Lecture 15: Regression to mediocrity

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.

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`

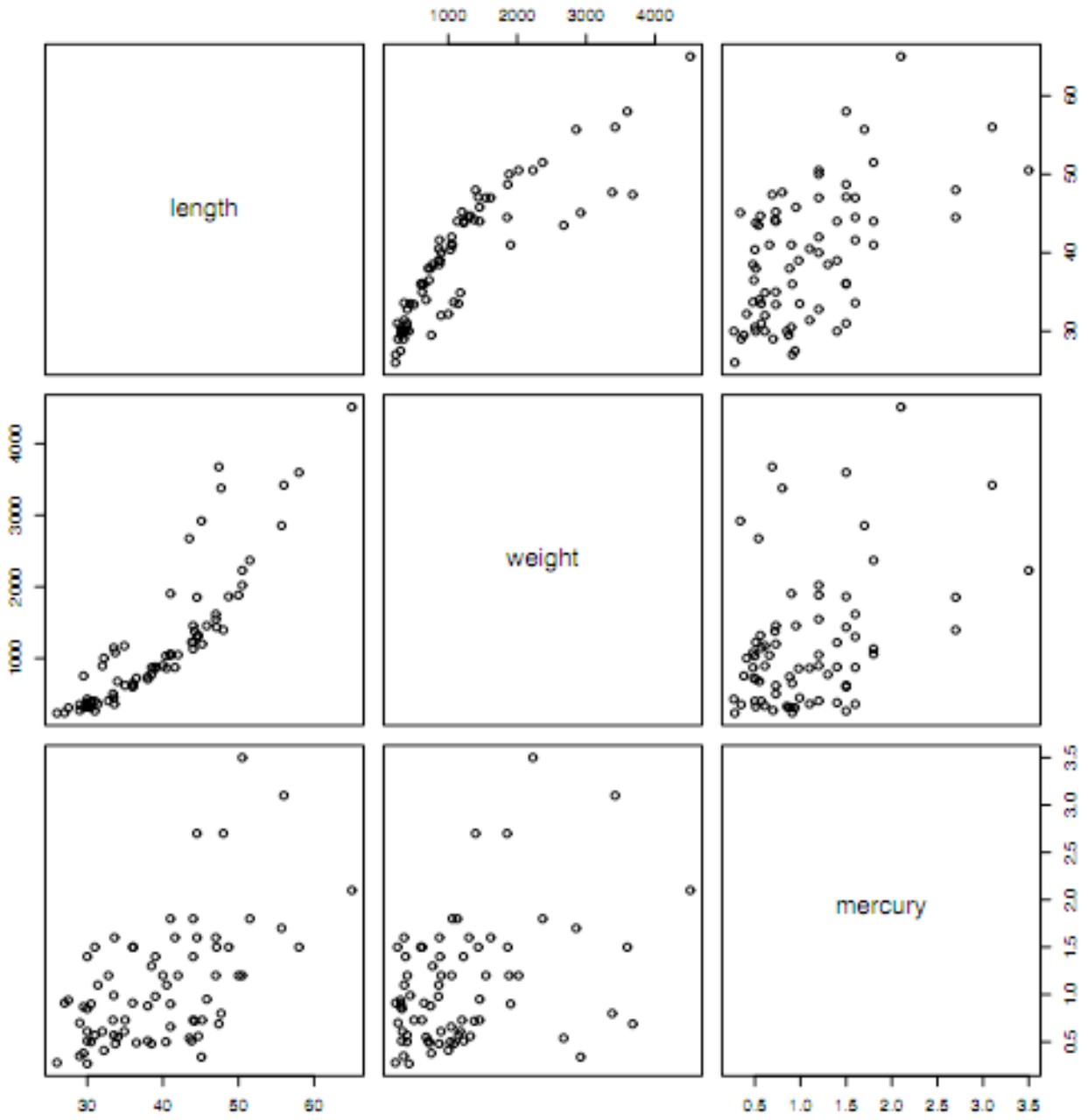
	river	stn	length	weight	mercury
1	0	0	47.0	1616	1.60
2	0	0	48.7	1862	1.50
3	0	0	55.7	2855	1.70
4	0	0	45.2	1199	0.73
5	0	0	44.7	1320	0.56
6	0	0	43.8	1225	0.51
7	0	0	38.5	870	0.48
8	0	0	45.8	1455	0.95
9	0	0	44.0	1220	1.40
10	0	0	40.4	1033	0.50
11	0	1	47.7	3378	0.80
12	0	1	45.1	2920	0.34
13	0	1	43.5	2674	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

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



Mercury contamination

From the EPA's perspective, it is natural to wonder how mercury content relates to the length of a fish

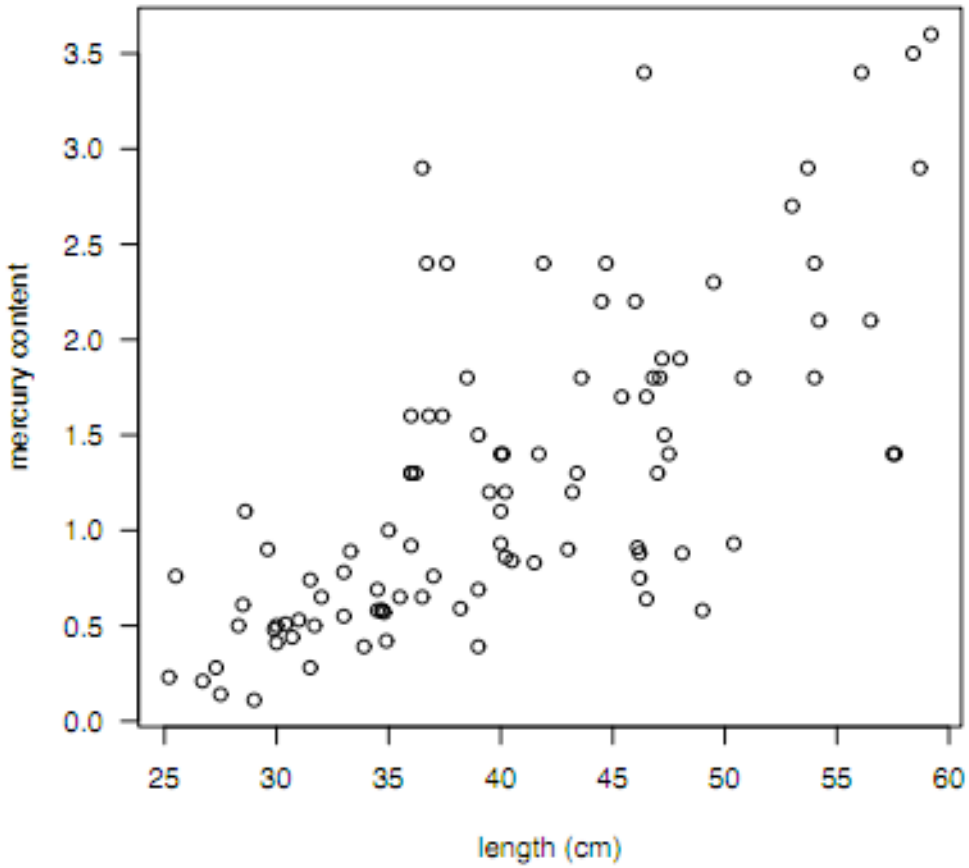
We are going to take fish length as a substitute measure for the age of the fish; it won't be precise, but we should still see some kind of relationship

We will first use data on fish taken from the Waccamaw

Mercury levels in water

Here we have length versus mercury content for the 98 fish; what do you notice?

How does mercury content "vary" as we look at fish with different lengths?



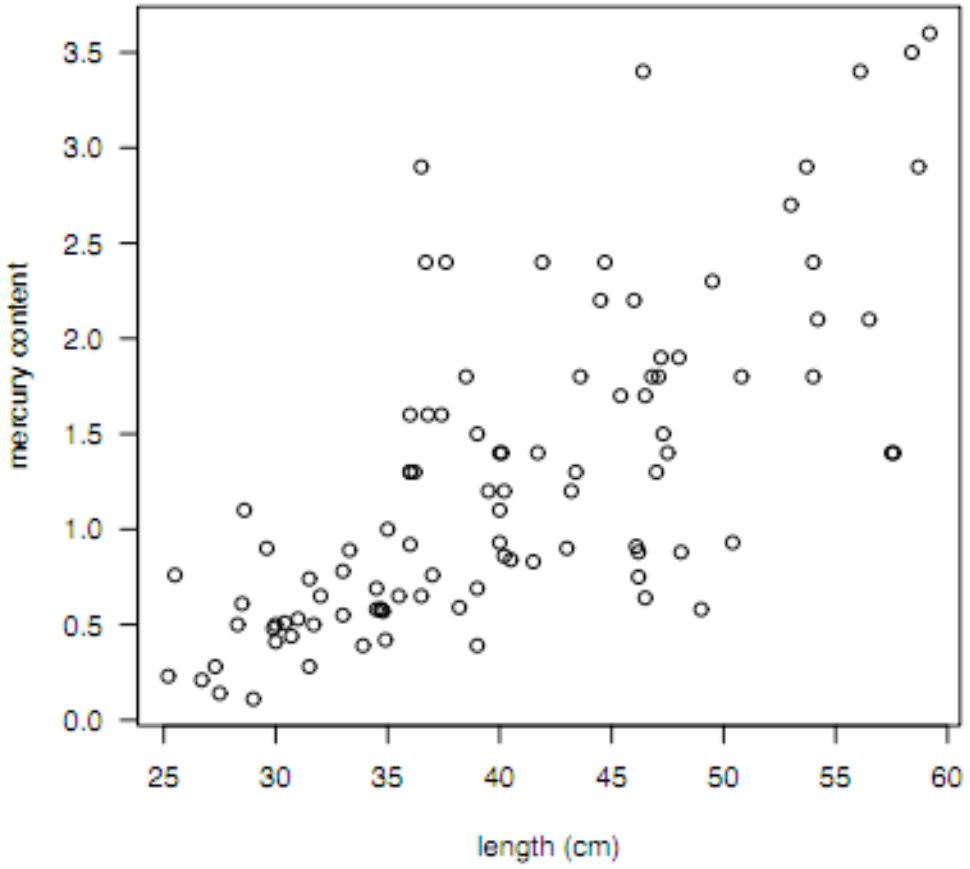
A linear model

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

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

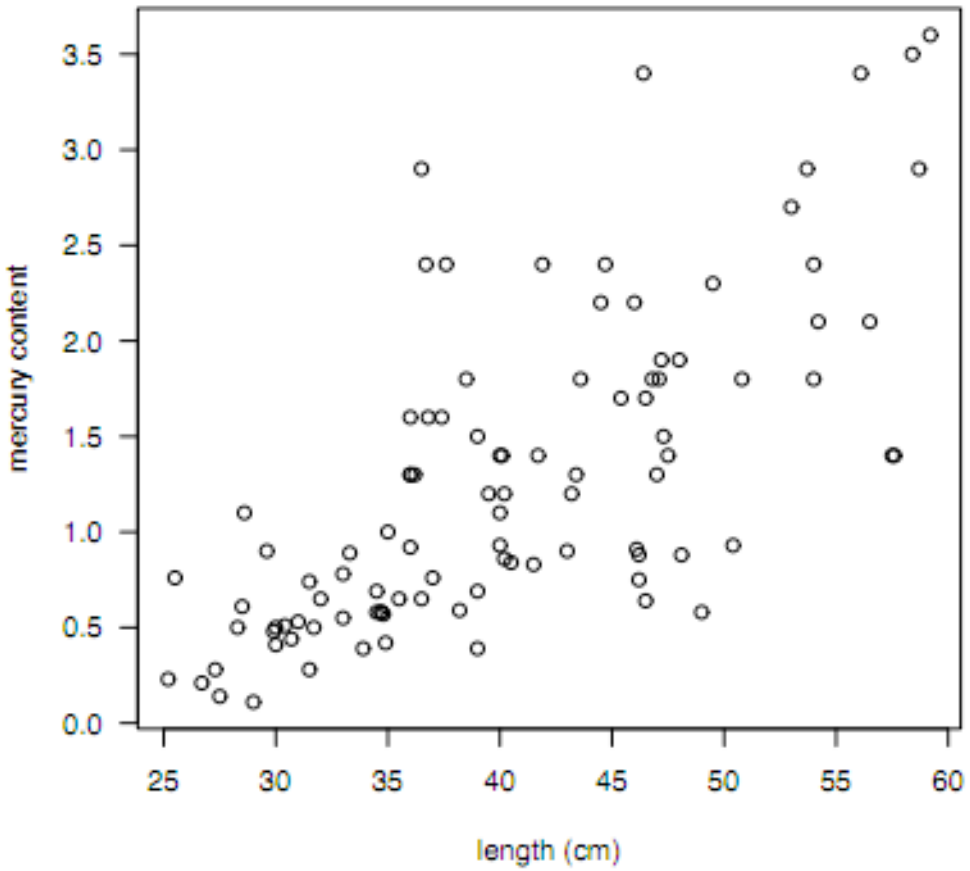
where β_0 and β_1 are parameters, the slope and intercept



