Lecture 12: Describing relationships

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

Is a county more than 20 percent black?

NO There are not many African-Americans in this county.

YES This county has a large African-American population.

Obama wins these counties 383 to 70.

And is the high school graduation rate higher than 78 percent?

NO This is a county with less-educated voters.

YES This is a county with more educated voters.

Clinton wins these counties 704 to 89.

And is the high school graduation rate higher than 87 percent?

NO 78 to 87 percent have a diploma.

YES This is a highly educated county.

Obama wins these counties 185 to 36.

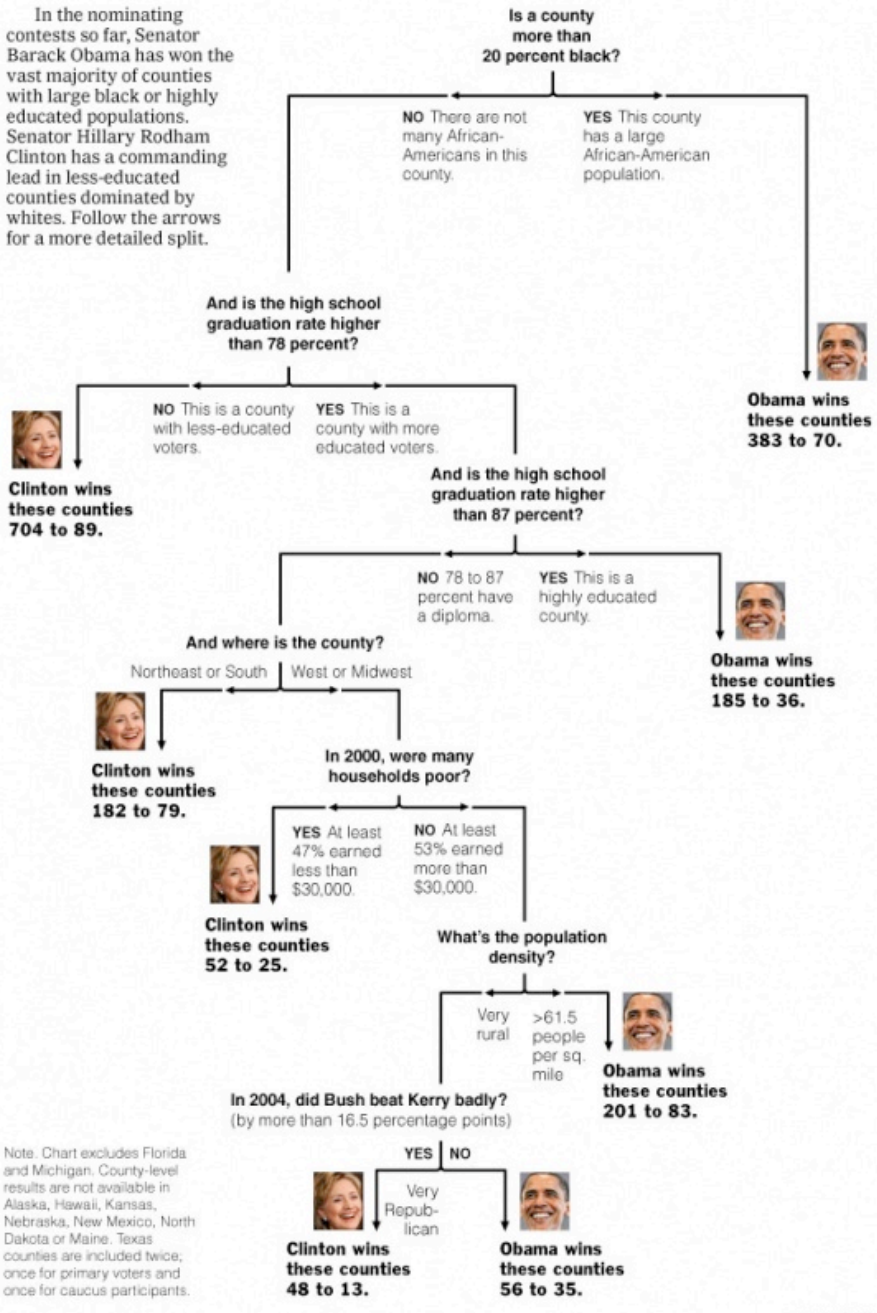
And where is the county?

Northeast or South | West or Midwest

In 2000, were many

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice, once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/
THE NEW YORK TIMES

Decision trees

The figure on the previous slide was made by **Amanda Cox**, a statistician working in the graphics department at the New York Times (and yes, the analysis was done in R)

The title “Decision Tree” is a pun, as the data refer to county-wide decisions for Obama or Clinton in the Democratic Primary; **the object is also a kind of statistical model known as a decision tree**

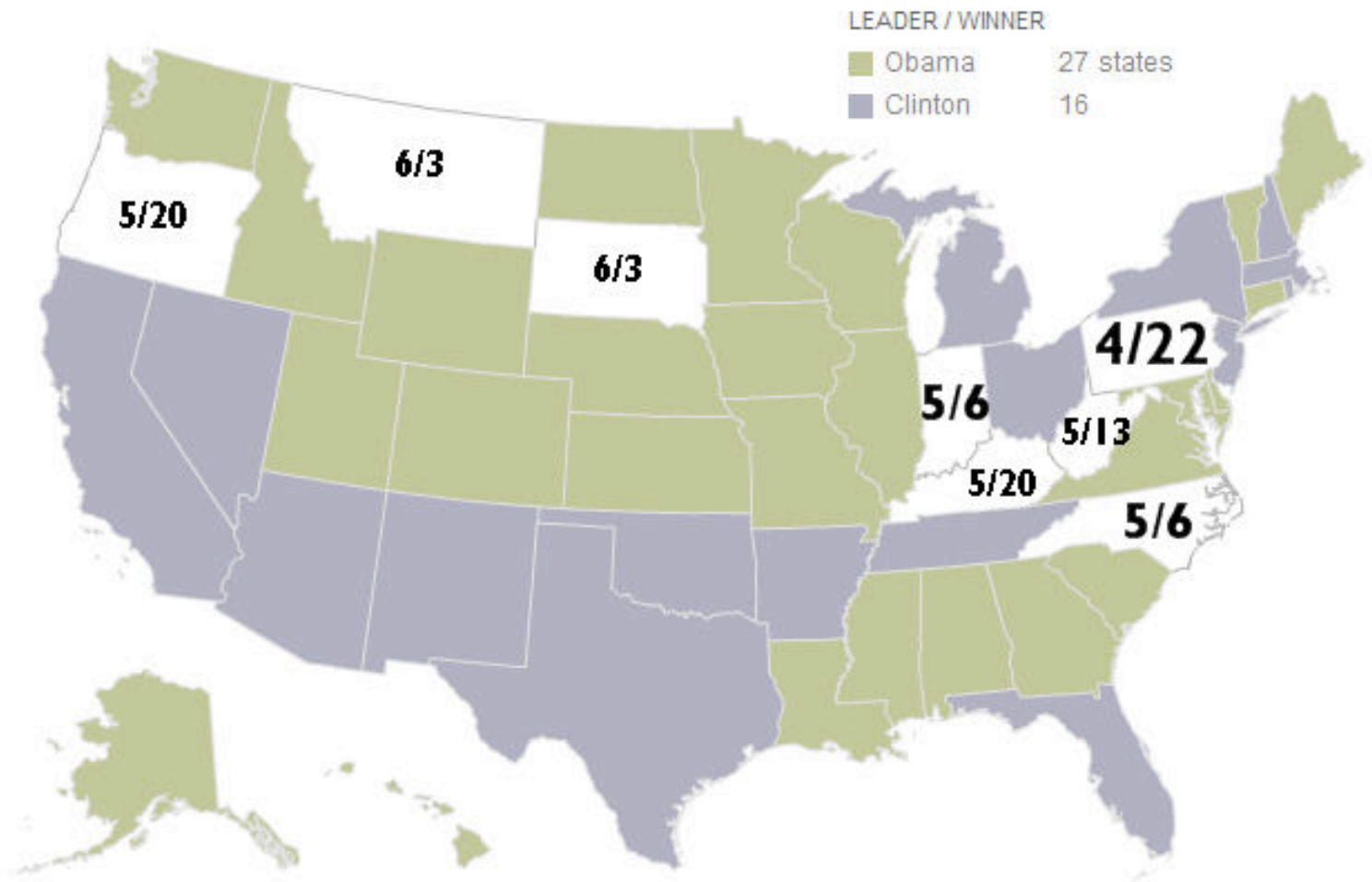
Coded in this elegant structure is **a pretty serious computational algorithm**; a relatively recent addition to the statistician’s toolbox

The Democratic Primary

Think back to April of last year, when it seemed Hillary Clinton's momentum was picking up again; the decision tree was built from a data set concerning **all the counties in states where primaries had already been held**

The unit of observation, then, is a county and variables included various **demographic measures** (age and ethnic makeup, education level, religious breakdown), **political measures** (did the county go to Bush or Kerry in 04) and **economic factors** (unemployment rate, the amount of construction in the county), and so on

```
primary = read.csv(url("http://www.stat.ucla.edu/~cocteau/primaries.csv"),head=T)

names(primary)
dim(primary)

# transformations

primary$black06pct = primary$black06/primary$pop06
primary$hisp06pct = primary$hisp06/primary$pop06
primary$white06pct = primary$white06/primary$pop06
primary$growth = 100*(primary$pop06/primary$pop00 - 1)

# drop some counties
primary = subset(primary, state_postal!="MI" )
primary = subset(primary, state_postal!="FL" )
primary = subset(primary, !(state_postal=="WA" & racetype=="Primary") )

table(primary$winner)

primary[primary$state=="CA",c("winner","county_name")]
```


LEADER / WINNER
Obama 27 states
Clinton 16
5/20
6/3
6/3
4/22
5/6
5/13
5/20
5/6

	winner	county_name
156	obama	Alameda
157	obama	Alpine
158	clinton	Amador
159	clinton	Butte
160	clinton	Calaveras
161	clinton	Colusa
162	clinton	Contra Costa
163	clinton	Del Norte
164	clinton	El Dorado
165	clinton	Fresno
166	clinton	Glenn
167	obama	Humboldt
168	clinton	Imperial
169	clinton	Inyo
170	clinton	Kern
171	clinton	Kings
172	clinton	Lake
173	obama	Lassen
174	clinton	Los Angeles
175	clinton	Madera
176	obama	Marin
177	clinton	Mariposa
178	obama	Mendocino
179	clinton	Merced
180	clinton	Modoc
181	obama	Mono
182	clinton	Monterey
183	clinton	Napa
184	obama	Nevada
185	clinton	Orange
186	clinton	Placer
187	obama	Plumas
188	clinton	Riverside
189	obama	Sacramento
190	clinton	San Benito
191	clinton	San Bernardino
192	clinton	San Diego
193	obama	San Francisco
194	clinton	San Joaquin
195	obama	San Luis Obispo
196	clinton	San Mateo
197	obama	Santa Barbara
198	clinton	Santa Clara
199	obama	Santa Cruz
200	clinton	Shasta
201	obama	Sierra
202	obama	Siskiyou
203	clinton	Solano
204	obama	Sonoma
205	clinton	Stanislaus
206	clinton	Sutter
207	clinton	Tehama
208	obama	Trinity
209	clinton	Tulare
210	clinton	Tuolumne
211	clinton	Ventura
212	obama	Yolo
213	clinton	Yuba

The Democratic Primary

Now, given all these potential predictor variables, **which ones "explain" county-level voting patterns?** Suppose, for example, we start simply, and consider whether or not a majority of the county voted for Bush in 04

	K04	B04
clinton	171	1039
obama	302	728

Therefore, Obama won about 64% of those counties not voting for Bush in 04, while Clinton won about 59% of those counties that did vote for Bush in 04

Now, consider the **simple prediction rule**: If a county voted for Bush in 04, we'll say that they will vote for Clinton in the primary, while if a county was mostly in favor of Kerry in 04, we'll assign the win to Obama

The Democratic Primary

Of course **this rule isn't perfect**; by applying it, we would make 171 + 728 = 899 mistakes (out of 2,240 counties, or about 40% error); we refer to these mistakes as having been "misclassified" by our simple rule

So the question becomes, **can we do any better?** There might be better indicators of Obama's success besides the vote in 2004 -- but how do we find them?

The decision tree encodes a large search across the complete data set; consider the top of the tree, "the root"...

The Democratic Primary

Decision trees work by repeatedly splitting the data into two parts; the root or first "split" is **the single division of the data into two pieces that produces the lowest misclassification error**

To investigate this a little further, let's consider the predictor that represents the percentage of a county that is African American; we now choose a breakpoint α that divides the data into two pieces (those counties with a greater percentage of African Americans and those with a smaller percentage)

We then form a table (as we did for counties that went for Bush or Kerry in 2004) and count the misclassification rate...

A technicality

Strictly speaking, the procedure we're describing today operates on the class probabilities produced by a split -- The misclassification rate is just one function that can be computed here

Other functions like the entropy or the so-called Gini index can also be used -- A couple decades of experience with three models suggests, in fact, that minimizing the misclassification error at each point has problems

Still, for pedagogical purposes, we'll use it -- But keep in mind that R is actually computing with a different function of the probabilities

The Democratic Primary

Suppose we take $\alpha = 0.1$; and we end up with the following table

	Pct Af Am > 0.1	
	FALSE	TRUE
clinton	989	221
obama	554	476

Of the 697 counties that are more than 10% African American, 221 went for Clinton and 476 went for Obama (68% in favor of Obama); since Obama won more counties in this group, we label the node “Obama” and we would be making 221 errors

In the remaining 1543 counties, 898 went for Clinton and 554 for Obama (64% in favor of Clinton); then we will label this group “Clinton” and we would make 554 errors

```
tmp = primary$winner[primary$black06pct>0.1]
table(tmp)

# clinton   obama
#     221     476

tmp = primary$winner[primary$black06pct<=0.1]
table(tmp)

# clinton   obama
#     989     554
```

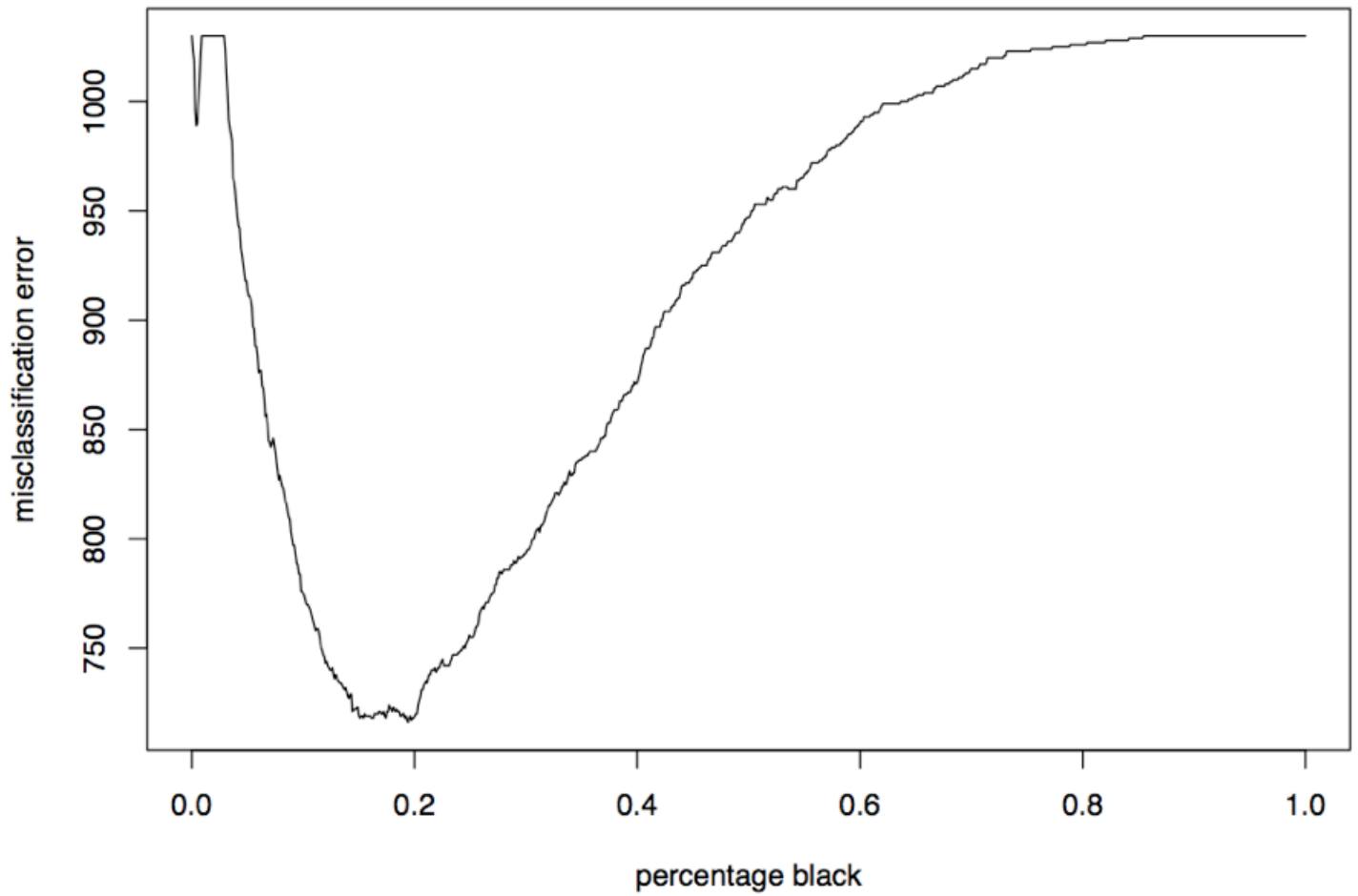
The Democratic Primary

The 20% figure at the top of the tree was obtained by finding the magic point $\alpha*$ that minimizes the misclassification errors

In fact, the search was conducted over **all the variables in the data set and all the possible splits**; and this choice produced the smallest error

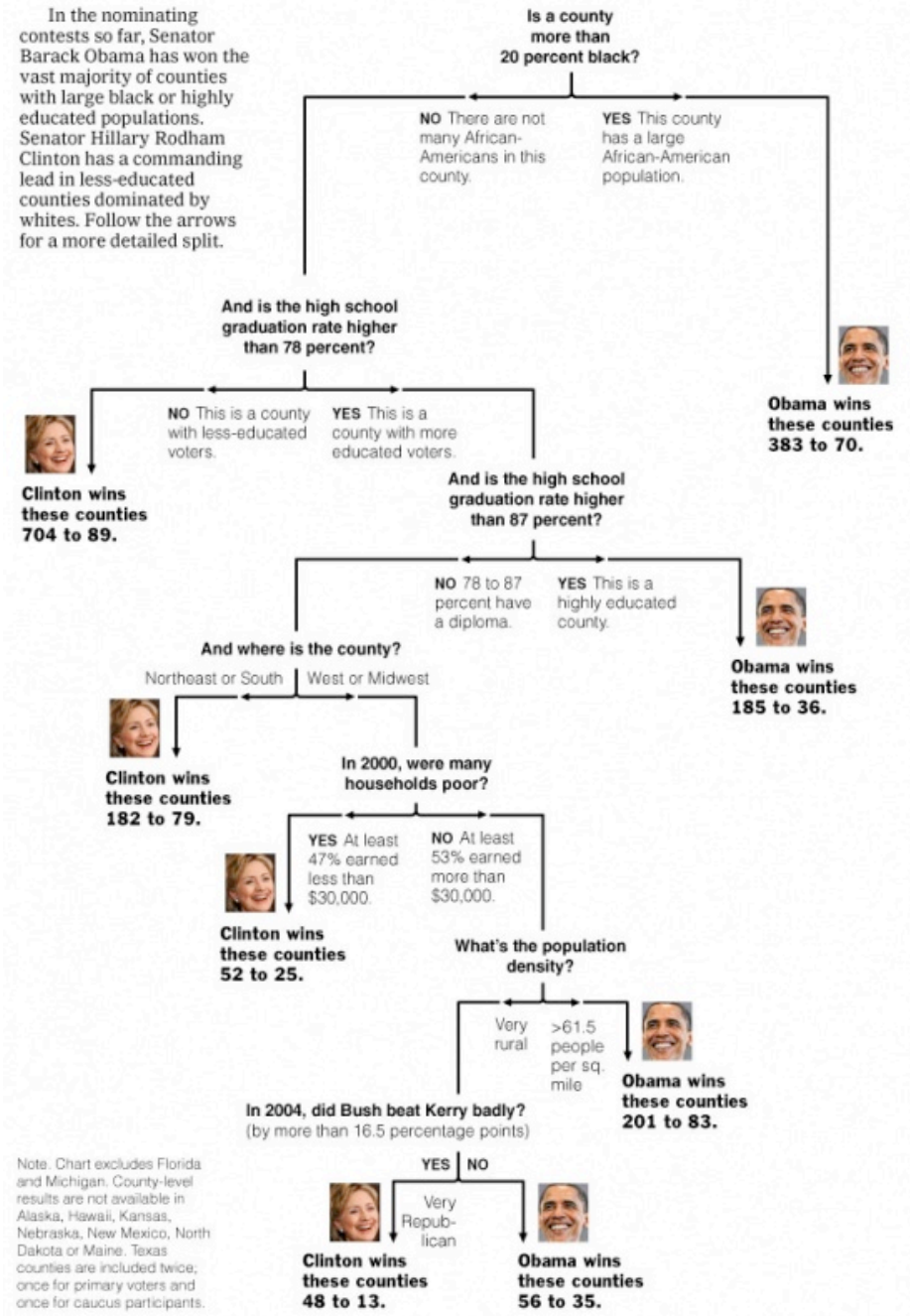
Once this node has been chosen, we work our way down the tree, conducting the same search but on the specified subsets of the data, **at each step attempting to minimize our errors**

```
num = 1000
error = rep(0,num)

fractions = seq(0,1,len=num)

for(i in 1:num){

  # right branch

  tmp = primary$winner[primary$black06pct>fractions[i]]
  error.right = min(table(tmp))

  # left branch

  tmp = primary$winner[primary$black06pct<=fractions[i]]
  error.left = min(table(tmp))
  error[i] = error.left+error.right
}

plot(fractions, error,
     xlab="percentage black",
     ylab="misclassification error",
     type="l")
```


misclassification error
750
800
850
900
950
1000
0.0
0.2
0.4
0.6
0.8
1.0
percentage black

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Note. Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/
THE NEW YORK TIMES

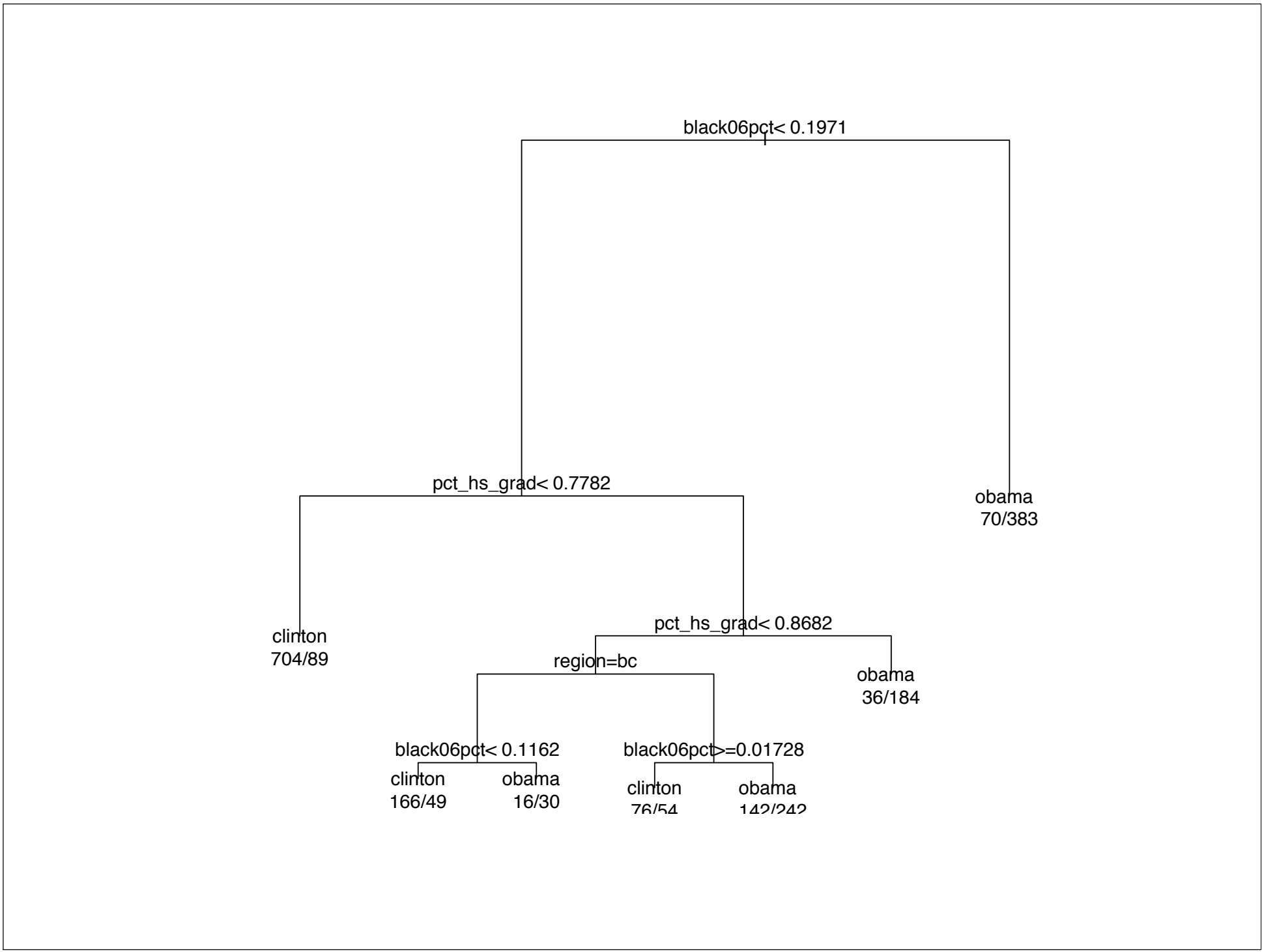
Tree-based models

Stop and think what this simple process has produced for us; we have a very **intuitive structure** (something akin to the game 20 questions) that makes evident **"important" variables that help "explain"** voting patterns

