Psych 10: Intro to Statistics

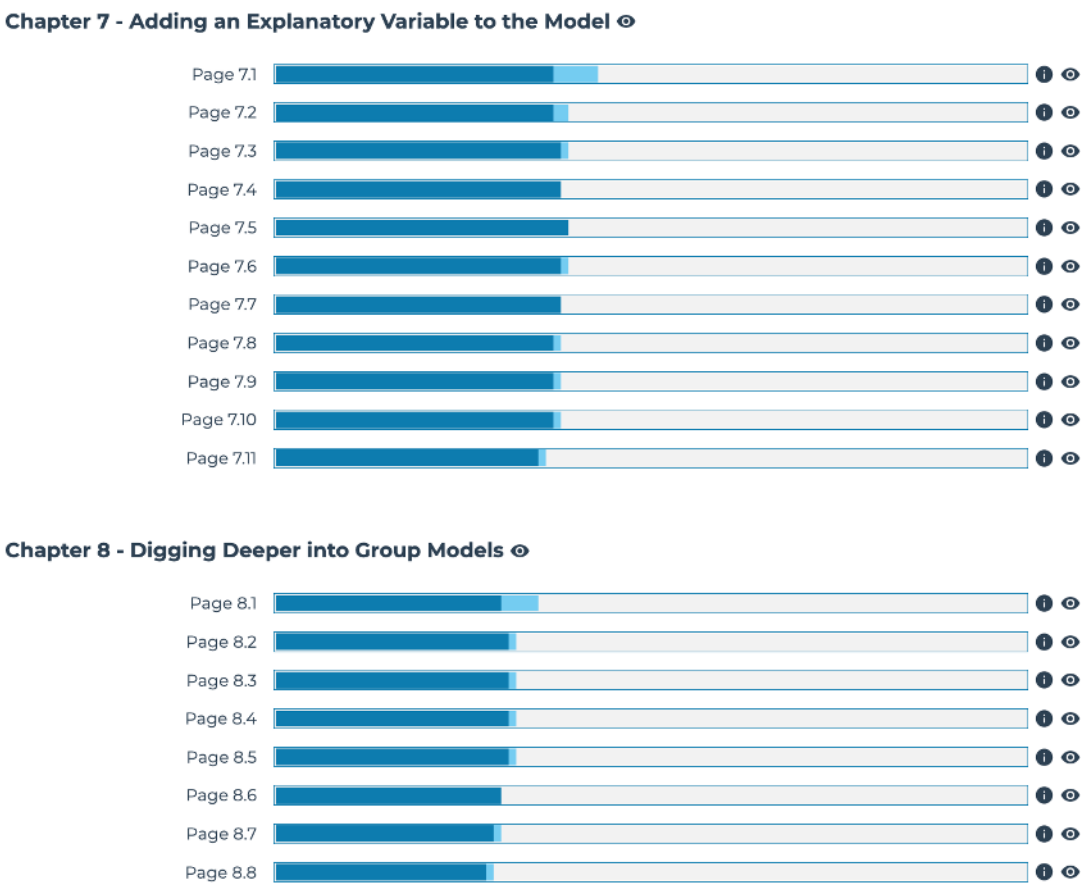
Lecture 10 · Linear regression · Mon Oct 26, 2026

Prof. Judith Fan · Fall 2026

Reminders & Announcements

- CourseKata **Chapter 9** is due this Friday.
- Section this week: **Milestone 5**, the preregistration, due Sunday, November 1 on Gradescope.
- Friday: **Case #5**, in your tutorial session. Check your discussion section's Slack channel for your time slot. Bring a pen.
- The **practice exam** is two weeks from today, Monday, November 9, in class. The exam is Wednesday, November 18.
- We'll be holding a **review session** on Wed, November 11th. Check your section's Slack channel for a link to a Google Form where you can suggest topics to cover in that session; I'll be collecting suggestions through November 4th.

Every part of this course is built on CourseKata



- Lecture, the worksheets, and the Friday cases are all built on the CourseKata chapters. That is how the parts of this course connect.
- Each bar is the share of the class that has completed that page of Chapters 7 and 8: a little over a third.
- If you are behind, catching up on CourseKata is the best preparation for the worksheets, the cases, and the exam.

Coming up this Wednesday

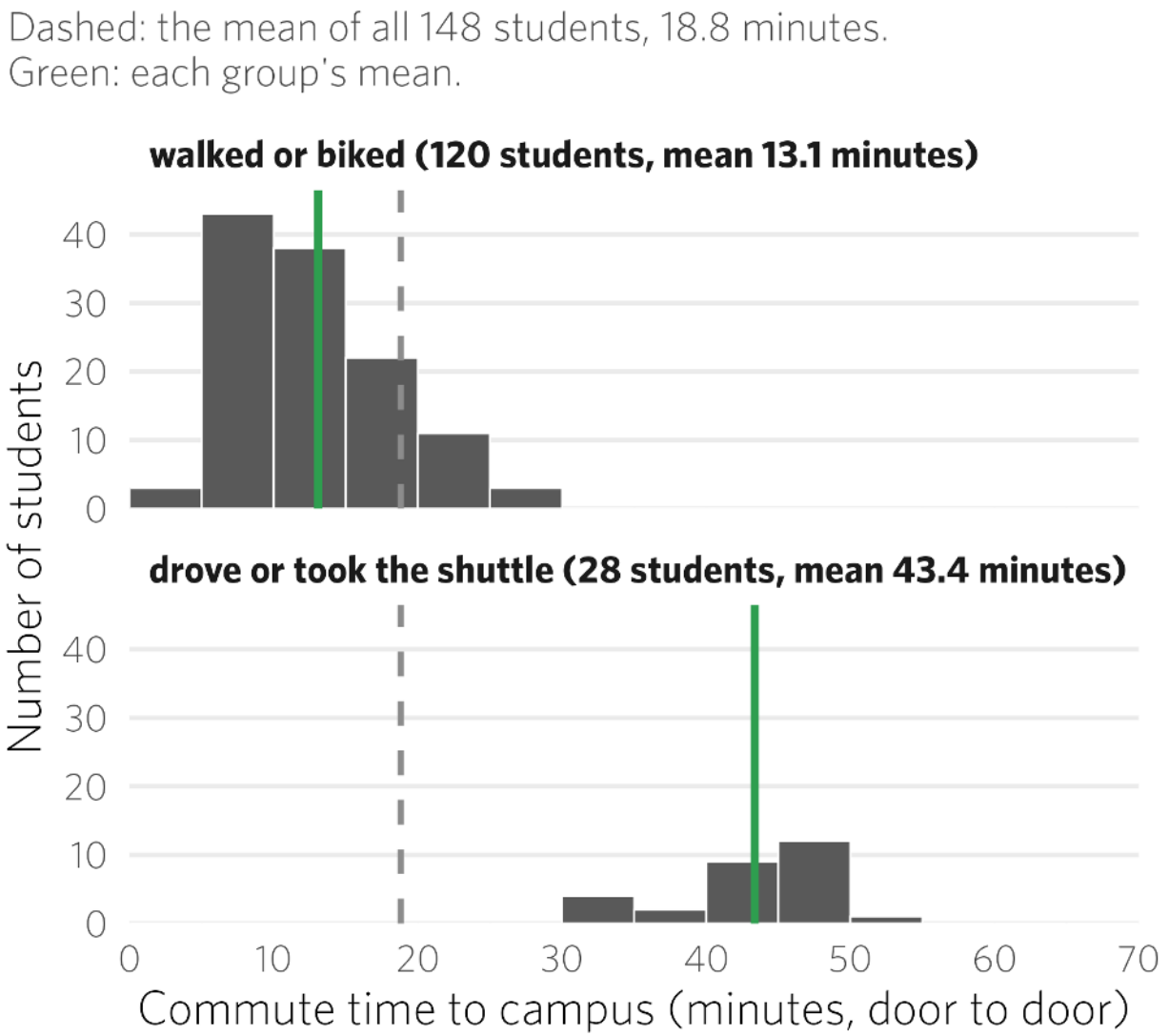
- This coming Wednesday, we'll have another guest lecture from a member of the teaching team. This time, Jane Stephenson, on tidying and wrangling data!
- I'll also be distributing a mid-quarter survey to collect feedback from you all about your experience in the course so far.
- And please suggest topics for the November 11th review session using the Google Form linked in your section's Slack channel. I'll be collecting suggestions through November 4th.