A linear model

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

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



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

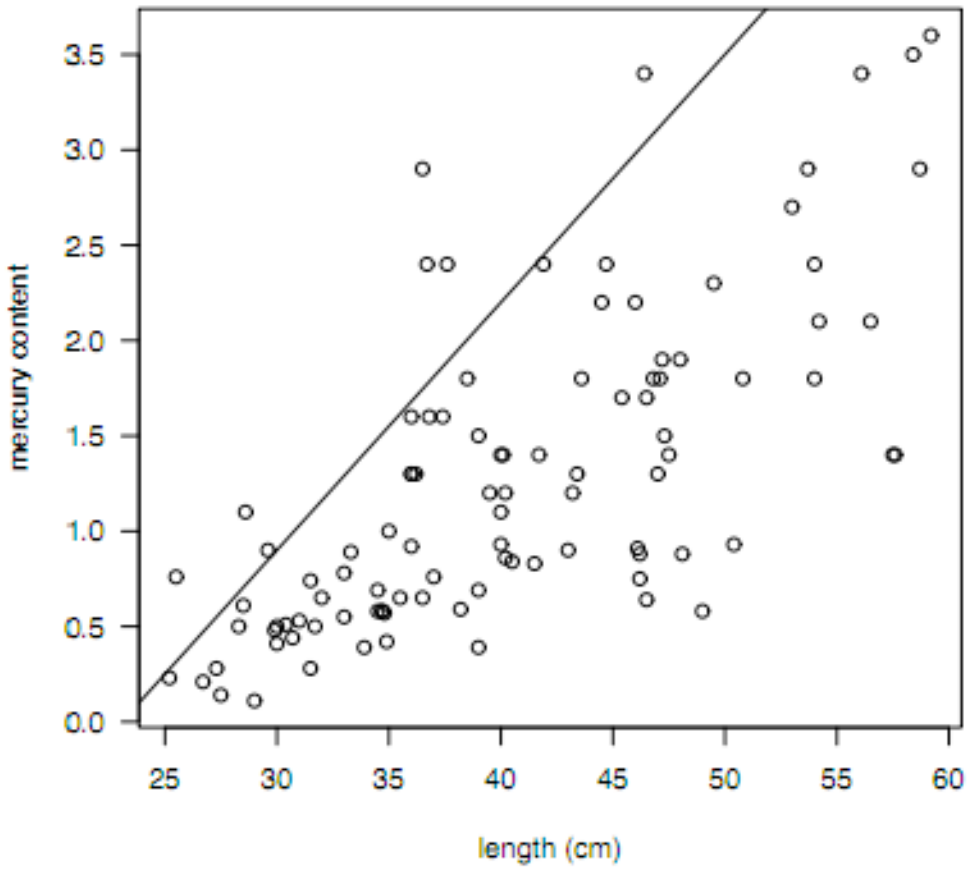
We then label our data set $(x_1, y_1), \ldots, (x_{98}, y_{98})$

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

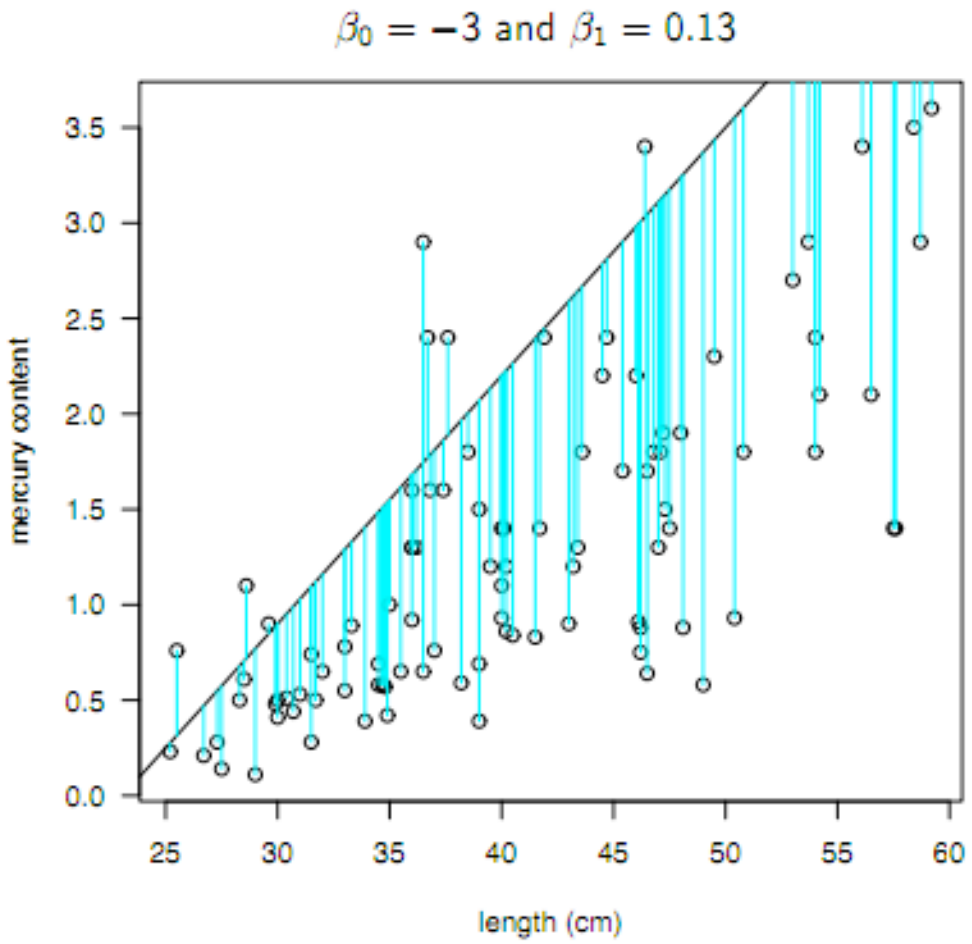
$\beta_0 = -3$ and $\beta_1 = 0.13$



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

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

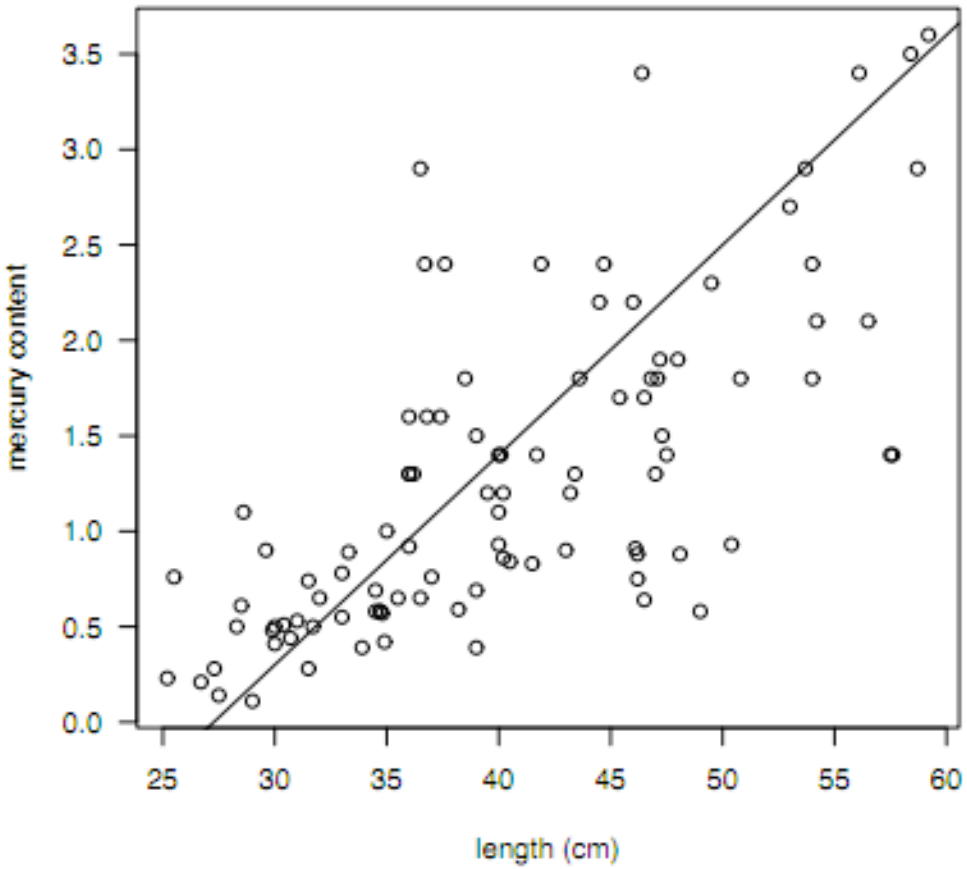
In this case, the squared error is 154.9

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 $\beta_1 = 0.11$

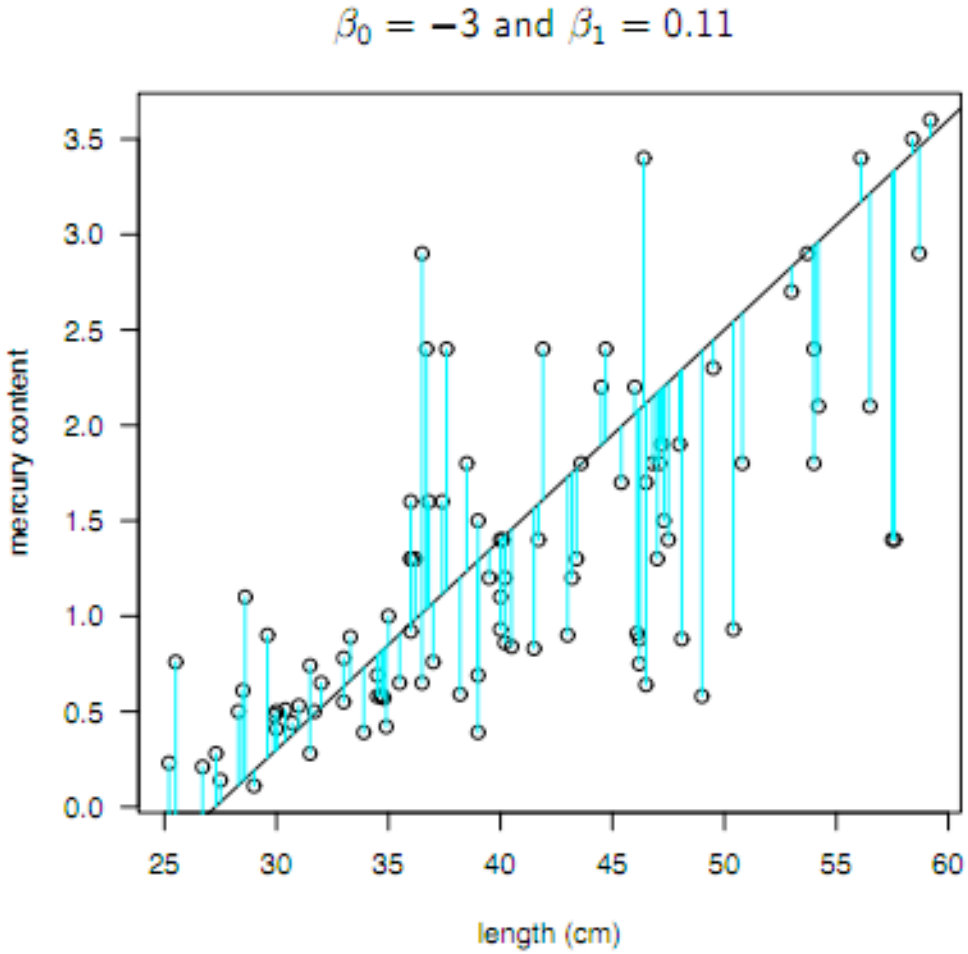


Least squares

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

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

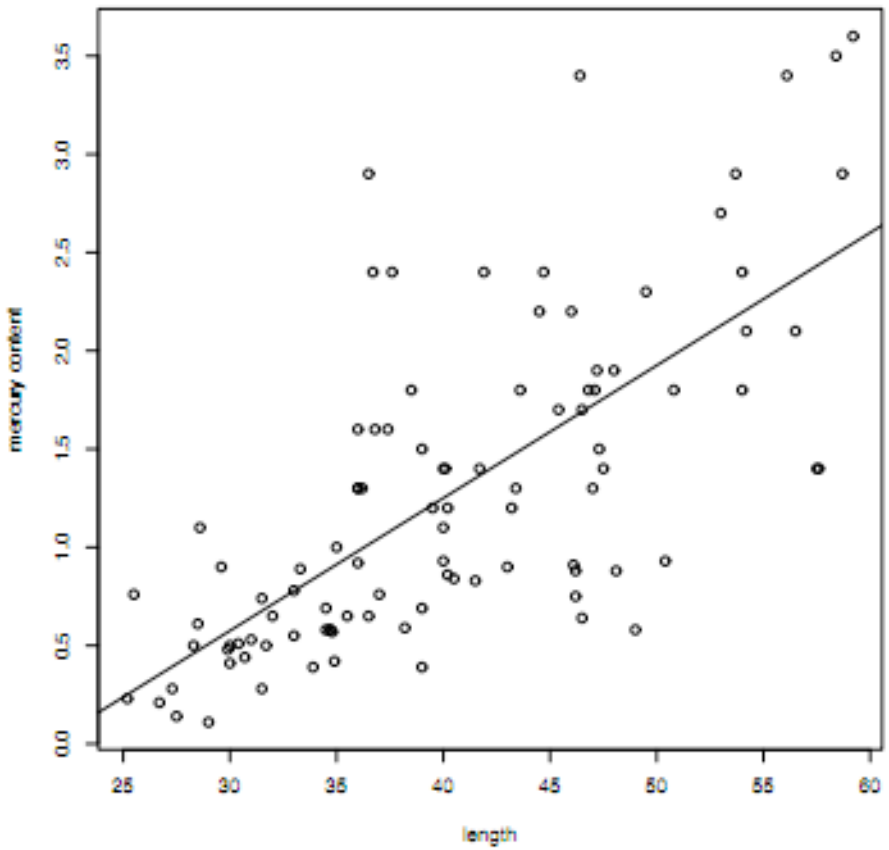
and the associated **sum of squares** is 33.4 (our simple trial and error approach was pretty far off!)