This kind of tool lives somewhere **between data analysis and modeling**; it is technically a model all by itself (making predictions) but is often used as a way to identify important predictors for another stage of model

This decision tree is part of a large class of methods called CART for Classification and Regression Trees and was developed in the 1980s as part of a move to deal with bigger and meaner data sets

```
# library rpart for recursive partitioning

library(rpart)

fit = rpart(winner~region+pct_hs_grad+black06pct,data=primary)

# look at the tree

plot(fit)
text(fit,use.n=T)

# how big should the tree be?
# cp here = complexity parameter not mallows' cp!

plotcp(fit)
```

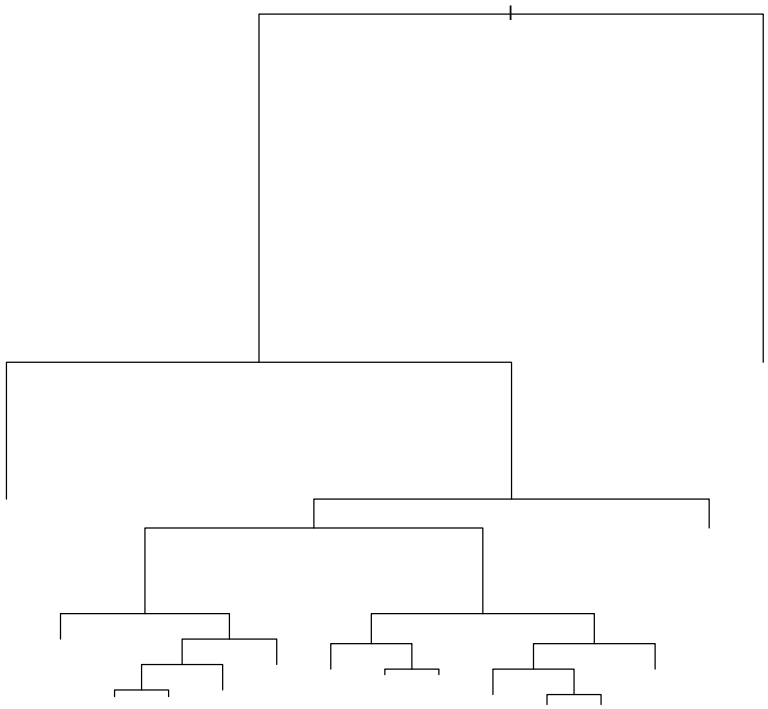

black06pct< 0.1971
pct_hs_grad< 0.7782
obama
70/383
clinton
704/89
pct_hs_grad< 0.8682
region=bc
obama
36/184
black06pct< 0.1162
black06pct>=0.01728
clinton
166/49
obama
16/30
clinton
obama

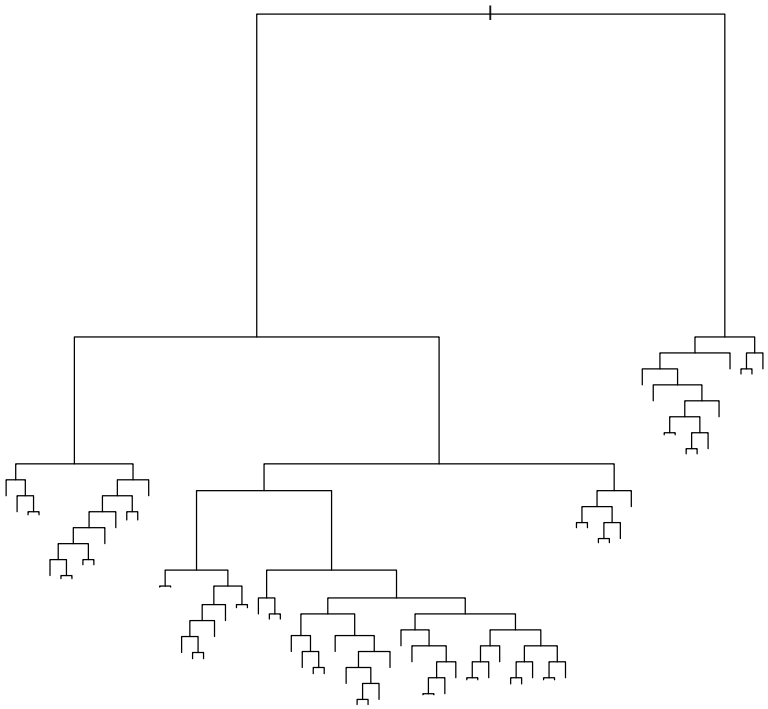
The Democratic Primary

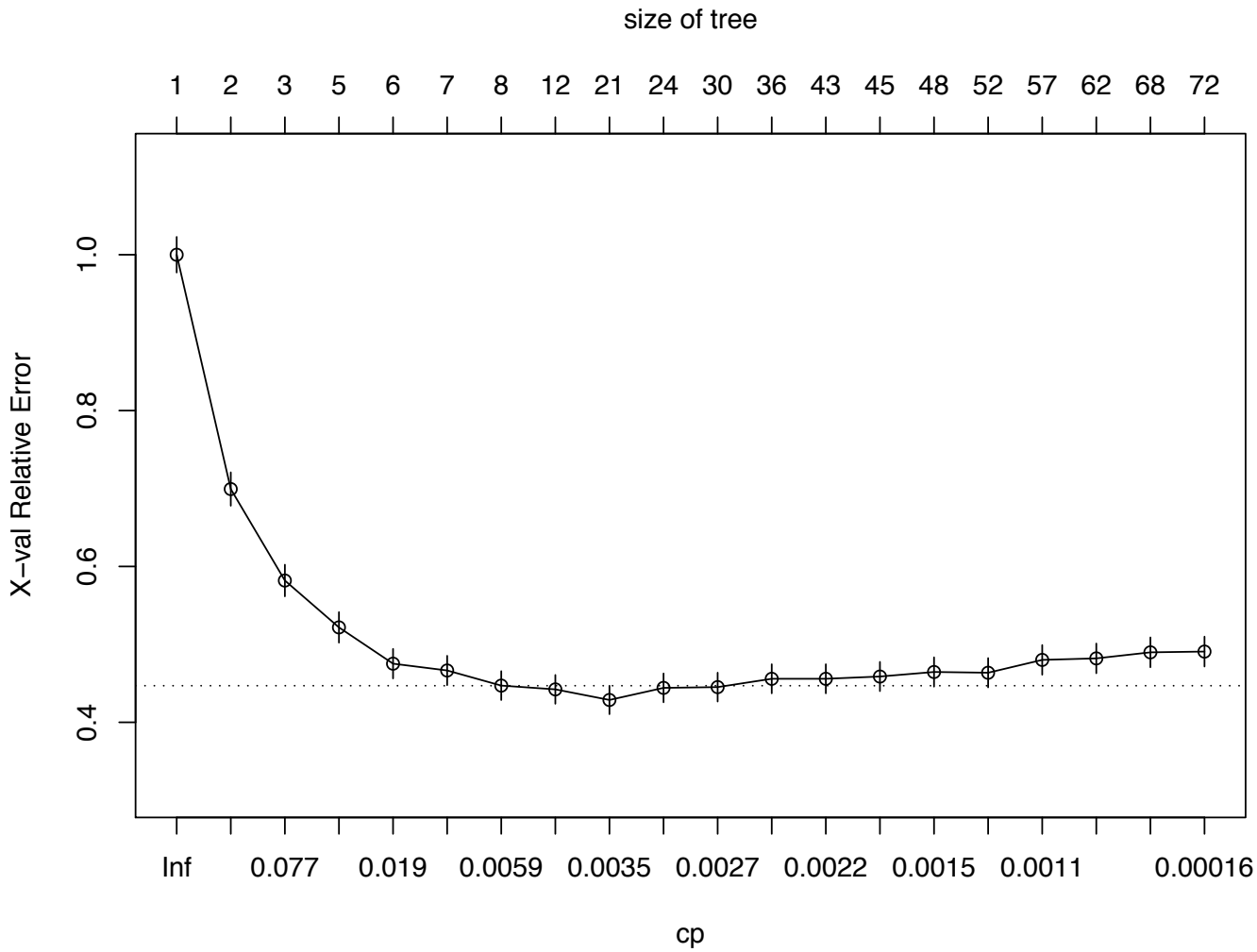
Let's think a bit more about the tree-growing process; with each split, **we cut our data at the node into two pieces** so that the "sample size" at each of the child nodes is lower than its parents

We then represent the data in the leaves with a simple model; for our 0/1 data (Obama or Clinton), we classify leaves according to majority vote

In principle we can grow trees until there's a single entry in each node -- What might be the problem with that? How do we decide to stop splitting? When we run out of data?






size of tree
1 2 3 5 6 7 8 12 21 24 30 36 43 45 48 52 57 62 68 72
X−val Relative Error
0.4 0.6 0.8 1.0
Inf 0.077 0.019 0.0059 0.0035 0.0027 0.0022 0.0015 0.0011 0.00016
cp

The Democratic Primary

In the R displays on the previous pages, **the heights of the branches correspond to the error in represented by the model**

So, the at the root, we are dealing with all 1240 counties; 1210 went for Clinton and 1030 for Obama -- therefore since Clinton won more counties, we would predict all future counties for Clinton, making 1030 mistakes

Last time we saw that the first split on the percentage of the county that is African American brought us down to 700 errors, a big drop; the next division based on education is another big drop, about half the size

As we continue to refine the tree, however, **the improvements diminish**...

The Democratic Primary

In the mid-1980s a fair bit of theoretical and methodological work was devoted to understanding the behavior of this kind of algorithm; **we are using it as a bit of data analysis** (how does a “response” relate to the potential “explanatory” variables?) but it can also be used as a tool for making predictions

The R command `rpart` (for recursive partitioning) employs a “pruning” algorithm that divides the original data into several parts; iteratively leaving one part out, building a tree and evaluating it on the part that was left out -- this is called cross validation

The Democratic Primary

The tradeoff between model complexity and uncertainty is a classic one and we will return to it as we dig deeper into the next topic -- For the moment, however, our interest is in how these tools can be used to highlight relationships between two or more variables in a data set

Specifically, we have a response (in this case the outcome of the Democratic primary) and we would like to understand how it is influenced by values of other variables in the data set, including a county's political, demographic and economic conditions

Our "model" makes predictions of an Obama win depending on the answers to a series of questions, each one dividing the "input space" into rectangles -- Why?

Linear models

We are going to step out of the trees for the moment and consider a much smaller data set and a somewhat simpler (at least conceptually) tool for uncovering relationships

Specifically, we will discuss the basic framework for the so-called linear model (regression analysis) -- We'll motivate the basic "loss function" and introduce some statistics that can be computed

A new example: Mercury contamination

Mercury is a naturally occurring element which is **usually only found in trace amounts in nature**; it is released into the environment, however, as a byproduct burning coal, for example, and the disposal of hazardous waste can contaminate soil and groundwater with mercury

Mercury in the soil and air eventually reach the oceans and groundwater, where aquatic microorganisms have the ability to convert it to methylmercury

Methylmercury in water then accumulates in the tissues of fish and marine animals; **the older the animal the greater the exposure**

Methylmercury in fish is a serious health hazard, especially for children and pregnant women, because it interferes with the developing nervous systems

Mercury contamination

A study was conducted to assess the extent of mercury contamination in two rivers in North Carolina

A total of 171 large mouth bass were caught in the Lumber and Waccamaw Rivers

Fish were caught at 15 different stations; the length, weight and mercury content of each fish was recorded

The New York Times
nytimes.com

March 16, 2005

New Rules Set for Emission of Mercury

By MATTHEW L. WALD

WASHINGTON, March 15 - The Environmental Protection Agency released its final rule on mercury emissions from power plants on Tuesday, asserting that allowing companies to buy and sell the right to pollute would encourage control of the biggest sources of mercury first.

Mercury from smokestacks poses a hazard, especially to children and developing fetuses, because it eventually ends up in rivers and lakes, where it is absorbed by fish that are then caught and eaten by people.

Some environmentalists said the agency should have simply required uniform emission limits, to reduce concentrations everywhere. They say the new rule means that some plants will end up doing nothing to curb emissions, allowing mercury "hot spots" to persist, affecting the health of people living nearby.

Summaries

Lumber (n=73)

Length: 39.41 (8.30)
Weight: 1197.16 (943.00)
Mercury: 1.07 (0.64)

Waccamaw River (n=98)

Length: 40.38 (8.68)
Weight: 1111.22 (824.75)
Mercury: 1.28 (0.83)

Mercury contamination

In this R dump of the 171 points, the first 73 observations correspond to fish from the Lumber River

`river = 0, stn=0,...,6`

The final 98 data points correspond to fish from the Waccamaw River

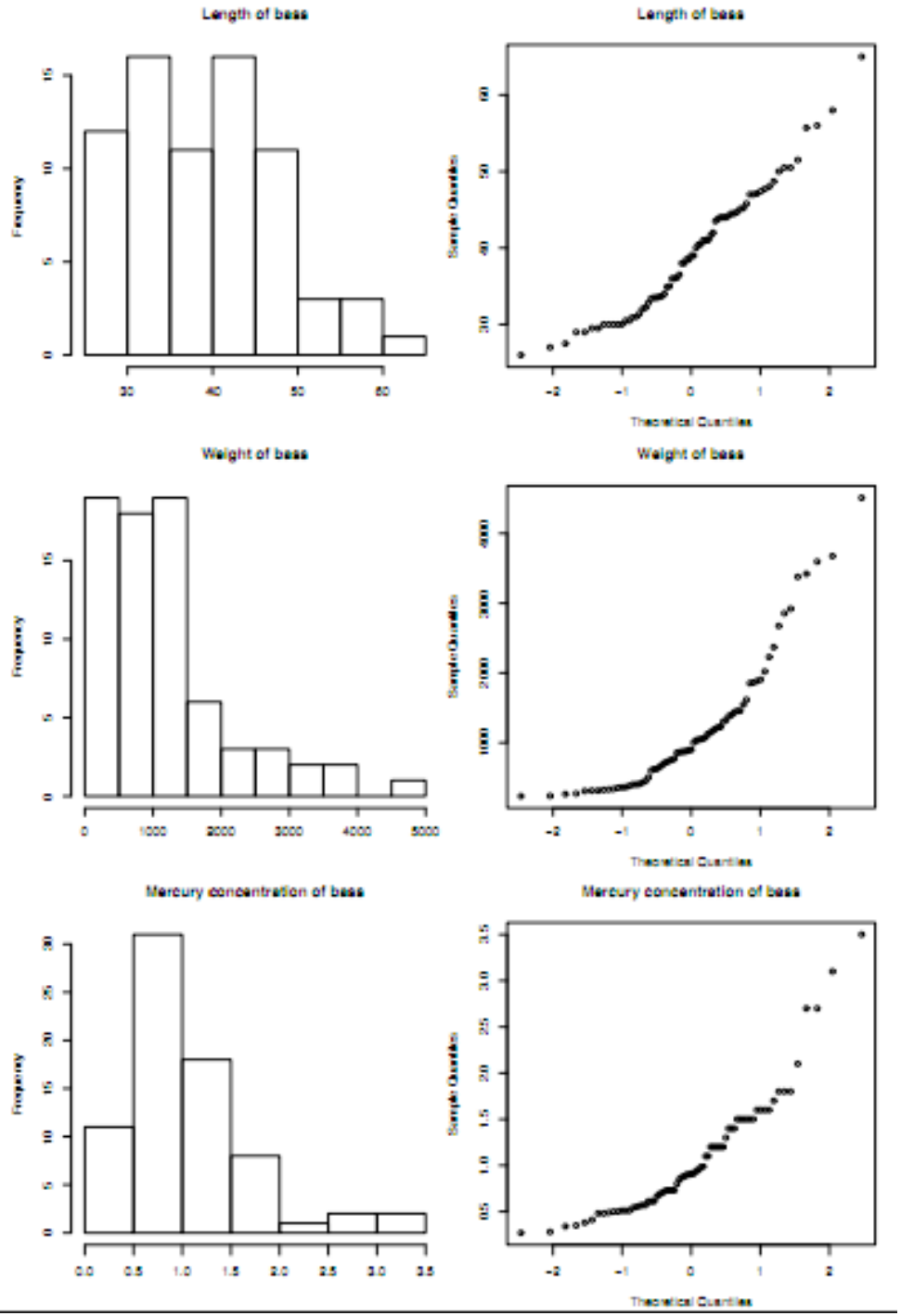
`river = 1, stn=7,...,15`

	river	stn	length	weight	mercury
1	0	0	47.0	1616	1.60
2	0	0	48.7	1862	1.50
3	0	0	55.7	2855	1.70
4	0	0	45.2	1199	0.73
5	0	0	44.7	1320	0.56
6	0	0	43.8	1225	0.51
7	0	0	38.5	870	0.48
8	0	0	45.8	1455	0.95
9	0	0	44.0	1220	1.40
10	0	0	40.4	1033	0.50
11	0	1	47.7	3378	0.80
12	0	1	45.1	2920	0.34
13	0	1	43.5	2674	0.54
14	0	1	47.4	3675	0.69
15	0	1	41.0	1904	0.90
...	...	...	...	...	...
157	1	14	40.0	869	1.40
158	1	14	37.4	879	1.60
159	1	14	46.5	772	1.70
160	1	14	36.0	724	1.30
161	1	15	50.4	1744	0.93
162	1	15	59.2	3524	3.60
163	1	15	58.4	2902	3.50
164	1	15	54.0	2709	2.40
165	1	15	53.7	2625	2.90
166	1	15	49.5	1924	2.30
167	1	15	47.5	1546	1.40
168	1	15	54.2	3164	2.10
169	1	15	45.4	1710	1.70
170	1	15	41.7	1255	1.40
171	1	15	36.0	702	0.92

Mercury contamination

At this point, we could consider various 1-dimensional summaries; we could look at the distribution of mercury content or lengths or weights

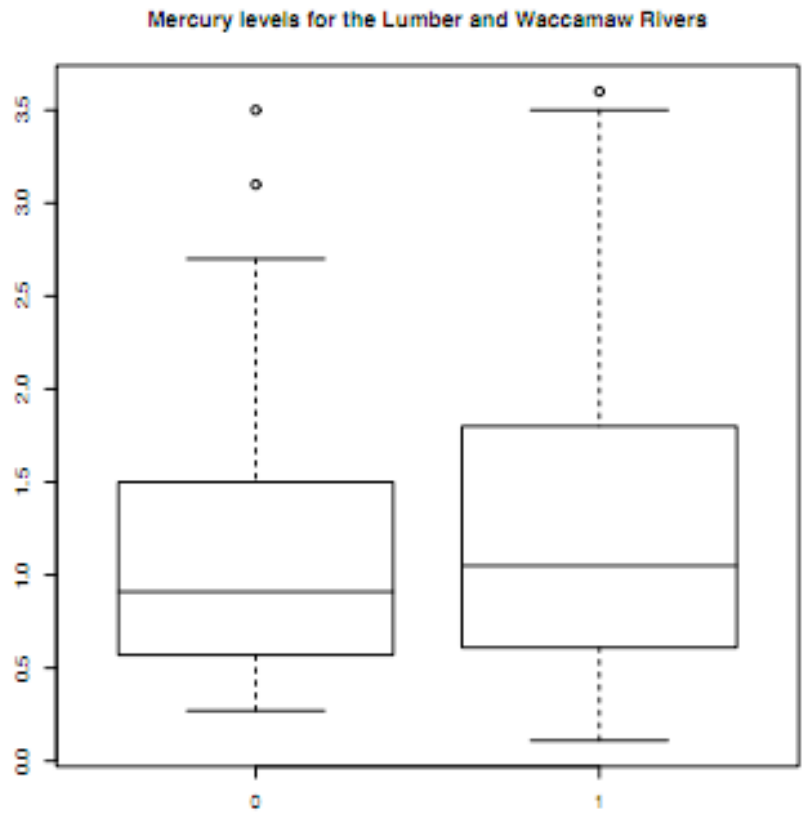
At the right we have our old friends the histogram and the boxplot



Relationships between variables

Comparisons between similar measurements taken from different populations could be made by overlaying simple 1-dimensional summaries

Here we have two boxplots for the Mercury levels of fish from the Lumber (0) and Waccamaw (1) rivers

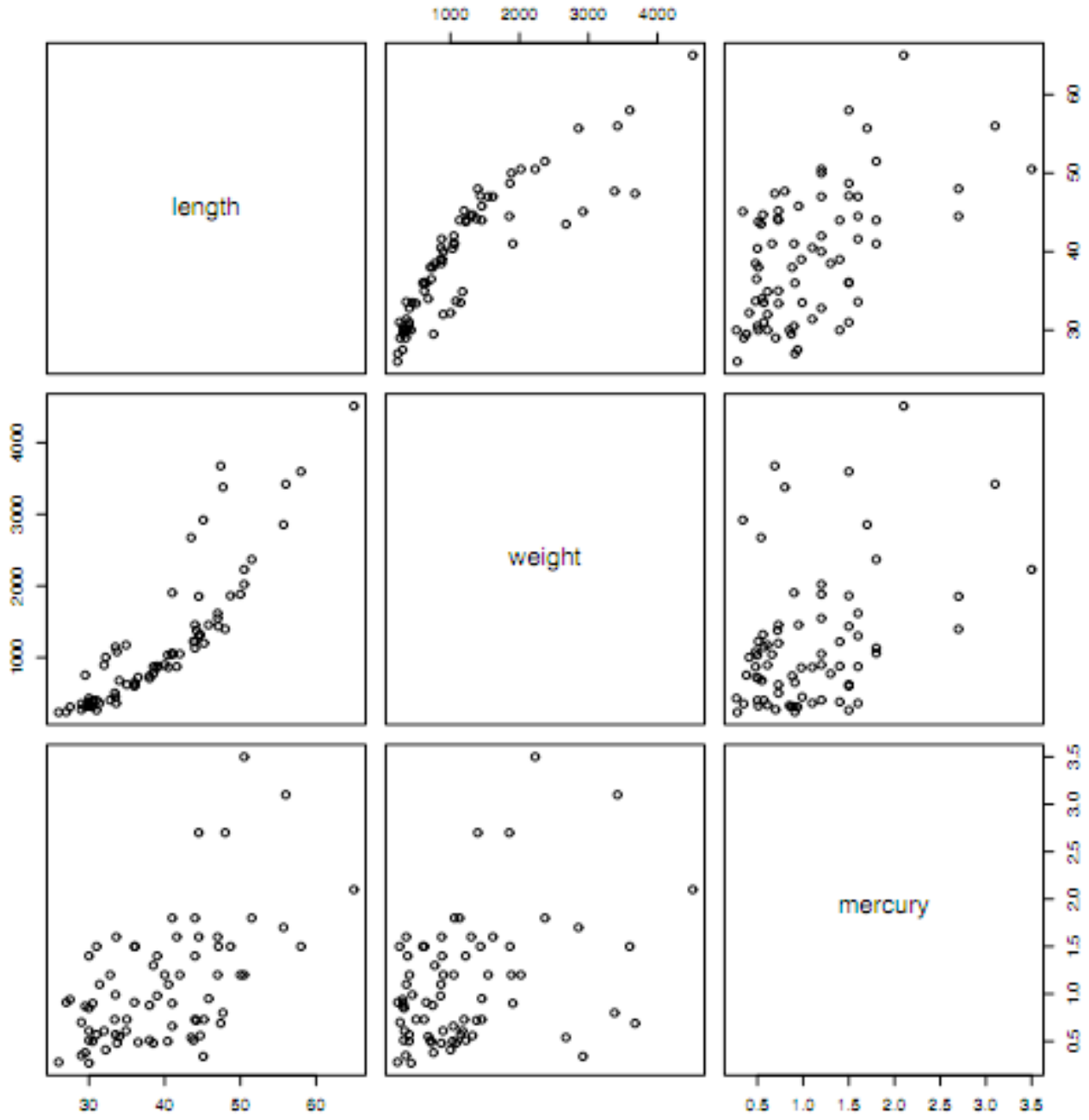


Mercury contamination

In this application, we want to understand how two or more variables relate directly to each other: What is the relationship between a fish's size and the amount of mercury that has built up in its system?