Today

Linear regression

1. The cost of a parameter
2. Could the difference be chance?
3. The slope and the intercept

Last Monday's model of commute times



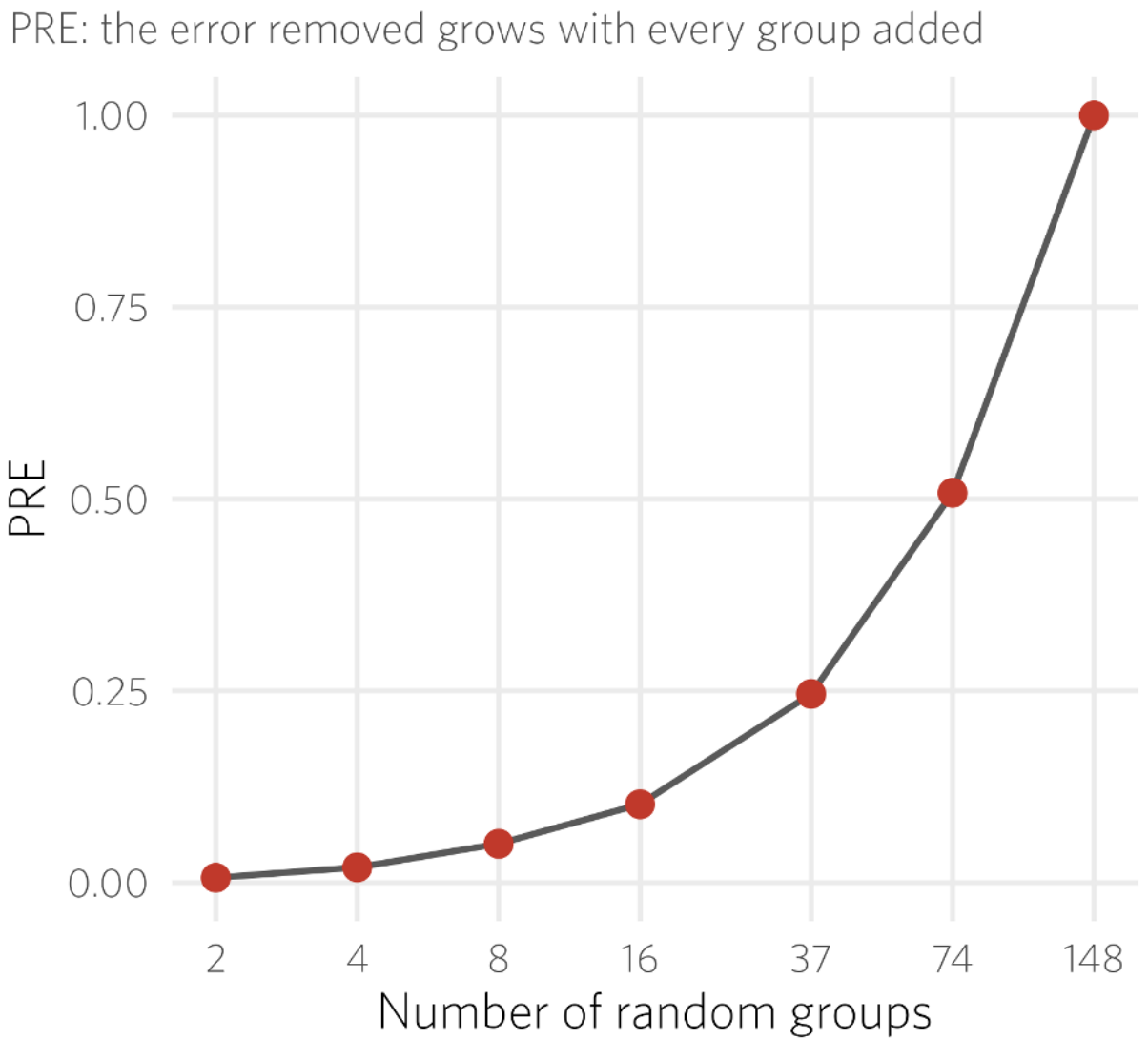
- The 148 commute times from the Week 2 survey, split by how each student got to campus: walkers and bikers average 13.1 minutes, drivers and shuttle riders 43.4.
- The model's prediction for each student is their group's mean; the residual is the rest.
- The two groups removed 82% of the error around the overall mean of 18.8 minutes: PRE = 0.82.

Would more groups make a better model?

- Two groups removed 82% of the error. Each further group adds a mean to the model.
- **Deal the 148 students at random into 16 groups, by drawing labels from a hat, one mean per group. Will that model's PRE be larger or smaller than 0.82?**



The error removed by random groups

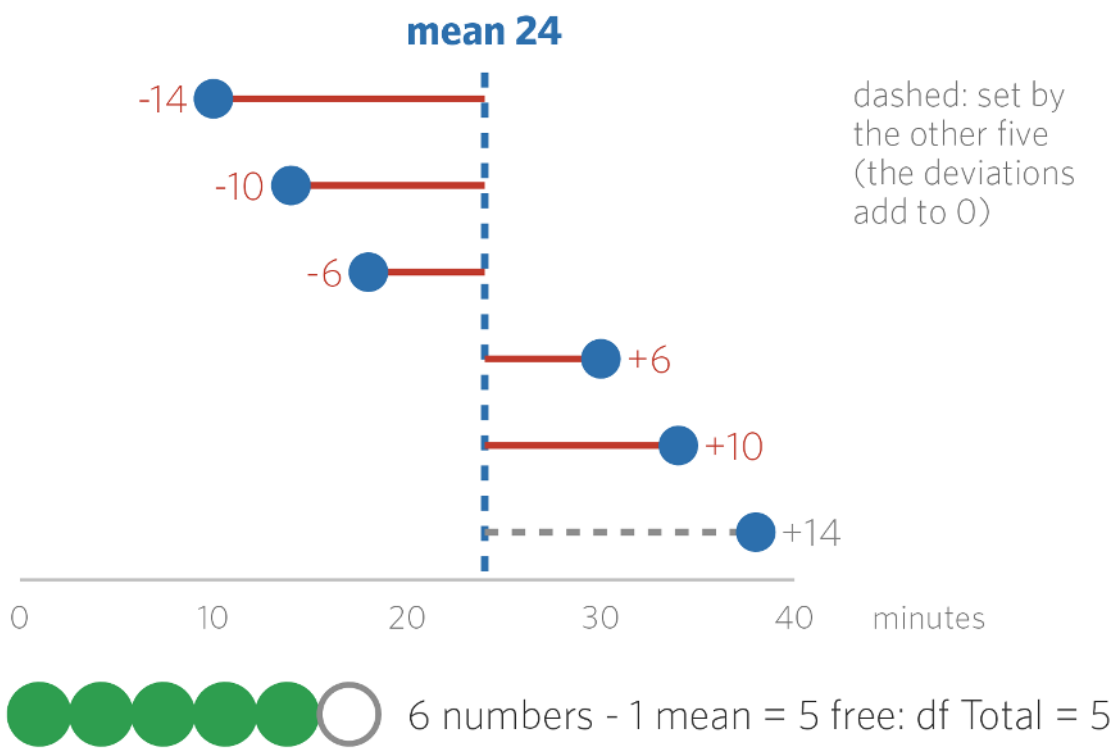


- The curve: PRE for the 148 students dealt at random into 2, 4, 8, up to 148 groups.
- PRE rises with every group added, even with meaningless labels: about 0.1 for 16 groups, past the real model's 0.82 at about 120, and 1 for 148.
- At 148 groups of one student each, the model says nothing about a 149th student.
- PRE rewards every parameter. A fair comparison charges for the parameters spent.

Why an added mean can only raise PRE

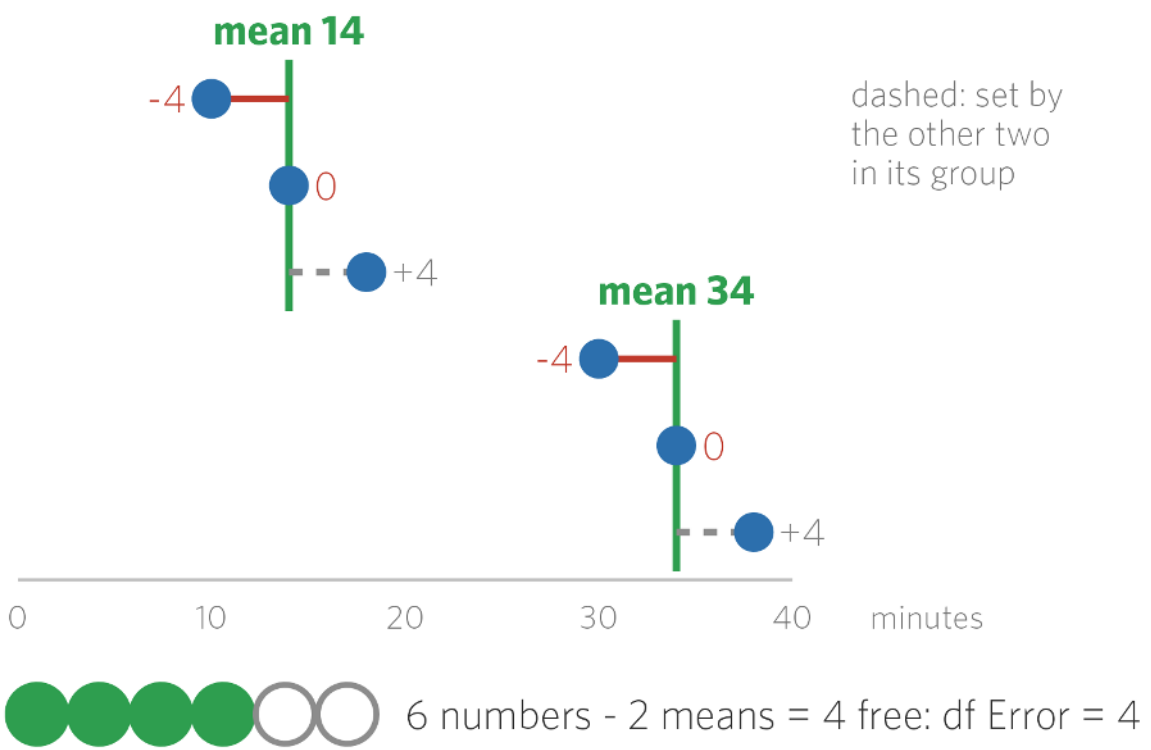
- A group's mean is the number with the smallest sum of squared residuals for that group, the fact from Lecture 5.
- Split a group in two and give each half its own mean. Each half is fit at least as well as before, so the error after the split is no larger than the error before the split.
- More groups mean less error and a higher PRE, even for labels drawn from a hat.

What a parameter costs



- Six commute times are six free numbers. The number of free numbers is called the degrees of freedom: 6 df.
- Fix their mean at 24. The first five deviations from the mean can be anything.
- The sixth is set by the other five, because the six deviations add to zero. Six numbers minus one fixed mean leaves 5 free: **df Total = 5**.

What two group means cost



- Fix the two group means, 14 and 34. Two deviations in each group can be anything.

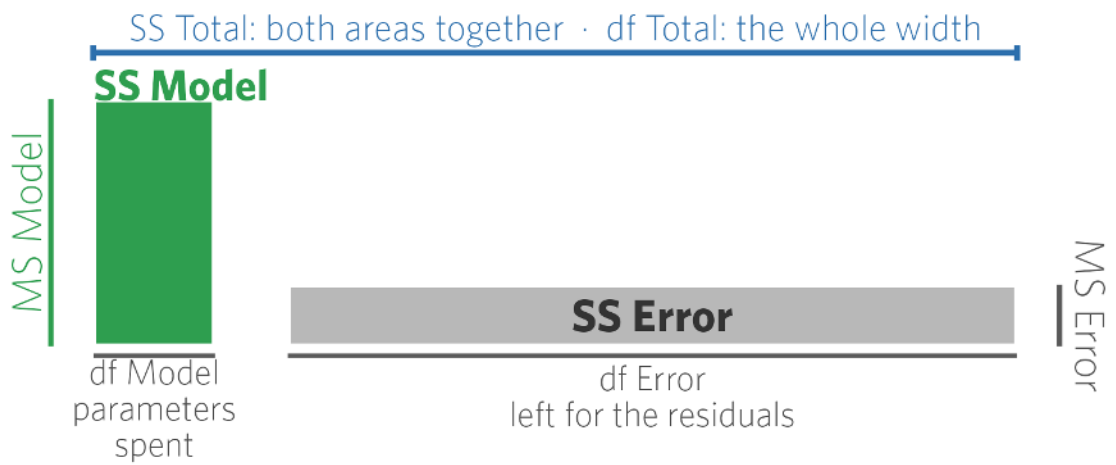
- The third in each group is set by the other two. Six numbers minus two fixed means leaves 4 free: **df Error = 4**. One more fixed mean than the empty model: **df Model = 1**.

