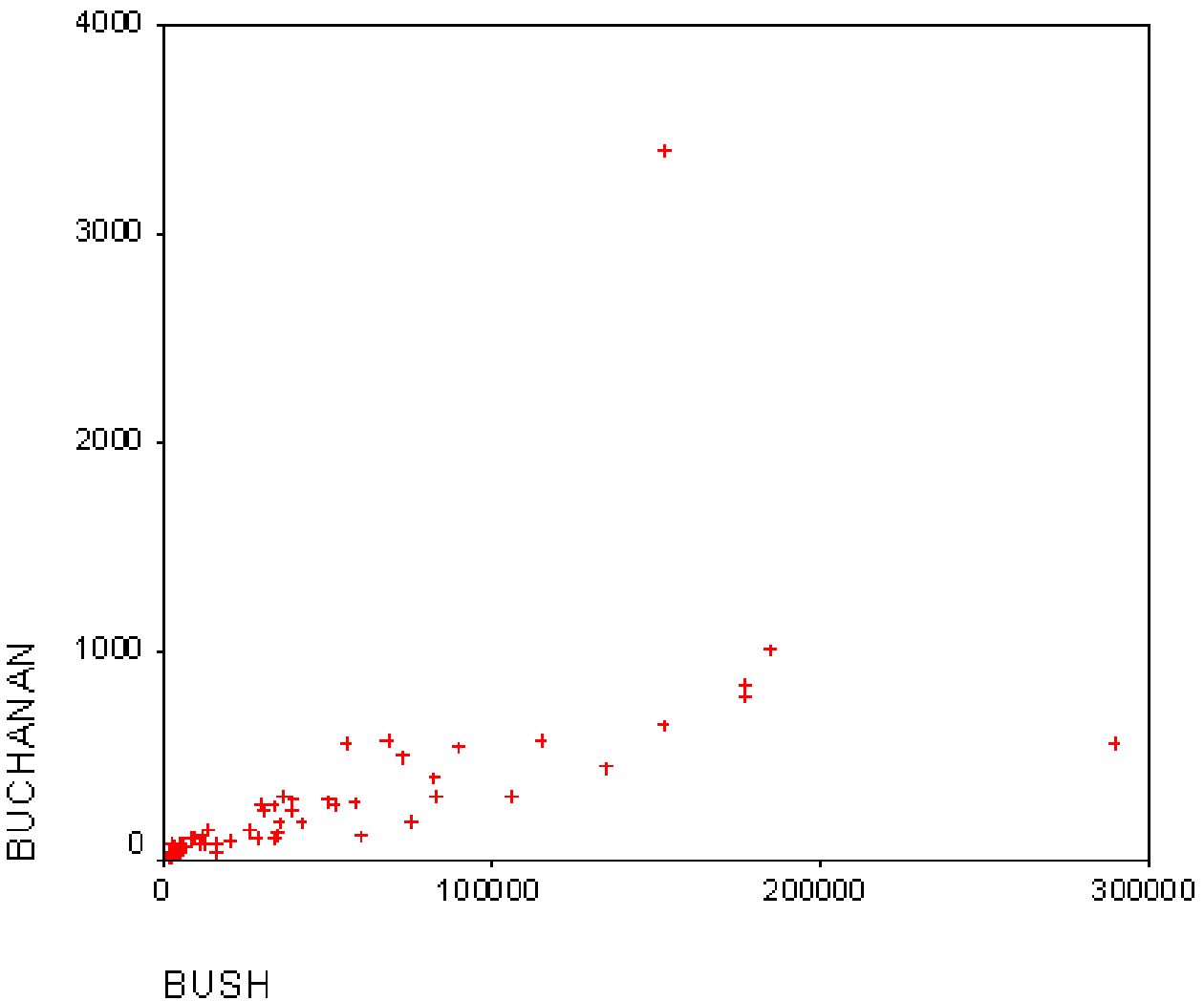

1
OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000
ELECTORS FOR PRESIDENT AND VICE PRESIDENT
(A vote for the candidates will actually be a vote for their electors.)
(Vote for Group)
(REPUBLICAN)
GEORGE W. BUSH - PRESIDENT
DICK CHENEY - VICE PRESIDENT
3
(DEMOCRATIC)
AL GORE - PRESIDENT
JOE LIEBERMAN - VICE PRESIDENT
5
(LIBERTARIAN)
HARRY BROWNE - PRESIDENT
ART OLIVIER - VICE PRESIDENT
7
(GREEN)
RALPH NADER - PRESIDENT
WINONA LaDUKE - VICE PRESIDENT
9
(SOCIALIST WORKERS)
JAMES HARRIS - PRESIDENT
MARGARET TROWE - VICE PRESIDENT
11
(NATURAL LAW)
JOHN HAGELIN - PRESIDENT
NAT GOLDHABER - VICE PRESIDENT
13
A
OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000
(REFORM)
4
PAT BUCHANAN - PRESIDENT
EZOLA FOSTER - VICE PRESIDENT
(SOCIALIST)
6
DAVID McREYNOLDS - PRESIDENT
MARY CAL HOLLIS - VICE PRESIDENT
(CONSTITUTION)
8
HOWARD PHILLIPS - PRESIDENT
J. CURTIS FRAZIER - VICE PRESIDENT
(WORKERS WORLD)
10
MONICA MOOREHEAD - PRESIDENT
GLORIA La RIVA - VICE PRESIDENT
WRITE-IN CANDIDATE
To vote for a write-in candidate, follow the directions on the long stub of your ballot card.