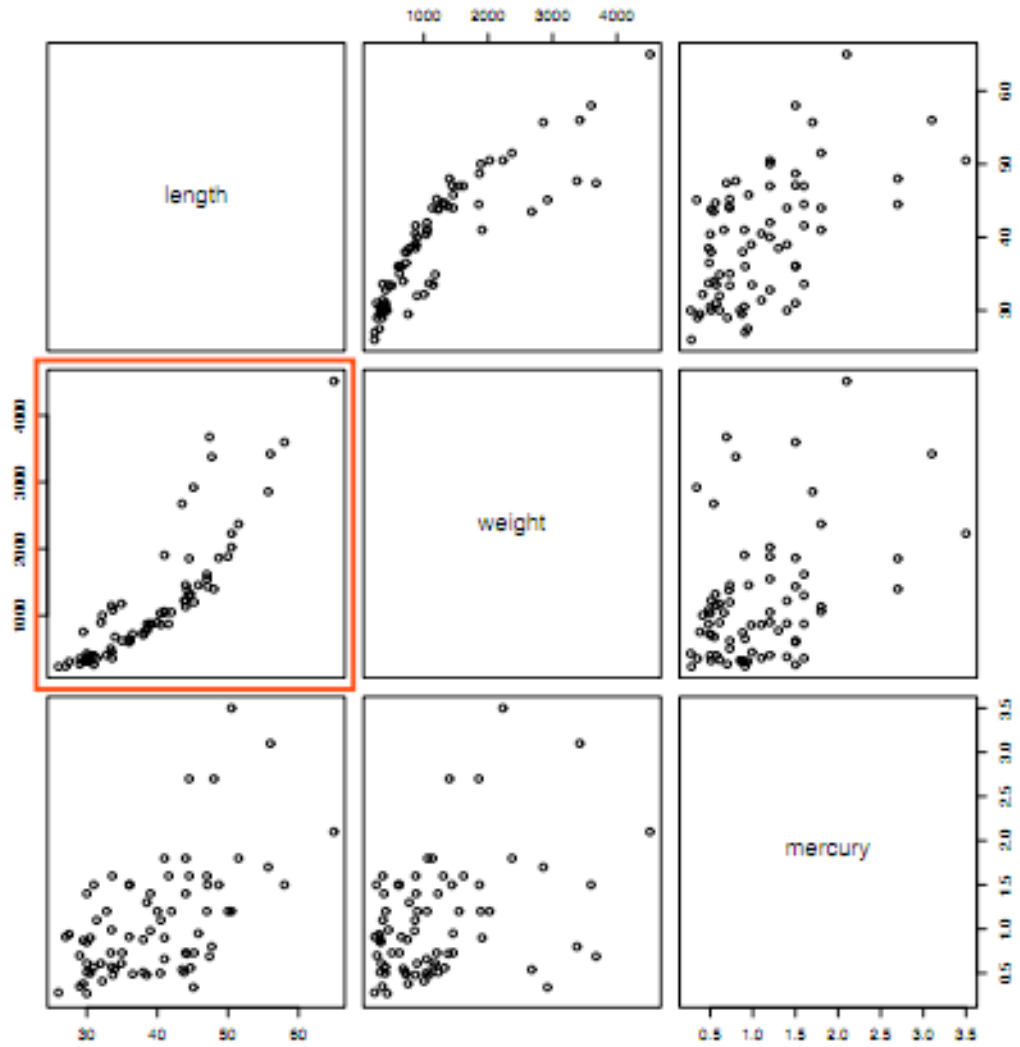
For this, we look at scatterplots; the scatterplot matrix allows us to look at several pairs of variables at one time

Lumber River



Lumber River

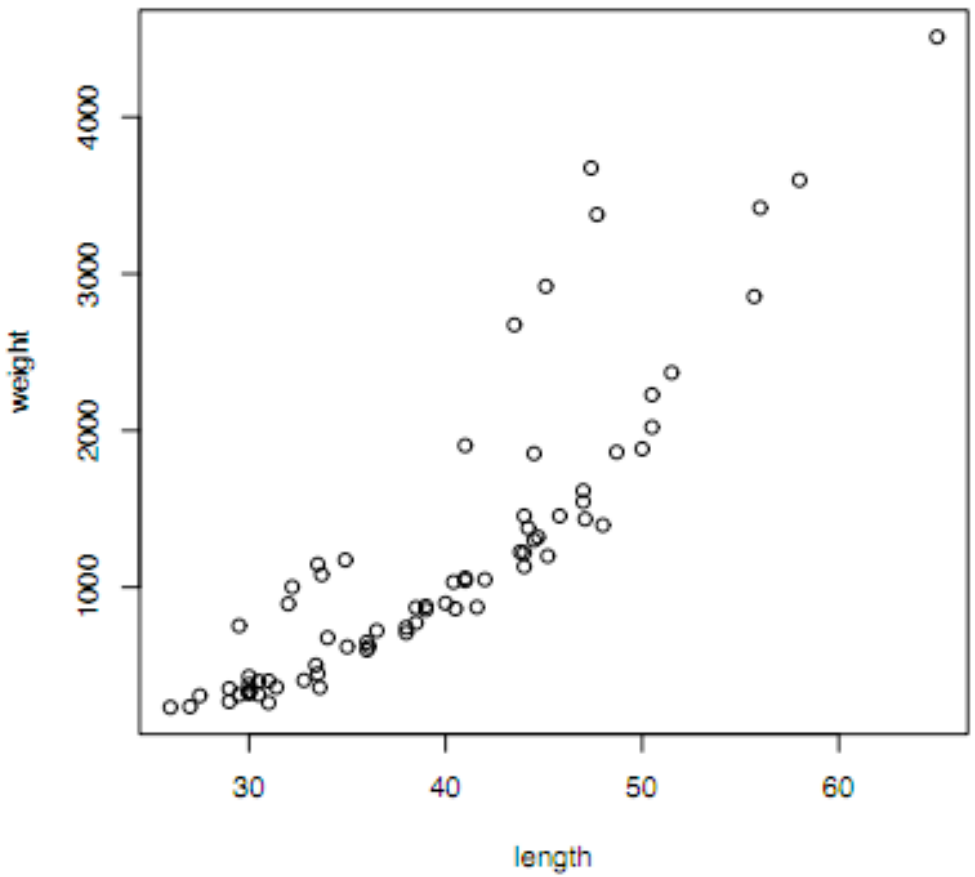
The plot in row 2 and column 1 is a scatter plot of weight on the y-axis and length on the x-axis



A scatterplot

Here we isolate just one comparison; the lengths and weights of fish in the Lumber river

What do we notice?



A scatterplot

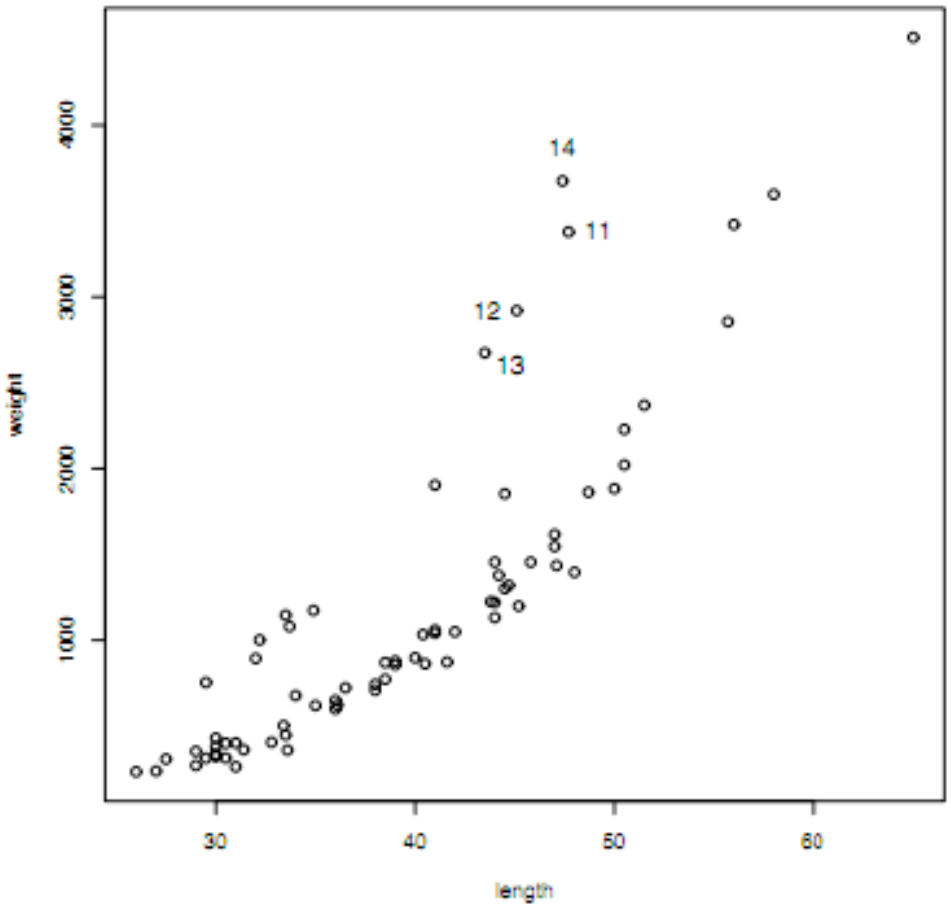
We can explore things a bit more using the `identify` command

```
> plot(fish$length, fish$weight)
> identify(fish$length,fish$weight)
```

A scatterplot

We can find the row number of any observation in a scatterplot using the `identify` command in R; after typing the last command, you can click in the plot

```
> plot(fish$len,fish$wei)
> identify(fish$len,fish$wei)
```



A scatterplot

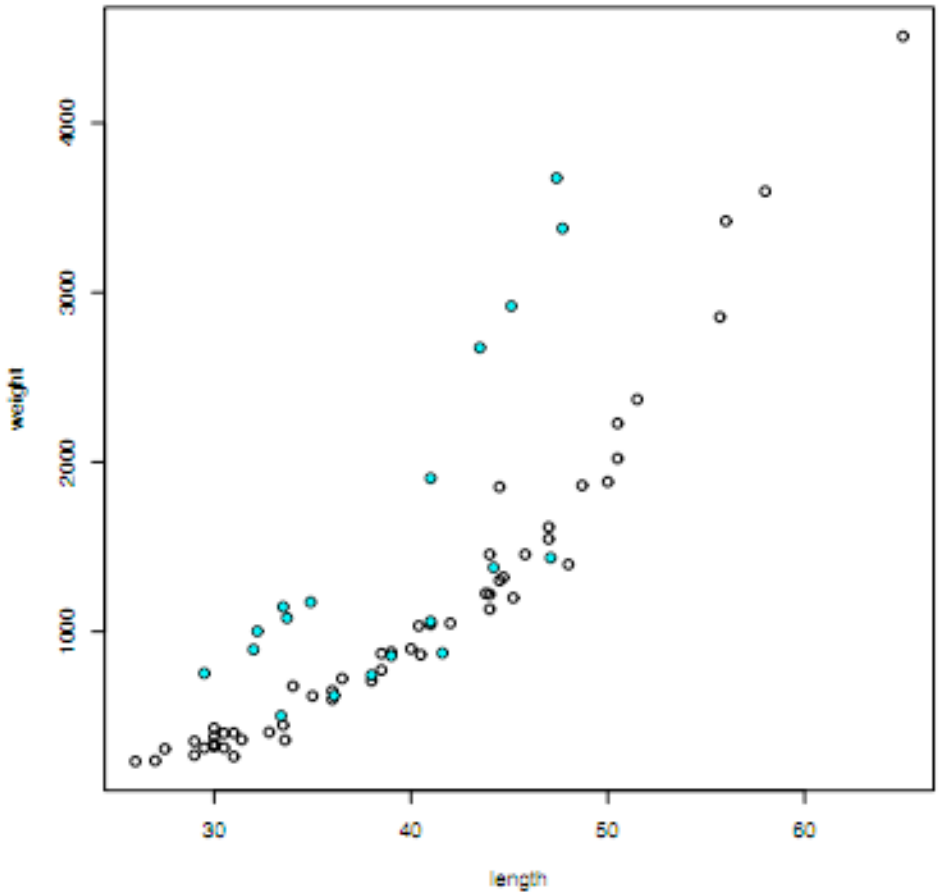
If we store the output of `identify` we can use it to subset our data matrix and visually inspect the outlying observations

```
> plot(fish$length, fish$weight)
> uu<- identify(fish$length,fish$weight)
> uu
 [1] 11 12 13 14 15 16 17 18 19 20 21
```

Mercury levels in water

```
> fish[uu,]

    stn length weight mercury
11    1   47.7   3378    0.80
12    1   45.1   2920    0.34
13    1   43.5   2674    0.54
14    1   47.4   3675    0.69
15    1   41.0   1904    0.90
16    1   33.7   1080    0.48
17    1   33.5   1146    0.57
18    1   32.2   1002    0.41
19    1   32.0    894    0.61
20    1   29.5    754    0.38
21    1   34.9   1174    0.61
```

Seems like they are all station 1; the cyan points mark fish coming from station 1

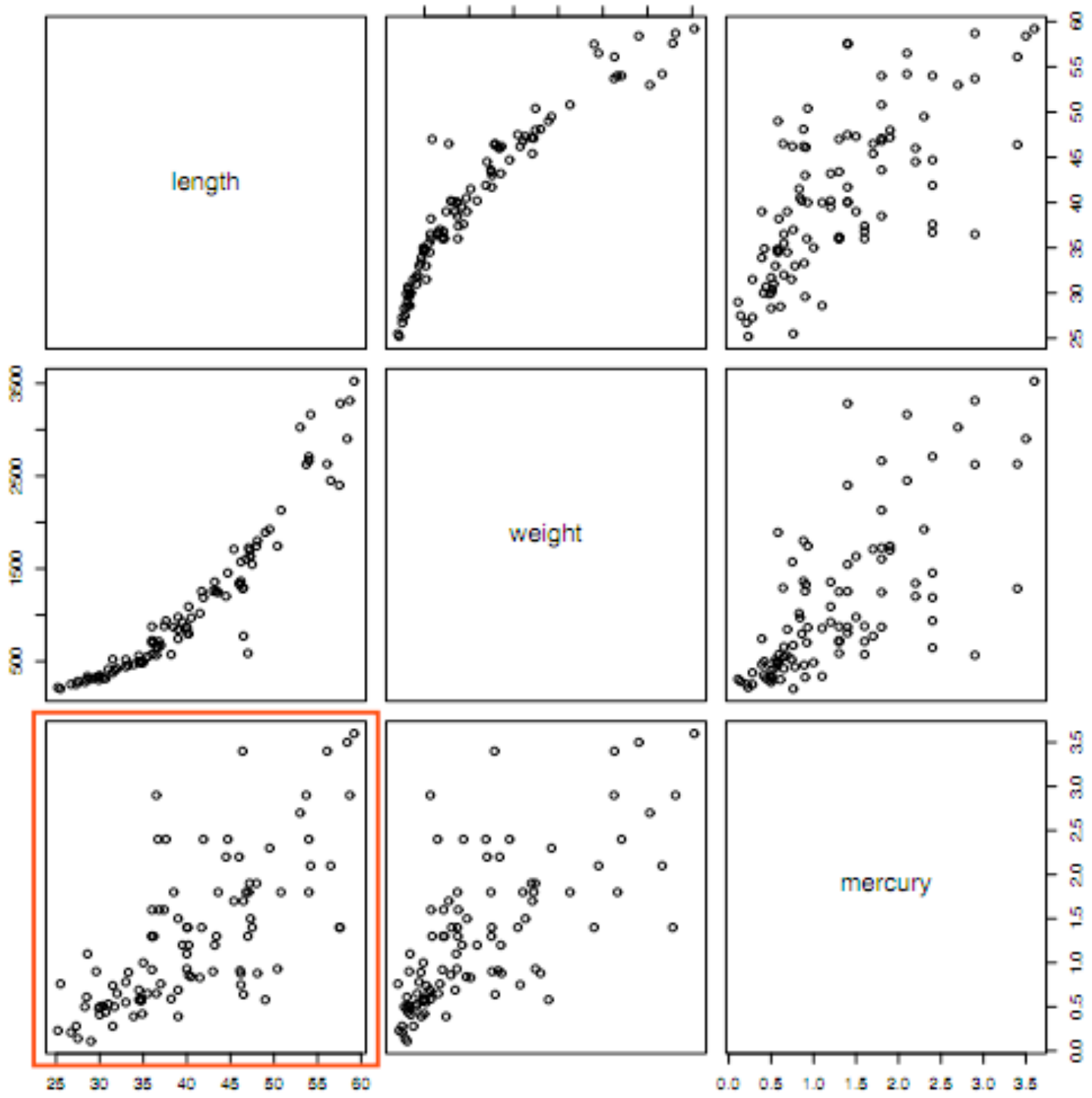


Mercury contamination

From the EPA's perspective, it is natural to wonder how mercury content relates to the length of a fish -- That is, when providing guidelines to people fishing along these rivers that they can use to assess the safety of a fish when they catch it

Obviously, age is the best predictor of mercury content, but that's not readily apparent to a fisherman -- We'll see to what extent length can be used as a proxy for age when it comes to predicting the amount of mercury in a fish

We'll first use data from the Waccamaw river...


length
weight
mercury
500
1500
2500
3500
25
30
35
40
45
50
55
60
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Modeling

Based on this plot, we might be tempted to describe the relationship between fish length and mercury content mathematically; that is, we construct a model relating length and mercury based on the data

The usefulness of the model depends on its ability to capture the major trends in the data

We might also be interested in making predictions: If we've just caught a fish, can we predict mercury content from data we can easily measure like length, and if so, how accurate are these predictions?

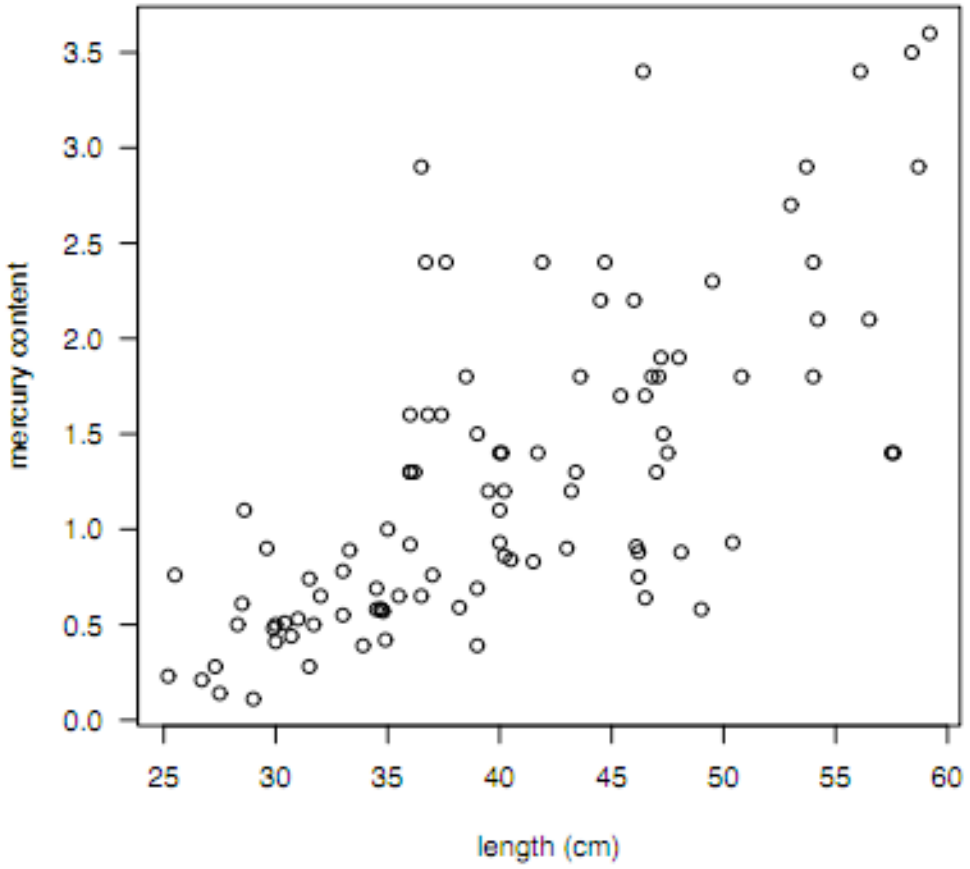
A linear model

To describe this relationship mathematically, we need to relate the input (length) to the output (mercury)

The simplest kind of model of this type is just a line

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

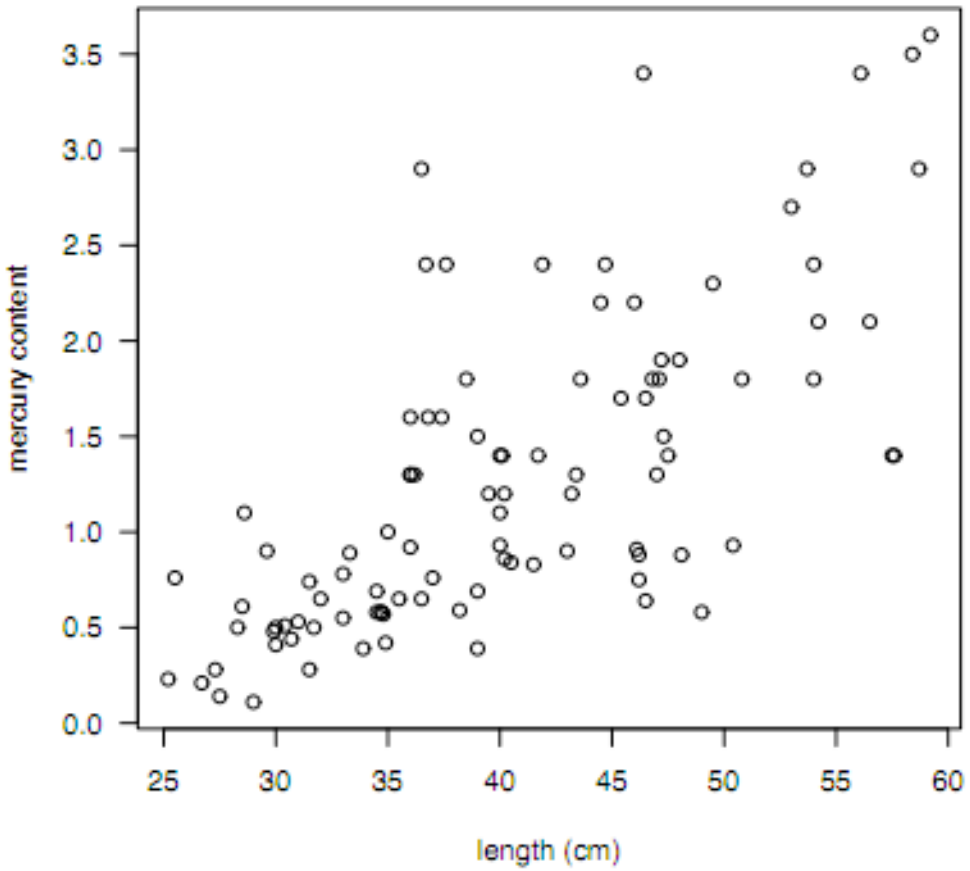
where β_0 and β_1 are parameters, the slope and intercept