- In general: df Total = n - 1, df Error = n - groups, df Model = groups - 1. For 148 students: 147, 146, 1.

Error per degree of freedom (MS)

- SS Model is the error the groups removed. Whether that is a lot depends on how many parameters they spent, so divide by df Model: **MS Model** is the error removed per parameter spent.
- SS Error is the error left. Spread it over the degrees of freedom the residuals keep, so divide by df Error: **MS Error** is the error left per degree of freedom, the variance of the residuals.
- For groups drawn from a hat, which have nothing to do with commute time, each added mean removes about as much error as remains per degree of freedom, so MS Model is about the size of MS Error. For the real groups, walkers and shuttle riders, the one added mean removes far more: MS Model is many times MS Error.

A sum of squares drawn as a block

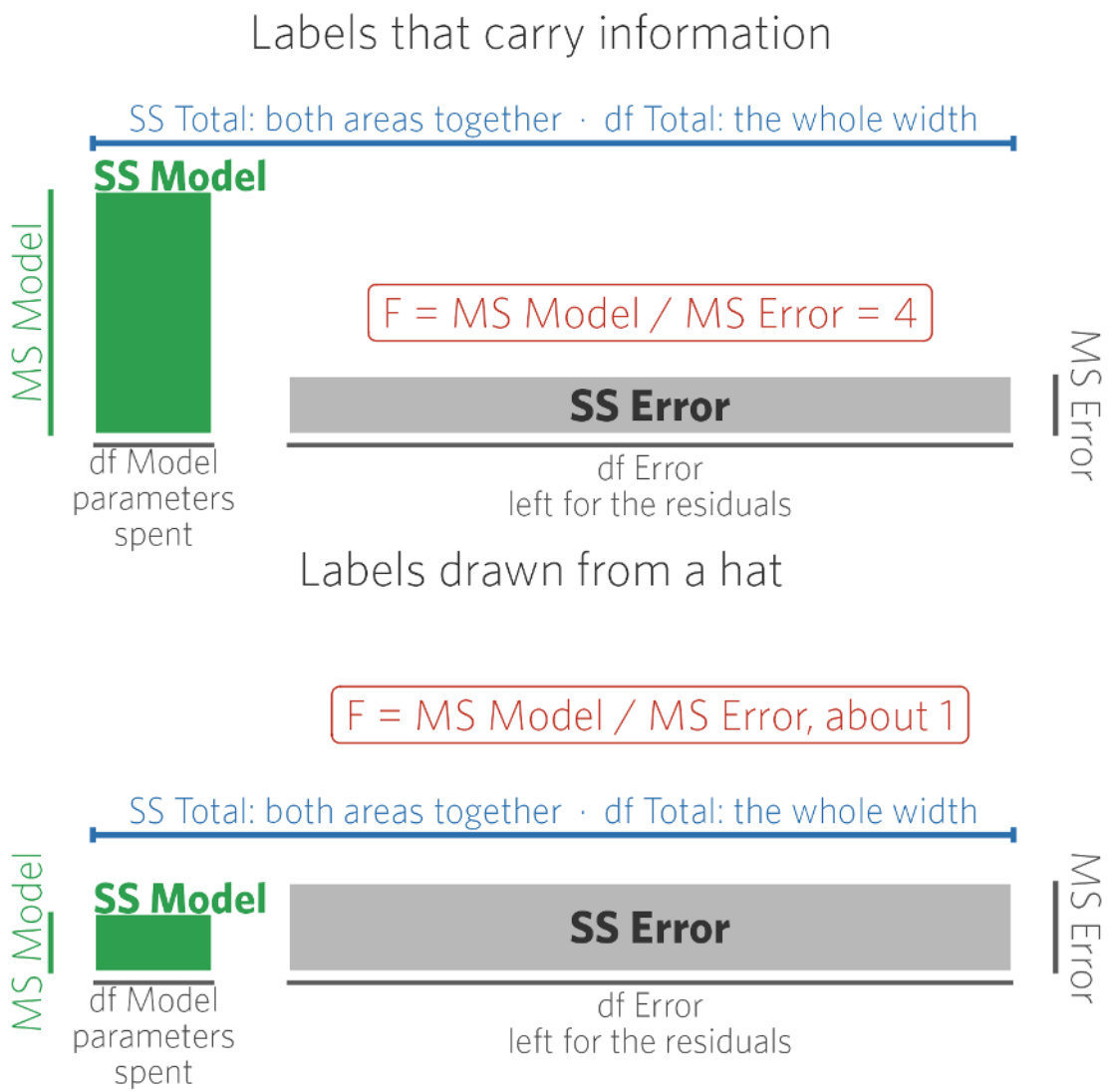


- SS Error, the error left, drawn as a block. Its width is df Error, the degrees of freedom left; its height is MS Error, the error left per degree of freedom.

- SS Model, the error removed, drawn the same way. Its width is df Model, the parameters spent; its height is MS Model, the error removed per parameter.

- Together the blocks are SS Total, the same area for every model of these data. A model decides how the area and the width are split.

What the F ratio compares

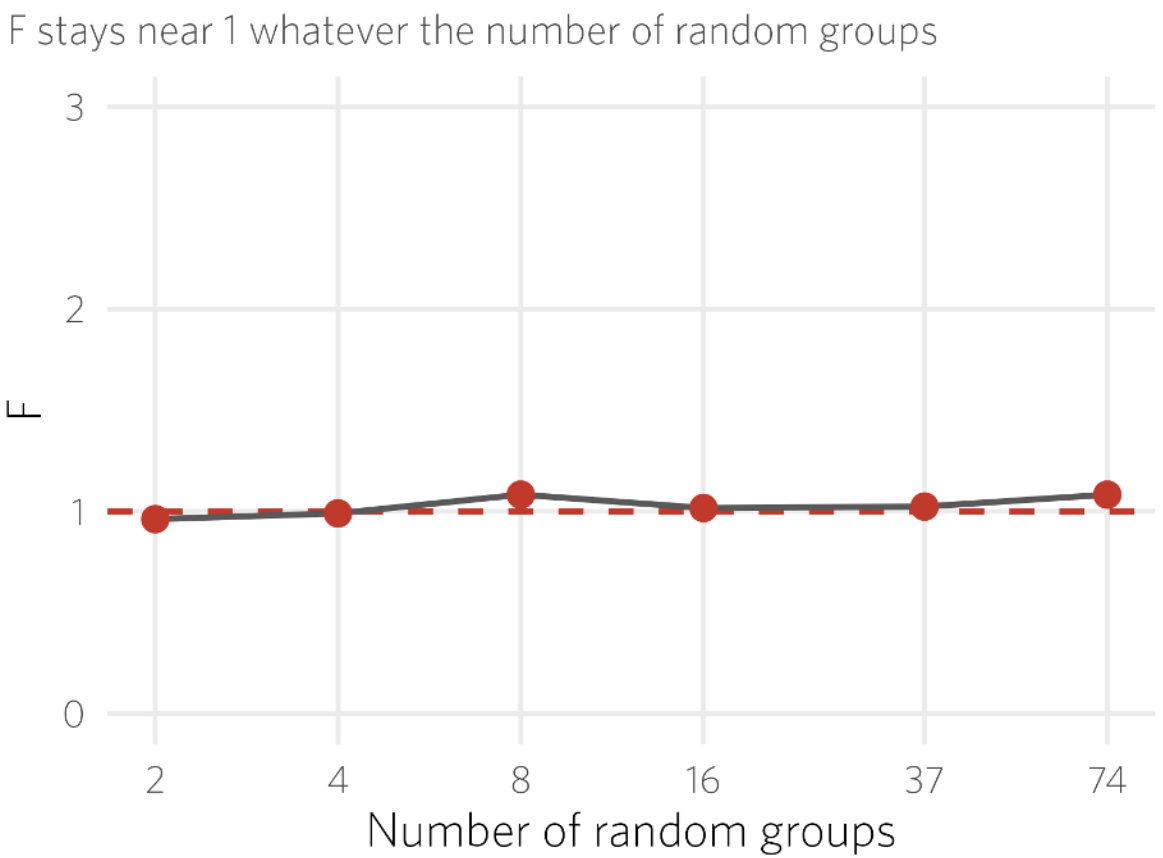


- F is the green height divided by the grey height, MS Model over MS Error. Here the one parameter removed four times as much error as is left per degree of freedom: F = 4.

- Groups drawn from a hat have the same total area, but their green block is no taller than the grey one, so F is about 1. Real groups, such as walkers and shuttle riders, make a tall green block and a large F.

The F ratio

$$F = \frac{MS_{\text{Model}}}{MS_{\text{Error}}} = \frac{\text{error removed per parameter spent}}{\text{error remaining per degree of freedom left}}$$



- For groups drawn from a hat, each added mean removes about as much error as remains per degree of freedom, so F is near 1.
- A large F means the model's parameters removed far more error per parameter than is left per degree of freedom.

The table for how students got to campus

	SS	df	MS	F	PRE
Model (how the student got to campus)	20,807	1	20,807	671	0.82
Error	4,528	146	31.0		
Total	25,335	147			

- The same quantities for last Monday's real model, commute time by how students got to campus, as R prints them.
- F = 671: the one parameter removed 671 times as much error as remains per degree of freedom.

Worksheet · Part 1

Eight students, four walkers and four shuttle riders, one mean per group: **SS Model = 72, SS Error = 36, SS Total = 108.** Fill in the table on your sheet as you go.

a. Fill in the df column: df Model, df Error, and df Total. (The empty model spends one of the eight degrees of freedom; the two-group model spends one more.)

b. Fill in the MS column: MS Model and MS Error, each SS divided by its df.

c. Compute F = MS Model / MS Error and write it in the Model row.