Speaker notes

One interesting (and controversial) application of residuals is in an analysis of election results in Florida during the presidential race of 2000.

Image taken from: http://en.wikipedia.org/wiki/File:Butterfly_large.jpg.

Residuals, 3



Speaker notes

There are several ways to look at the data, and they all show that the vote for Buchanna in Palm Beach County was unusual. Here is a graph looking at the votes for Bush in each county of Florida on the X-axis and the votes for Buchanan in that same county.

Notice that there is a strong positive trend. Counties with lots of votes for Bush tended to produce lots of votes for Buchanan. Not one to one of course, because votes for Bush were in the hunderds of thousands, while votes for Buchanan were two orders of magnitude smaller.

The relationship occurs in part because both Bush and Buchanan were candidates that appealed to the same type of voter, someone on the conservative side of the voting spectrum. More importantly, though, is that bigger counties would tend to produce more votes for both candidates.

Notice the one point very hight in the middle of the graph. This is Palm Beach County. Buchanan got over 3,000 votes there and

Residuals, 4

- Votes (Buchanan) = 45.3 +0.0049 * Votes (Bush)
  - The estimated average number of votes for Buchanan increase by 1/200 for every increase of one vote for Bush.

Residuals, 5

- In Palm Beach County
  - Votes (Bush) = 152,846
  - Predicted Votes (Buchanan) = 797
    - 45 + 0.0049 * 152,846
  - Actual Votes (Buchanan) = 3,407
- Residual = 3,407 - 797 = 2,610

Speaker notes

We can compute a predicted number of votes for Buchanan for each county by using the above equation. Palm Beach County had 152,846 votes for Bush. So the regression model would predict that Buchanan should get:

45 + 0.0049 * 152,846 = 797.

Thus, if the relationship observed across the entire state held exactly in Palm Beach County, then we would estimate the vote count for Buchanan to be 797.

There were actually 3,407 votes recorded for Buchanan, which is quite a discrepancy from what we predicted. The residual, the difference between what we observed and what would be predicted by the regression model is:

3,407 - 797 = 2,610.

One possible interpretation is that this discrepancy represents an estimate of the number of people who voted incorrectly for Buchanan. Such an interpretation would have to consider other possibilities, though. Is there something unique about Palm Beach County that would cause that county to vote in disproportionate numbers for Buchanan? Buchanan does indeed have some relatives in the area, and although they do not number in the thousands, perhaps they exerted some influence on their community.

Other information might tend to corroborate that a large number of votes were cast erroneously for Buchanan. Some of the highest vote counts for Buchanan were in precincts that were most heavily Democratic. There were also a large number of complaints received by the election board prior to anyone knowing how close the vote count in Florida would be.

There are other models that have been considered for the Palm Beach County vote, and most of them show a similar size discrepancy between the observed vote and the vote that would be predicted the regression model. It would set a dangerous precedent, of course, to use a statistical model to adjust vote counts, so this example is more for understanding what might have gone wrong and the magnitude of the error made.

The general lesson here is that the residual represents the discrepancy between what the actual data value and what a linear regression model predicts. When the residual is large, there is reason to investigate. Please, please, please don’t toss out a data value, though, just because it has an extreme residual.

Residuals, 6

```
# A tibble: 4 × 5
  .rownames mom_age age_stop .fitted .resid
  <chr>       <dbl>    <dbl>   <dbl>  <dbl>
1 8              20        5    13.7  -8.70
2 40             20       29    13.7  15.3 
3 44             20        6    13.7  -7.70
4 67             20       12    13.7  -1.70
```

Break #1

- What you have learned
  - Interpretation of linear regression coefficients
- What’s coming next
  - Computing linear regression in R

Location of data dictionary and code

- breast-feeding-preterm.yaml
- simon-5501-05-bf.qmd

Break #2

- What you have learned
  - Computing linear regression in R
- What’s coming next
  - The least squares principle

The population model

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, N$
  - ϵ_i is an unknown random variable
    - Mean 0, standard deviation, σ
    - Often assumed to be normal
  - β_0 and β_1 are unknown parameters
  - b_0 and b_1 are estimates from the sample