A linear model

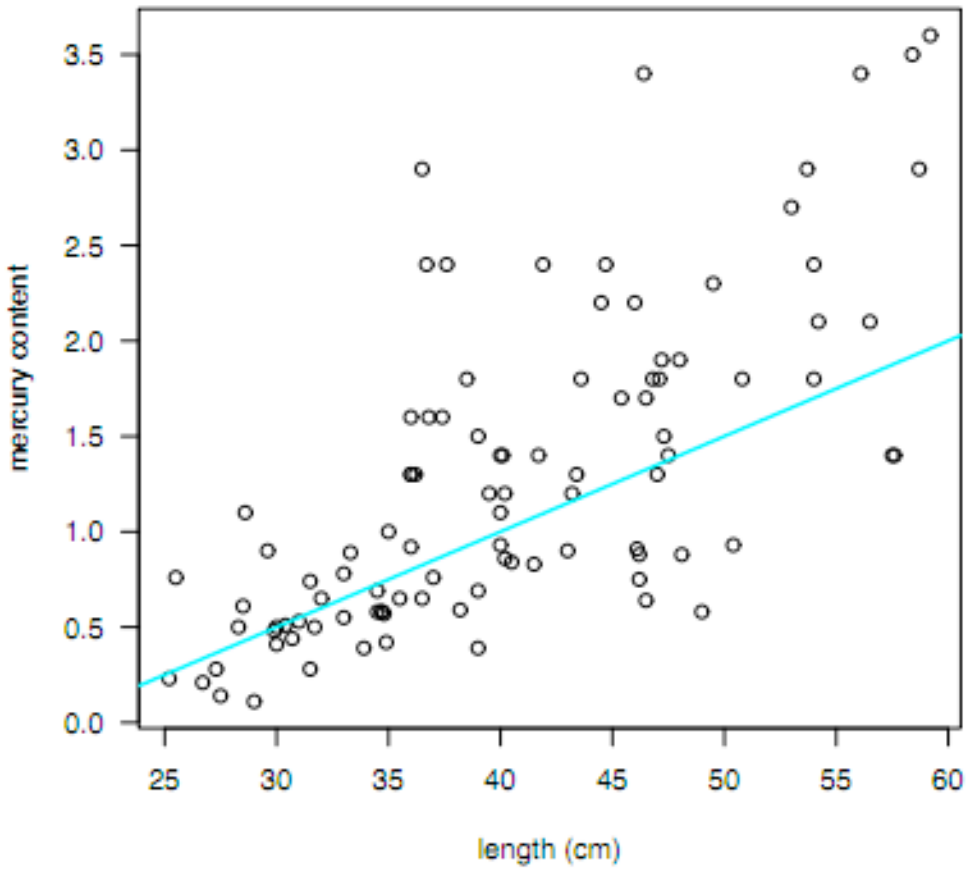
In terms of our data, we might posit a model of the form

$$(\text{mercury}) = \beta_0 + \beta_1(\text{length}) + (\text{error})$$



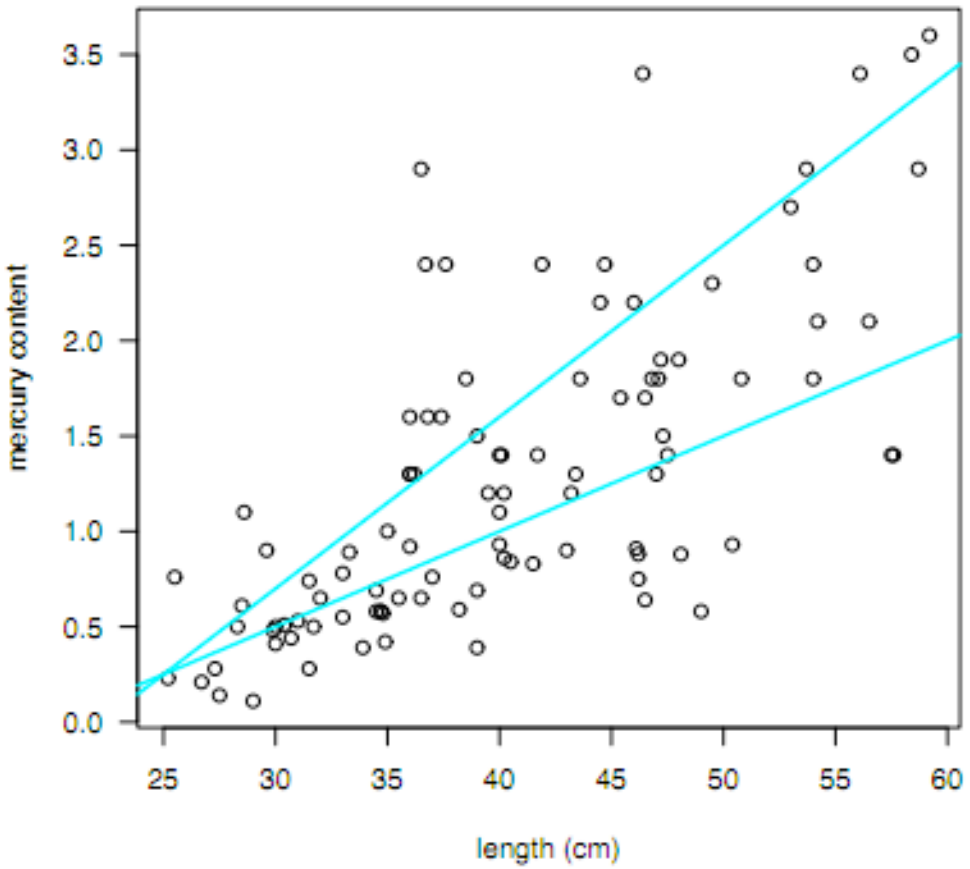
Linear models

There are many lines one could draw through the data... which one is the “best”?



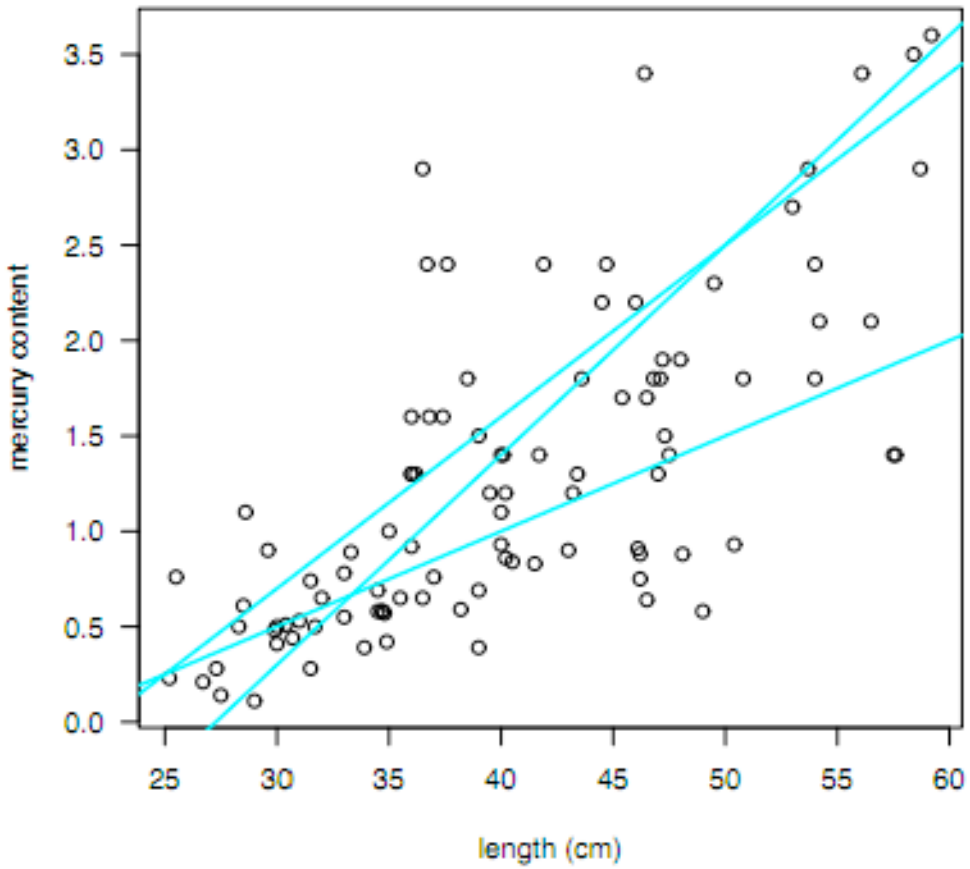
Linear models

There are many lines one could draw through the data... which one is the “best”?



Linear models

There are many lines one could draw through the data... which one is the “best”?



Least squares

The method of least squares provides us with a way to select the slope and intercept: For simplicity (and ultimately, generality) define the following two variables for each of the 98 fish in the Waccamaw river data set

x = fish length and y = mercury content

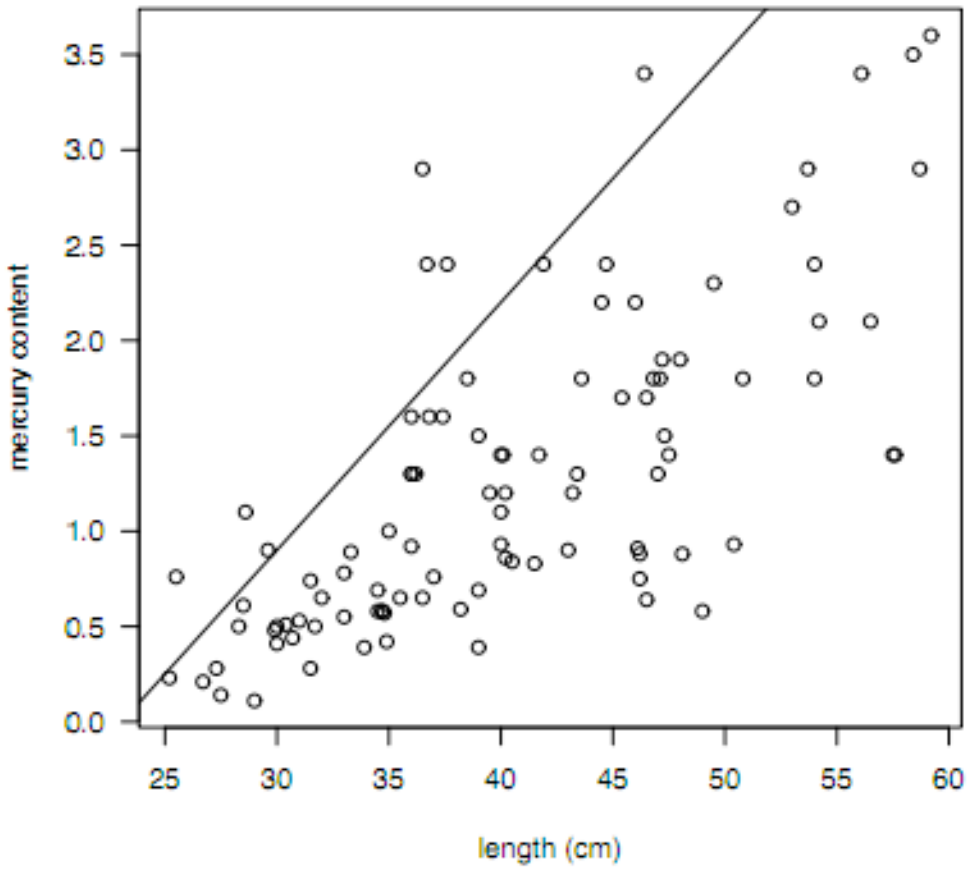
We then label our data set $(x_1, y_1), \ldots, (x_{98}, y_{98})$

Least squares

Specify a choice for the slope and intercept

Here we have selected an intercept of -3 and a slope of 0.13; or in terms of our parameters

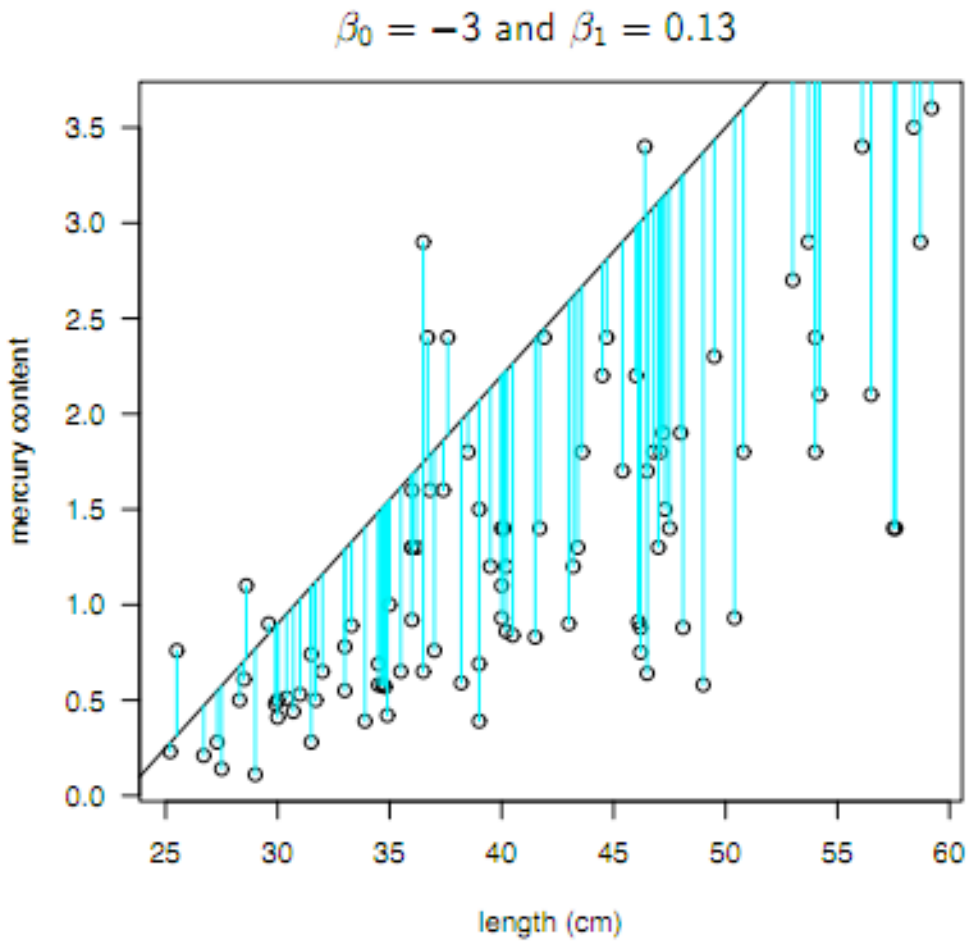
$\beta_0 = -3$ and $\beta_1 = 0.13$



Least squares

We then measure the distance from each data point to the line

If we were to use our line to "predict" the value of mercury at each weight in the data set, then these are the errors we would make



Least squares

We then consider the sum of squared errors from the predicted values (points on the line) and the actual observations

$$\sum_{i=1}^{98}[y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{98}[y_i - (-3 + 0.13x_i)]^2$$

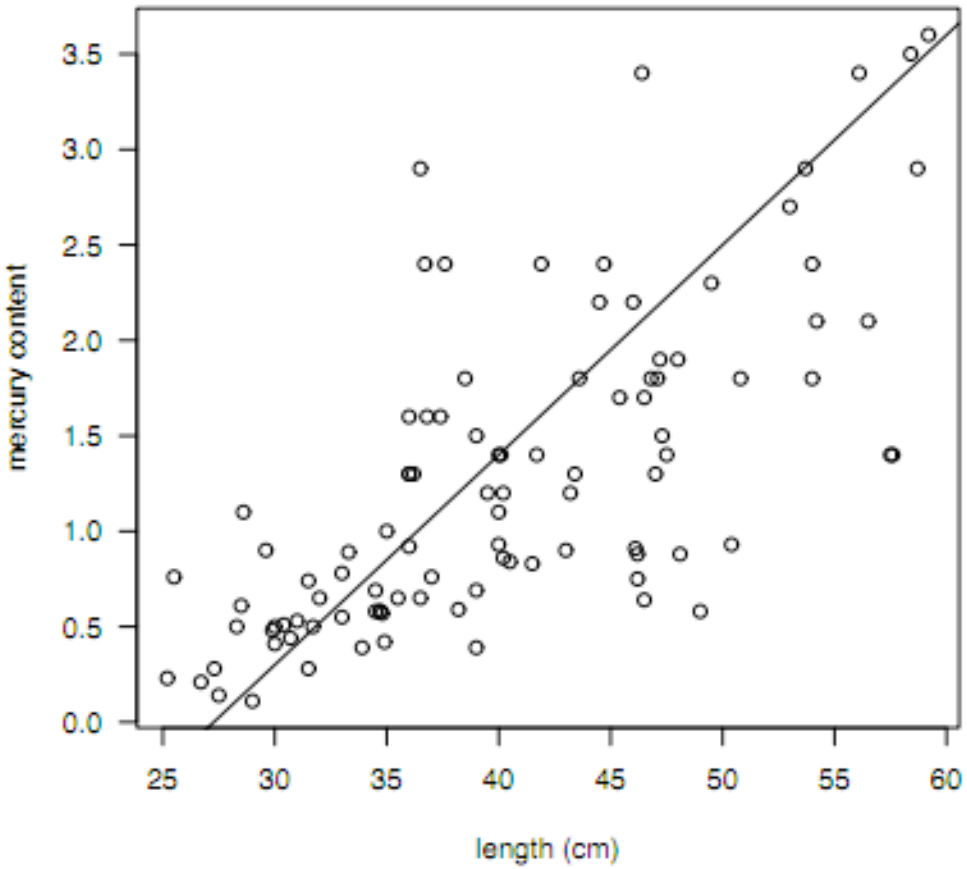
In this case, the squared error is 154.9

Least squares

Let's try another line

Here we have selected an intercept of -3 and a slope of 0.11; or in terms of our parameters

$\beta_0 = -3$ and $\beta_1 = 0.11$

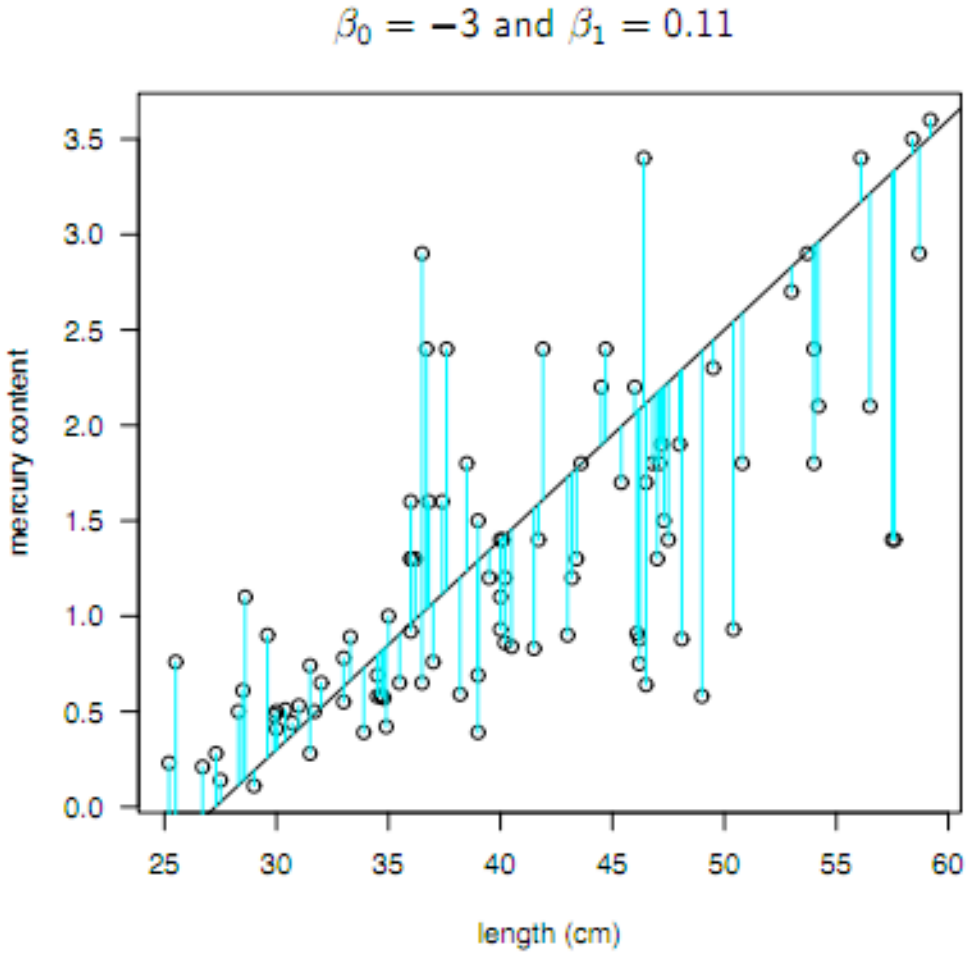


Least squares

We then measure the distance from each data point to the line

If we were to use our line to "predict" the value of mercury at each weight in the data set, then these are the errors we would make

In this case, the squared error sum to 49.26; how did we do?



Least squares

We define the "best" choice of the intercept β_0 and β_1 slope to be the ones that minimize the sum of squares

$$\sum_{i=1}^{98} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

The values that make this quantity the smallest are unique (assuming some things about the data; but we'll ignore that for now)

We use $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to denote them, and refer to them as "least squares estimates"

Least squares

For our mercury data, **the least squares fit** corresponds to

$$\widehat{\beta}_0 = -1.45 \text{ and } \widehat{\beta}_1 = 0.068$$

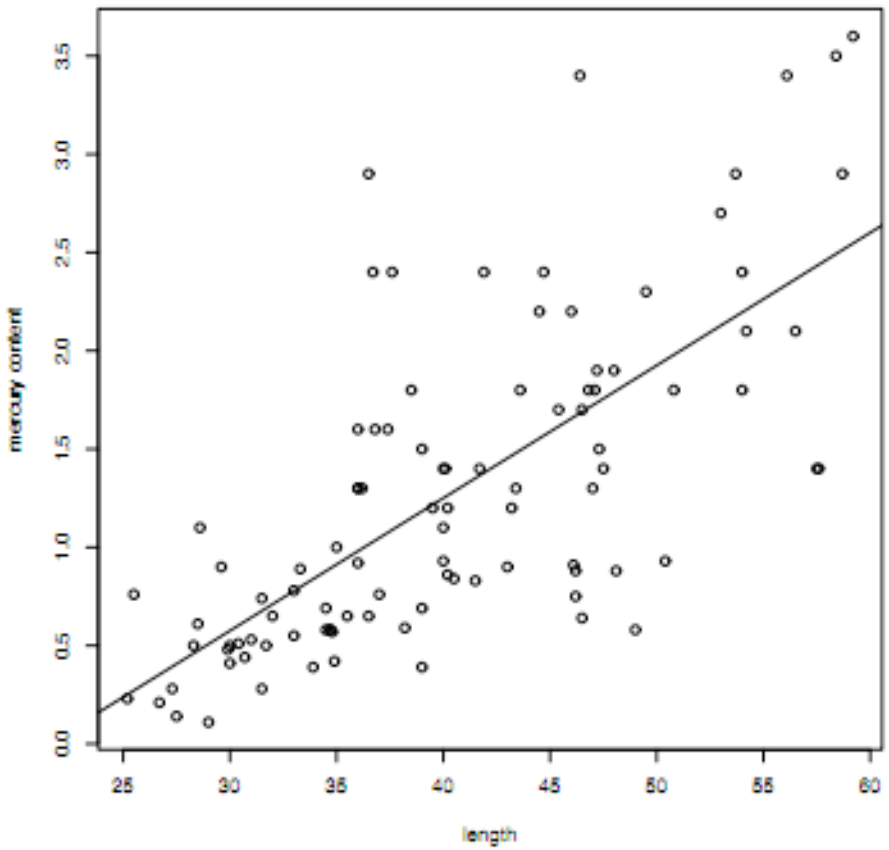
and the associated **sum of squares** is 33.4 (our simple trial and error approach was pretty far off!)

The least squares fit is often called **the regression line**, and the difference between the fitted and observed values

$$r_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$$

are called **residuals**

The sum of squares associated with the least squares line is referred to as **the residual sum of squares**



Least squares

For this simple model (and by "simple" we mean a linear equation with just a single input variable -- in this case, length) we can **write down the least squares fit exactly**

$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \text{and} \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$$

Downstairs in the expression for $\widehat{\beta}_1$ we have a quantity that looks an awful lot like the standard deviation of the x-values; **if for some reason this is zero, we no longer have a unique solution for the least squares line** -- Does this make sense intuitively?