	SS	df	MS	F
Model (walked or took the shuttle)	72			
Error	36			
Total	108			

Formulas for Part 1

$$SS_{\text{Total}} = SS_{\text{Model}} + SS_{\text{Error}}$$

$$PRE = \frac{SS_{\text{Model}}}{SS_{\text{Total}}}$$

$$\text{df}_{\text{Total}} = n - 1$$

$$\text{df}_{\text{Model}} = \text{number of groups} - 1$$

$$\text{df}_{\text{Error}} = n - \text{number of groups}$$

$$\text{df}_{\text{Total}} = \text{df}_{\text{Model}} + \text{df}_{\text{Error}}$$

$$MS_{\text{Model}} = \frac{SS_{\text{Model}}}{\text{df}_{\text{Model}}}$$

$$MS_{\text{Error}} = \frac{SS_{\text{Error}}}{\text{df}_{\text{Error}}}$$

$$F = \frac{MS_{\text{Model}}}{MS_{\text{Error}}}$$

n is the number of cases, here eight students. MS Model is the error removed per parameter spent; MS Error is the error left per degree of freedom.

Worksheet · Part 1, continued

d. Now suppose each of the eight students is in a group of their own, so the model has eight means. Write that model's SS Error, PRE, and df Error, one number each.

e. A ninth student from the same section is about to report a commute time. Which model gives a usable prediction for that student: the two-group model or the eight-group model? Write your choice, and one sentence giving your reason.

Part 1a–c · The table for the eight students

a. $\text{df} = 1,\ 6,\ 7$

b. $MS_{\text{Model}} = 72/1 = 72$

$MS_{\text{Error}} = 36/6 = 6$

c. $F = 72/6 = 12$

- **a.** *The df column.* df Model = 1, one mean beyond the empty model's. df Error = 8 - 2 = 6. df Total = 8 - 1 = 7.
- **b.** *The MS column.* MS Error = 6 square minutes: the variance of the eight residuals.
- **c.** *F.* F = 12: the one parameter removed twelve times as much error as remains per degree of freedom.

Part 1d–e · One mean per student

d. $SS_{\text{Error}} = 0$

$PRE = 108/108 = 1$

$\text{df}_{\text{Error}} = 8 - 8 = 0$

e. $F = \dfrac{108/7}{0/0}$ undefined

- **d.** *One mean per student: SS Error, PRE, df Error.* Every prediction equals the student's own time: no error left, PRE = 1, and no degrees of freedom left.
- **e.** *Which model predicts a ninth student?* The two-group model. The eight-group model has eight means, and nothing says which one a ninth student gets.
- With df Error = 0 there is no estimate of the error around a prediction, and F cannot be computed.
- A perfect fit to these eight students is not a model of the ninth.

Today

Linear regression

1. The cost of a parameter
2. Could the difference be chance?
3. The slope and the intercept

The sample's b_1 and the process's β_1

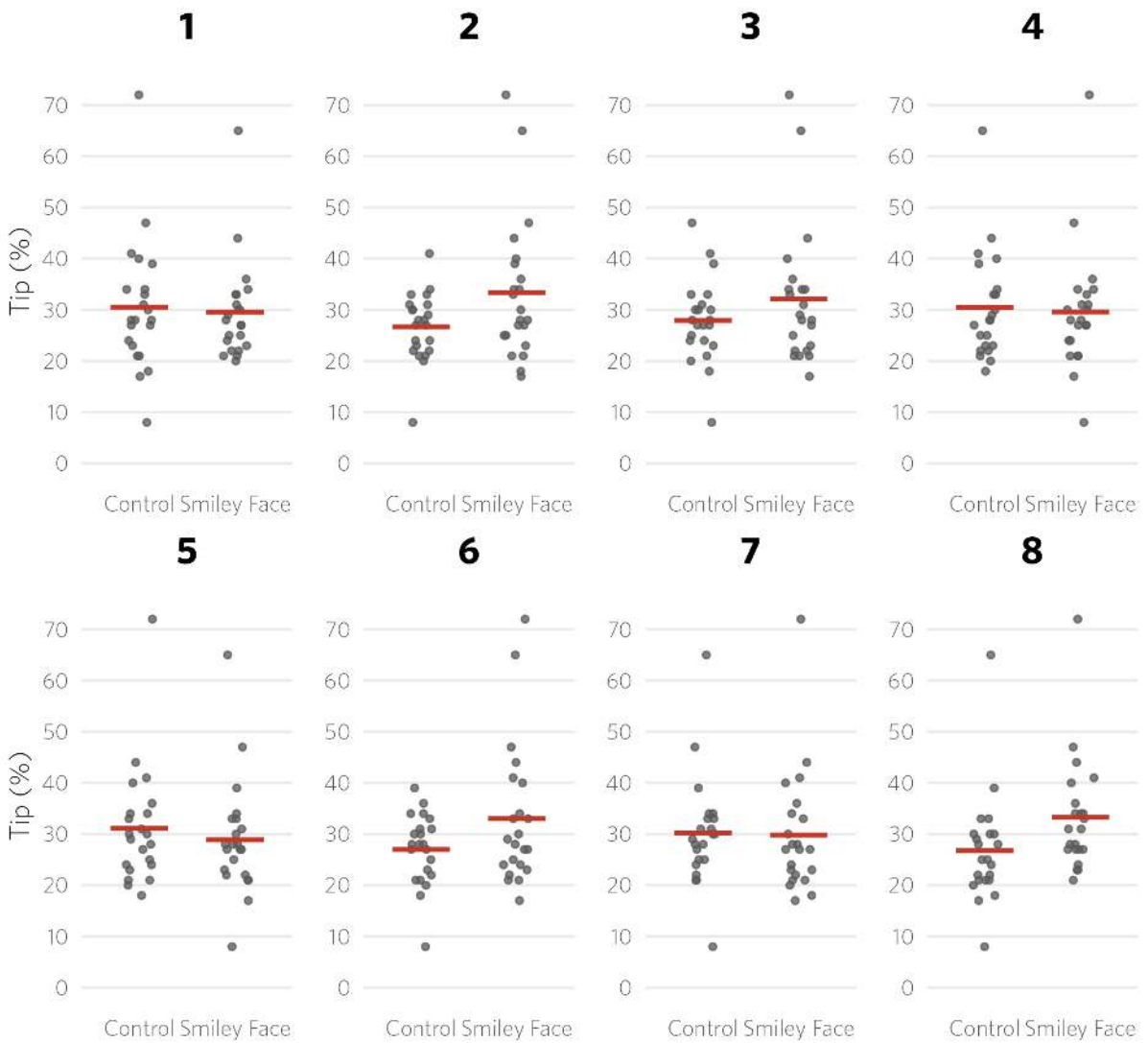
$$\text{process: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{empty model: } Y_i = \beta_0 + \varepsilon_i$$

- A dataset is one sample from a process: another 148 students would give other times.
- Last Monday's b_1 = 30.3 minutes, the shuttle riders' mean minus the walkers', estimates β_1, the gap the process itself produces.
- The empty model is the same model with β_1 set to 0, as if how a student got to campus made no difference.
- Even if β_1 were 0, a sample's b_1 would rarely be exactly 0. The question: how large a b_1 can a sample produce when β_1 is 0?

The smiley-face tipping experiment

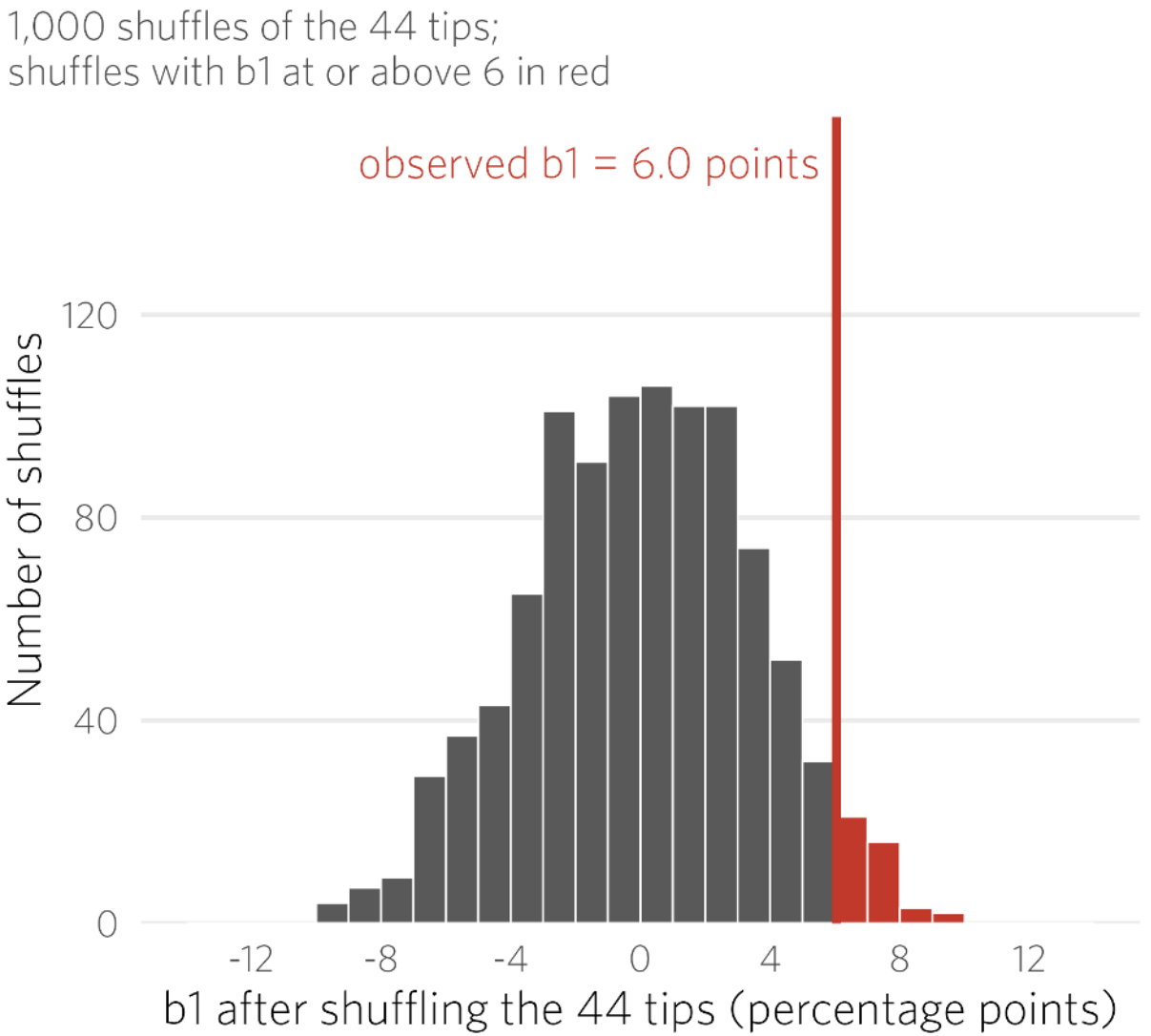
- In a 1996 study, a server drew a smiley face on the checks of 22 tables, chosen at random out of 44, and left the other 22 checks plain.
- Smiley-face tables tipped 33% of the bill on average and plain-check tables 27%: $b_1 = 6$ percentage points.
- The question: is a 6-point gap more than chance would produce between two random halves of the same 44 tables?