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)

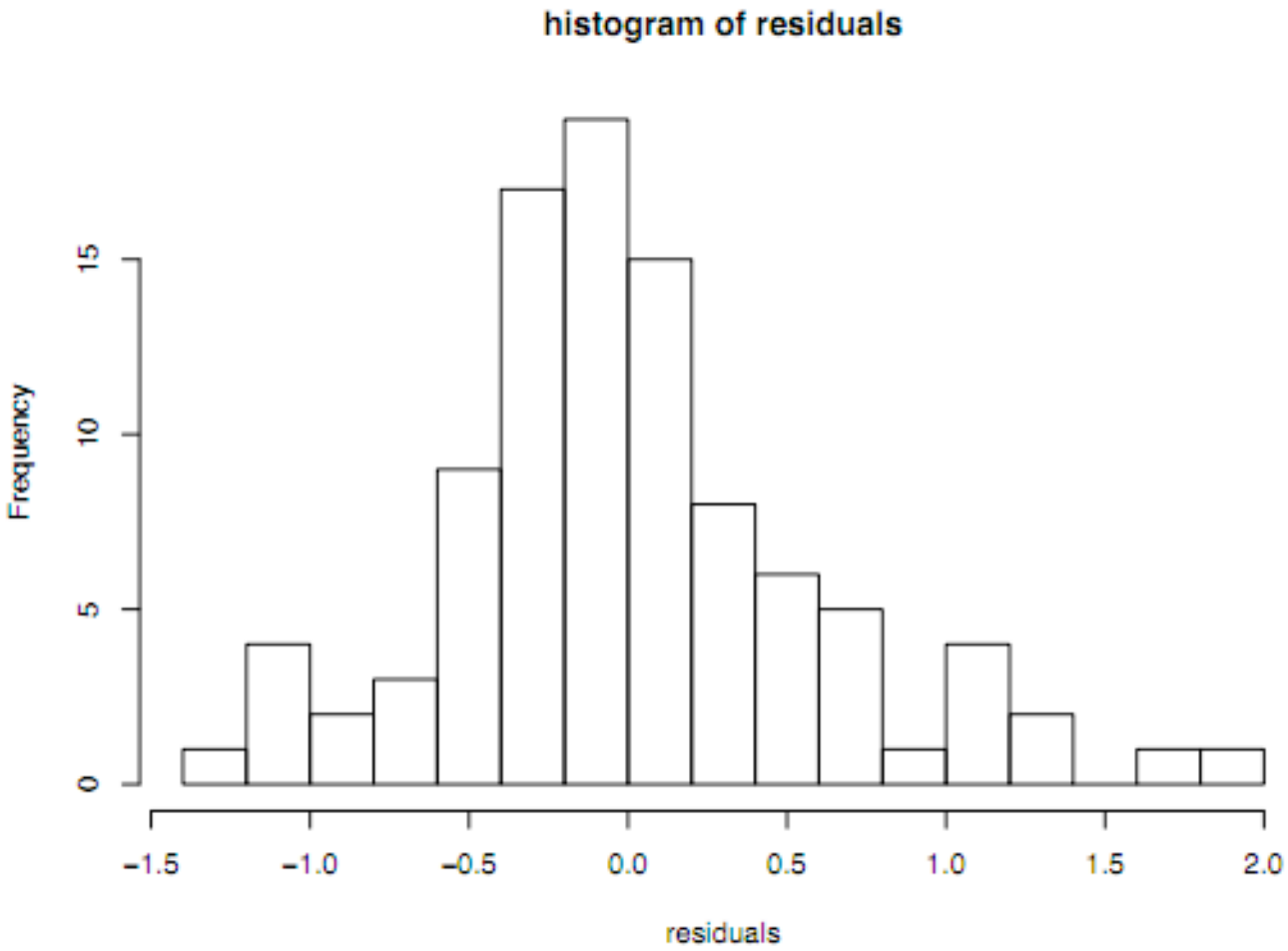
An example

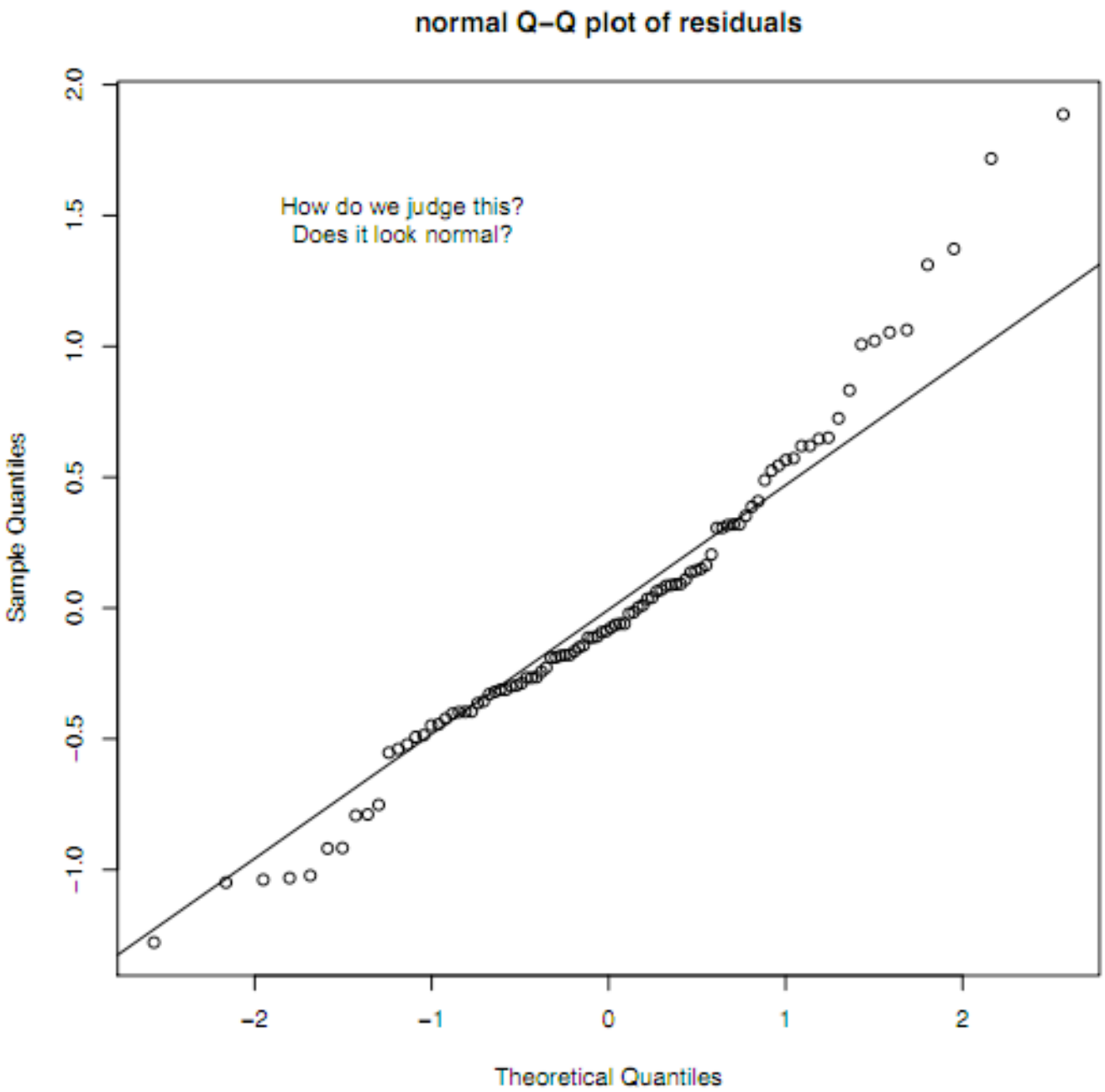
On the next two pages, we plot the residuals from the least squares fit to the fish data; the regression relating fish length and mercury content

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

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

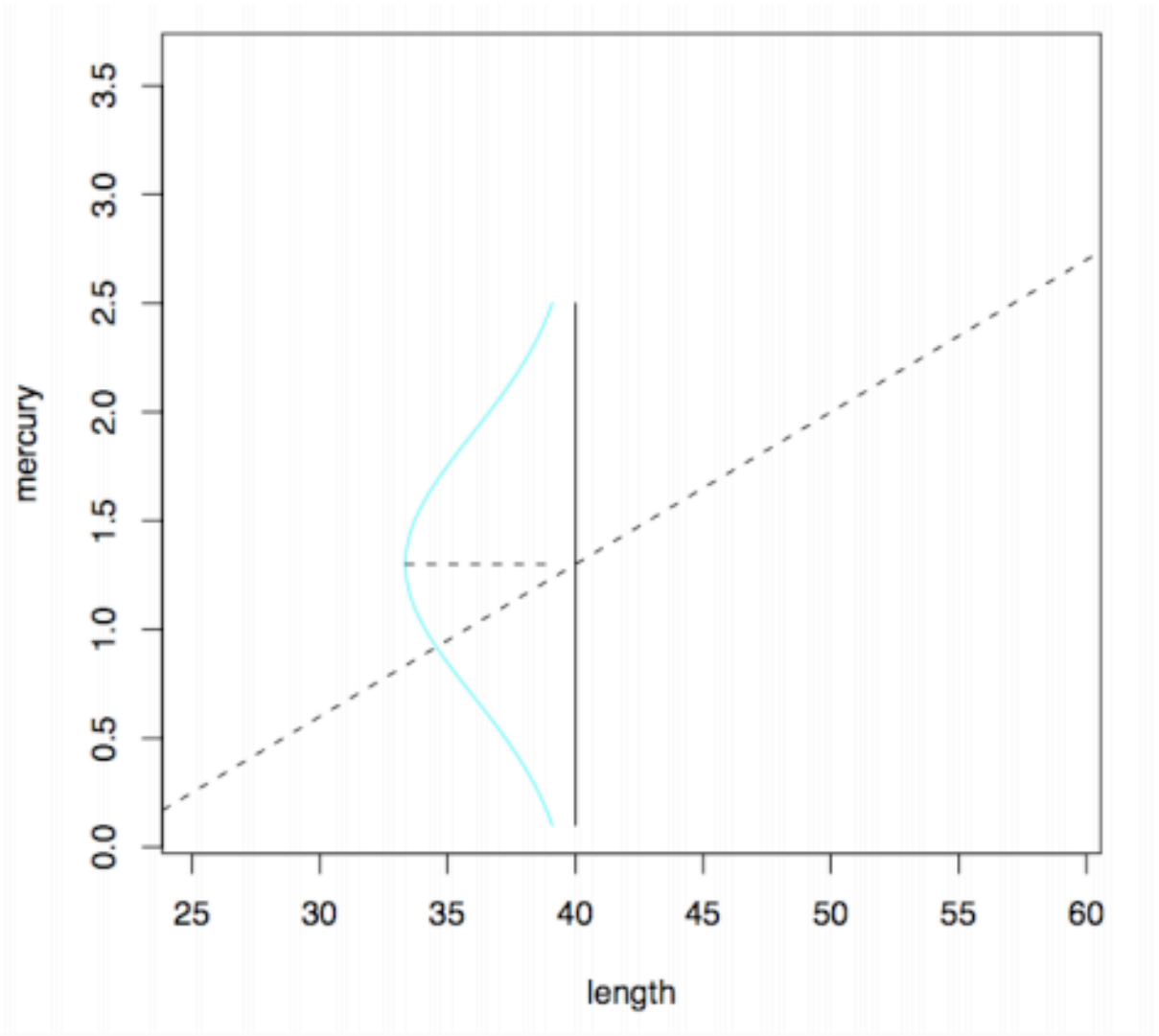
In its simplest form, the model we wrote for descriptive purposes

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

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

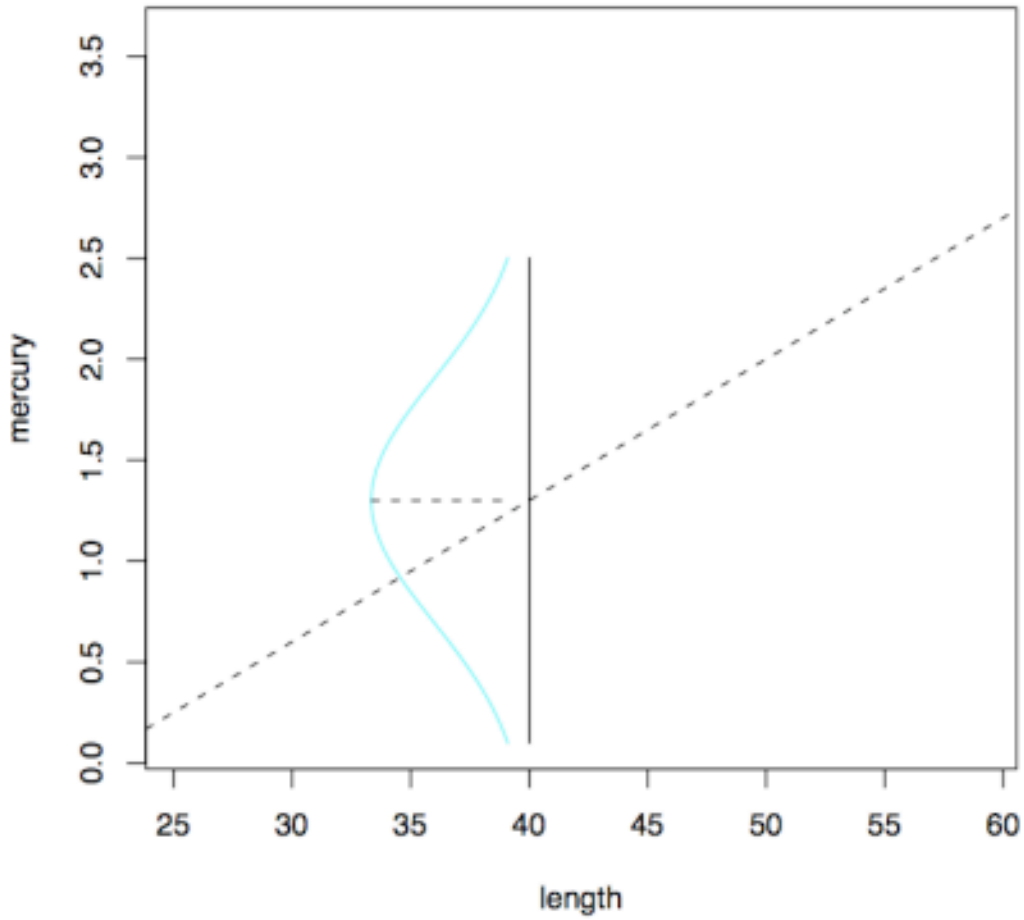
A population model

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!