Some interpretation

The magnitude of the slope $\hat{\beta}_1$ represents, in an average sense (with respect to the errors around the line), **the rate of change of Mercury content with length; it has units of ppm/cm**

Since $\hat{\beta}_1$ = 0.068 ppm/cm, the least squares summary says that **for each centimeter of length, fish in our sample contain, on average, 0.068 ppm Mercury**

A useful comparison

While this idea of minimizing squared differences might seem new, **we've seen an example of this before**

What can you say about the value of b that minimizes

$$\sum_{i=1}^{98} [y_i - b]^2$$

Flashback: The sample mean and standard deviation

The value of b that minimizes $\sum(y_i - b)^2$ is **the sample mean** $\bar{y}$

Recall that the sum of squared deviations, or in our current terminology "residuals," from this "fit" is the main ingredient in the **sample standard deviation**

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

which measures the spread of the data around the mean $\bar{y}$

We divided by n-1 because the expression involved a single estimate, the sample mean $\bar{y}$ (we showed that this meant that sum of the deviations was zero and so we didn't have n independent pieces of information in the sum)

Residual standard deviation

By analogy with this simple setup, we will define the **residual standard deviation** to be

$$s_{y|x} = \sqrt{\frac{1}{n-2}\sum r_i^2} = \sqrt{\frac{1}{n-2}\sum \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2}$$

where we have now divided by n-2 because we have two estimates in our expression, $\hat{\beta}_0$ and $\hat{\beta}_1$

One can also show that the residuals from the least squares line satisfy two constraints

$$\sum r_i = 0 \quad \text{and} \quad \sum x_i r_i = 0$$

meaning that we have **n-2 independent pieces of information in this sum**

More interpretation: Residuals

From the first of these constraints, $\sum r_i = 0$, we can conclude that the **residuals from the least squares fit have an arithmetic mean of 0;** their spread is captured by the residual standard deviation

The second constraint has to do with the correlation between the residuals and the input data, the predictor variable; we'll make this precise in the next lecture

More interpretation: Residuals

Before we leave this minimization idea, we want to comment on the two minimization problems

$$\underset{\text{over } b}{\text{minimize}} \sum [y_i - b]^2 \qquad \text{and} \qquad \underset{\text{over } b_0, b_1}{\text{minimize}} \sum [y_i - (b_0 + b_1 x_i)]^2$$

Notice that by setting $b_1 = 0$ in the second expression, the two are really the same problem; because we let b_1 vary in the second expression, however, it stands to reason that **its minimum value will be at least as large as that for the first expression** -- In other words

$$\sum [y_i - \bar{y}]^2 \geq \sum \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

You can think of the gap as a measure of the usefulness of the variable x (in our case, fish length) in describing our data

More interpretation: Residuals

We capture the gap through **the coefficient of determination**

$$R^2 = 1 - \frac{\sum \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2}{\sum [y_i - \bar{y}]^2}$$

This expression takes values between 0 and 1; with **1 meaning the least squares line is a perfect fit** (all zero residuals) and 0 meaning the variable we introduced (in our case, fish length) was of no help in describing the relationship between x and y (the coefficient $\hat{\beta}_1$ is zero)

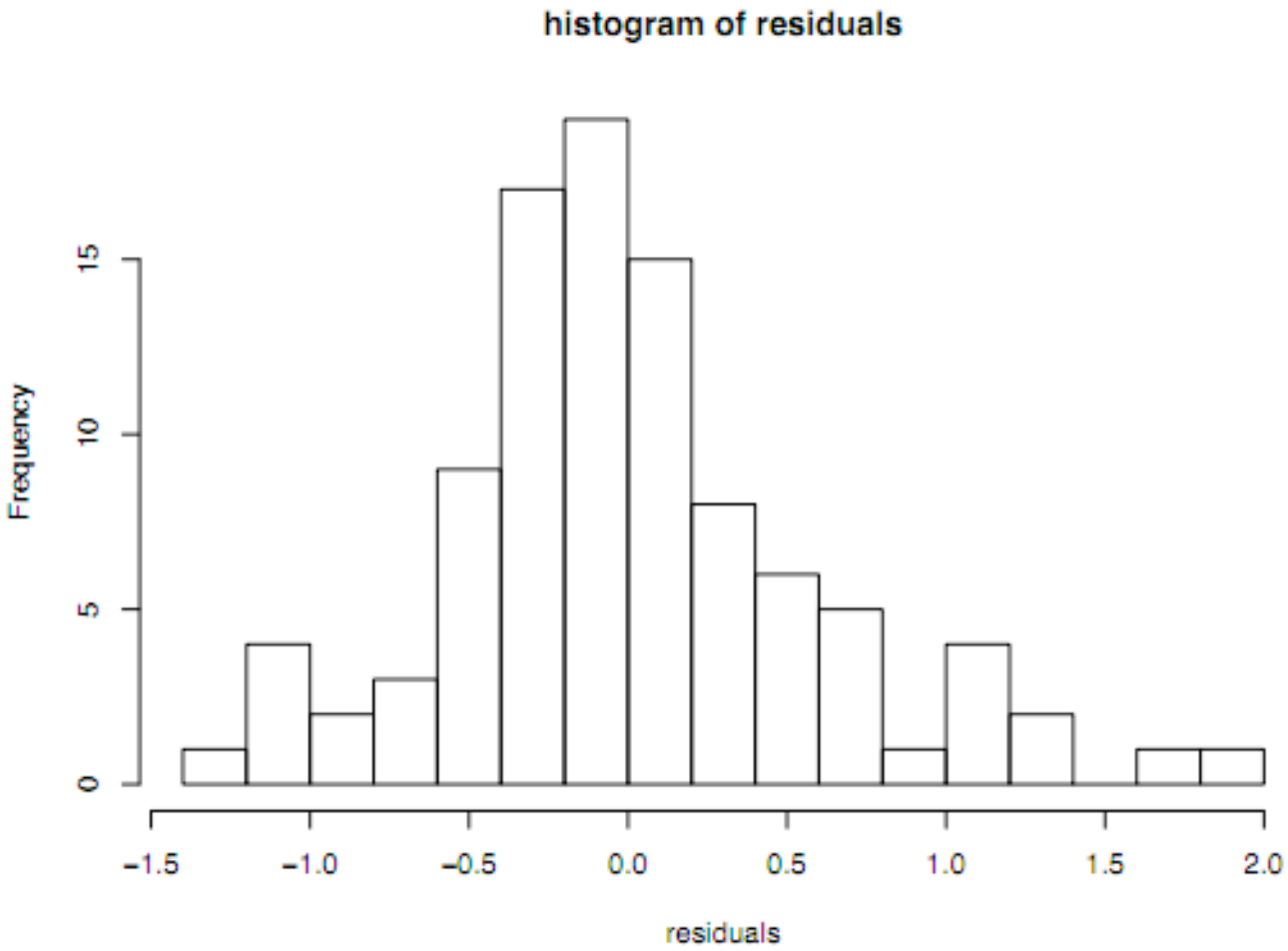
An example

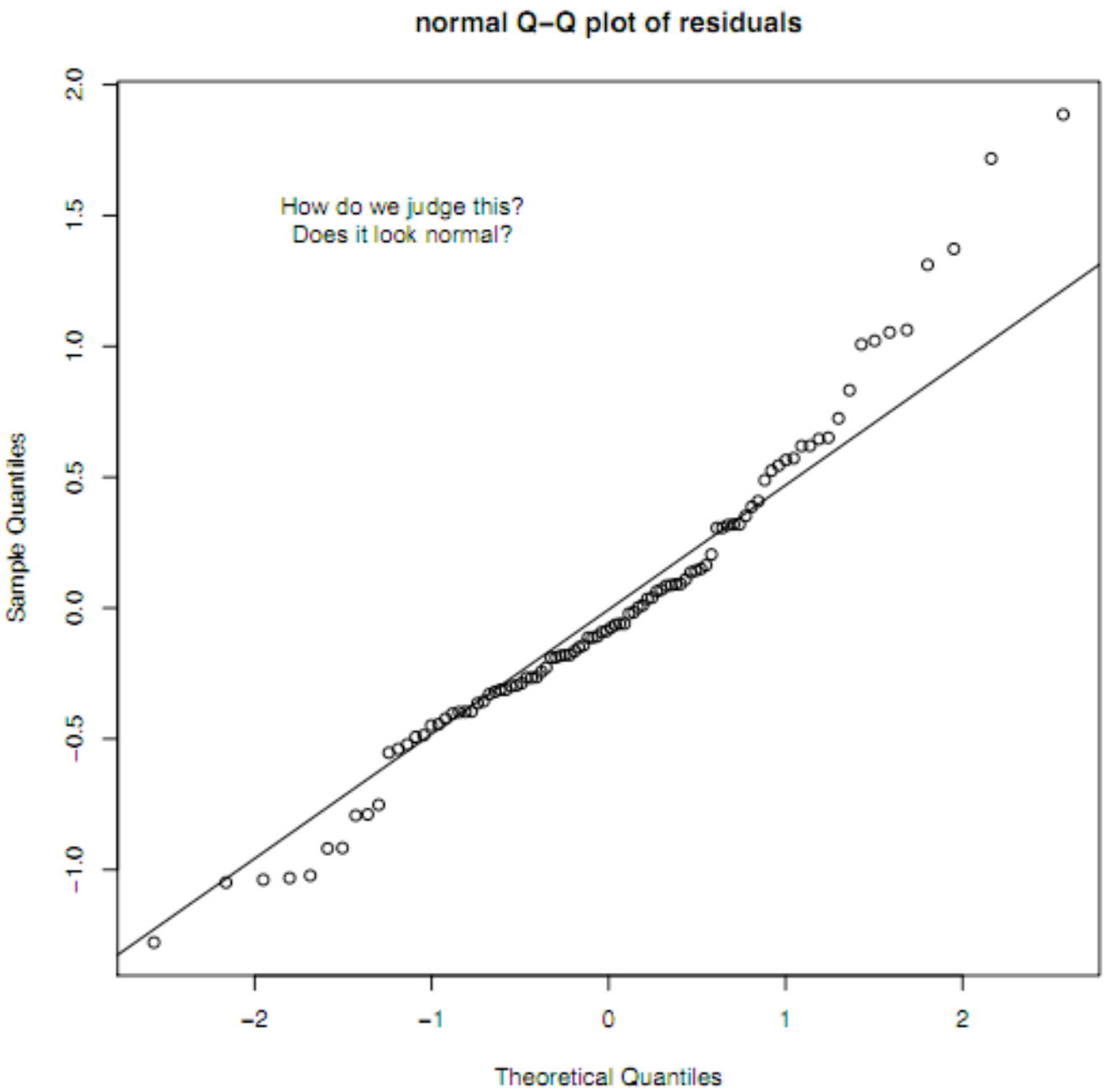
On the next two pages, we plot the residuals from the least squares fit to the fish data; the regression relating fish length and mercury content

Since the sum of squared residuals is 33.4 with n=98, the residual standard deviation is given by $\sqrt{33.4/96} = 0.59$

The mean Mercury level for fish in our sample is 1.28 and the sum of squares around this value is 66.7; therefore, the coefficient of determination is 1-33.4/66.7 = 0.50 -- the relationship is not perfect, but fish length seems useful in describing Mercury levels

Does this view of the residuals match your expectations?


histogram of residuals
Frequency
0
5
10
15
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
residuals


normal Q–Q plot of residuals
How do we judge this?
Does it look normal?
Sample Quantiles
Theoretical Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

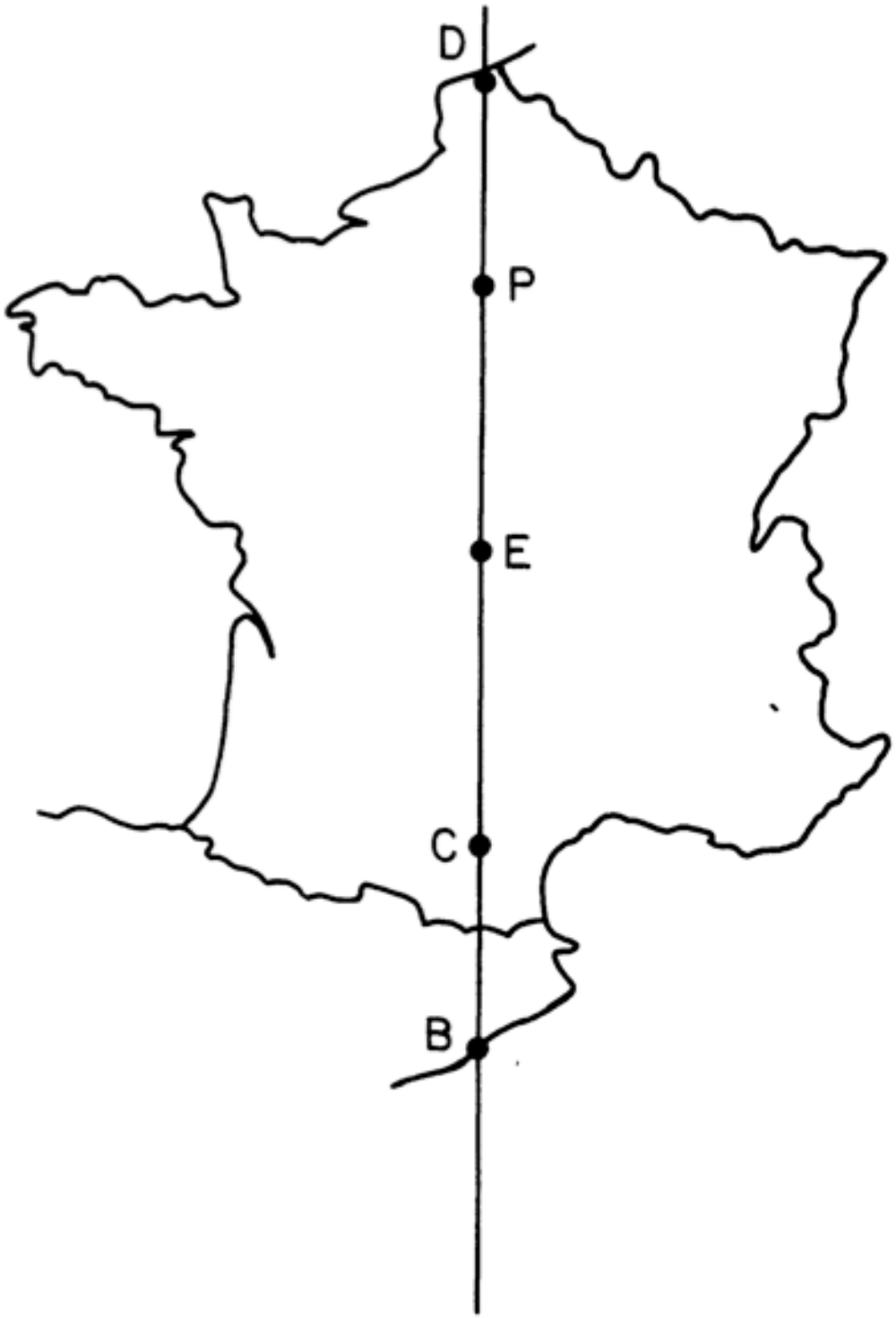
Historical detour

The contested origins of least squares

Stephen Stigler, a well-known statistician who writes extensively on the history of our field, begins a 1981 article on least squares with the sentence “The **most famous priority dispute** in the history of statistics is that between Gauss and Legendre, over the discover of the method of least squares.”

Legendre is undisputedly the first to publish on the subject, laying out the whole method in an article in 1805 -- **Gauss claimed to have used the method** since 1795 and that it was behind is computations of the “Meridian arc” published in 1799

In that paper, Gauss used a famous data set collected **to define the first meter** -- In 1793 the French had decided to base the new metric system upon a unit, the meter, equal to one 10,000,000th part of the meridian quadrant, the distance from the north pole to the equator along a parallel of latitude passing through Paris...


D
P
E
C
B

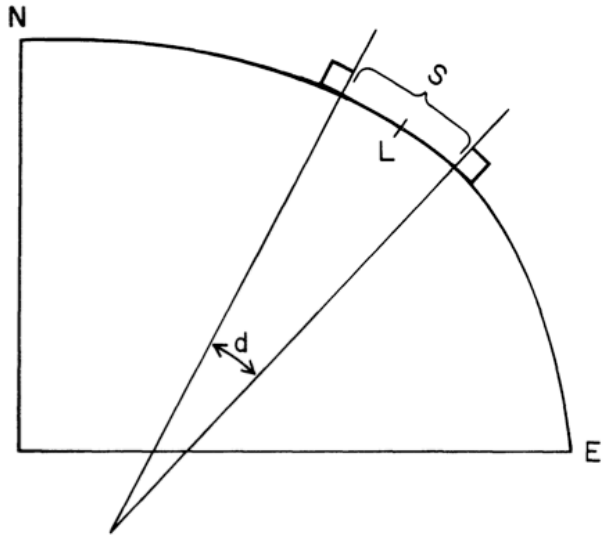

N
s
L
d
E

TABLE 1.

French arc measurements, from Allgemeine Geographische Ephemeriden, **4,** 1799, *page* xxxv. *The number* 76545.74 *is a misprint; the correct number is* 76145.74. *The table gives the length of four consecutive segments of the meridian arc through Paris, both in modules* S *(one module* ≅ 12.78 *feet) and degrees* d *of latitude (determined by astronomical observation). The latitude of the midpoint* L *of each arc segment is also given.*

	Modules S	Degrees d	Midpoint L
Dunkirk to Pantheon	62472.59	2.18910	49° 56′ 30″
Pantheon to Evaux	76545.74	2.66868	47° 30′ 46″
Evaux to Carcassone	84424.55	2.96336	44° 41′ 48″
Carcassone to Barcelona	52749.48	1.85266	42° 17′ 20″
Totals	275792.36	9.67380	

Least squares

The relationships between the variables in question (arc length, latitude, and meridian quadrant) are all nonlinear -- But for short arc lengths, a simple approximation holds

$$a = (S/d) = \alpha + \beta \sin^2 L$$

Having found values for α and β , one can estimate the meridian quadrant via

$$\text{meridian quadrant } = 90(\beta + \alpha/2)$$

Label the four data points in the previous table

$$(a_1, L_1), (a_2, L_2), (a_3, L_3) \text{ and } (a_4, L_4)$$

and apply **the method of least squares** -- That is, we identify values for α and β such that the sum of squared errors is a minimum

$$\sum_{i=1}^{4} (a_i - \alpha - \beta \sin^2 L_i)^2$$

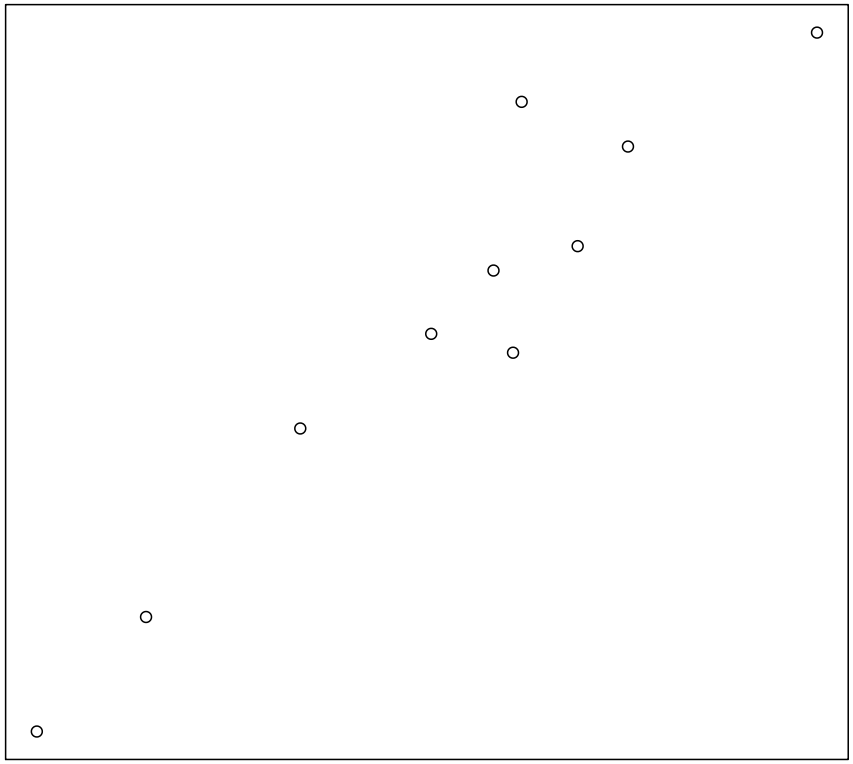
Least squares

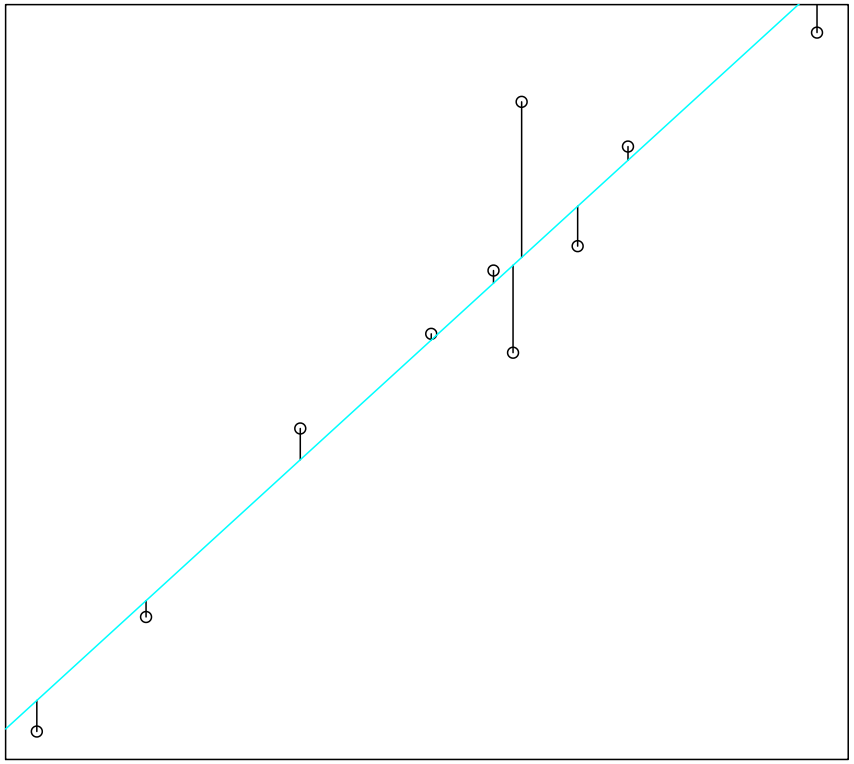
Given a set of predictor-response pairs $(x_1, y_1), \ldots, (x_n, y_n)$, we can write the ordinary least squares (OLS) criterion (as opposed to a weighted version that we'll get to) as

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Least squares

Graphically, in this simple case, we are doing nothing more than hypothesizing a linear relationship between the x and y variables and choosing that line that minimizes the (vertical) errors between model and data





Gauss and least squares

Stigler attempts to reproduce Gauss's calculations, but cannot given the simple linearization (and a couple not-so-simple linearizations) on the previous slide

Ultimately, he reckons that because Gauss was a mathematician and not a statistician, he might have derived a more elaborate expansion -- No matter what form was used, **Stigler seems convinced that something like least squares was required**

Gauss eventually publishes on least squares in 1809, and his account of the method is much more complete than Legendre's -- **Linking the method to probability and providing computational approaches**

Galton and regression

While least squares, as a method, was developed by several people at around the same time (often ideas are “in the air”), regression as we have come to understand it, was almost entirely the work of one man

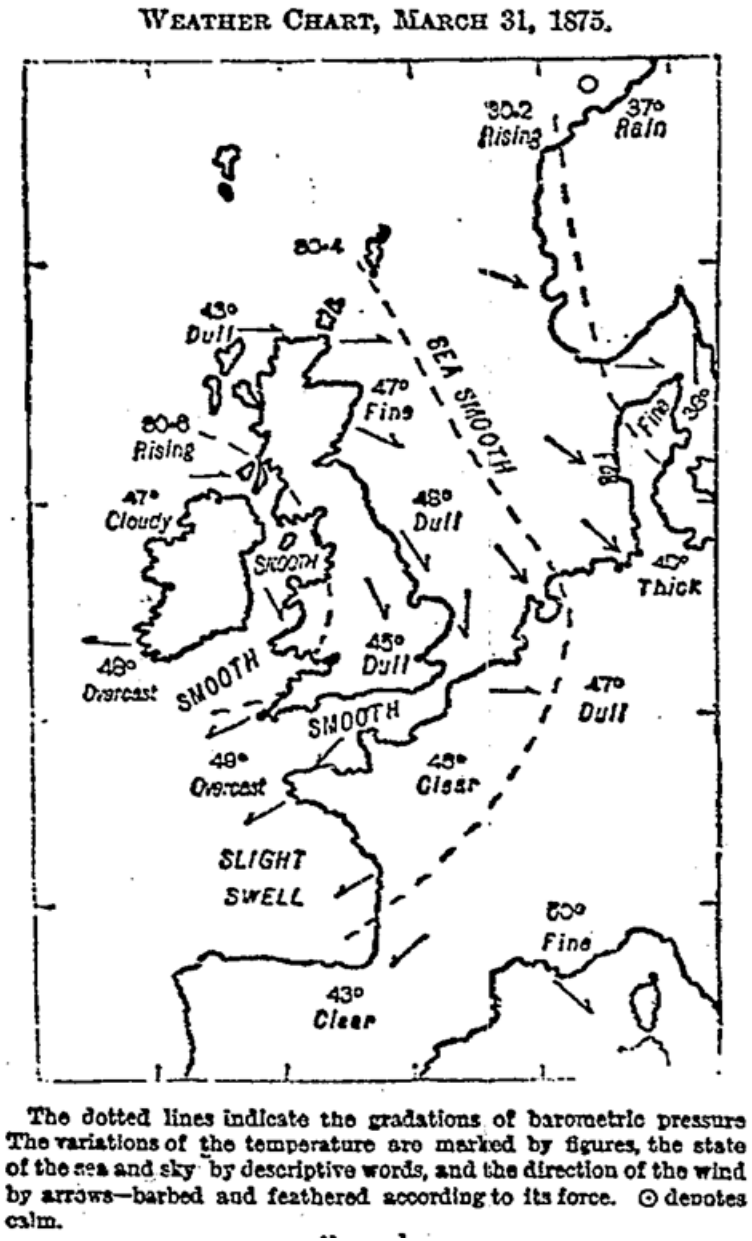
Stigler writes “Few conceptual advances in statistics can be as unequivocally associated with a single individual. Least squares, the central limit theorem, the chi-squared test -- all of these were realized as the culmination of many years of exploration by many people. Regression too came as the culmination of many years’ work, but in this case **it was the repeated efforts of one individual**.”