Shuffles compared by eye



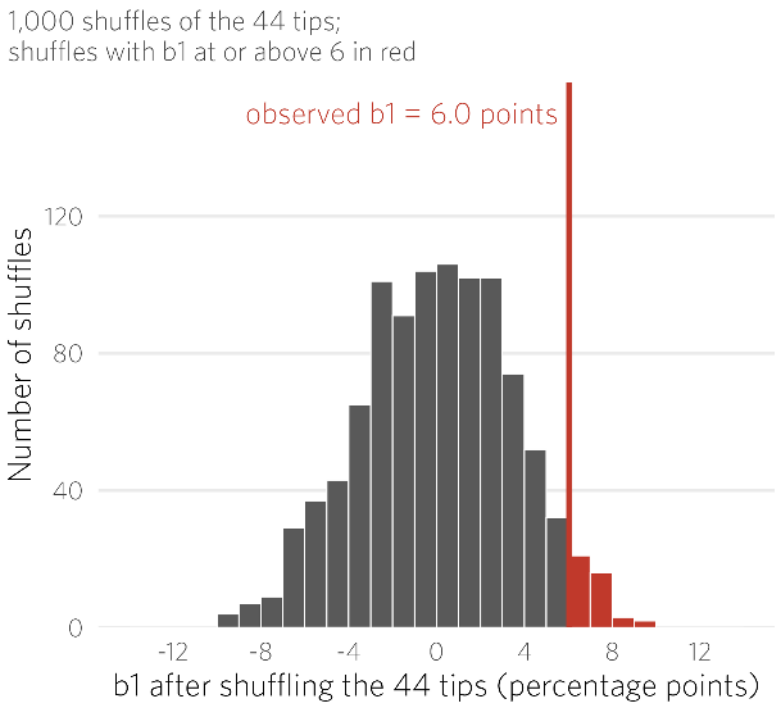
- If the smiley face had no effect, which table left which tip would be chance. Shuffling imitates that: deal the 44 tips to the 44 tables at random.
- In Lecture 5 we shuffled seven times, plotted each beside the real data, and tried to pick out the real plot by eye: panel 6.
- Each shuffle also yields a number, b_1. A thousand shuffles give the distribution of b_1 when the smiley face has no effect.

The 44 tips shuffled 1,000 times



- Each bar counts the shuffles that produced a given b_1; the red line is the real gap, 6.0 points.
- Most shuffles give a b_1 within a few points of 0, in either direction.
- Red bars: shuffles with b_1 at or above 6.

Worksheet · Part 2



In 1996 a server drew a smiley face on the checks of 22 tables chosen at random out of 44, and left the other 22 plain. Smiley-face tables tipped 33% of the bill on average and plain-check tables 27%, so b_1 = 6 percentage points. The histogram shows b_1 for 1,000 shuffles of the 44 tips across the two groups.

a. Add the heights of the bars to the right of the line: in about how many of the 1,000 shuffles was b_1 at least 6? Write a whole number.

b. If the smiley face had no effect, in about how many of every 100 shuffles would a difference of 6 points or more arise?

c. If the smiley face had no effect, would a difference of 6 points or more be common or rare? Write *common* or *rare*, and one sentence giving your reason, using your answer to b.

Part 2a–b · The shuffles at or above 6 points

a. 42 of $1{,}000$

b. $\dfrac{42}{1{,}000} = 4.2\% \approx 4$ in 100

- **a.** *How many of the 1,000 shuffles reached 6?* 42 of the 1,000; the red bars read as about 40.
- **b.** *In about how many of every 100?* About 4 of every 100.
- In Lecture 5 we shuffled only seven times, and two of the seven came out above 6 points. Seven shuffles are too few to estimate a proportion this small.

Part 2c · Common or rare?

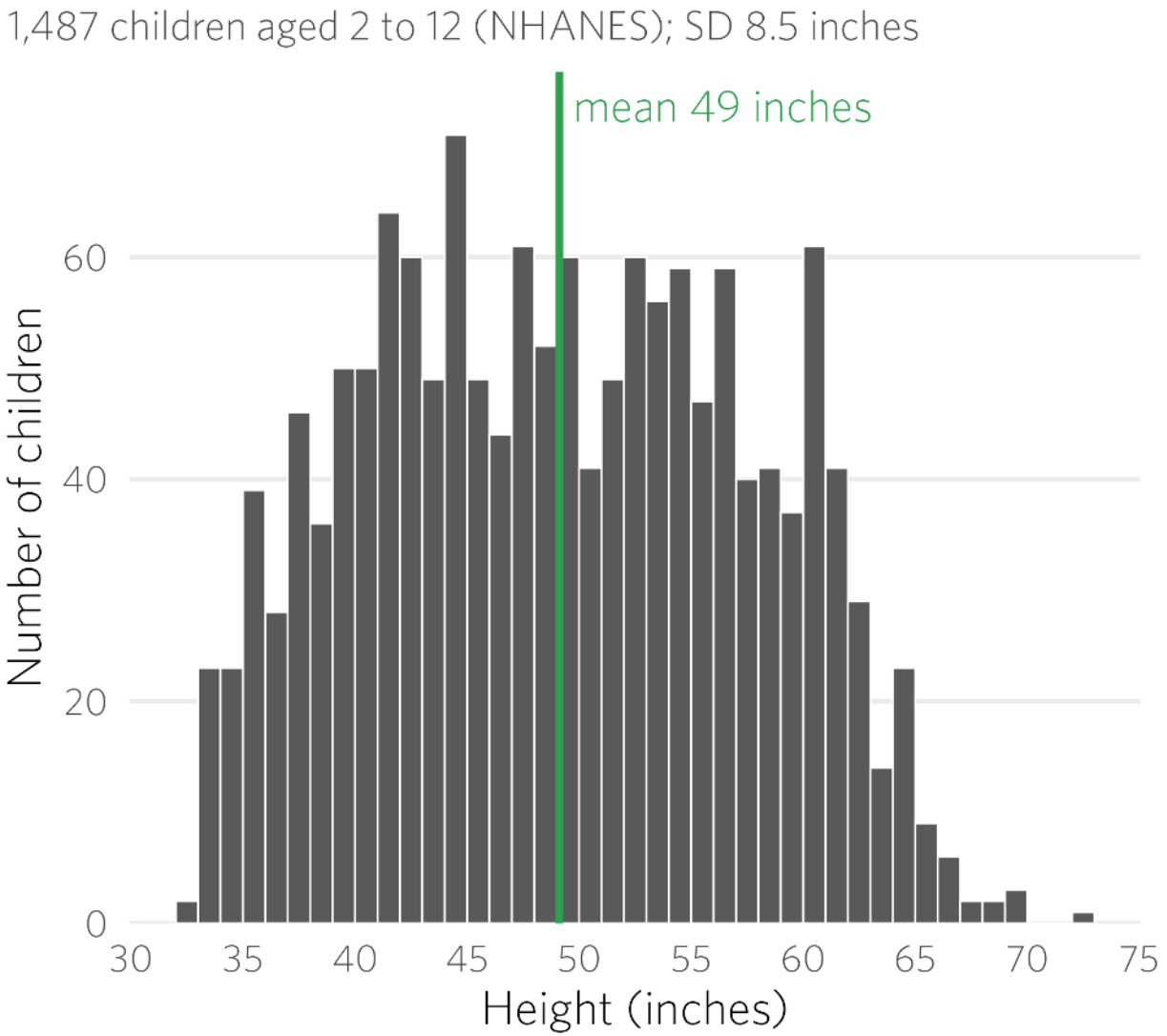
- **c.** *Common or rare?* Rare. A 6-point difference arises in about 4 of every 100 shuffles with no smiley face: uncommon, but chance is not ruled out.
- The 5% convention from Lecture 7: an outcome that chance produces in fewer than 5 of 100 tries counts as unusual. Four in 100 is just inside that line, so the difference deserves a closer look; 44 tables rule out neither chance nor an effect.