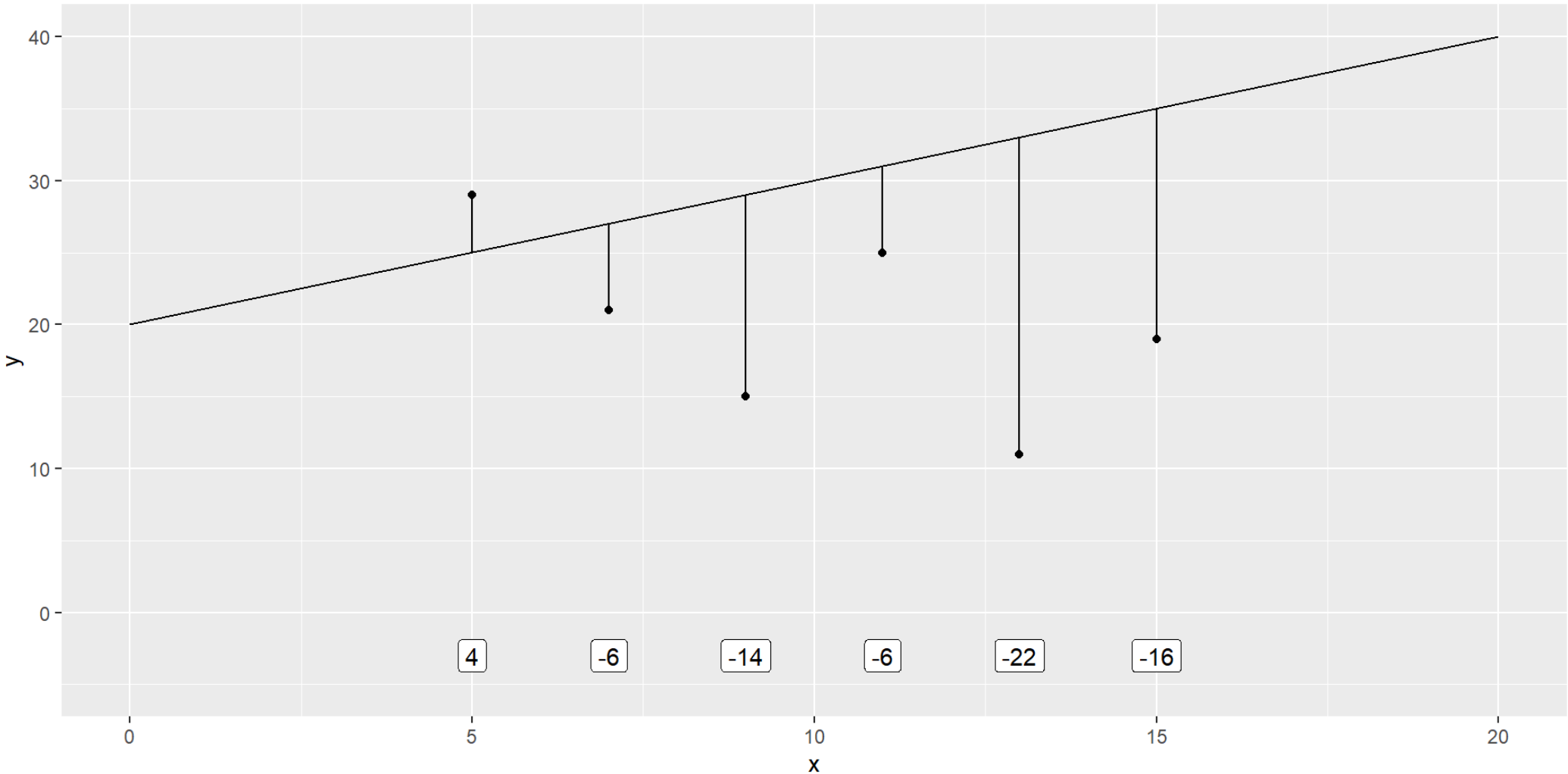
Speaker notes

Add note.

Least squares principle, 1

- Collect a sample
  - $(X_1, Y_1),\ (X_2, Y_2),\ \ldots\ (X_n, Y_n)$
- Compute residuals
  - $e_i = Y_i - (b_0 + b_1 * X_i)$
  - Choose b_0 and b_1 to minimize Σe_i^2