Galton and regression

Francis Galton (1822-1911) was at various points in his career an inventor, an anthropologist, a geographer, a meteorologist, a statistician and even **a tropical explorer** -- The latter gig paid quite well as his book “The art of travel” was a best seller

Among his many innovations, was **the first modern weather map**, appearing in The Times in 1875 -- To draw it, Galton requested data from meteorological stations across Europe

He also developed the use of **fingerprints as a means of identification** -- This work is just one small part of his larger interest how human characteristics (physical or even mental) varied across populations



The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

Galton and regression

Galton was also half-cousins with Charles Darwin (sharing the same grandfather) and took a strong interest in how physical and mental characteristics move from generation to generation -- **Heredity**

His work on regression started with a book entitled Hereditary Genius from 1869 in which he studied **the way "talent" ran in families** -- The book has lists of famous people and their famous relatives (great scientists and their families, for example)

He noted that there was a rather dramatic reduction in awesomeness as you moved up or down a family tree from the great man in the family (the Bachs or the Bernoullis, say) -- And thought of this as a kind of **regression toward mediocrity**

Galton and regression

In some sense, his work builds on that of Adolphe Quetelet -- Quetelet saw **normal distributions in various aggregate statistics** on human populations

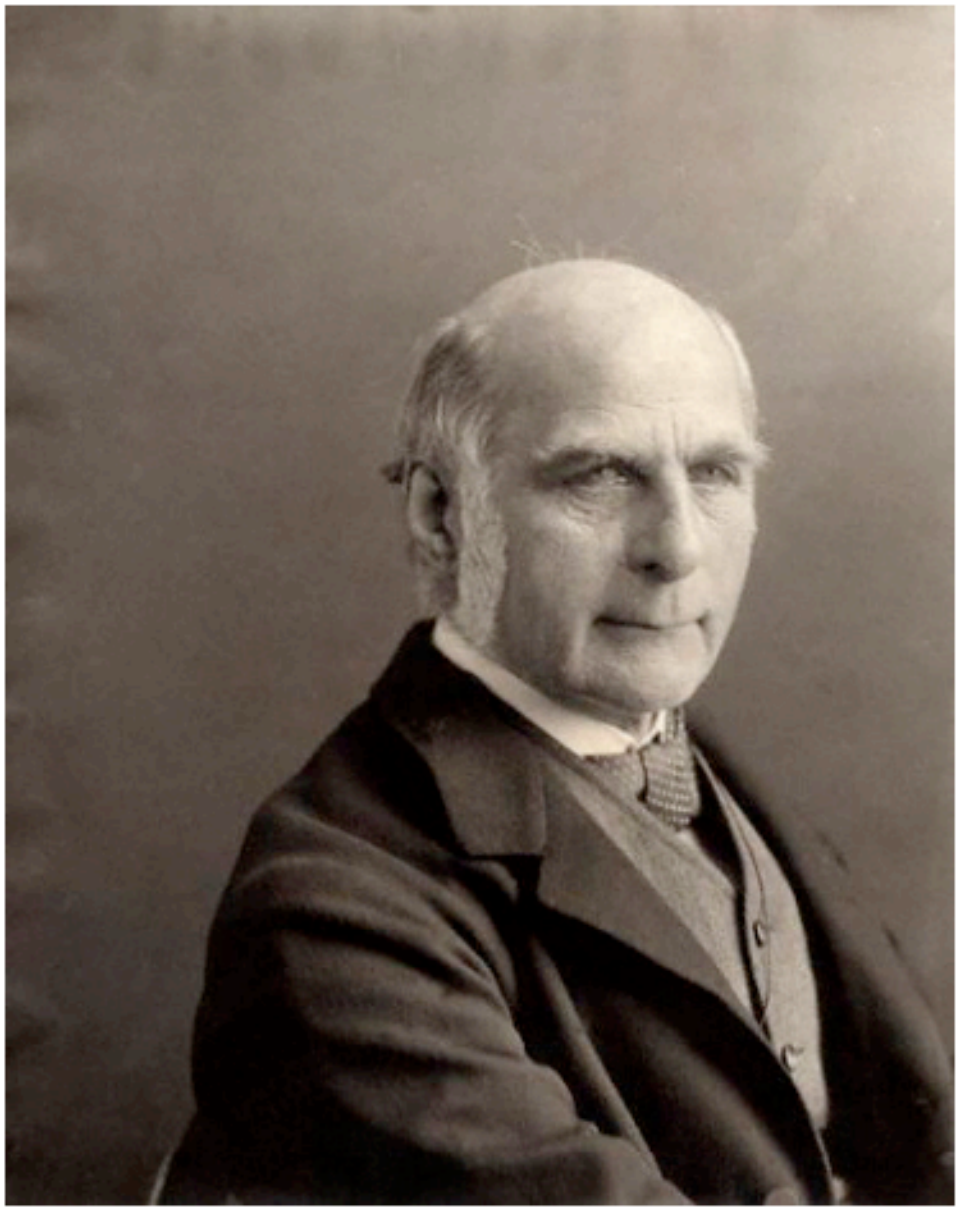
Galton writes “Order in Apparent Chaos -- I know of scarcely anything so apt to impress the imagination as the wonderful cosmic order expressed by the Law of Frequency of Error. The law would have been **personified by the Greeks and deified**, if they had known of it.”

Galton and regression

Relating the normal curve (and the associated central limit theorem) to heredity, however, proved difficult for Galton -- He could not **connect the curve to the transmission abilities** or physical characteristics from one generation to the next, writing

"If the normal curve arose in each generation as the aggregate of a large number of factors operating independently, no one of them overriding or even significant importance, what opportunity was there for a single factor such as parent to have a measurable impact?"

So at first glance, the normal curve that Galton was so fond of in Quetelet's work was at odds with the possibility of "inheritance" -- Galton's solution to the problem would be **the formulation of regression and its link to the bivariate normal distribution**



http://www.npg.org.uk

Some history

Galton collected data **928 children** (a large sample size compared to Gosset's n=4 experiments motivating the t-statistic), recording, among other things, **their heights and the heights of their parents**

He then "transmutes" the heights of girls and women in his data set, multiplying these heights by 1.08; finally, he forms a table of the heights of the mid-parents (the average height of the father and mother) by child heights

Here are some "views" of his data...

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on February 9th, 1877. It appeared from these experiments that the offspring did *not* tend to resemble their parent seeds in size, but to be always more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small. The point of convergence was considerably below the average size of the seeds contained in the large bagful I bought at a nursery garden, out of which I selected those that were sown, and I had some reason to believe that the size of the seed towards which the produce converged was similar to that of an average seed taken out of beds of self-planted specimens.

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings, conducted for me by friends living in various parts of the country, from Nairn in the north to Cornwall in the south, during one, two, or even three generations of the plants, that I could entertain no doubt of the truth of my conclusions. The exact ratio of regression remained a little doubtful, owing to variable influences; therefore I did not attempt to define it. But as it seems a pity that no

2 FATHER ...

1. Date of birth. August 7th 1838. Birthplace. Neath, Glamorganshire.

2. Occupation. Clerk in Holy Orders. Residences.

3. Age at marriage.	The place for this entry is at 4 in next page.	23.
4. do. of wife	The place for this entry is at 3 in next page	23.

5. Mode of life so far as affecting growth or health.

6. Was early life laborious? why and how? No.

7. Adult height. 5ft. 6in. Colour of hair when adult. Dark brown. Colour of eyes. Blue.

8. General appearance. Slender.

9. Bodily strength and energy, if much above or below the average. During 22 years from Ordination, have preached 3600 times. Yearly average unable to preach (from temporary indisposition) only 4 Sundays during these 22 years.

10. Keenness or imperfection of sight or other senses. Have always possessed good sight, both for near & distant objects. No failure as yet. (age 46).

11. Mental powers and energy, if much above or below the average. Rapid reader.

12. Character and temperament. Cool, cautious, methodical.

2 FATHER

1. Date of birth. August 7th 1838. Birthplace. North Glamorganshire

2. Occupation. Clerk in Holy Orders. Residence.

3. Age at marriage. — The place for this entry is at 4 in next page. — 23.

4. do. of wife — The place for this entry is at 3 in next page — 23.

5. Mode of life so far as affecting growth or health.

6. Was early life laborious? why and how? No.

7. Adult height. 5 ft. 6 in. Colour of hair when adult. Dark brown. Colour of eyes. Blue.

8. General appearance. Slender.

9. Bodily strength and energy, if much above or below the average. During 22 years from ordination have preached three times [illegible] (from temporary indisposition) only [illegible] Sundays during these 22 years.

10. Keenness or imperfection of sight or other senses. Have always possessed good sight, both for near & distant objects. No failure as yet (age 46).

11. Mental powers and energy, if much above or below the average. Rapid reader.

12. Character and temperament. Cool, cautious, methodical.

13. Favorite pursuits and interests. Artistic aptitudes. From earliest childhood a decided inclination to the [illegible]. Fond of music. [illegible] but have no [illegible] musical composition (not very difficult) at first sight, as if I had [illegible]

14. Minor ailments to which there was special liability
- in youth. Colds in the head. Sore headaches.
- in middle age. Very rarely suffer now from these attacks.

15. Graver illnesses
- in youth. None, excepting measles & whooping cough, but mild.
- in middle age. None, excepting lumbago occasionally.

16. Cause and date of death, and age at death. Still living.

17. General remarks.

26

MOTHER... 3

1. Date of birth. Oct. 18, 1838. Birthplace. [illegible], Huddersfield.

2. Occupation. Residence. [illegible]

3. Age at marriage. 23 — Total No. of sons / — No. of sons deceased. None. ages

4. do. of husband 23 — daus. / — No. of daus. deceased. None. ages

5. Mode of life so far as affecting growth or health. Educated at private schools.

6. Was early life laborious? why and how? No.

7. Adult height. 5 ft. 4 in. Colour of hair when adult. Dark brown. Colour of eyes. Brown.

8. General appearance. Slender frame. Dark complexion.

9. Bodily strength and energy, if much above or below the average.

10. Keenness or imperfection of sight or other senses. Sight & other senses good.

11. Mental powers and energy, if much above or below the average.

12. Character and temperament. Nervous & active.

13. Favorite pursuits and interests. Artistic aptitudes.

14. Minor ailments to which there was special liability
- in youth. Bilious attacks, with vomiting & headache.
- in middle age. Indigestion.

15. Graver illnesses
- in youth.
- in middle age. [illegible]

16. Cause and date of death, and age at death. Still living.

17. General remarks.

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.													Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

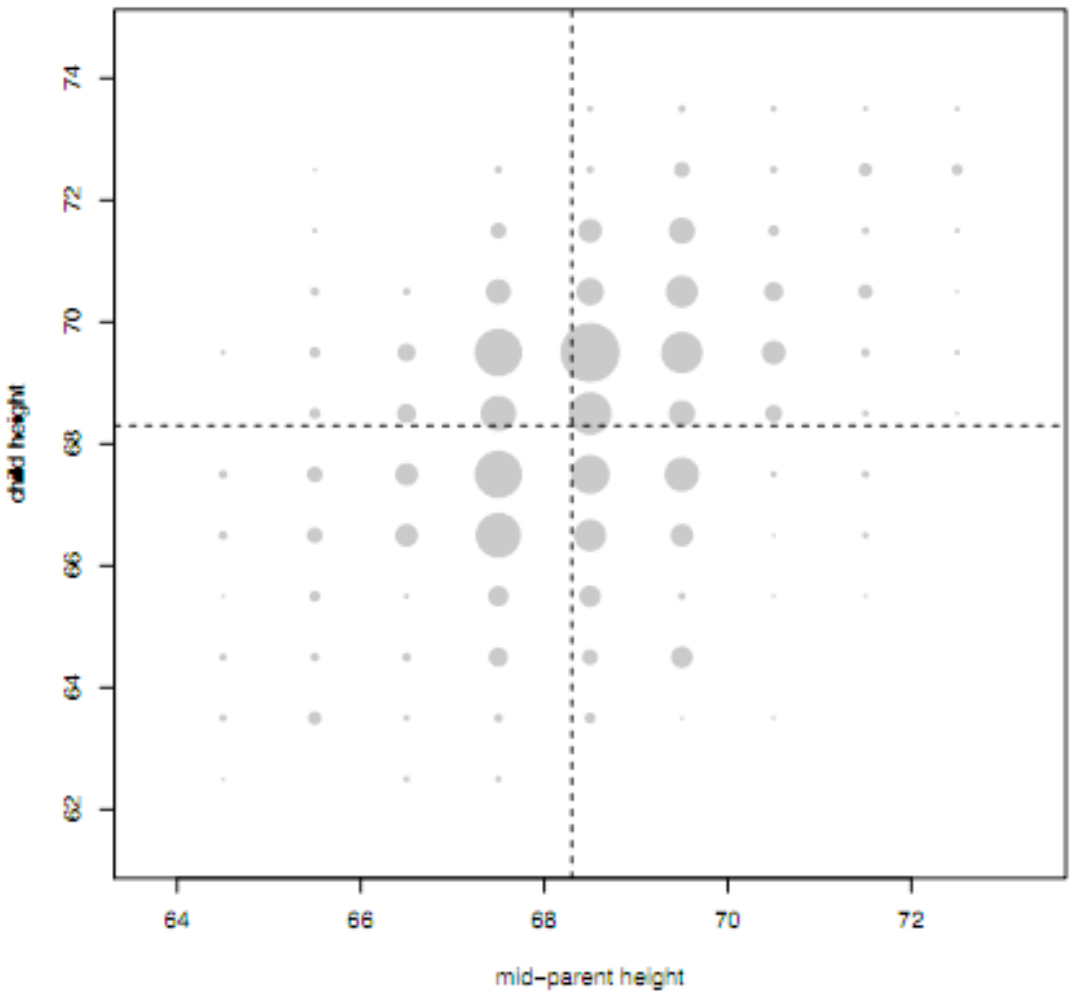
NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Some history

Here is another look at Galton's table; in this case the different cells in the table are represented by circles, sized according to the counts

The dashed lines mark the **mean of the parents' and children's heights** (both about 68.3 inches)

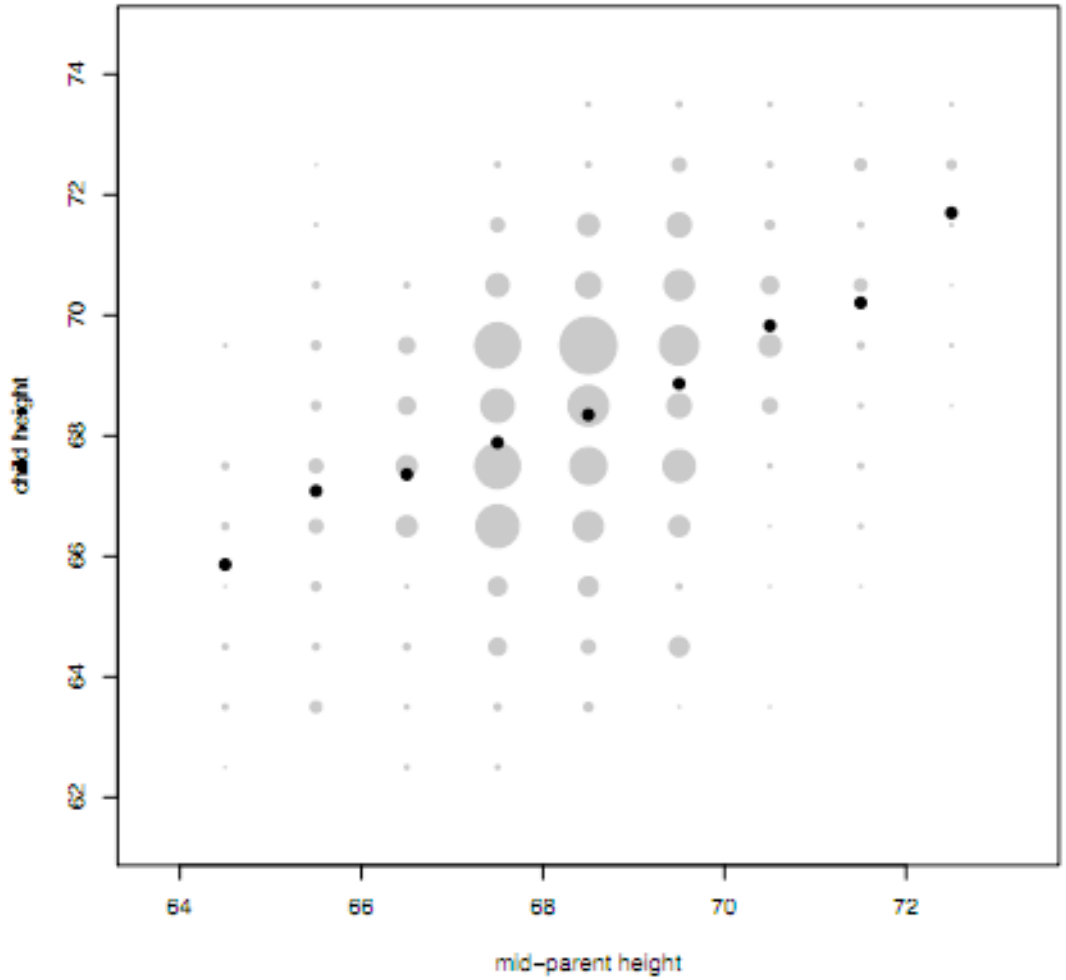
What do we notice about this pattern?



Some history

Now, consider parents who are between 69 and 70 inches tall; the average height of their children is 68.9 and is marked with a dark circle on the right

We repeated this process for the full range of mid-parents' heights -- What do you notice?

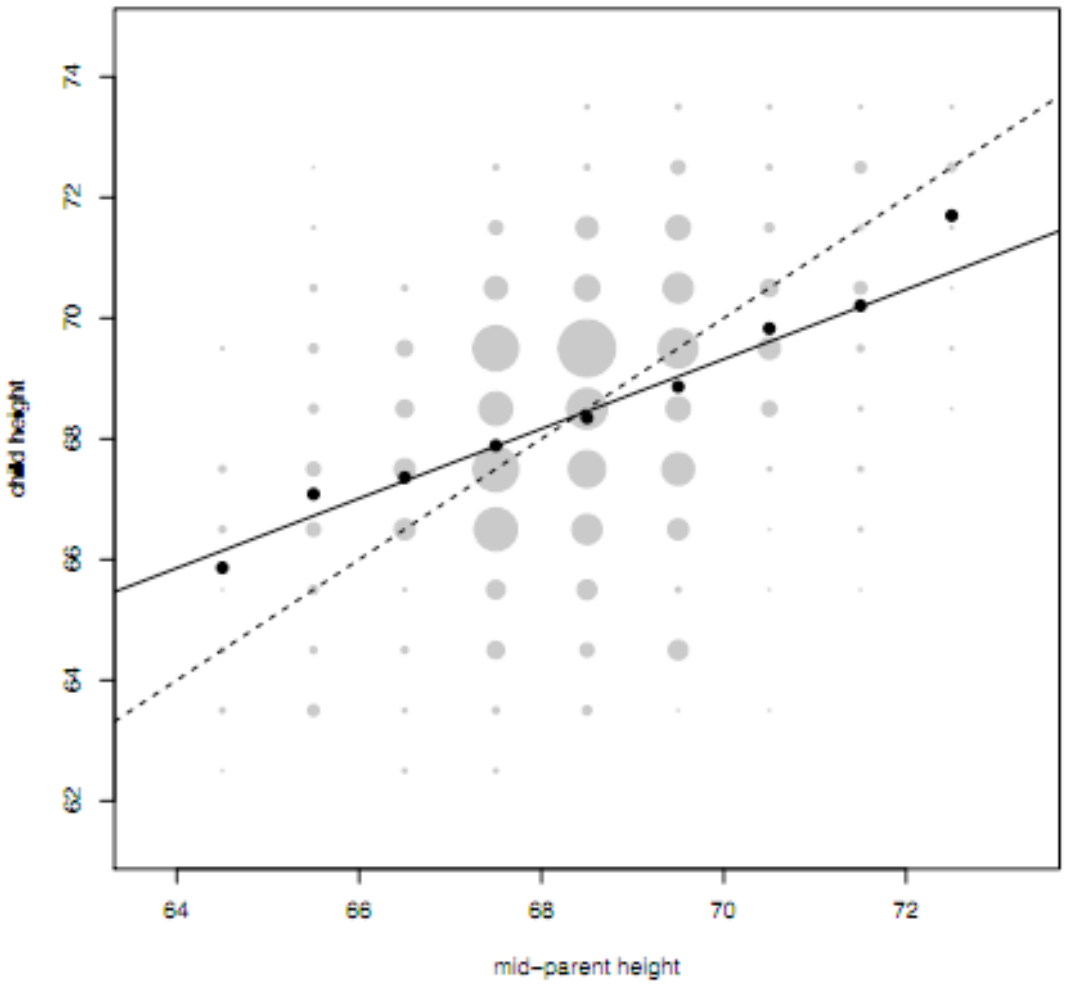


Some history

The solid line represents the least squares fit to the data and the dashed line is just y=x

In general, if we define a subgroup of children based on their parents' mid-heights, their mean height is closer to the mean of all children's heights than the mean height of the subgroup of parents is to the mean height of all parents

Galton referred to this as **regression toward mediocrity** or **regression to the mean**

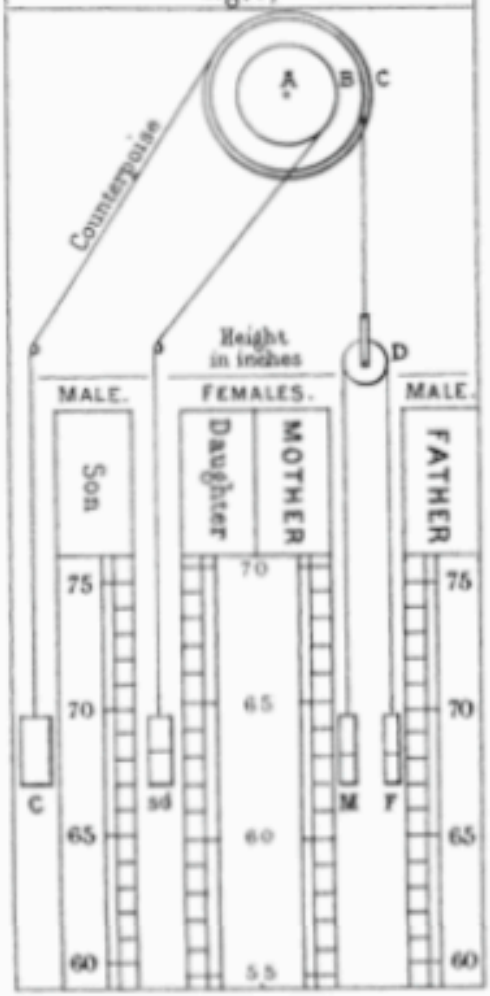


RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)

FORECASTER OF STATURE

Fig. (b)



J.T. & W.B. Lowdon, lith.