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 next 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.

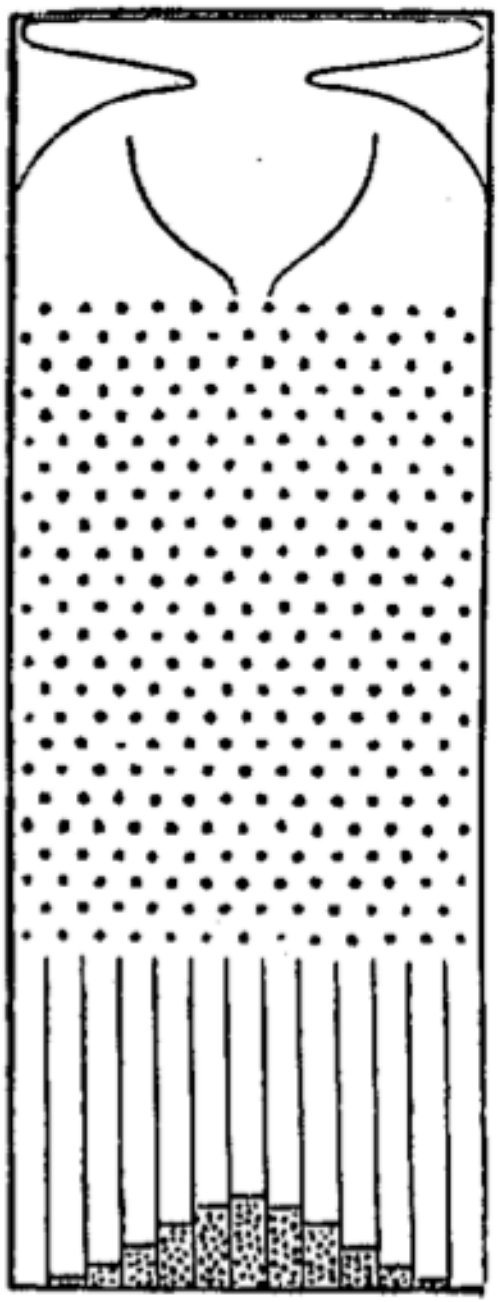


FIG. 8.

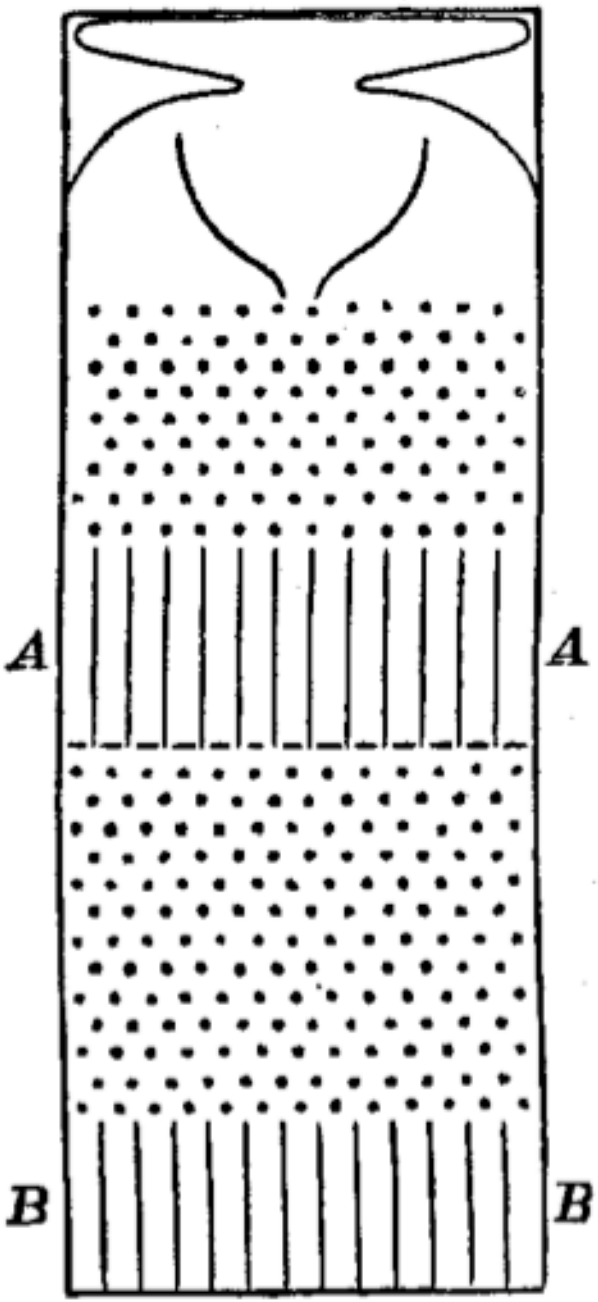
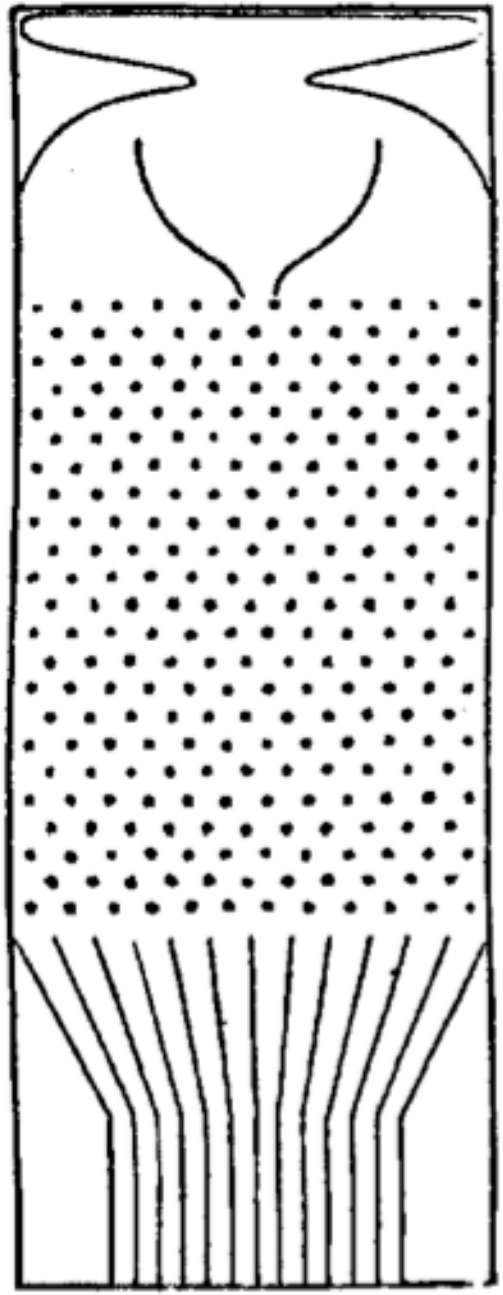


FIG. 9.

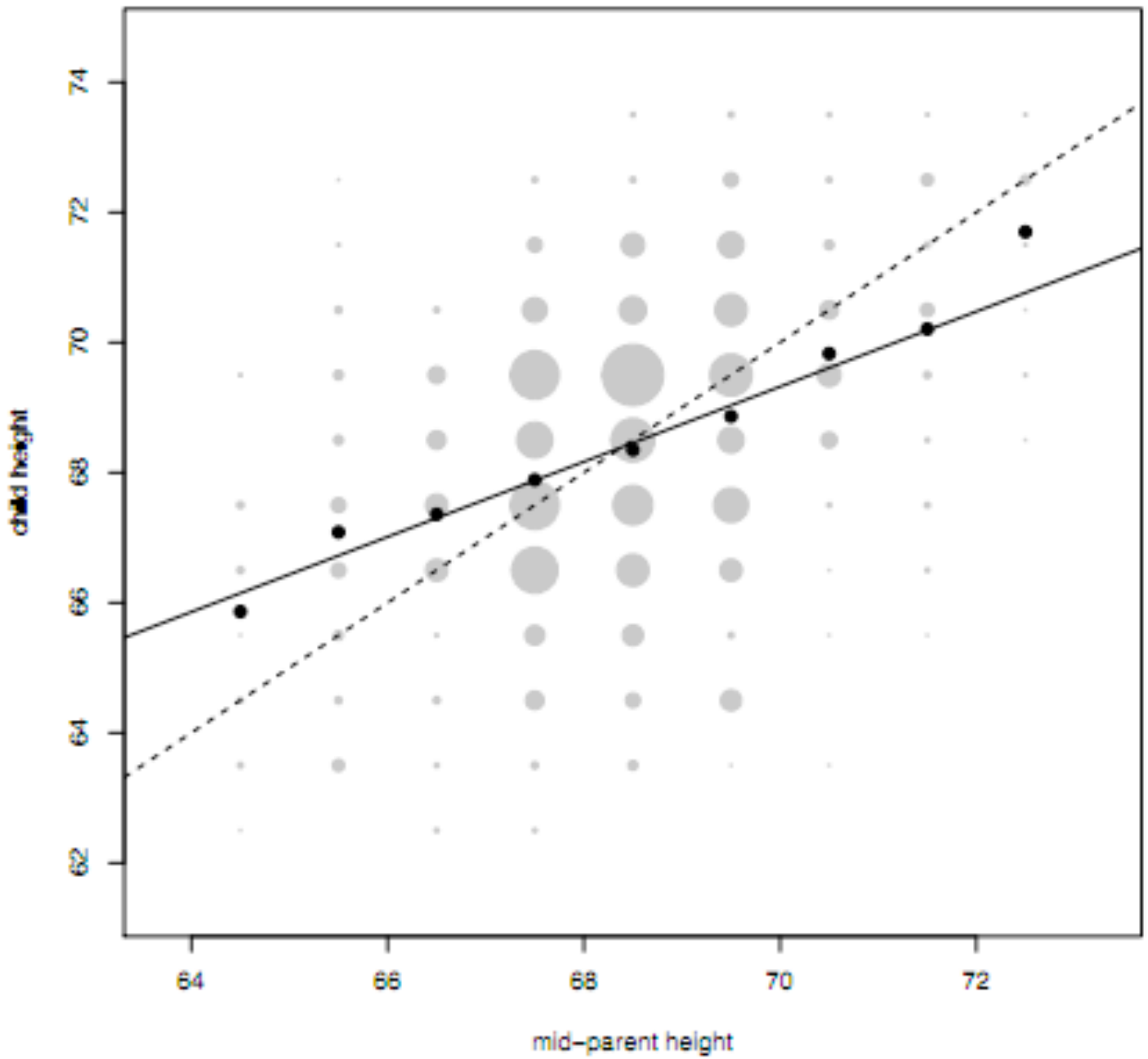


Galton and regression

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

This idea can also be seen in his treatment of the height data we looked at last time -- Around each value of mid-parent heights we had some spread...