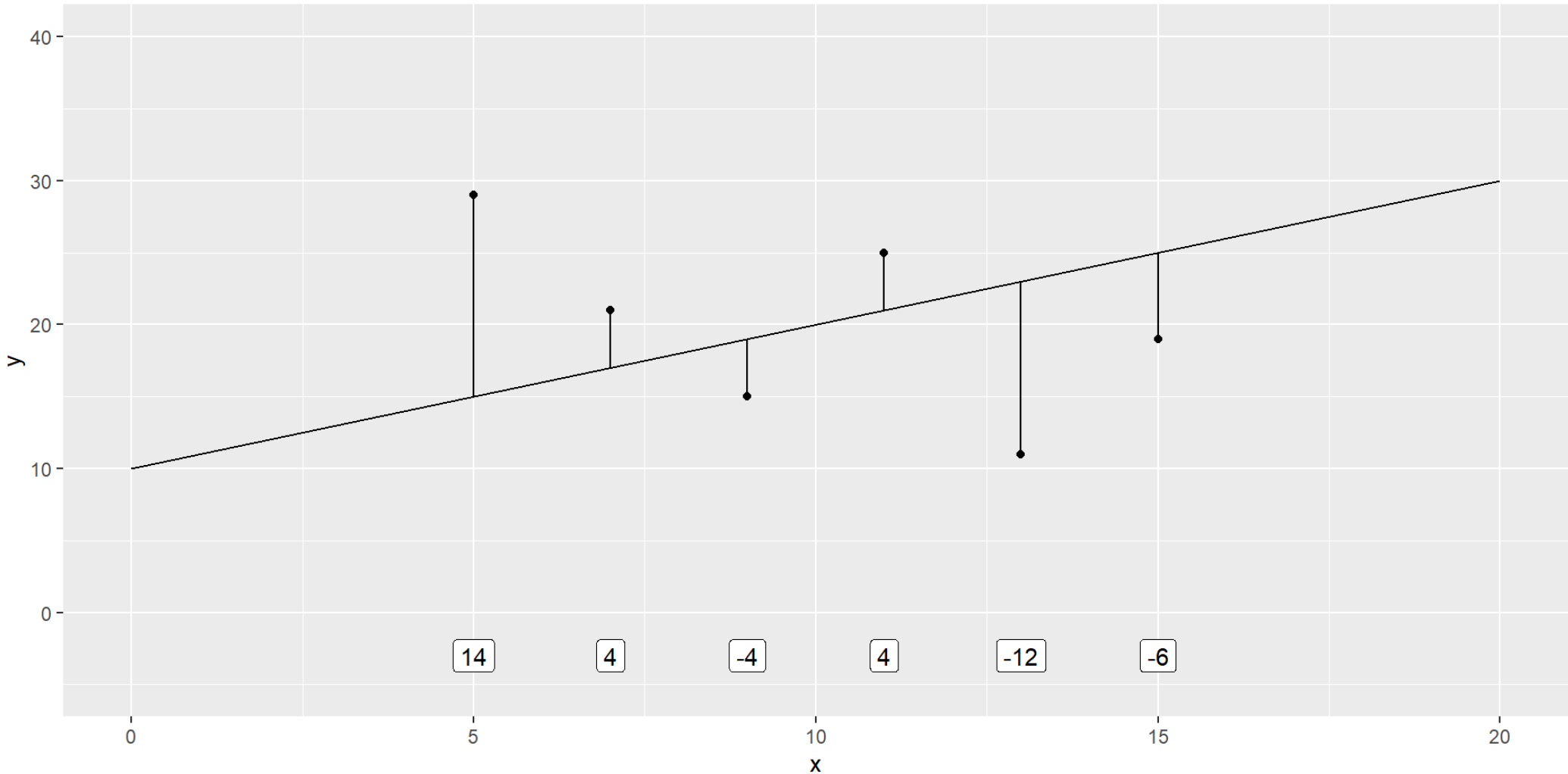
Least squares principle, 2



Speaker notes

Here’s one possible choice for the regression line, an intercept of 20 and a slope of 1. It does an okay job for the leftmost two datapoints, but really overpredicts for the rest of the data.

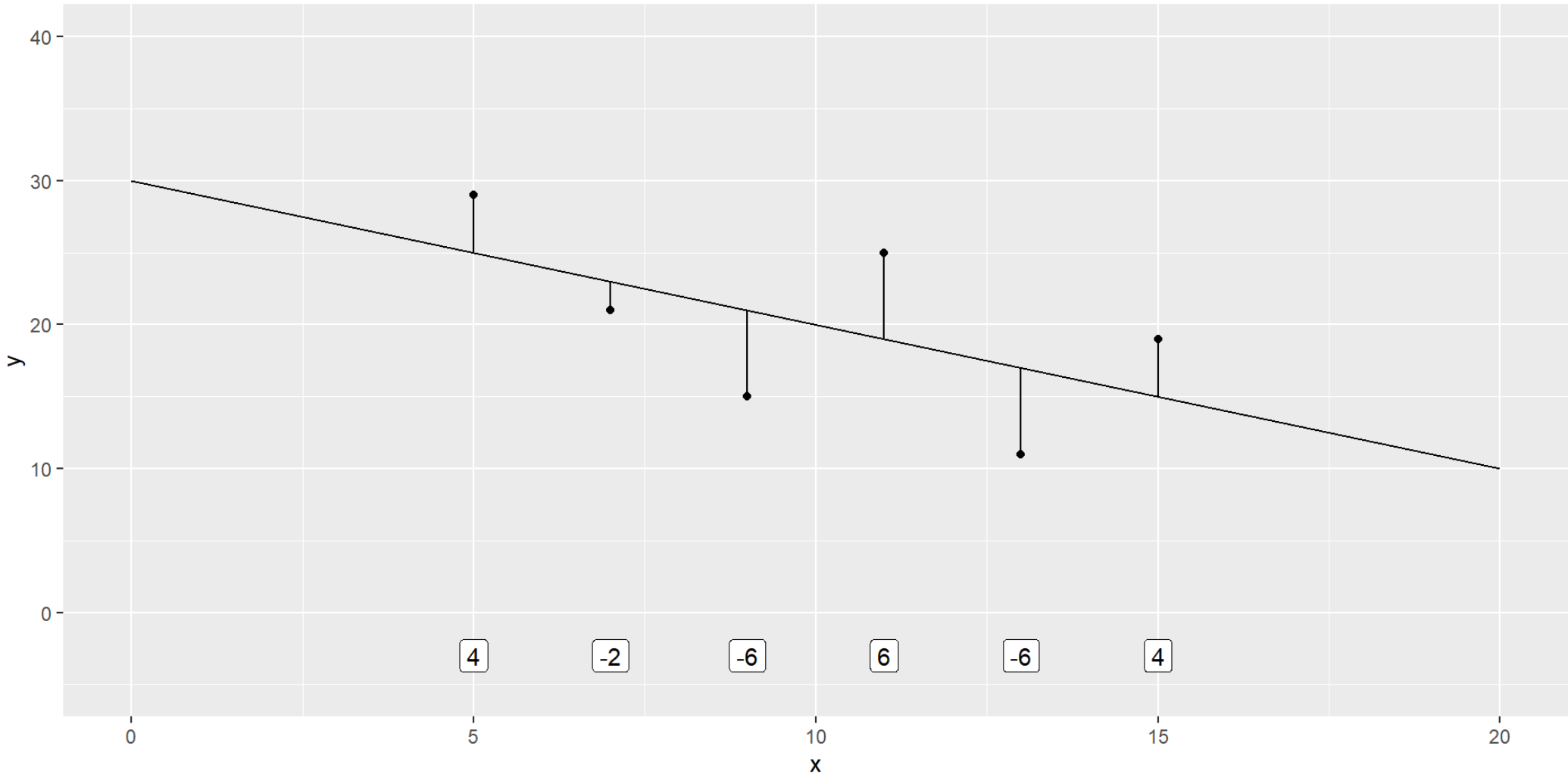
Least squares principle, 3



Speaker notes

Here’s an improvement, using an intercept of 10 instead of 20. You can see that there might still be some room for improvement, if we could get the line a bit closer to the first and the fifth data points.