Galton and regression

In his text Natural Inheritance, he approached a table like this by first examining the **heights of the mid-parents** and noted that it appeared to be normal -- He then looked at the **marginal distribution of child heights** and found them to also be normally distributed

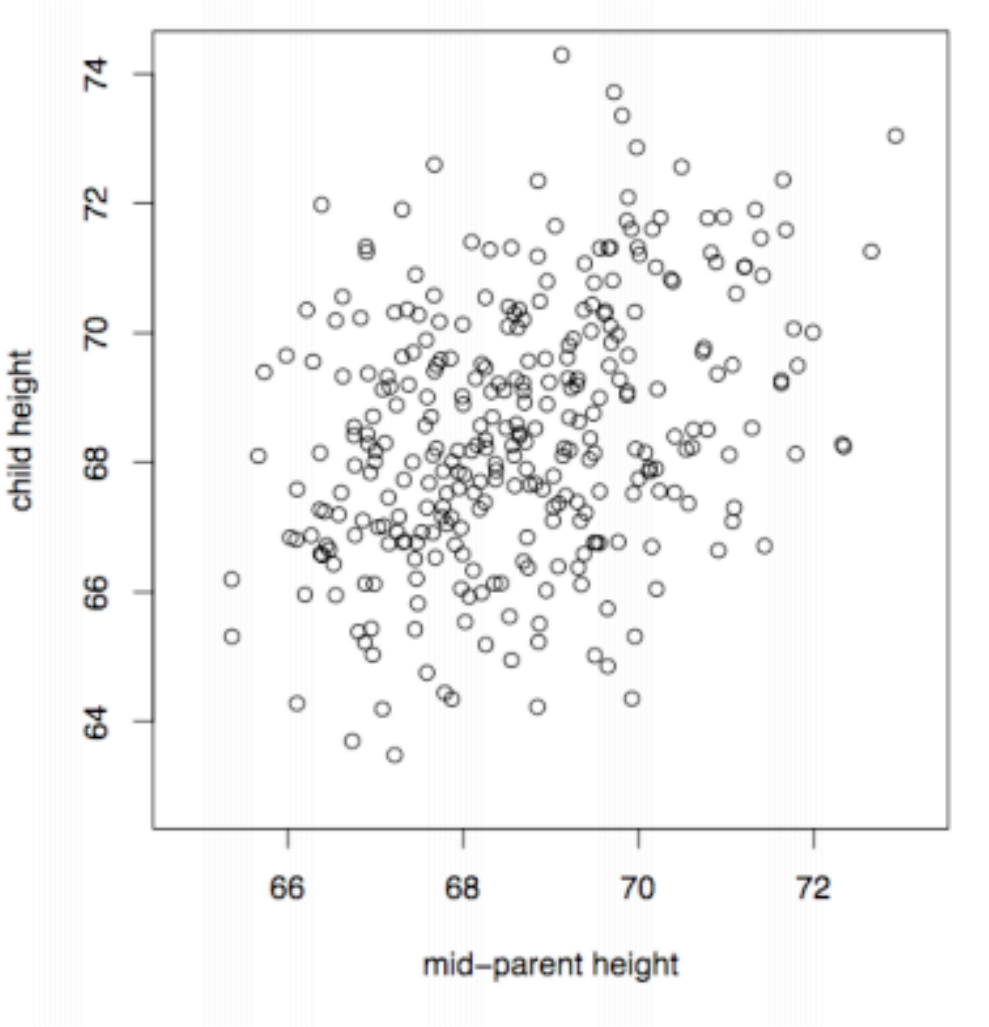
He then considered the heights of the children associated with different columns in his table, plotting median values against mid-parental height and finding a straight line (which he fit by eye)

He found that the slope was about 2/3 -- If children were on average as tall as their parents, he'd expect a slope of 1, leading him to coin the phrase "regression toward mediocrity"

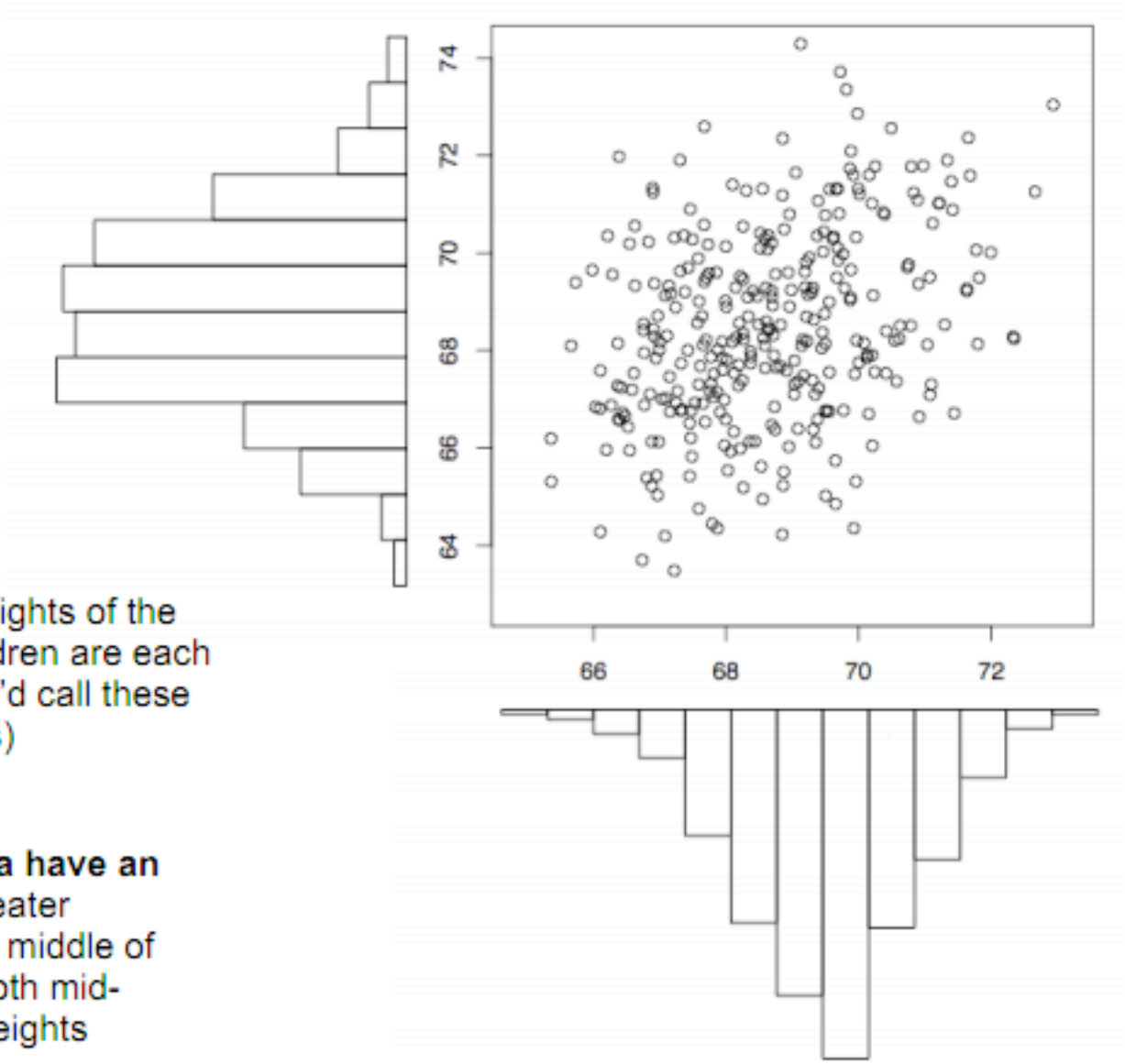
The bivariate normal

Gosset's data and Galton's table have **a common "elliptical" shape**; there is a central portion with greater density and then things spread out as you go toward the edges

At the right we have a sample of a bivariate normal distribution, selected to "match" the data from Galton's table



The bivariate normal



The distribution of the heights of the mid-parents and the children are each **individually normal** (we'd call these the marginal distributions)

Viewed as pairs, **the data have an elliptical shape**, with greater concentration toward the middle of the cloud, the mean of both mid-parents' and children's heights

Galton and regression

What Galton found through essentially geometric means was the following relationship (which we'll see later)

$$\frac{y - \bar{y}}{sd(y)} = r \frac{x - \bar{x}}{sd(x)}$$

where we might take x to be the heights of mid-parents and y to be the heights of their adult offspring -- The quantity r is the correlation coefficient between x and y (another Galton innovation)

This gives a precise meaning to his phrase "regression to the mean"

Galton and regression

In 1873, Galton had a machine built which he christened **the Quincunx** -- The name comes from the similarity of the pin pattern to the arrangement of fruit trees in English agriculture (quincunxial because it was based on a square of four trees with a fifth in the center)

The machine was originally devised to **illustrate the central limit theorem** and how a number of independent events might add up to produce a normal distribution -- Lead shot were dropped at the top of the machine and piled up according to the binomial coefficients at the bottom

The other panels in the previous slide illustrate a thought experiment by Galton (it's not clear the other devices were ever made) -- The middle region (between the A's) in the central machine, could be closed, **preventing the shot from working their way down the machine**

NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

"HEREDITARY GENIUS," "INQUIRIES INTO HUMAN FACULTY," ETC.

FIG. 7.

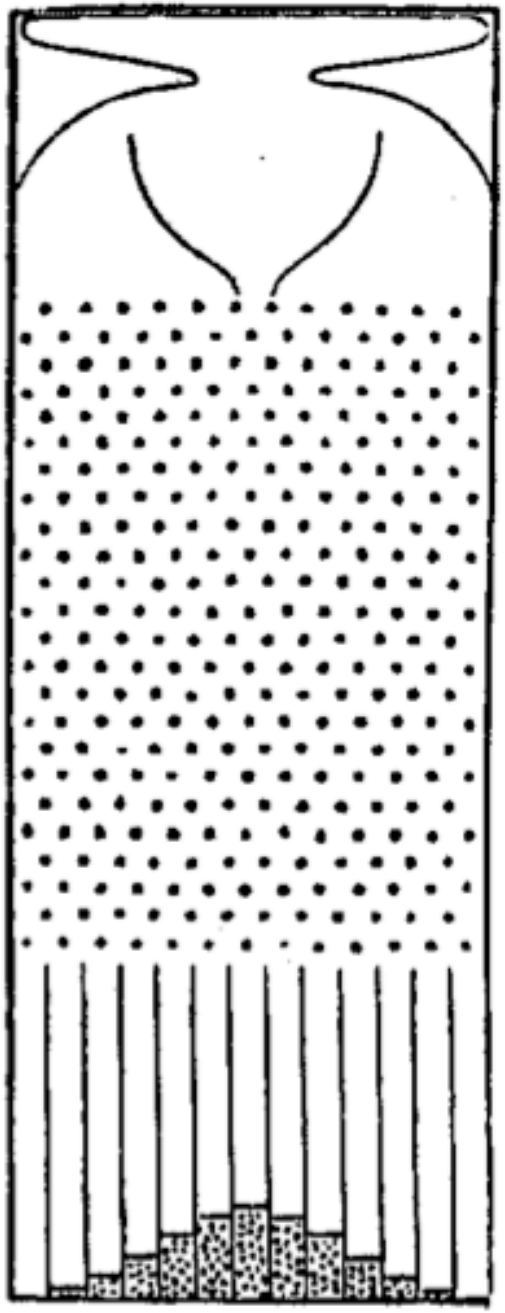


FIG. 8.

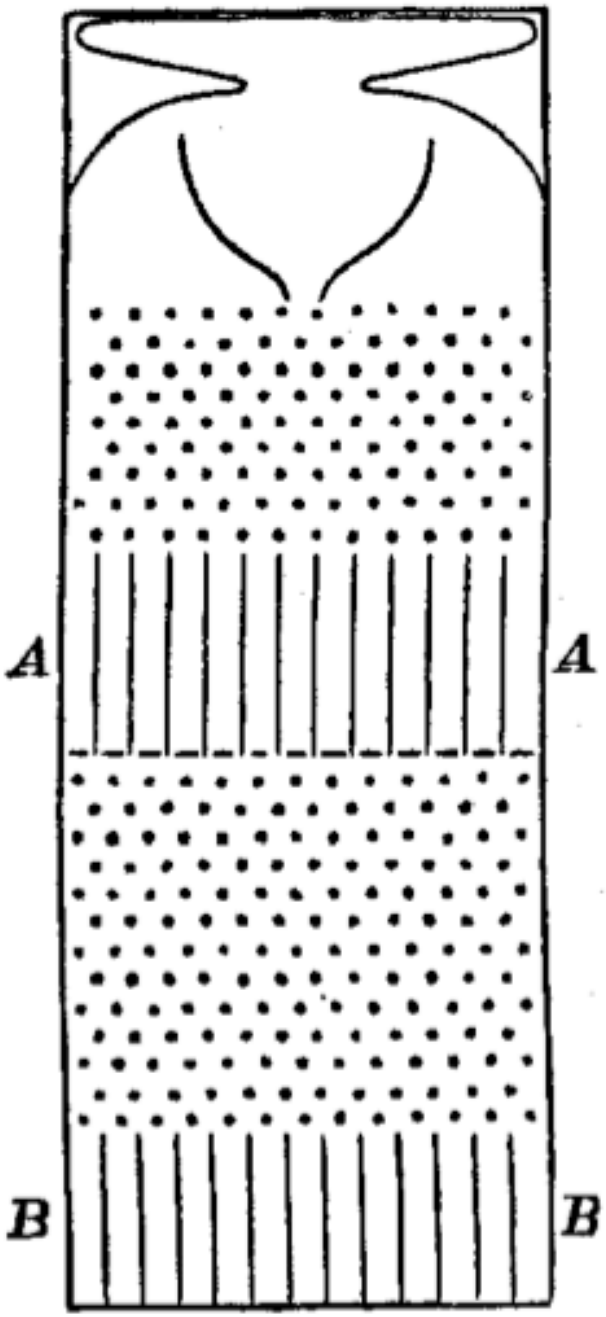
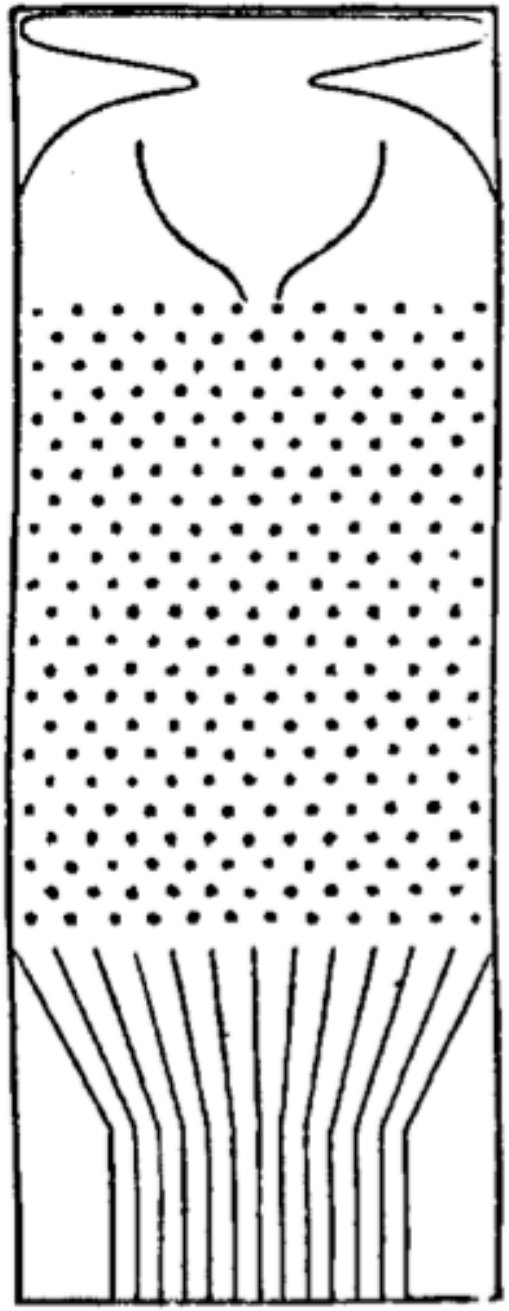


FIG. 9.



Galton and regression

By imagining holding back a portion of the shots, Galton expected to still see a normal distribution at the bottom of the machine, but one with less variation -- As he opened each barrier, t**he shot would deposit themselves according to small normal curves**, adding to the pattern already established

Once all the barriers had been opened, you'd be left with the original normal distribution at the bottom -- Galton, in effect, showed how the normal curve **could be dissected into components** which could be traced back to the location of the shot at A-A level of the device

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Table 13 (Special Data).

Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	...	...	...	...	...	1	1	...	5	3	12	24	
73·5	...	...	...	...	...	1	3	4	8	3	3	2	3	27	
72·5	...	...	...	...	1	1	6	5	9	9	8	3	5	47	71·1
71·5	...	1	...	1	2	8	11	18	14	20	9	4	...	88	70·2
70·5	...	...	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5	...	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5	...	1	5	9	18	38	46	36	30	11	6	3	...	203	68·7
67·5	2	4	8	26	35	38	38	20	18	8	1	1	...	199	67·7
66·5	4	3	10	33	28	35	20	12	7	2	1	...	...	155	67·0
65·5	3	3	15	18	33	36	8	2	1	1	...	...	...	110	66·5
64·5	3	8	12	15	10	8	5	2	1	...	...	...	...	64	65·6
63·5	5	2	8	3	3	4	1	1	...	1	...	...	1	20	
Below 63.....	5	5	3	3	4	2	...	...	...	...	...	...	1	23	
Totals.....	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	

Galton and regression

Looking at these tables, we see the Quincunx at work -- The righthand column labeled “Total number of Adult Children” being the **counts of shot at the A-A level**, while the row marked “Totals” can be thought of as **the distribution one would see at the bottom of the device** when all the barriers are opened and the **individual counts in each row as the corresponding normal curves**

By 1877, Galton was starting to examine these ideas mathematically -- He essentially **discovered the important properties of the bivariate normal distribution** (the bivariate normal had been derived by theorists unknown to Galton, but they did not develop the idea of regression, nor did they attempt to fit it from data as Galton did)

End: Historical detour

Generalization

While the least squares line and the associated concepts of residuals and residual standard deviation **are interesting summaries or descriptors** of the relationship between length and Mercury for fish in our sample, the EPA or state regulatory agencies will want to know what an be concluded about **the population of fish in the Waccamaw river** -- What can we say?

For guidance, we can again, look to the sample mean -- Just as our view of the sample mean shifted from **a descriptive statistic to an estimate of a population mean**, we can interpret our least squares fit as more than just a description, but as **an estimate of population-level quantities**

A population model

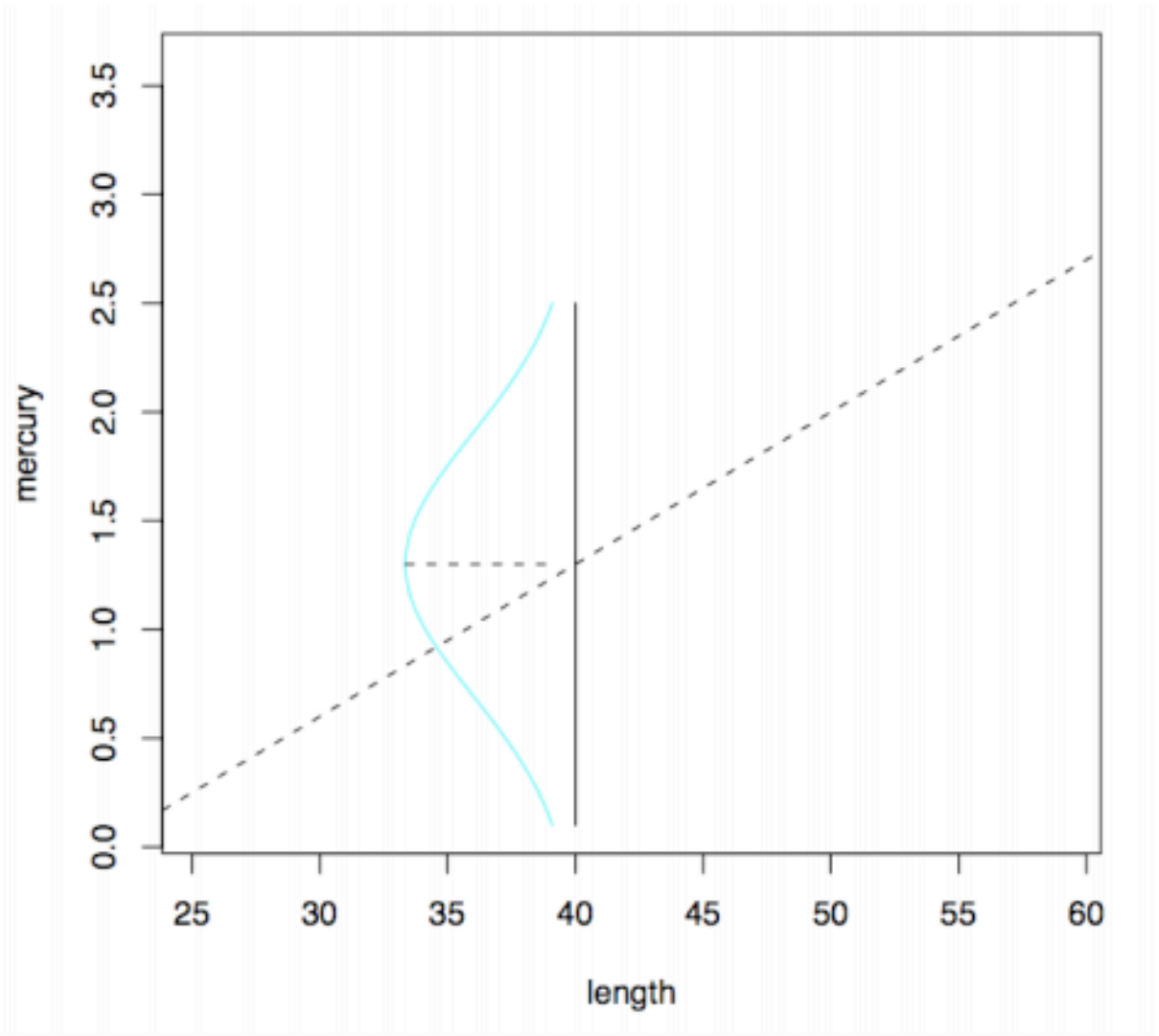
In its simplest form, the model we wrote for descriptive purposes

$$(\text{mercury}) = \beta_0 + \beta_1(\text{length}) + (\text{error})$$

is assumed **to hold for all the largemouth bass in the Waccamaw river** (the coefficients β_0 and β_1 being relabeled to indicate they are now population parameters)

For the moment, we will assume that **the errors follow a normal distribution** with mean zero and some unknown standard deviation σ, another parameter to be estimated

Another way to view the model specified above is that for some fixed value of length, x, the distribution of Mercury levels in fish of that length **in the population** has a normal distribution with mean $\beta_0 + \beta_1 x$ and standard deviation σ


mercury
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
25
30
35
40
45
50
55
60
length

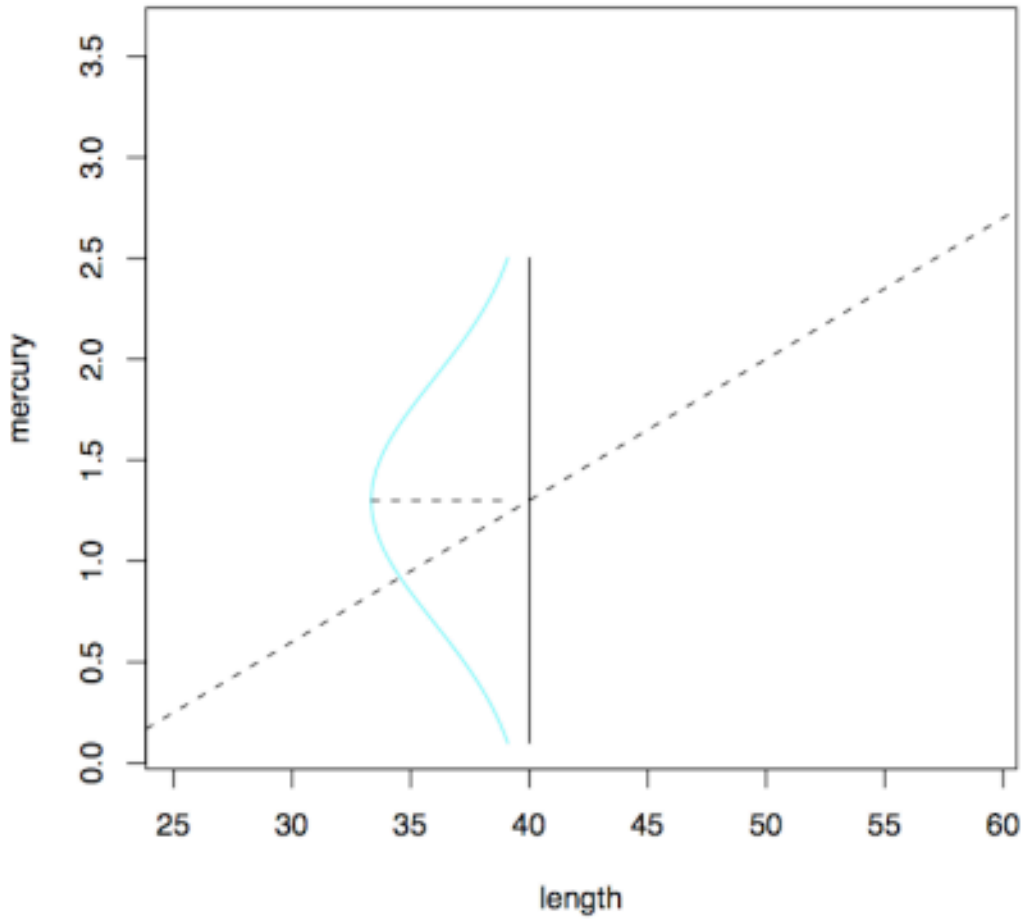
A population model

For each value of the variable length (here, x=40cm), we imagine the distribution of Mercury content for fish of that length **in the population** as a little normal curve

The parameters of this model are **the slope and intercept of the line** as well as the **unknown standard deviation of the error**

β_0, β_1 and σ

Note that σ is the same everywhere!