child height
74
72
70
68
66
64
62
64
66
68
70
72
mid−parent height

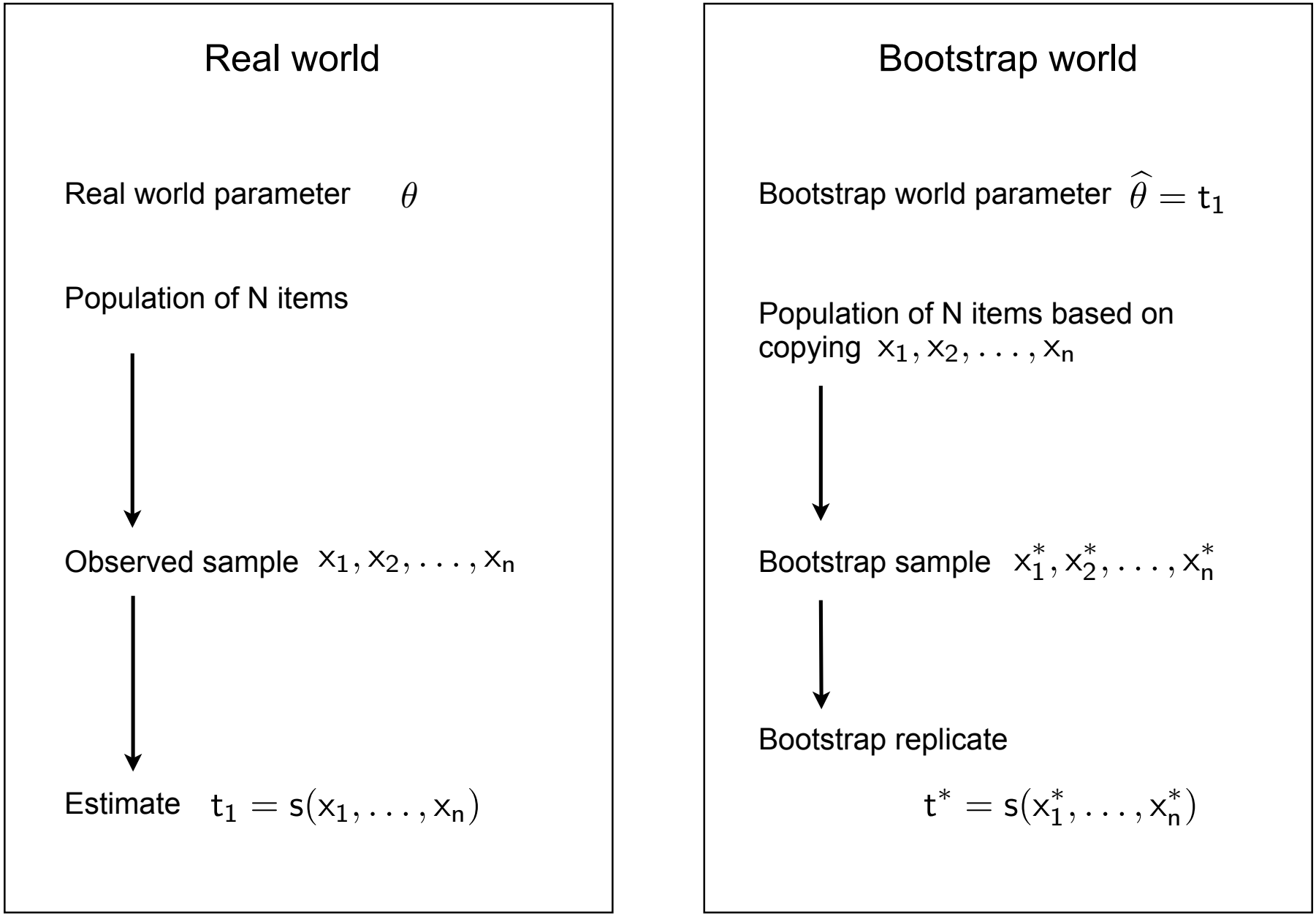
Inference

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 $x_1, x_2, \ldots, x_n$
Estimate $t_1 = s(x_1, \ldots, x_n)$
Bootstrap world
Bootstrap world parameter $\widehat{\theta} = t_1$
Population of N items based on copying $x_1, x_2, \ldots, x_n$
Bootstrap sample $x_1^*, x_2^*, \ldots, x_n^*$
Bootstrap replicate
$t^* = s(x_1^*, \ldots, x_n^*)$

The bootstrap: Regression (I)

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

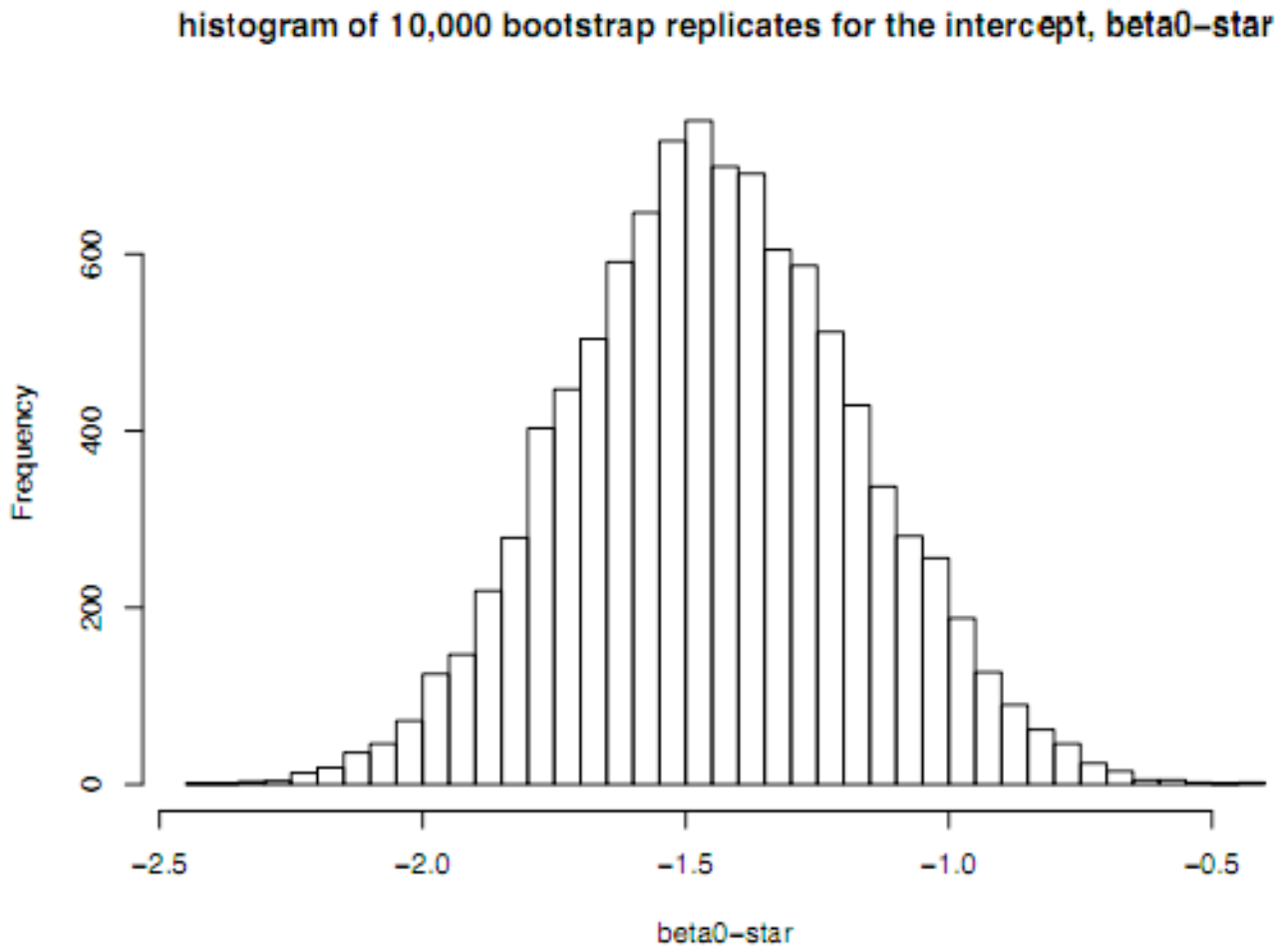
1. Create a "population" consisting of our 98 pairs $(x_1, y_1), \ldots, (x_{98}, y_{98})$
2. We then **draw 98 pairs with replacement** to form a bootstrap sample $(x_1^*, y_1^*), \ldots, (x_{98}^*, y_{98}^*)$
3. Next, we compute a least squares fit to bootstrap sample, producing a bootstrap replicate 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

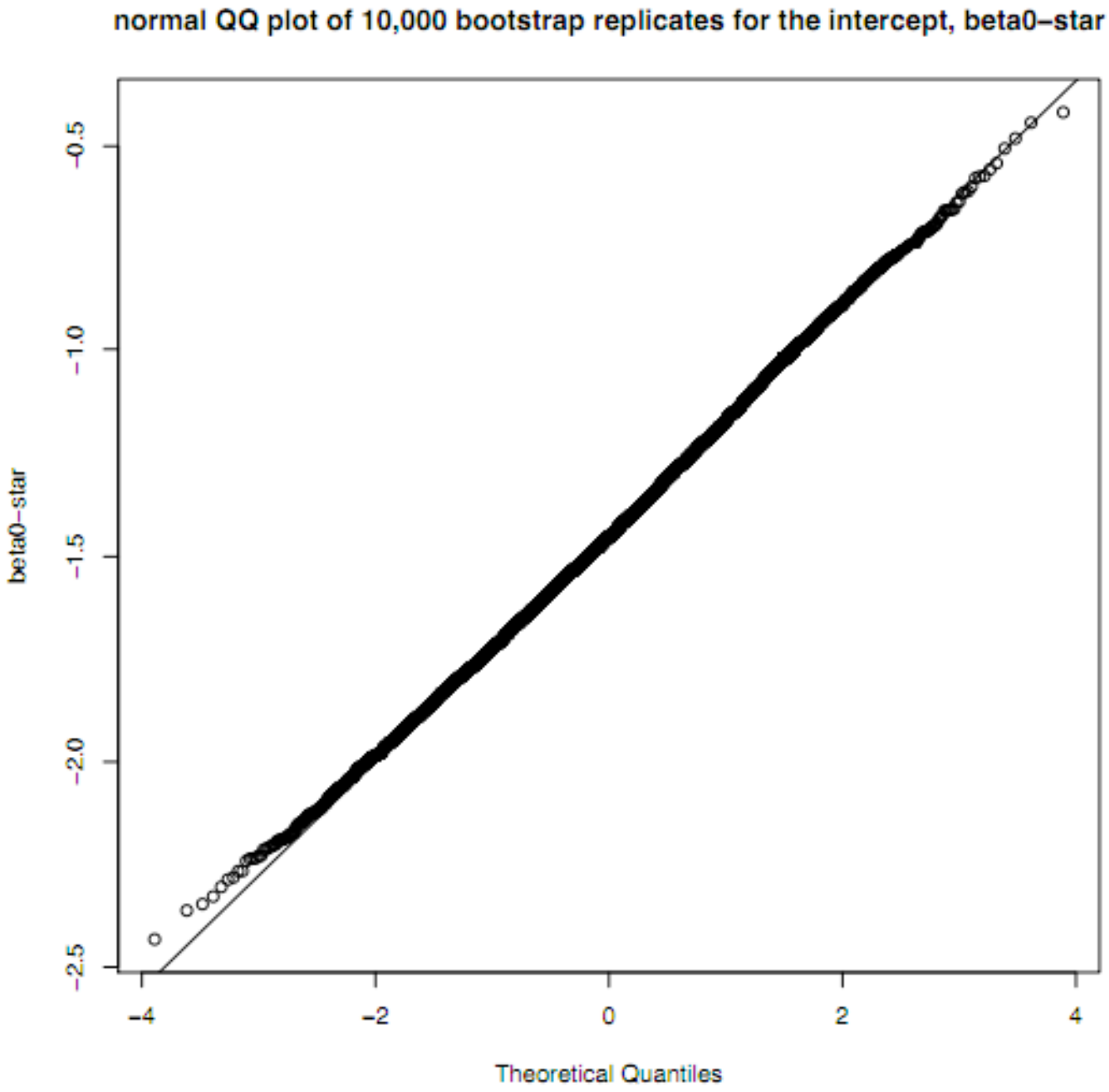
The bootstrap

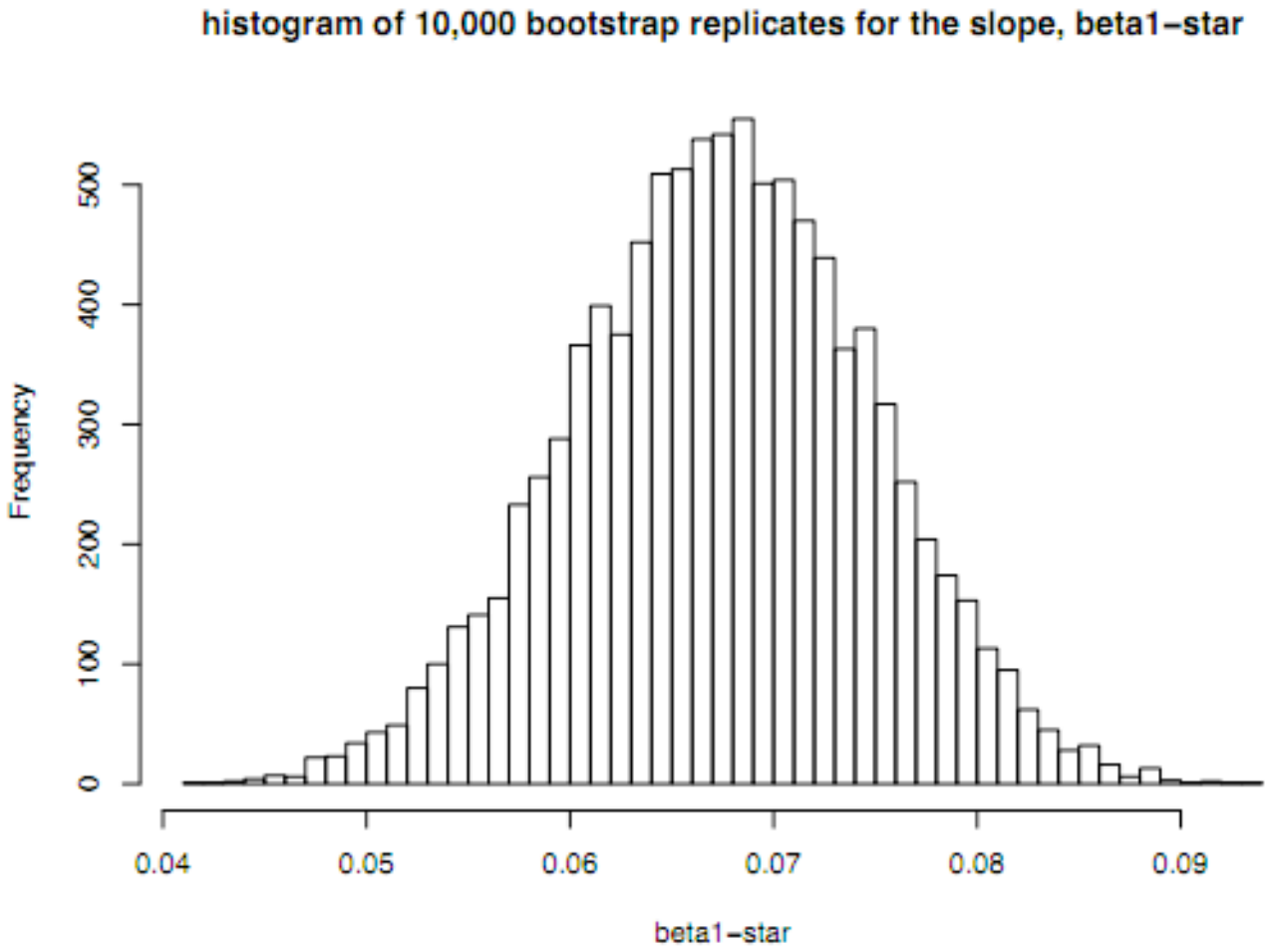
The bootstrap distribution for $\widehat{\beta}_0^*, \widehat{\beta}_1^*$ again is an estimate of their sampling distribution and we can use it to estimate the standard error of each, together with confidence intervals