Least squares principle, 4



Speaker notes

Twist the line so the slope is -1 instead of +1 and shift the intercept all the way up to 30. This does even better than the first two lines.

It turns out that you can’t do any better than this. You might be the line to be a bit closer to one of the data points, but you’d do a lot worse across all of the other data points.

Least squares principle, 5

- General solution
  - $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$
  - $b_0 = \bar{Y} - b_1\bar{X}$
- Notice the similarity between b_1 and r
  - $b_1 = r\frac{S_Y}{S_X}$

Relationship to the correlation coefficient

- Recall from the previous module
  - $Cov(X, Y) = \frac{1}{n-1}\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$
  - $r_{XY} = \frac{Cov(X,Y)}{S_X S_Y}$
- This implies that
  - $b_1 = r_{XY}\frac{S_Y}{S_X}$

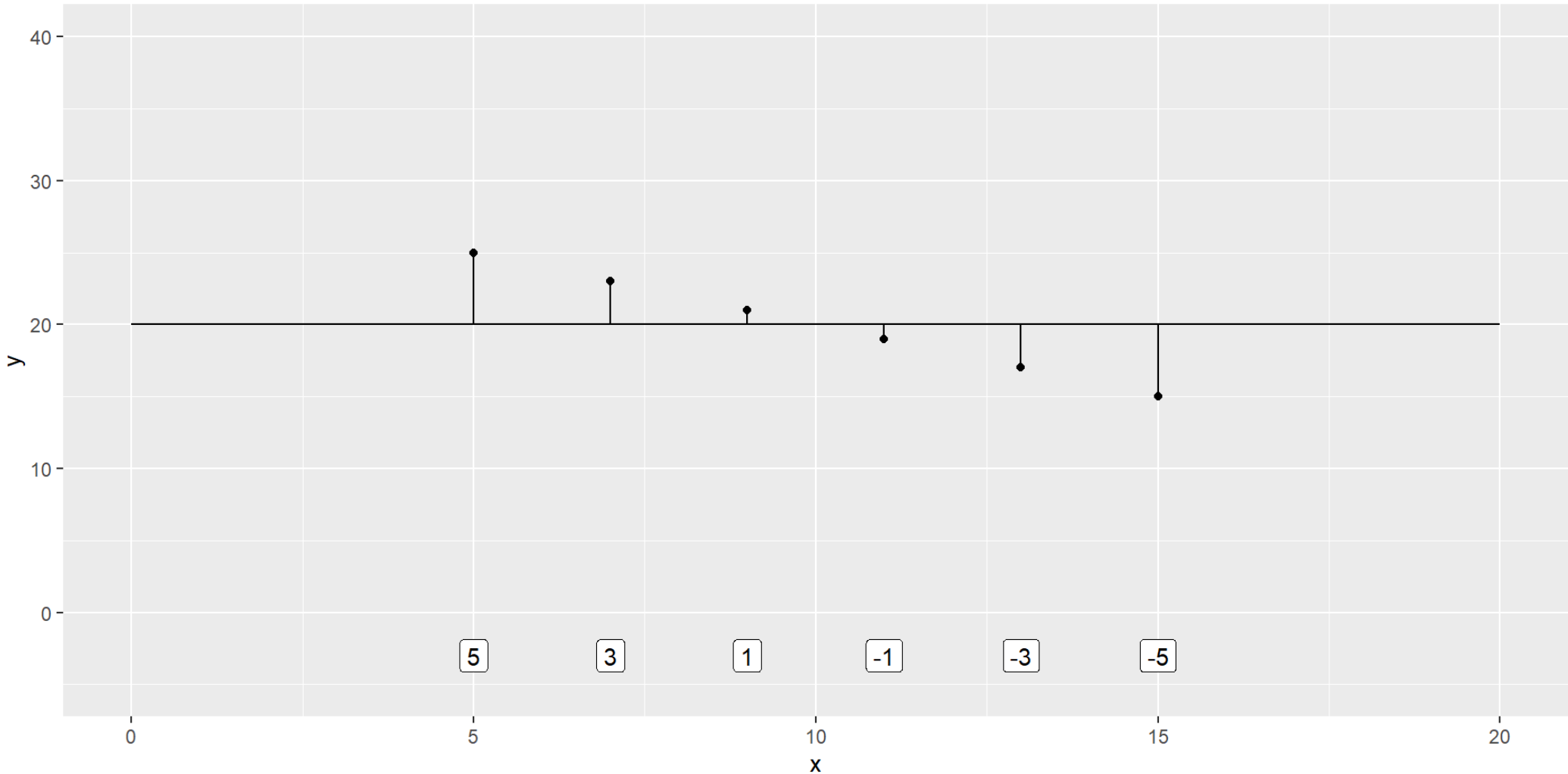
Important implications

- r_{XY} is unitless, b_1 is Y units per X units
- $r_{XY} > 0$ implies $b_1 > 0$
- $r_{XY} = 0$ implies $b_1 = 0$
- $r_{XY} < 0$ implies $b_1 < 0$
  - and vice versa