Today

Linear regression

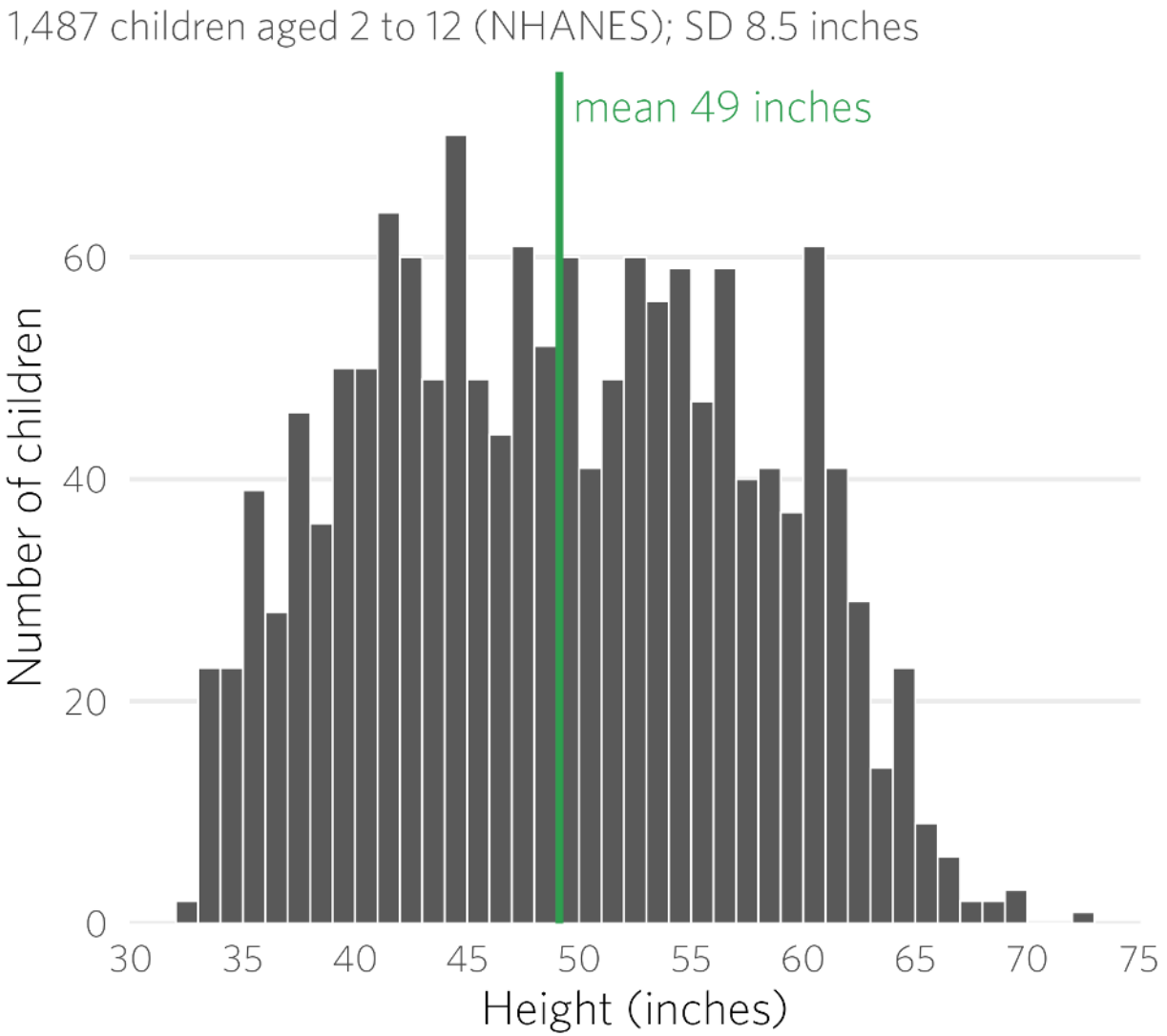
1. The cost of a parameter
2. Could the difference be chance?
3. The slope and the intercept

The heights of 1,487 children



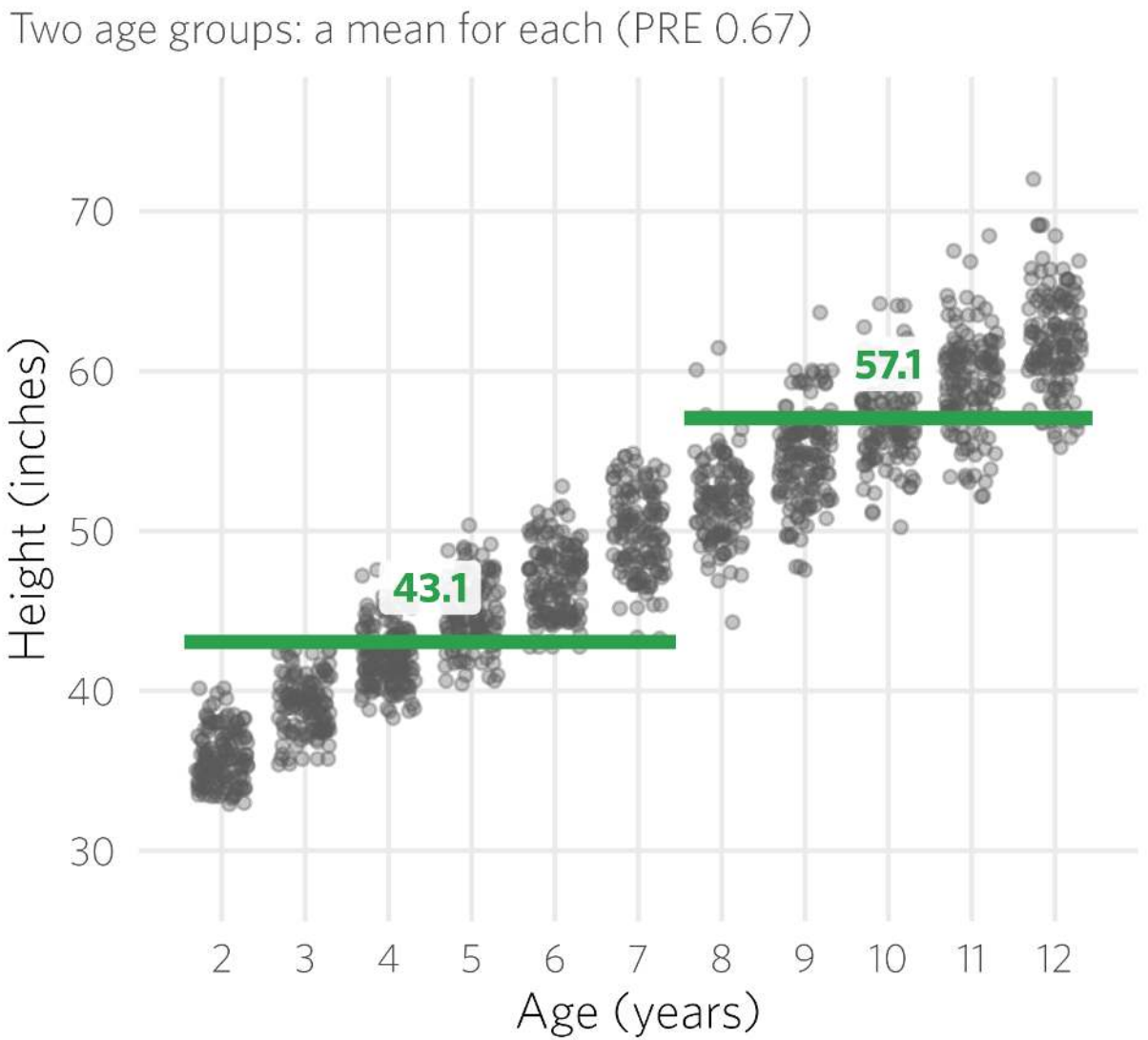
- Children from the NHANES health survey, the dataset from Lecture 5, now the 1,487 aged 2 to 12.
- With nothing else known about a child, the mean, 49 inches, is the prediction for every child, as in Lecture 5.
- The SD is 8.5 inches: a typical child's height is about 8.5 inches from the mean.

Predict a child's height



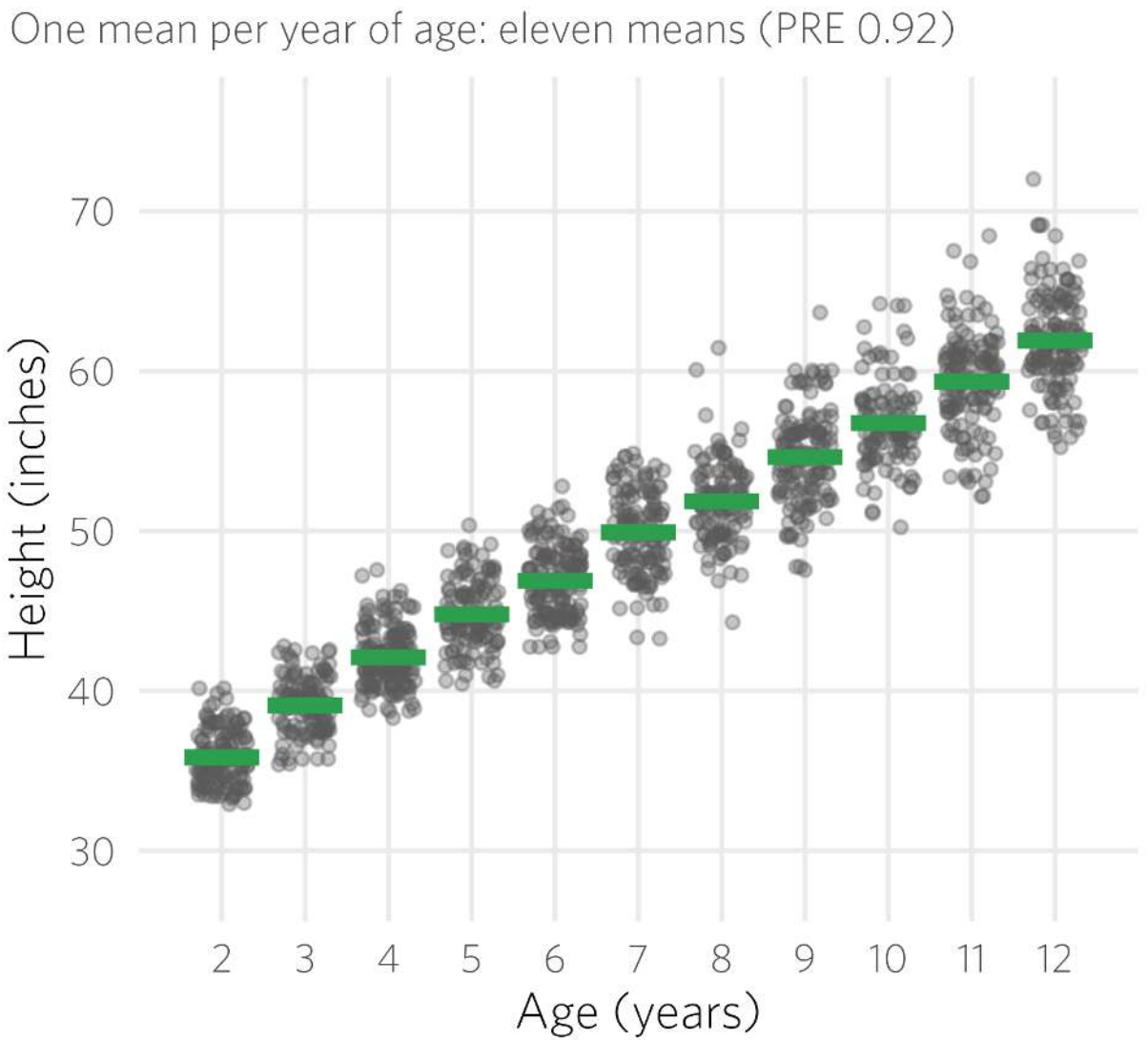
- A child is picked at random from these 1,487. **What height do you predict, in inches?**
- Now you learn that the child is 10 years old. **What height do you predict for this child?**