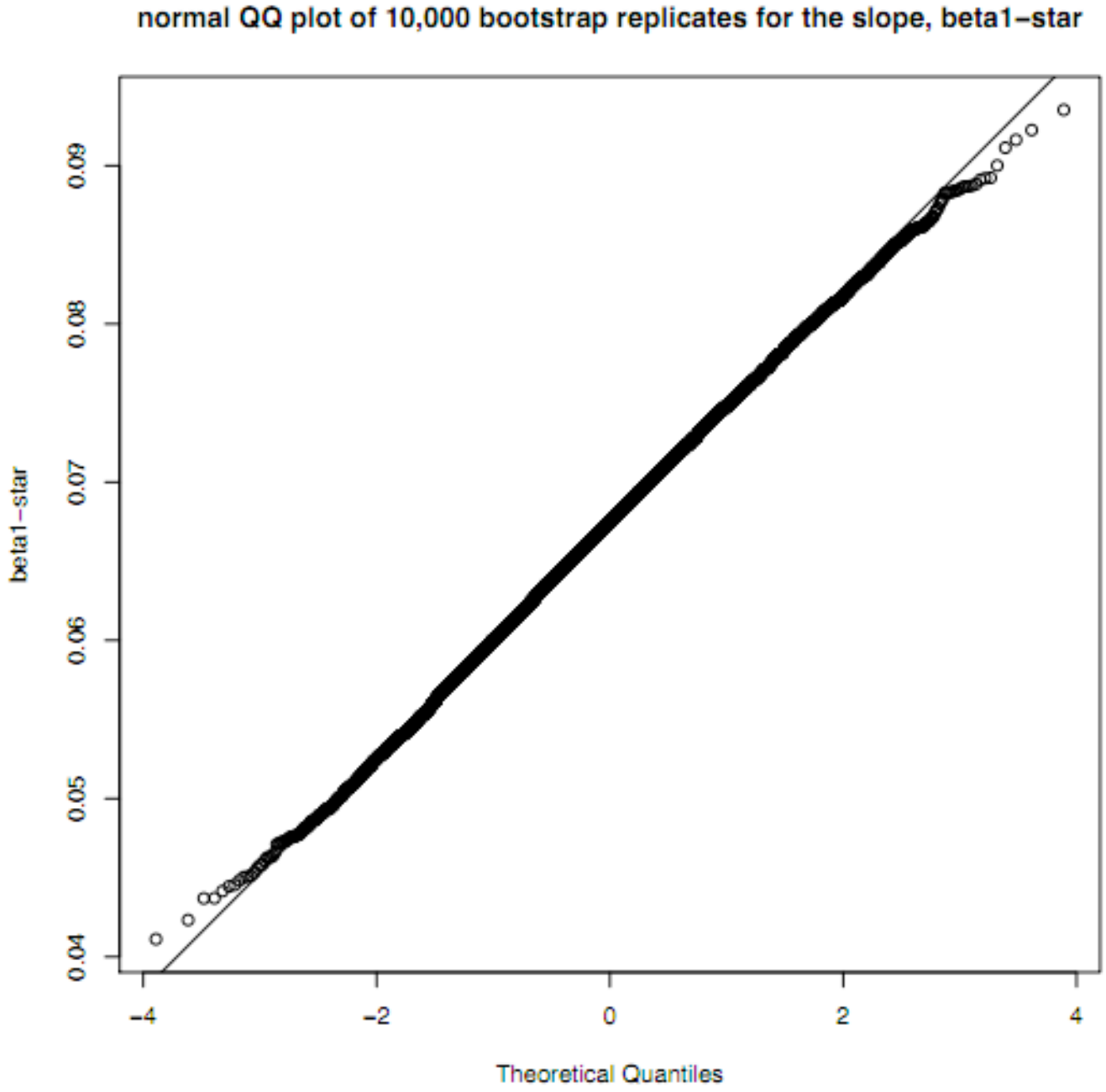
First, $\widehat{\beta}_0 = -1.45$ and the standard deviation of the bootstrap replicates is 0.28; a confidence interval using the 0.025 and 0.975 quantiles agrees with $-1.45 \pm 1.96 * 0.28 = [-2.00, -0.90]$

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

Prelude

In Lab this week, you will fit a linear model, and possibly a decision tree; we specify statistical models in R using its formula language -- Below we given an example for the linear model relating Mercury levels and length

```
> names(waccamaw)
[1] "river"   "station" "length"  "weight"  "mercury"

> fit = lm(mercury~length,data=waccamaw)
> summary(fit)

Call:
lm(formula = mercury ~ length, data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27784 -0.32696 -0.08177  0.31462  1.88604

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.450166   0.284687  -5.094 1.75e-06 ***
length       0.067510   0.006893   9.794 4.12e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5895 on 96 degrees of freedom
Multiple R-squared: 0.4998,   Adjusted R-squared: 0.4946
F-statistic: 95.92 on 1 and 96 DF,  p-value: 4.118e-16
```

Comparing the classics

The R output refers to t-statistics and tests of hypothesis for each of the coefficients in the regression equation (we even see multiple *'s to denote significance)

There are **close connections between the analytical approach** to the sampling distribution we mapped out for $\bar{x}$ and that for $\widehat{\beta}_0, \widehat{\beta}_1$; in both cases, we end up with a t-distribution when the data (or for least squares, when the errors) are normal or an approximate t for large samples

Comparing the classics

Sample mean $\bar{x}$

Approximately normal sampling distribution for large n

$\frac{\bar{x} - \mu}{\widehat{SE}}$ has a t-distribution with n-1 degrees of freedom when the data are normal

95% confidence intervals are of the form $\bar{x} \pm t^* \widehat{SE}$

$\frac{\bar{x}}{\widehat{SE}}$ can be used to test the null hypothesis that $\mu = 0$

Least squares estimates $\hat{\beta}_0, \hat{\beta}_1$

Approximately normal sampling distributions for large n

$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}}$ has a t-distribution with n-2 degrees of freedom when the errors are normal

95% confidence intervals are of the form $\hat{\beta}_1 \pm t^* \widehat{SE}$

$\frac{\hat{\beta}_1}{\widehat{SE}}$ can be used to test the null hypothesis that $\beta_1 = 0$

The standard error, classically

There are three parameters in our population model -- The two regression coefficients (slope and intercept) $\beta_1,\ \beta_0$ and the error standard deviation σ

We estimate the regression coefficients by least squares giving us $\widehat{\beta}_1$ and $\widehat{\beta}_0$ and we can assess sampling variability using the bootstrap

The classical tools, however, follow what we did for the mean -- The estimate of the error standard deviation is simply

$$\widehat{\sigma} = \sqrt{\frac{\sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}}$$

and the classical estimate of the standard error of $\widehat{\beta}_1$, say, is

$$\frac{\widehat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

Correlation

Galton quantified the extent to which his regression effect would take place; let $\bar{x}$ and s_x be the mean and standard deviation of the parent's height and let $\bar{y}$ and s_y be the mean and standard deviation of the children's heights

If we consider all parents whose heights are k standard deviations from the mean or $k = (x - \bar{x})/s_x$, then we expect children's heights y to be r*k standard deviations s_y from the overall average height of the children $\bar{y}$

Galton coined the term co-relation for r; today we've blurred things together to just call it **correlation**

Correlation

Put another way, Galton defined r so that

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

If we unpack that a little, we see that r is related to the slope of the regression line

$$y = (r\, s_y/s_x)\, x - (r\, s_y/s_x)\, \bar{x} \quad \text{or} \quad \widehat{\beta}_1 = r\, s_y/s_x$$

You can also write it out explicitly

$$r = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{s_x\, s_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

While this may seem a bit daunting, the correlation has a fairly intuitive interpretation, describing how “tight” the linear relationship between two variables is

The bivariate normal

To make this concrete, consider first **the data model that Galton was working with**; it is a two-dimensional version of the normal distribution -- a “bell shape” that defines the scatter of two variables

We have seen examples of this before...

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).

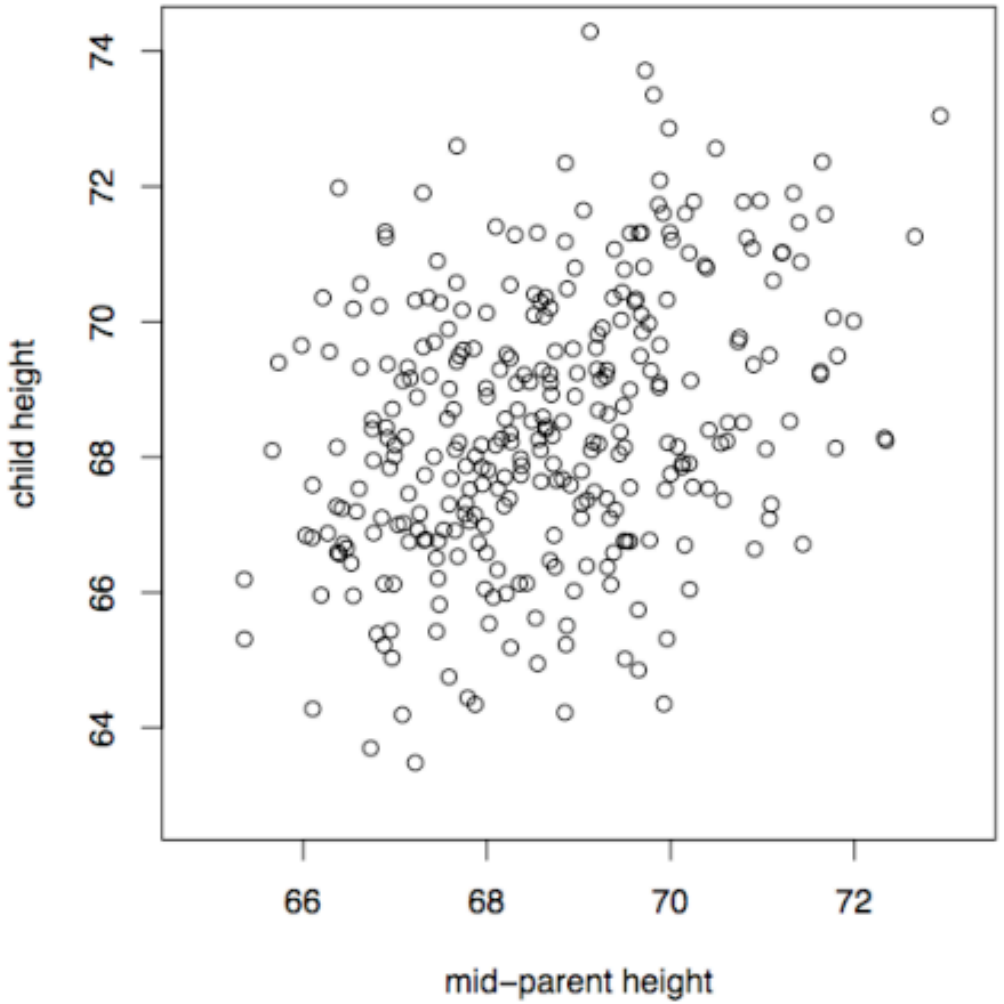
Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