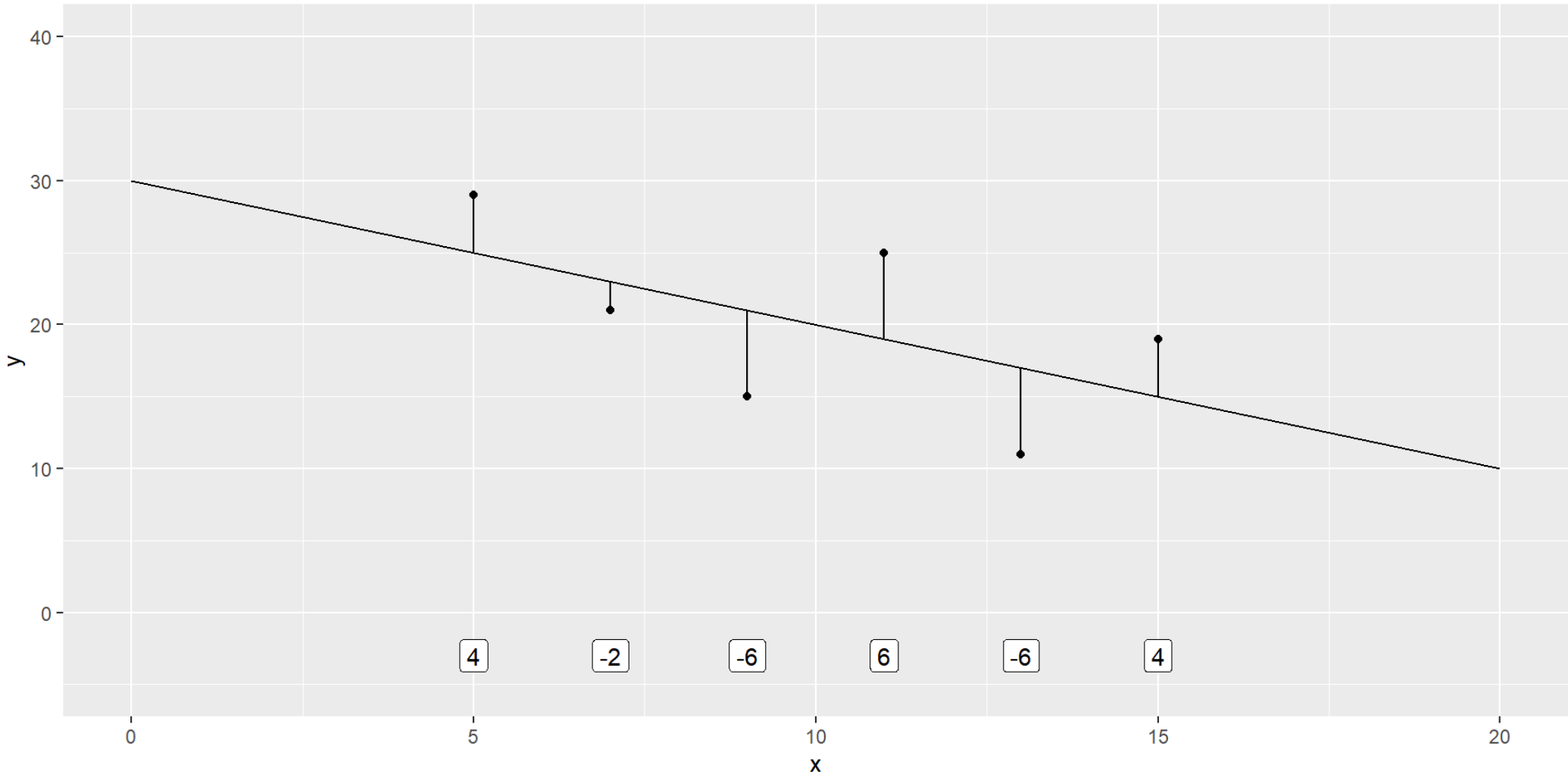
Break #3

- What you have learned
  - The least squares principle
- What's coming next
  - The analysis of variance table