Age as two groups



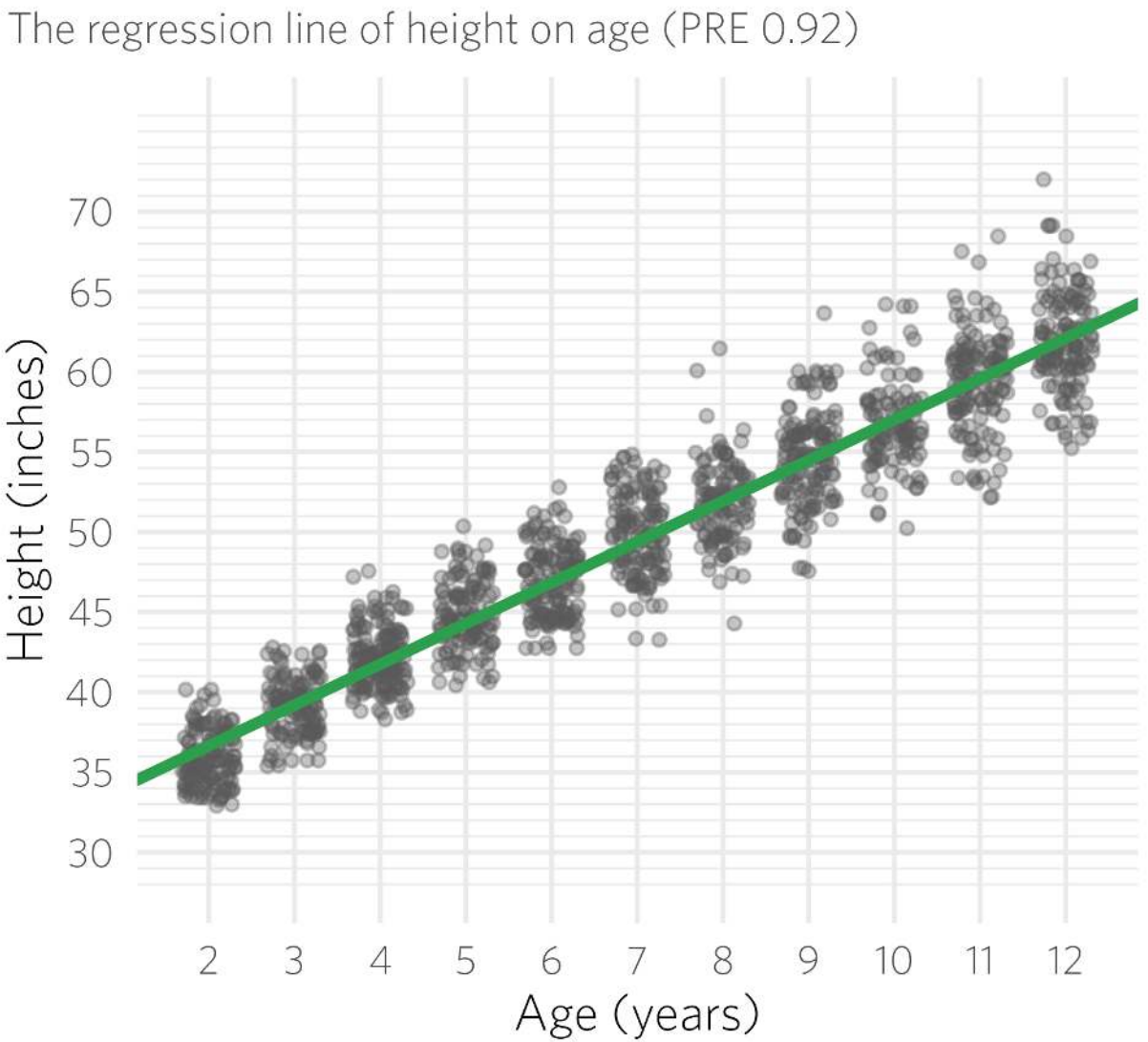
- Ages 2 to 7 average 43 inches; ages 8 to 12 average 57.
- Each group's mean is the prediction for every child in the group, as in last week's group models.
- The proportional reduction in error (PRE) is 0.67: the two groups remove 67% of the error around the mean.

One mean per year of age



- Eleven means and ten parameters beyond the empty model: PRE 0.92.
- The means rise about 2.5 inches a year, nearly along a straight line.

A line in place of eleven means



- The line predicts height from age with two numbers: an intercept and a slope.
- Its PRE is 0.92, the same as eleven means, with nine fewer parameters.

$$Y_i = b_0 + b_1 X_i + e_i$$

- This is the same equation as last week's two-group model.
- There, $\mathbf{X_i}$ was 0 or 1. Here, $\mathbf{X_i}$ is the child's age in years.

The line as R reports it

```
lm(inches ~ Age, data = children)

(Intercept)          Age
      31.56         2.55
```

$$\text{height} = 31.6 + 2.55 \times \text{age}$$

- The slope: for each additional year of age, the predicted height is **2.55 inches** greater.
- Its units are inches per year: the outcome's units per unit of the explanatory variable.
- Last week, b_1 was the difference between two group means. Now it is the change in the predicted height per year of age.

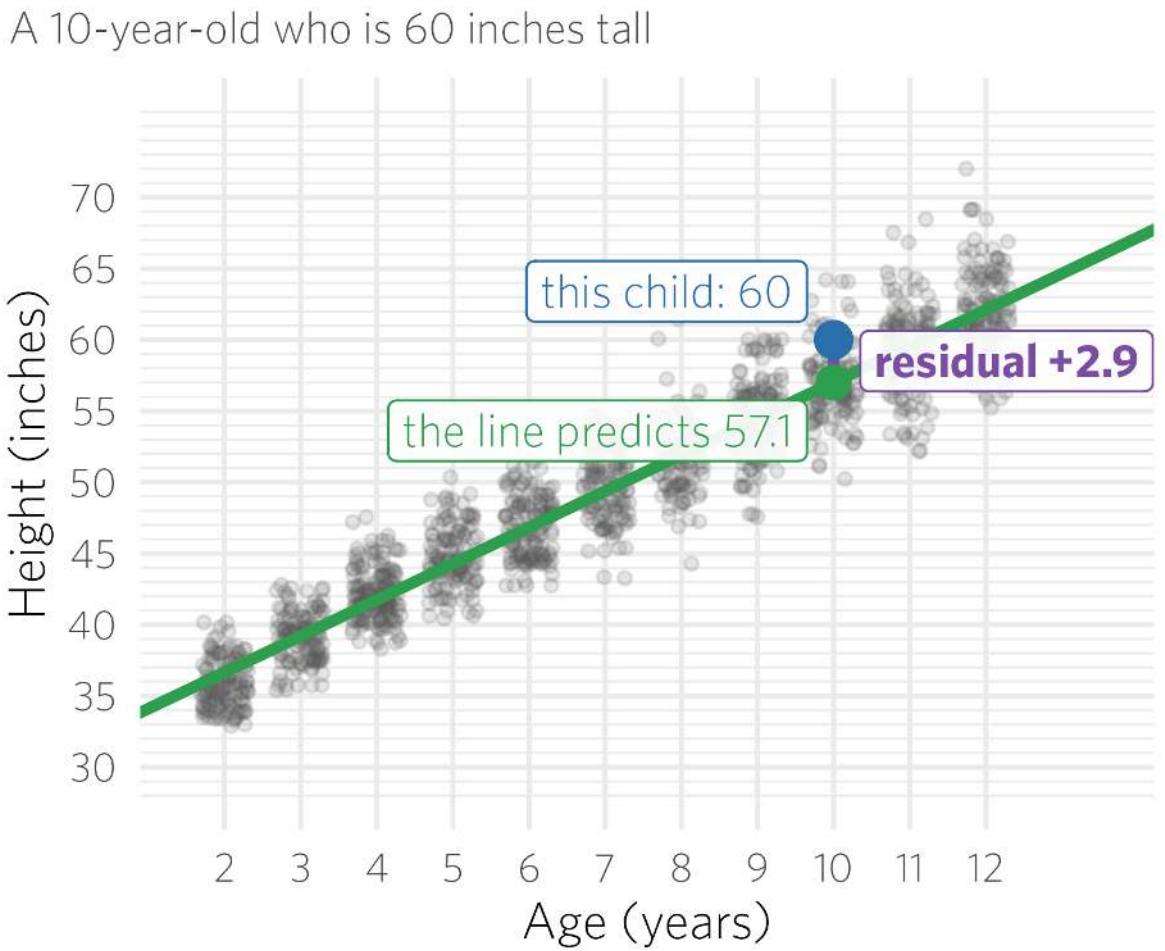
The intercept is the line's value at age zero