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	

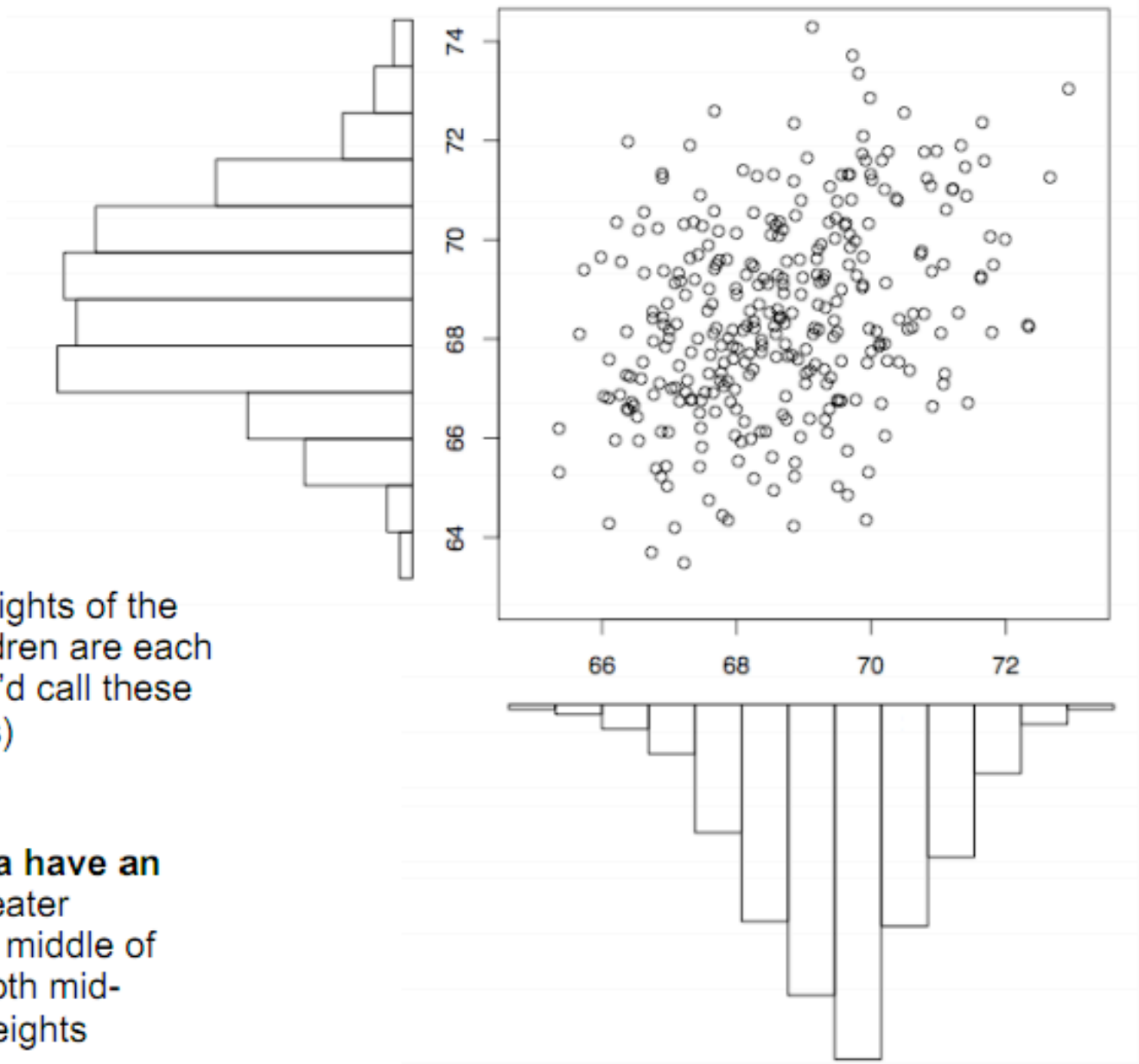
The bivariate normal

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



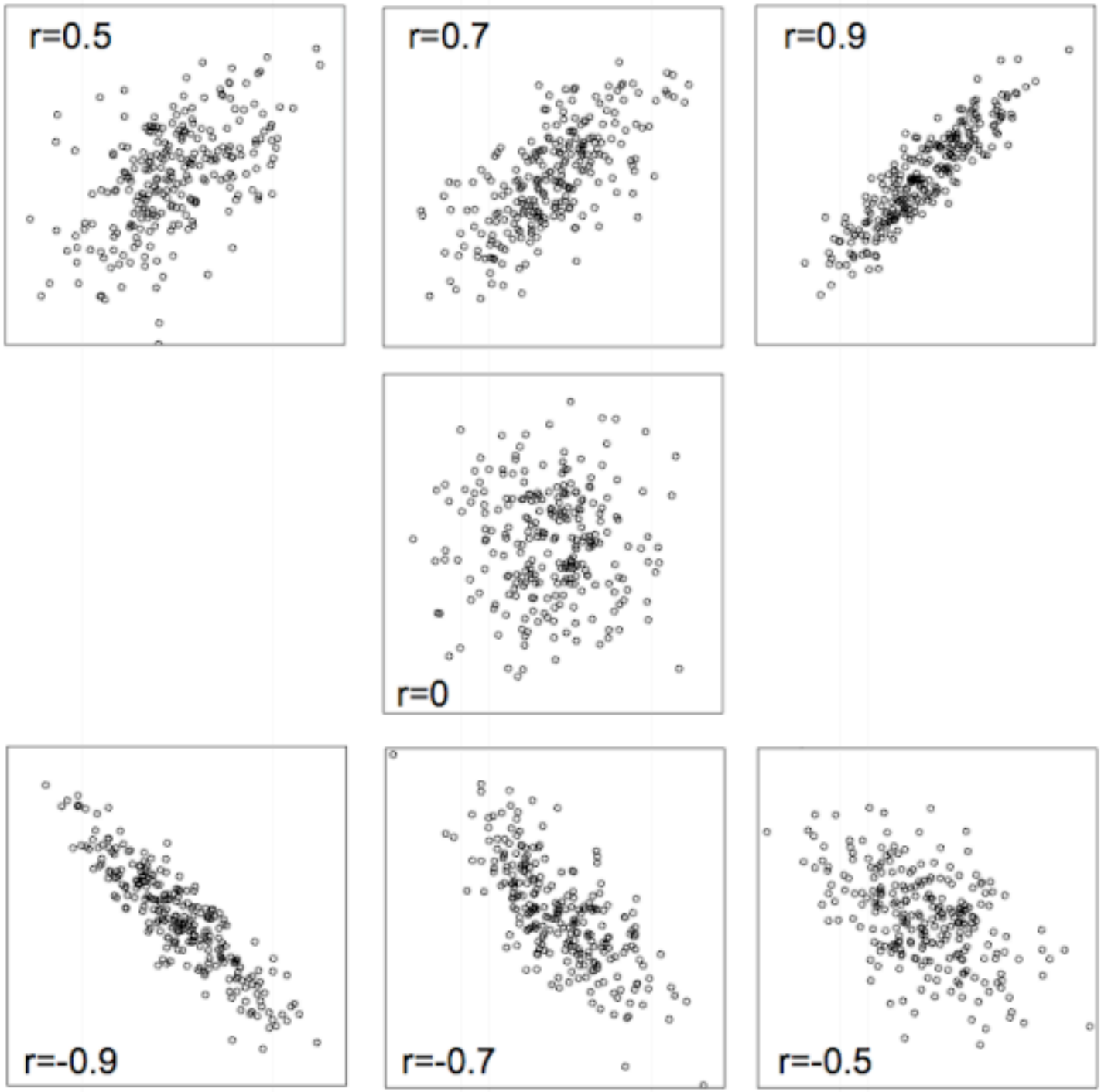
The distribution of the heights of the mid-parents and the children are each **individually normal** (we'd call these the marginal distributions)

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

The bivariate normal

While the mean and standard deviation are enough to specify a regular normal distribution, for a bivariate normal we need a little more

In addition to the means and standard deviations of each individual variable, we also need the correlation coefficient; it describes how tight the relationship is


r=0.5
r=0.7
r=0.9
r=0
r=-0.9
r=-0.7
r=-0.5

Correlation

Therefore, as a measure, the correlation coefficient has a couple nice properties

1. It is bounded between -1 and 1; with either -1 or 1 indicating a "perfect relationship" between the two variables

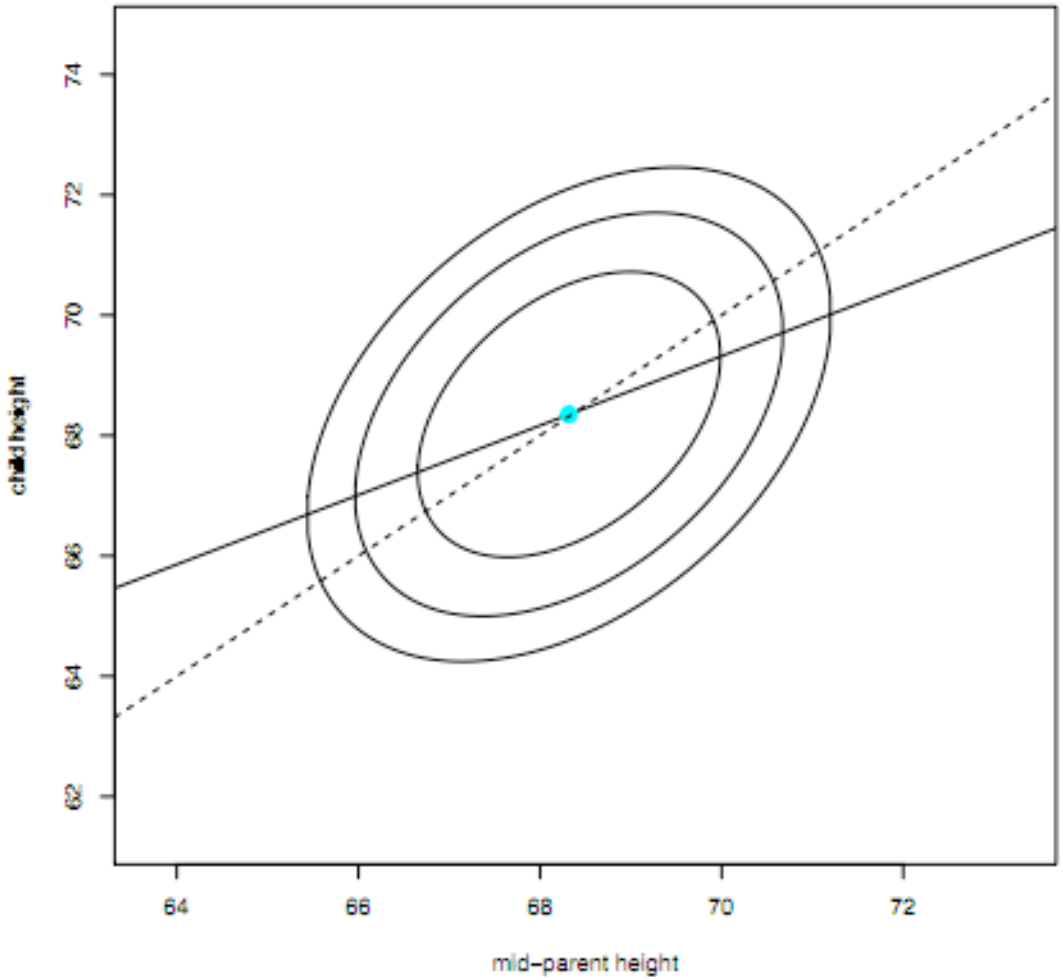
2. If it is negative, then as one variable increases, the other tends to decrease; if it is positive, they vary in the same direction

Finally, the squared correlation coefficient r^2 is also known as the coefficient of determination and is related to the sum of squares computations we performed last time (although in that lecture, I think we used R^2)

Pulling it all together

At the right we have a diagram of **the population model that Galton studied**; the elliptical contours represent a bivariate normal distribution (imagine a surface such that as you moved out from the center, each ellipse was at lower value)

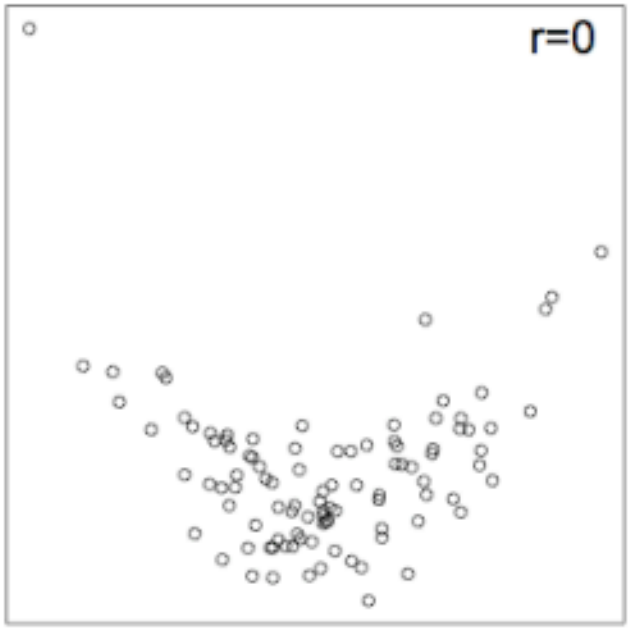
The solid line is the regression line of child height on mid-parent height and the dashed line is y=x

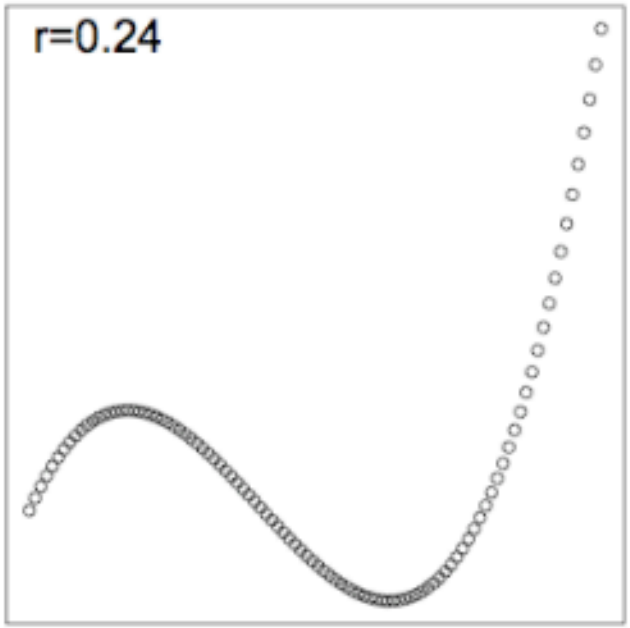


Correlation

It's important to keep this simple data model in mind when you interpret your own regression results; the nice properties of the correlation coefficient, for example, can break down if the underlying data are not bivariate normal