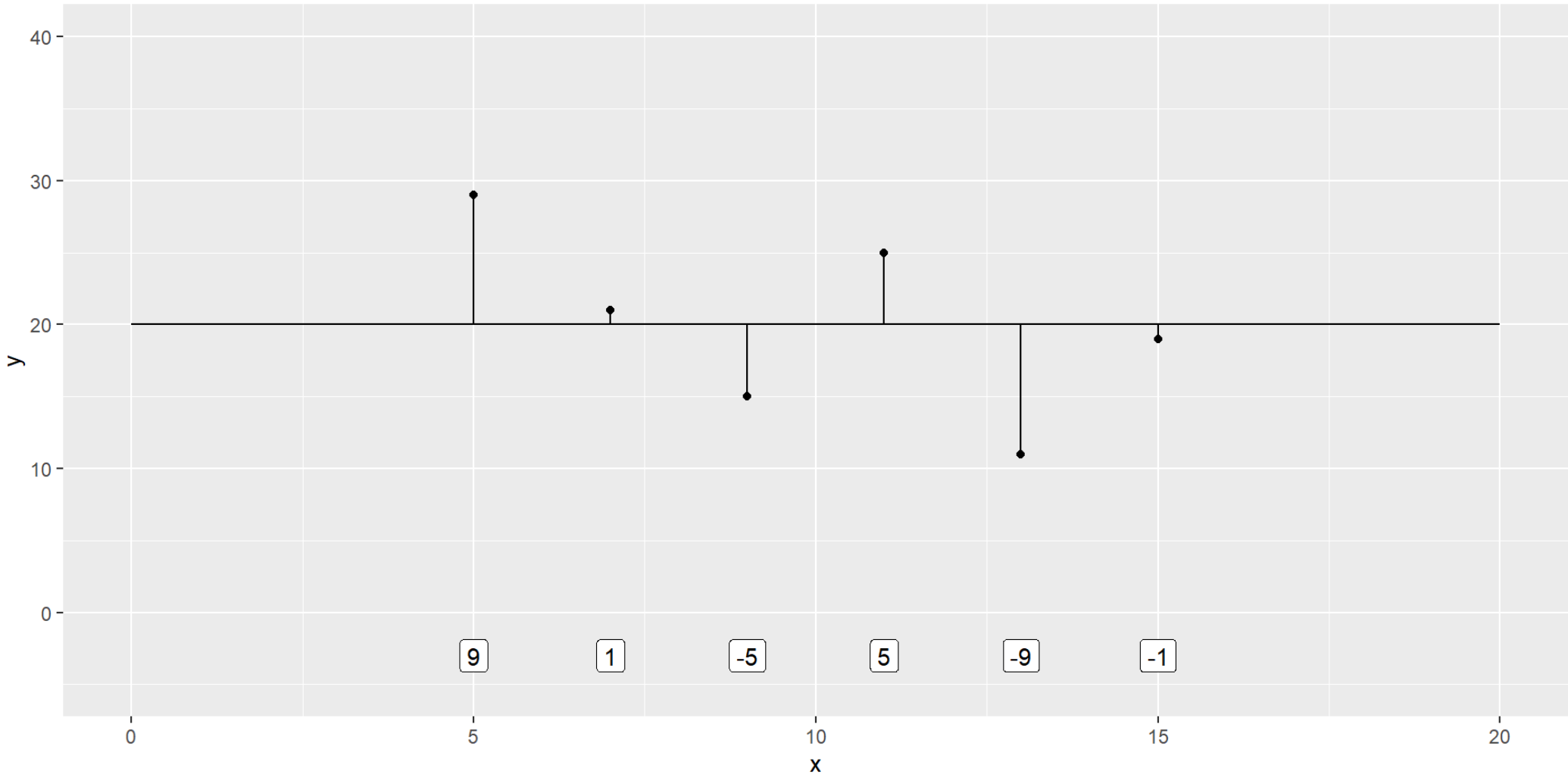
Sum of squares regression



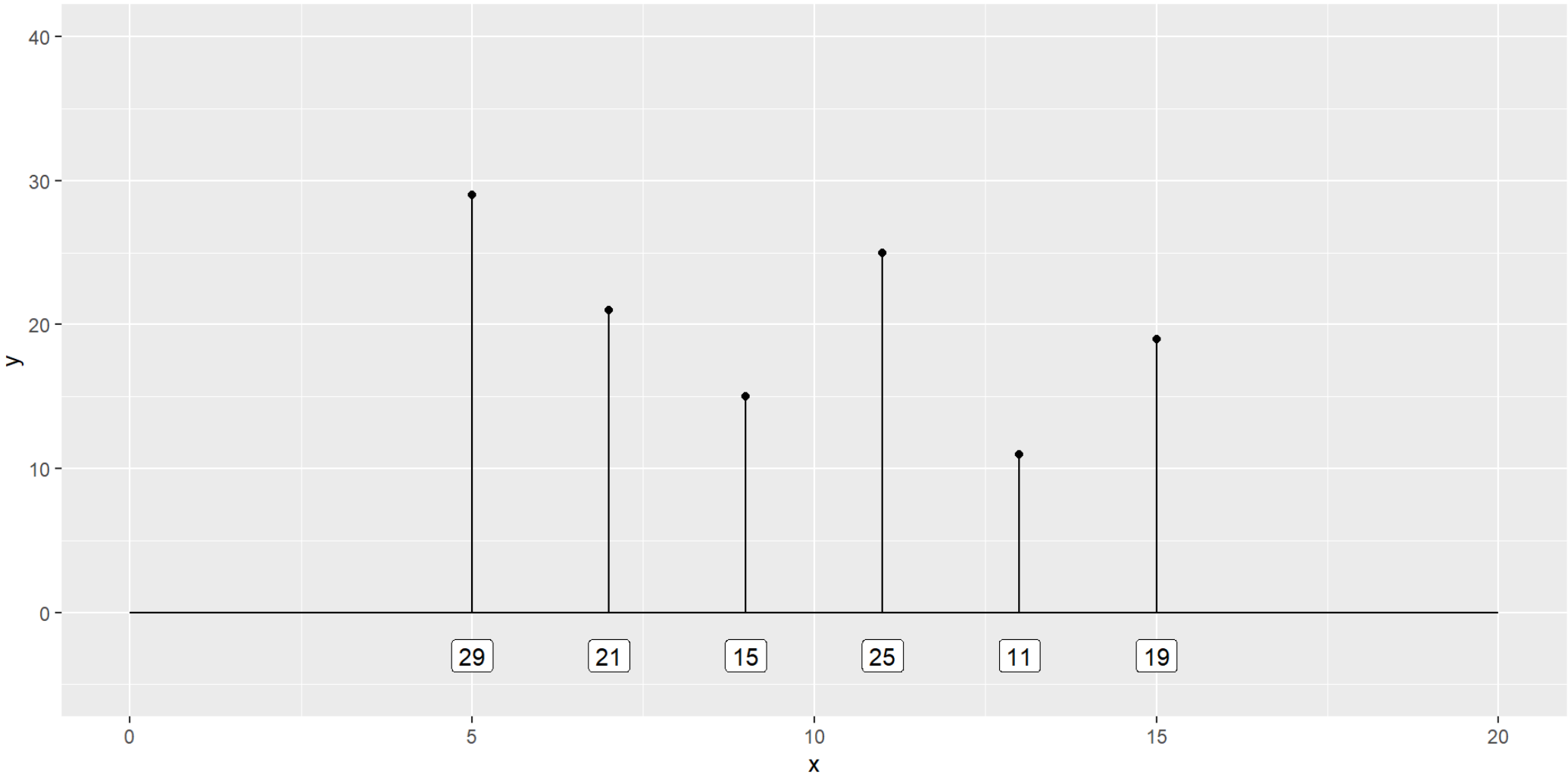
Sum of squares error



Sum of squares total / corrected total



Sum of squares total (uncorrected)