$$\text{age } 0: \quad 31.6 + 2.55 \times 0 = 31.6 \text{ in}$$

$$\text{age } 2: \quad 31.6 + 2.55 \times 2 = 36.7 \text{ in}$$

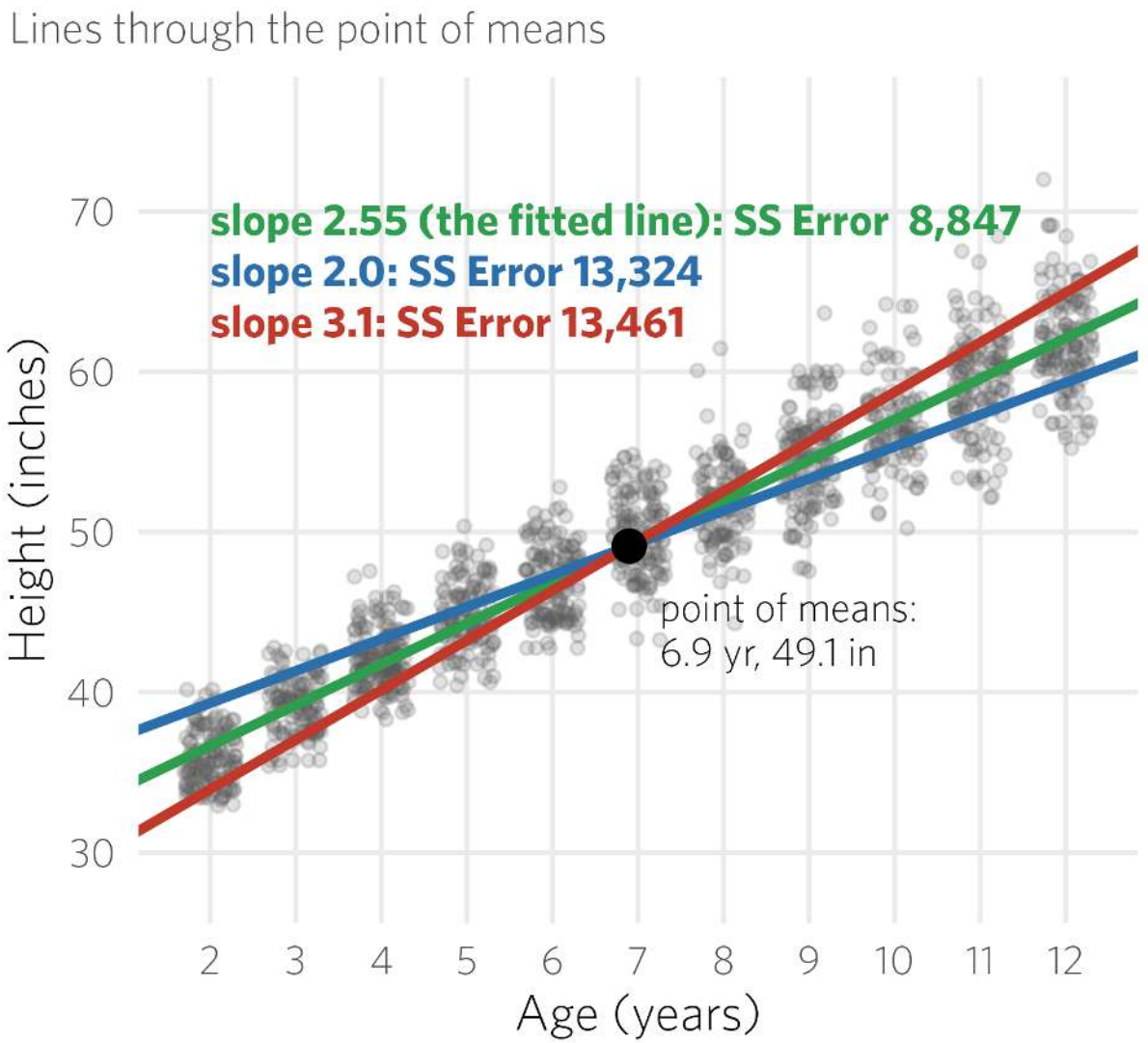
- At age 0 the line gives 31.6 inches. Newborns average about 20 inches.
- The youngest children are 2. The intercept places the line; it is no one's height.
- The line describes the data between ages 2 and 12.

Predictions and residuals



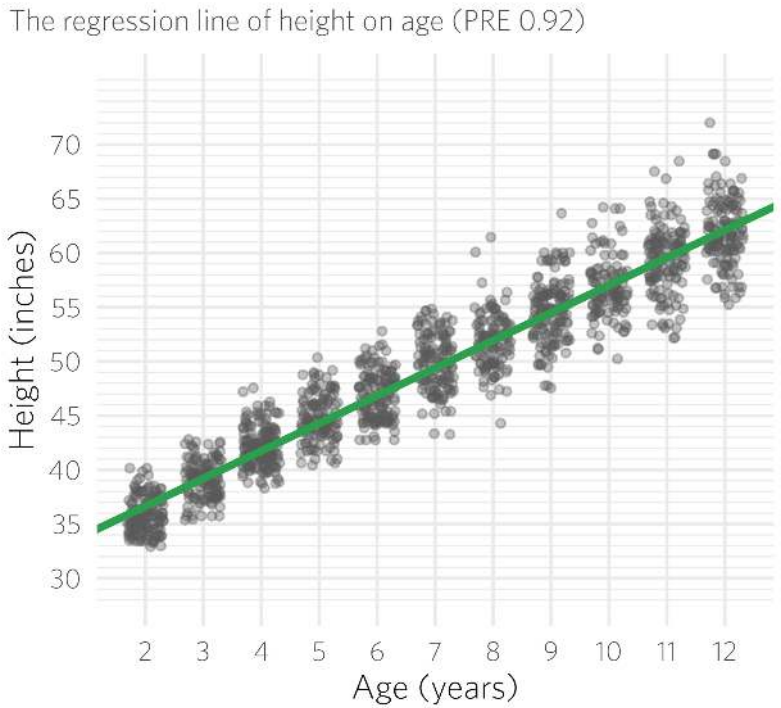
- Take one child from the survey, a 10-year-old who is 60 inches tall. For a 10-year-old the line predicts 31.6 + 2.55 x 10 = 57.1 inches.
- Residual = actual minus predicted: 60 - 57.1 = +2.9 inches, the vertical gap from the point to the line.
- A positive residual: the child is taller than the line predicts for their age.

The best-fitting line



- Of all the lines one could draw through these points, the fitted line has the smallest sum of squared residuals: 8,847.
- A flatter line through the same point of means: 13,324.
- A steeper one: 13,461.
- The fitted line passes through the point of means, 6.9 years and 49.1 inches, and its residuals add to zero, like deviations from a mean.

Worksheet · Part 3



The figure shows the heights of 1,487 children aged 2 to 12 from the NHANES health survey, with the regression line of height on age. Rounded to simple numbers, the line is **height = 32 + 2.5 x age**, with height in inches and age in years.

a. Compute the line's predicted height for a 4-year-old and for an 8-year-old, in inches.

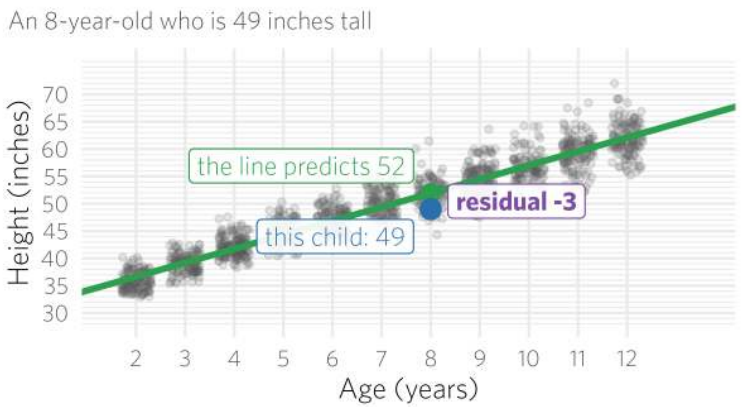
b. An 8-year-old in the survey is 49 inches tall. Compute this child's residual, in inches, with its sign.

c. In one sentence: is this child taller or shorter than the line predicts, and by how many inches?

d. The oldest children in these data are 12. Compute the line's predicted height for a 24-year-old, in inches.

e. Is that prediction trustworthy? Write *yes* or *no*, and one sentence giving your reason.

Part 3a–c · Predictions and a residual



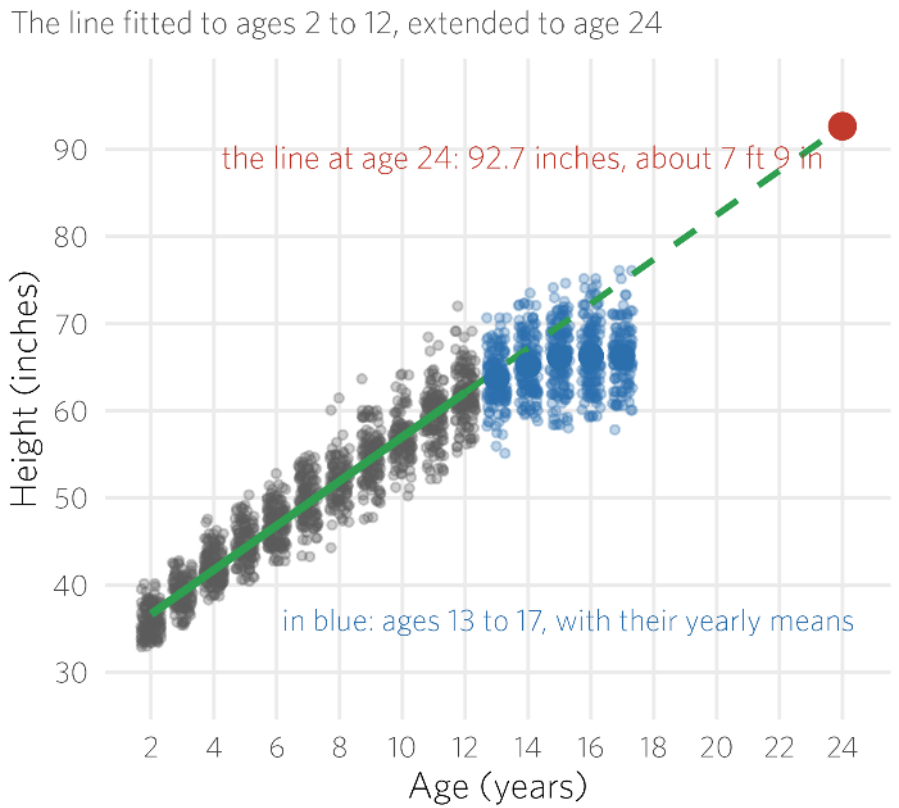
a. $32 + 2.5 \times 4 = 42$ in

$32 + 2.5 \times 8 = 52$ in

b. $49 - 52 = -3$ in

- **a.** *The line at ages 4 and 8.* 42 and 52 inches. The survey's mean heights at those ages are 42.1 and 51.9.
- **b.** *The 49-inch 8-year-old's residual.* Actual minus predicted: 49 - 52 = -3 inches.
- **c.** *Taller or shorter, and by how much?* Shorter, by 3 inches.

Part 3d–e · The line at age 24



- **d.** *The line at age 24.* 32 + 2.5 x 24 = 92 inches, about 7 ft 8 in. The unrounded line gives 92.7.
- **e.** *Trustworthy?* No. The data stop at age 12.
- The survey's 13-to-17-year-olds fall farther below the line each year: growth slows after about 14.
- At age 24 there are no data to support the prediction.

Before Friday

- Finish CourseKata **Chapter 9**.
- Friday: **Case #5**, in your tutorial session. Check your discussion section's Slack channel for your time slot.
- Milestone 5 is due Sunday.

Hand in your worksheet on the way out

On the last line, write down one thing you learned today.

Thanks!