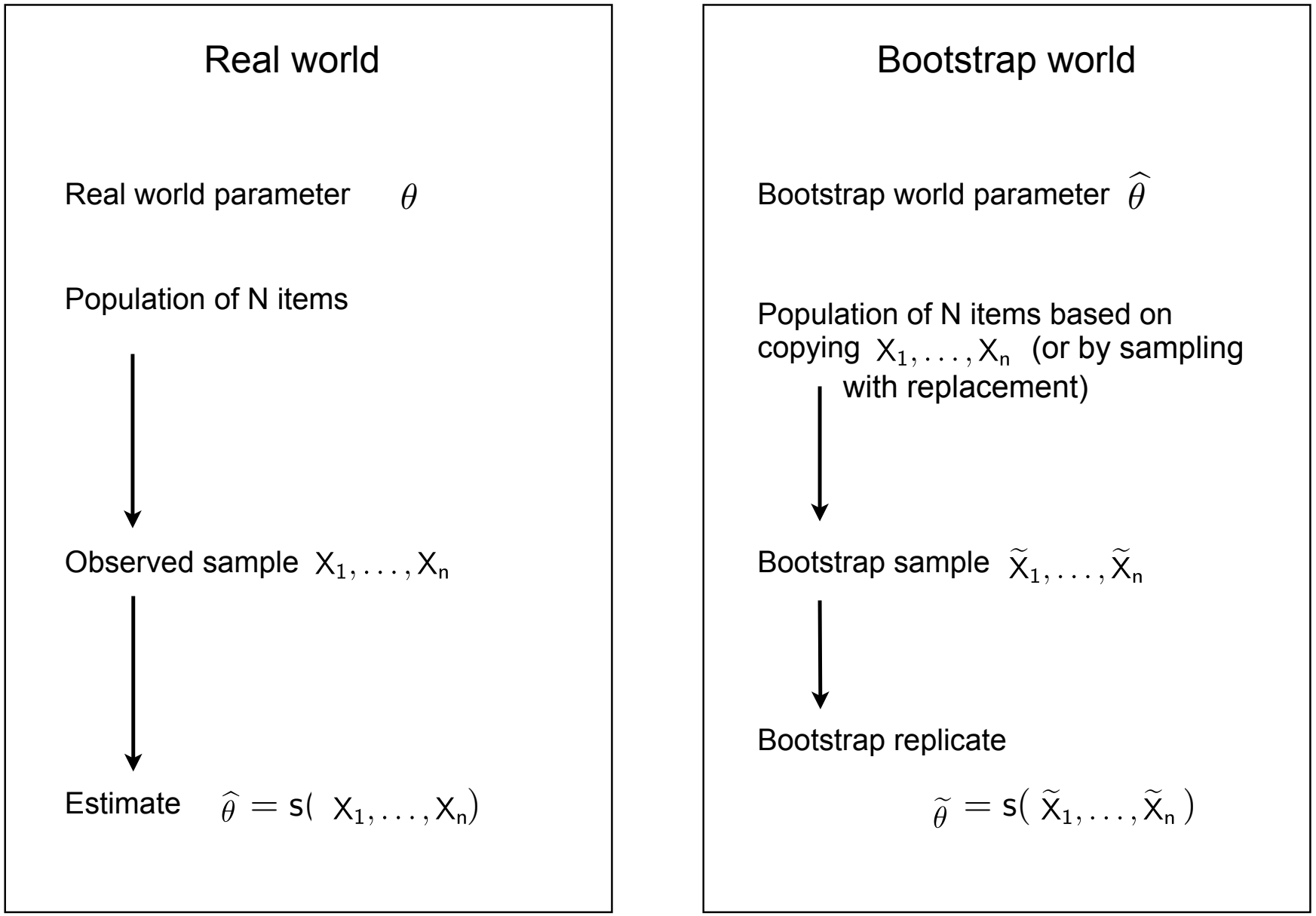
Inference

By analogy with the sample mean, we know that **producing an estimate is not the end of the story**; we needed to describe **the variability in our estimate**, an assessment which ultimately led to constructions like **confidence intervals**

In the case of linear regression, we are also subject to sampling variability -- We have a sample of fish taken from the Waccamaw river and certainly **if we repeated the trial** and caught another sample of fish, **our least squares estimates would be different**

What can we say about the variability in our estimate? How would we estimate its precision?*

* Hint: The answer rhymes with jute-frappe


Real world
Real world parameter θ
Population of N items
Observed sample $X_1, \ldots, X_n$
Estimate $\widehat{\theta} = s(X_1, \ldots, X_n)$
Bootstrap world
Bootstrap world parameter $\widehat{\theta}$
Population of N items based on copying $X_1, \ldots, X_n$ (or by sampling with replacement)
Bootstrap sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$
Bootstrap replicate
$\widetilde{\theta} = s(\widetilde{X}_1, \ldots, \widetilde{X}_n)$

The bootstrap: Regression (I)

Following our motto "analyze as you randomized", we can simulate the process of drawing random samples via the bootstrap

1. Create a "population" consisting of our 98 pairs $(x_1, y_1), \ldots, (x_{98}, y_{98})$

2. We then **draw 98 pairs with replacement** to form a bootstrap sample $(x_1^*, y_1^*), \ldots, (x_{98}^*, y_{98}^*)$

3. Next, we compute a least squares fit to bootstrap sample, producing a bootstrap replicate for the intercept and slope, $\widehat{\beta}_0^*$ and $\widehat{\beta}_1^*$

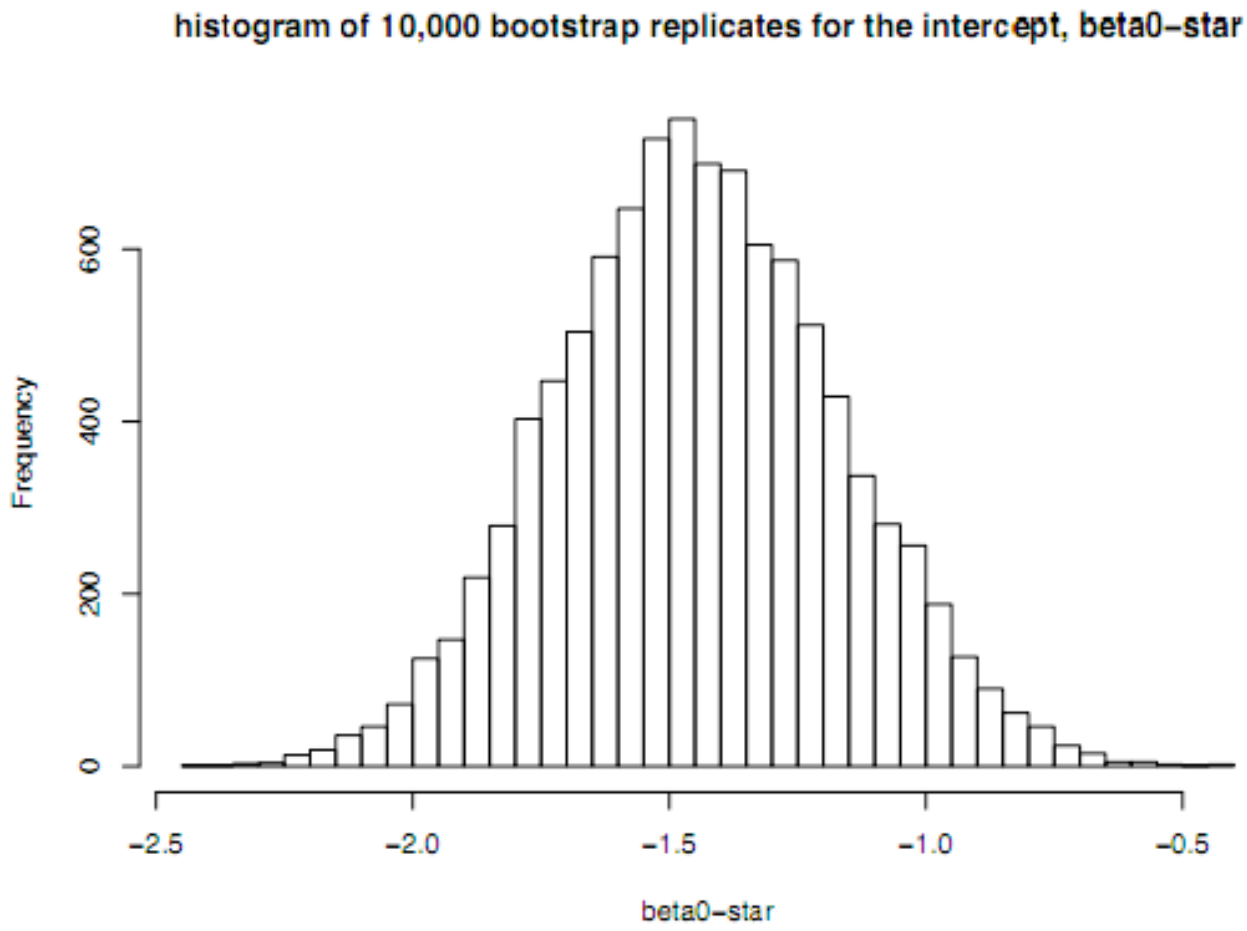
4. Repeat steps 1-3 a large number of times, say 10,000, to obtain a set of bootstrap replicates

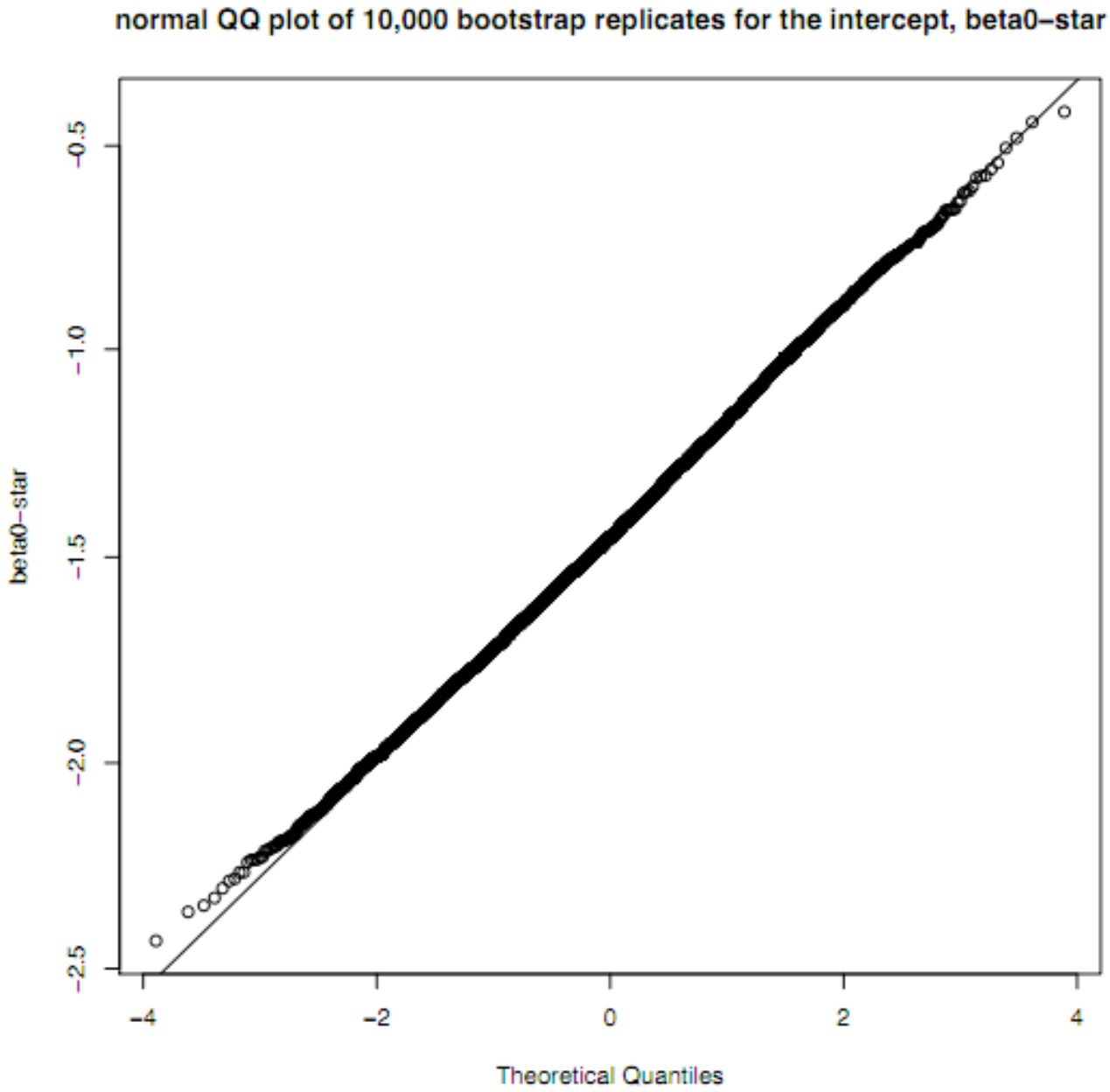
The bootstrap

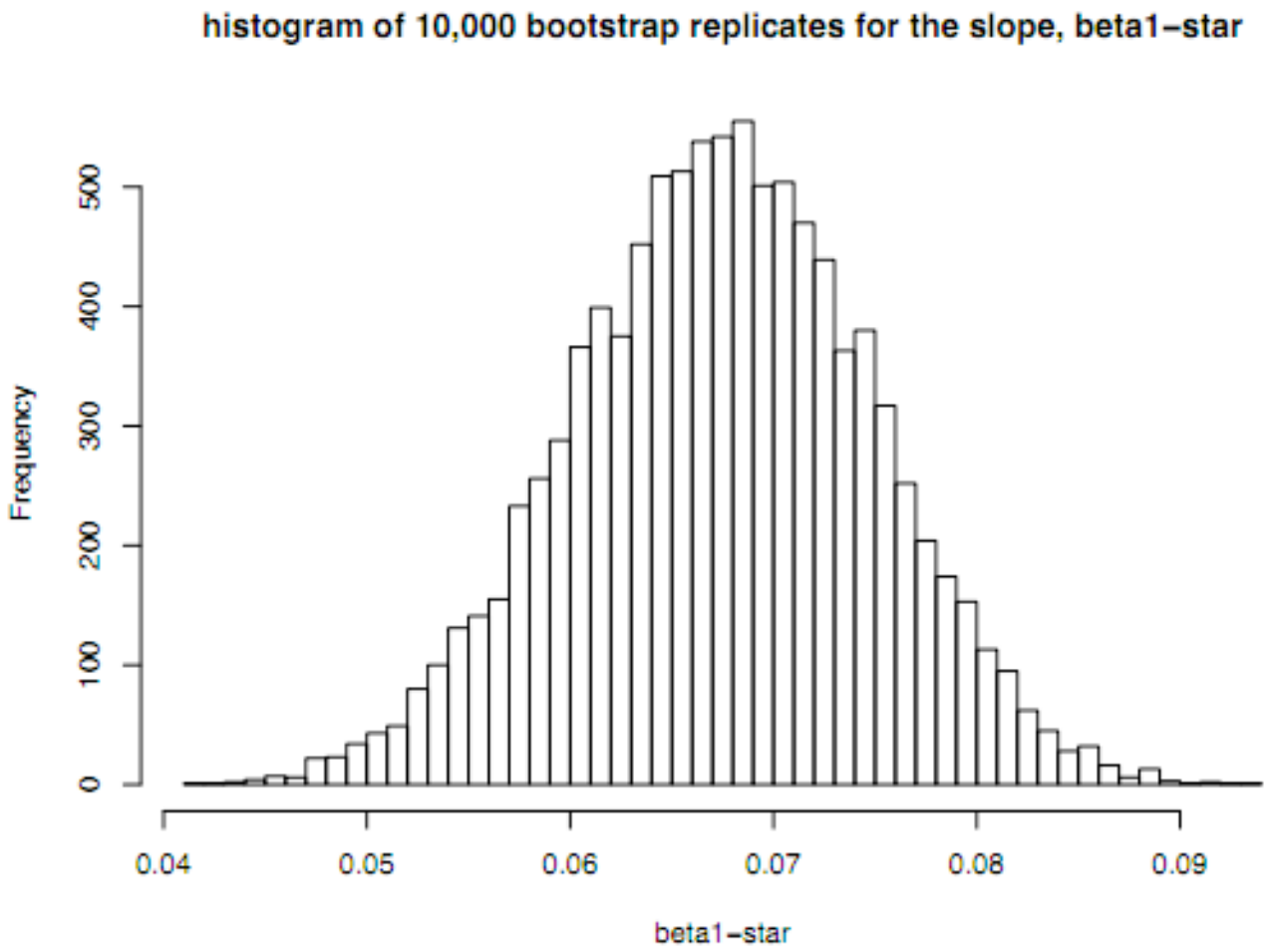
The bootstrap distribution for $\widehat{\beta}_0^*, \widehat{\beta}_1^*$ again is an estimate of their sampling distribution and we can use it to estimate the standard error of each, together with confidence intervals

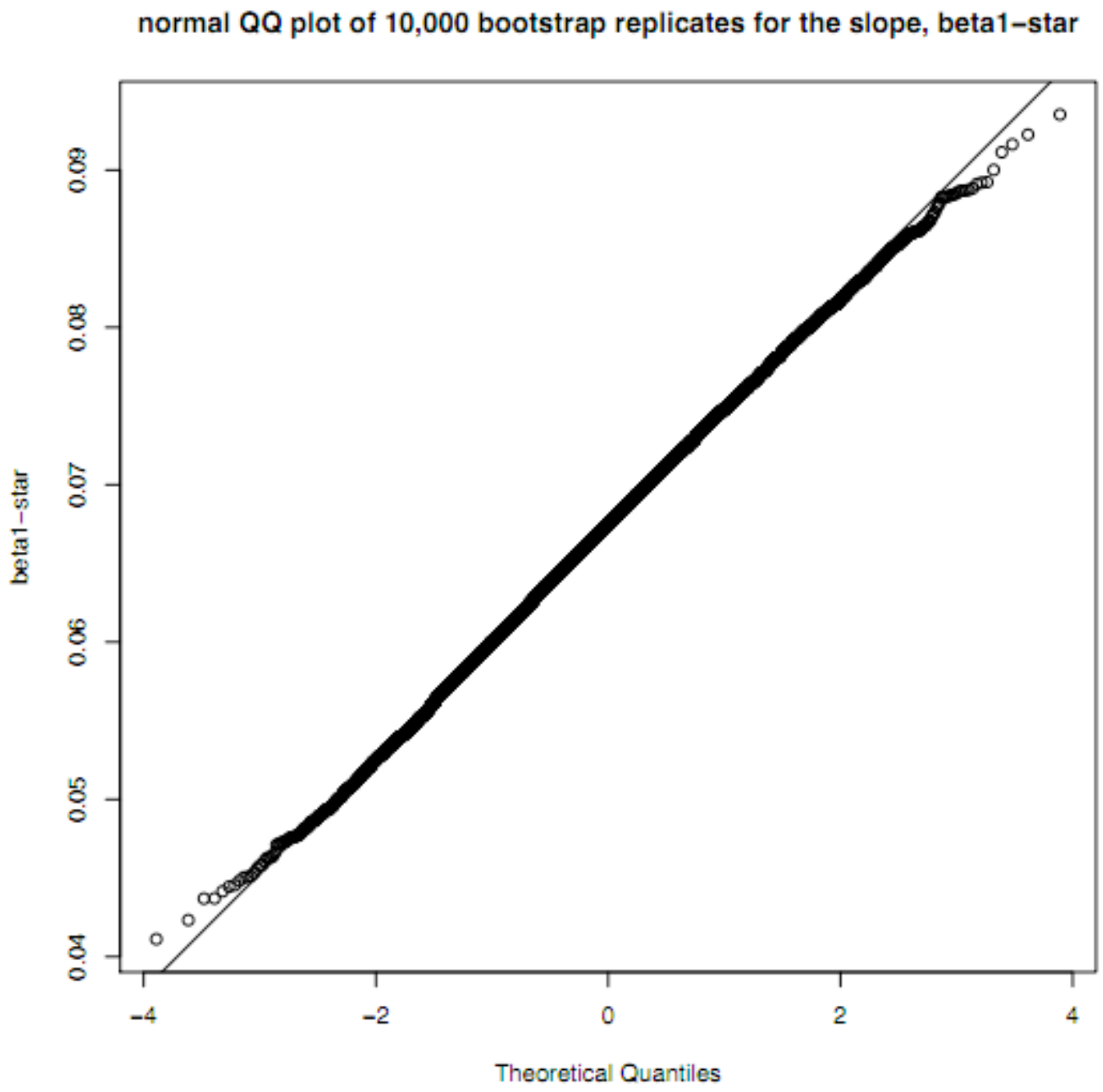
First, $\widehat{\beta}_0 = -1.45$ and the standard deviation of the bootstrap replicates is 0.28; a confidence interval using the 0.025 and 0.975 quantiles agrees with $-1.45 \pm 1.96 * 0.28 = [-2.00, -0.90]$

Next, $\widehat{\beta}_1 = 0.068$ and the standard deviation of the bootstrap replicates is 0.007; a confidence interval using the 0.025 and 0.975 quantiles agrees with $0.068 \pm 1.96 * 0.007 = [0.054, 0.082]$


histogram of 10,000 bootstrap replicates for the intercept, beta0−star
Frequency
0
200
400
600
−2.5
−2.0
−1.5
−1.0
−0.5
beta0−star


normal QQ plot of 10,000 bootstrap replicates for the intercept, beta0−star
beta0−star
−0.5
−1.0
−1.5
−2.0
−2.5
−4
−2
0
2
4
Theoretical Quantiles


histogram of 10,000 bootstrap replicates for the slope, beta1−star
Frequency
0
100
200
300
400
500
0.04
0.05
0.06
0.07
0.08
0.09
beta1−star


normal QQ plot of 10,000 bootstrap replicates for the slope, beta1−star
beta1−star
0.04
0.05
0.06
0.07
0.08
0.09
−4
−2
0
2
4
Theoretical Quantiles

Inference

Recall that when studying the relative risk, a major concern was whether or not 1 was a plausible value (1 meaning there was no difference in the population between treatment and control)

When thinking about a regression line, what values of our population parameters are special or distinguished in this way?

Inference

For the coefficient $\widehat{\beta}_1$, a value of zero would mean that **a fish's length is unrelated to its Mercury content**

In our case, $\widehat{\beta}_1 = 0.068$ and we used the bootstrap to estimate its standard error to be 0.007; therefore, the estimated regression coefficient is about 10 standard errors away from zero -- making it very unlikely to be the result of chance

So, while 0.068 seems small as a number, it is statistically quite far from 0; and in terms of practical importance, keep in mind that the EPA has a safety threshold of 1ppm

Testing and confidence intervals

Reasoning this way reminds us a bit of the logic behind hypothesis tests -- That is, we are scanning the confidence interval for important values (like 0 for a regression coefficient)

Formally, the two constructions are looking for consistency between samples and population parameters, but they are coming at it from slightly different perspectives

Confidence intervals: Fix the (sample) statistic and ask what values of the population parameter are consistent with the fixed statistic

Hypothesis tests: Fix the population parameter value and ask what (sample) statistics are consistent with that fixed value

It turns out there is a one-to-one correspondence between tests and confidence intervals

Testing and confidence intervals

Let [lo,hi] be a 95% confidence interval, say, for a population parameter θ -- Then for any θ_0 we can test the null hypothesis that $H_0 : \theta = \theta_0$, rejecting the null if θ_0 is not contained in [lo,hi]

The resulting test has significance level 0.05 -- In general, any $100(1-\alpha)$ percent confidence interval is equivalent to a test with significance level α

The logic works in reverse if we start with a hypothesis test and consider the set of values θ_0 for which we would fail to reject the null hypothesis that $H_0 : \theta = \theta_0$ -- These values form a confidence interval for the population parameter