ANOVA table for linear regression

	SS	df	MS	$F-ratio$
$Regression$	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
$Error$	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
$Total$	SST	$n-1$		

Analysis of variance table in R

```
Analysis of Variance Table

Response: age_stop
           Df Sum Sq Mean Sq F value  Pr(>F)
mom_age     1  570.0  569.99  5.7531 0.01879 *
Residuals 80 7925.9   99.07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

The analysis of variance table looks different in different software programs and it differs from publication to publication. Notice for example that R does not include a row for total. SPSS, in contrast includes both a corrected total and an uncorrected total.

R-squared

- SST, total variation, is split into
  - SSR, explained variation, and
  - SSE, unexplained variation
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - $0 < R^2 < 1$
  - Proportion of explained variation
    - $R^2 > 0.5$ strong
    - $0.1 < R^2 < 0.5$ weak
- $R^2 = r^2$

Speaker notes

You can compare the three sources of variation. SSR is explained variation, variation in the predicted values. SSE is unexplained variation. These add up to total variation. Either explained variation divided by total variation or 1 minus unexplained variation divided by total variation is a measure of the strength of the relationship. If you can explain half of the variation (R2=0.5), that's evidence of a strong relationship. Anything between 10% and 50% is evidence of a weak relationship.

There is a simple mathematical relationship. Square the correlation between X and Y to get R-squared.

R-squared calculation in R

```
[1] 0.06708949
```

Speaker notes

Here is the value of R-squared for the breast feeding example. This is a very weak relationship.

F-ratio, 1

- $F = \frac{MSR}{MSE}$

Speaker notes

The F-ratio, MSR divided by MSE. If this value is close to 1, you would accept the null hypothesis. The signal (MSR) is comparable to the noise (MSE). If you see a large positive ratio, that implies that the signal is much stronger than the noise. A large positive ratio would cause you to reject the null hypothesis.

Alternately look at the p-value. If the p-value is large, you should accept the null hypothesis.

I'll talk a lot more about p-values in several of the upcoming modules.

Break #4

- What you have learned
  - The analysis of variance table
- What’s coming next
  - Computing the analysis of variance table in R

Location of code

- simon-5501-05-bf.qmd

Break #5

- What you have learned
  - Computing the analysis of variance table in R
- What's coming next
  - Confidence interval for the slope parameter

Confidence intervals, 1

- Population model
  - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ i = 1, \ldots, N$
- Sample estimates
  - $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$
  - $b_0 = \bar{Y} - b_1 \bar{X}$

Confidence intervals, 2

- Standard error (se)
  - $se(b_1) = \sqrt{\frac{MSE}{(n-1)S_x^2}}$
- Confidence interval for β_1
  - $b_1 \pm t\ se(b_1)$

Confidence intervals, 3

```
                  2.5 %    97.5 %
(Intercept) -3.19546976 15.035265
mom_age      0.06625878  0.711827
```

Hypothesis test, 1

- $H_0: \ \beta_1 = 0$
- $H_1: \ \beta_1 \neq 0$
  - Accept H_0 if $T = \frac{b_1}{se(b_1)}$ is close to zero
  - or Accept H_0 if the confidence interval includes zero
  - or Accept H_0 if the p-value is large

Speaker notes

The hypotheses involve the population parameter, β_1. To test this hypothesis, compare the sample statistic, b_1, to its standard error. If that ratio is close to zero, then you would accept the null hypothesis. If you see extreme values, very large negative or very large positive values, then you should reject the null hypothesis.

Hypothesis test, 2

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    5.92      4.58       1.29  0.200
2 mom_age        0.389     0.162      2.40  0.0188
```

Speaker notes

Here is the output for the t-test in R.

Why two different ways to test?

- $F = T^2$
  - Accept H_0 if F is close to one
  - Accept H_0 if T is close to zero
- For both tests
  - Accept H_0 if p-value is large
- F and T differ in more complex settings
  - F is a global test of all variables
  - T is series of separate tests of individual variables

Speaker notes

Although it seems redundant to have two different ways to test the same hypothesis, it actually becomes important for more complex settings where you are trying to predict an outcome using two or more independent variables.

Break #6

- What you have learned
  - Confidence interval for the slope parameter
- What's coming next
  - Computing confidence intervals in R

Location of code

- simon-5501-05-bf.qmd

Break #6

- What you have learned
  - Computing confidence intervals in R
- What’s coming next
  - Your homework

Location of programming assignment

- simon-5501-05-directions.md

Summary

- What you have learned
  - Interpretation of linear regression coefficients
  - Computing linear regression in R
  - The least squares principle
  - The analysis of variance table
  - Computing the analysis of variance table in R
  - Confidence interval for the slope parameter
  - Computing confidence intervals in R
  - Your homework