At the right we have two data sets that arguably have fairly "tight" relationships; but because they are not linear



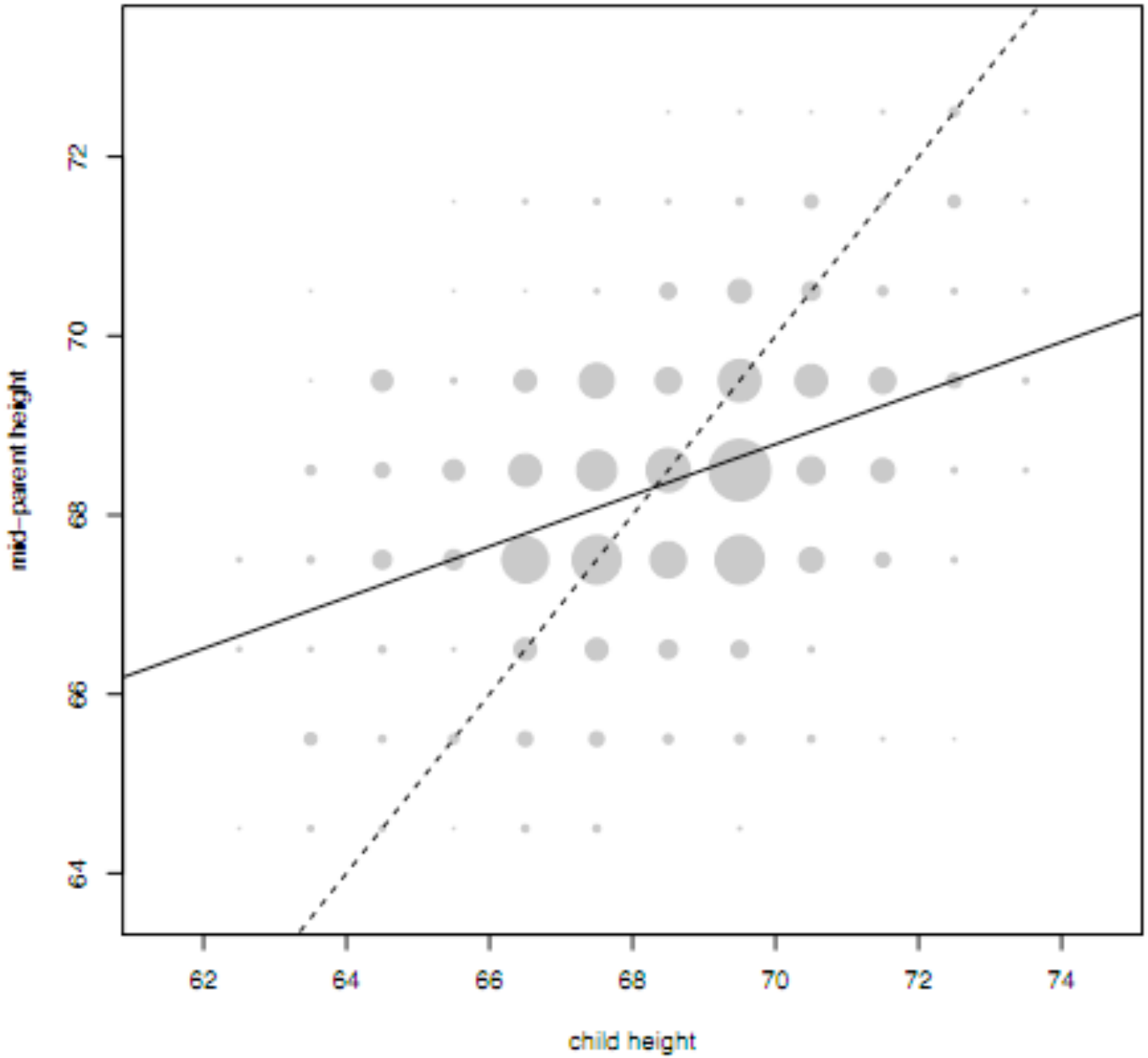


A final comment

As a term, **"regression analysis"** has come to represent more than just the stature studies Galton pursued; it has become synonymous with the linear model and its extensions

Regression toward the mean is also a general phenomenon not restricted to heights; it is a fact about regression in general

For example, we can flip Galton's figure, regressing mid-parent height on child height and find the same effect....


mid−parent height
72
70
68
66
64
62
64
66
68
70
72
74
child height

A final comment

For experimenters, we have to be careful **not to interpret this effect as the result of some kind of intervention**; in the life sciences, for example, it occurs in situations of repeated measurements when extremely large or small values are followed by measurements of the same subjects that on average are closer to the mean of the basic population

Such changes are likely to be interpreted as a real drift, although they just might be artificial coming from the fact that the sampling of values was not random but selected

For example, we can flip Galton's figure, regressing mid-parent height on child height and find the same effect....

A final comment

So far, we have dealt with regression in the context of **random sampling**; fish were drawn from a river, children sampled at random **-- Here our data appear in (x,y) pairs**

The mechanics behind regression is the same if we **conduct an experiment** -- In that case we are typically have **control over the x's, assigning treatments or varying conditions,** and we observe the y's

In the first case, we are merely drawing conclusions about associations between variables, while in the latter case we can make statements about causation

Back to the river

Last time, we performed a simple regression analysis to assess the relationship between Mercury levels and fish length for 98 fish taken from the Waccamaw river in North Carolina

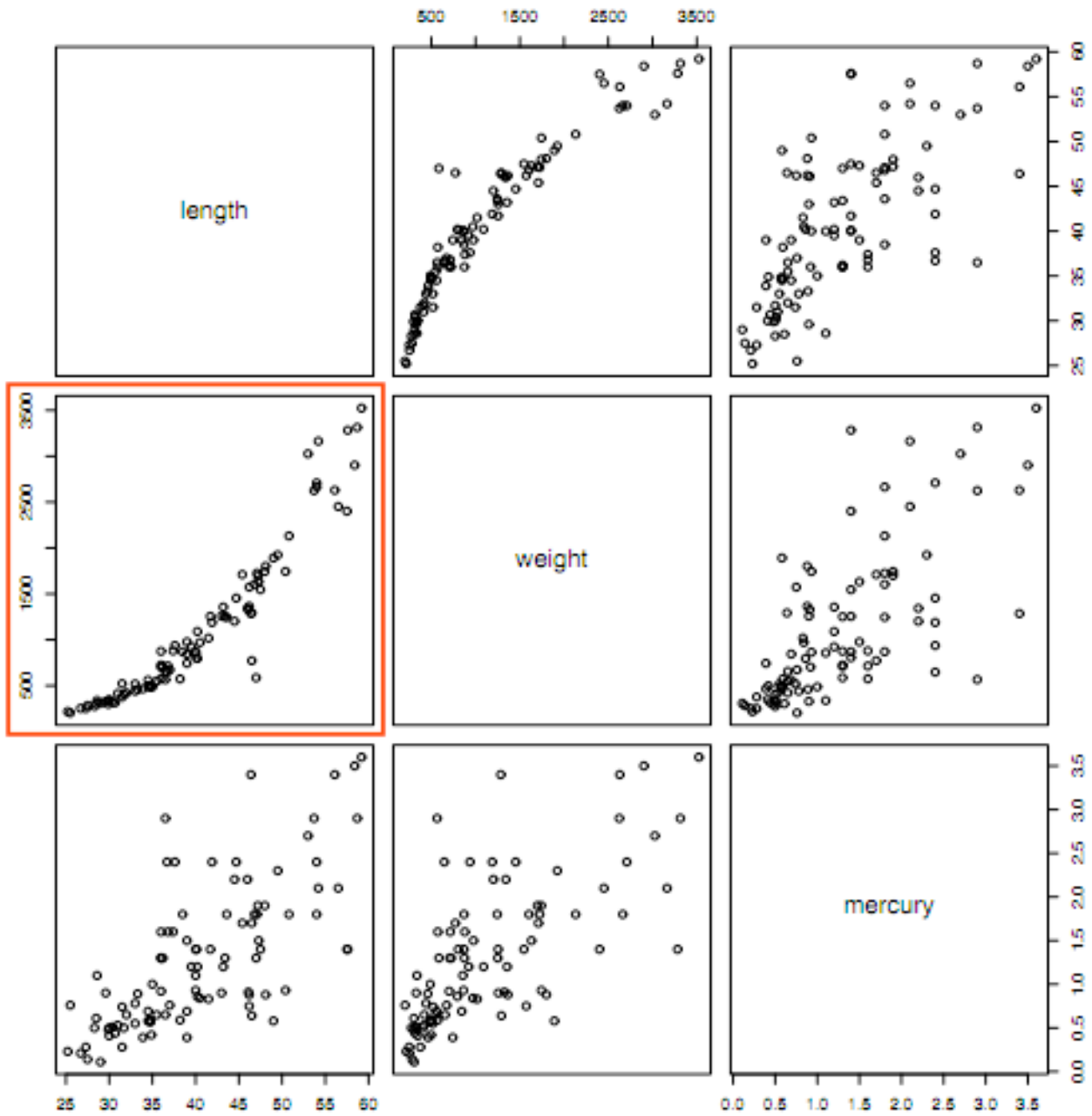
We had a look at the R commands to perform the fit and even assessed the sampling variability using the bootstrap; we saw that the computationally intensive approach gave results that were close to what the classical formula told us (based on the t-distribution)

There are two dangling ideas I wanted to follow up on before we branched out into a different topic...

Back to the river

We're going to consider a second pair of variables now; it will introduce a small complication that will provide us with a richer view of regression analysis

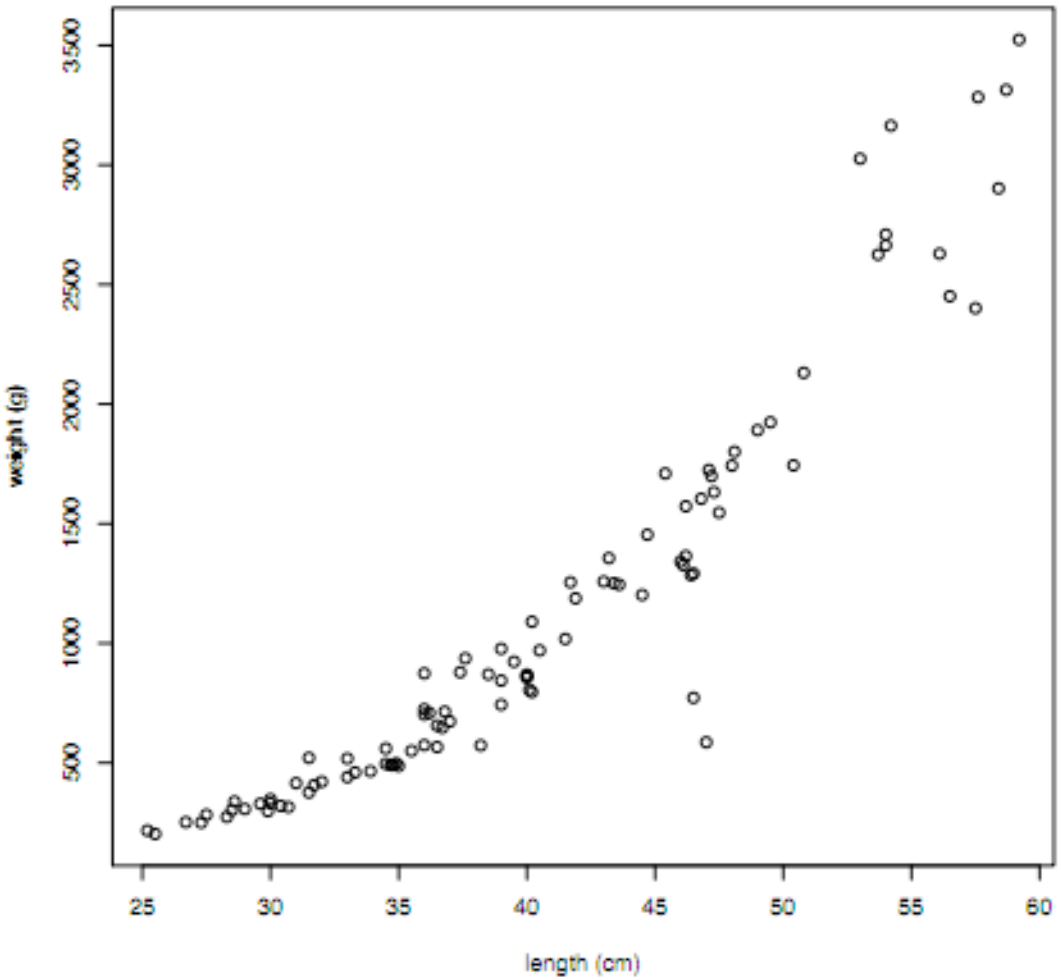
Let's have a look at the relationship between fish length and fish weight; recall our scatterplot matrix...

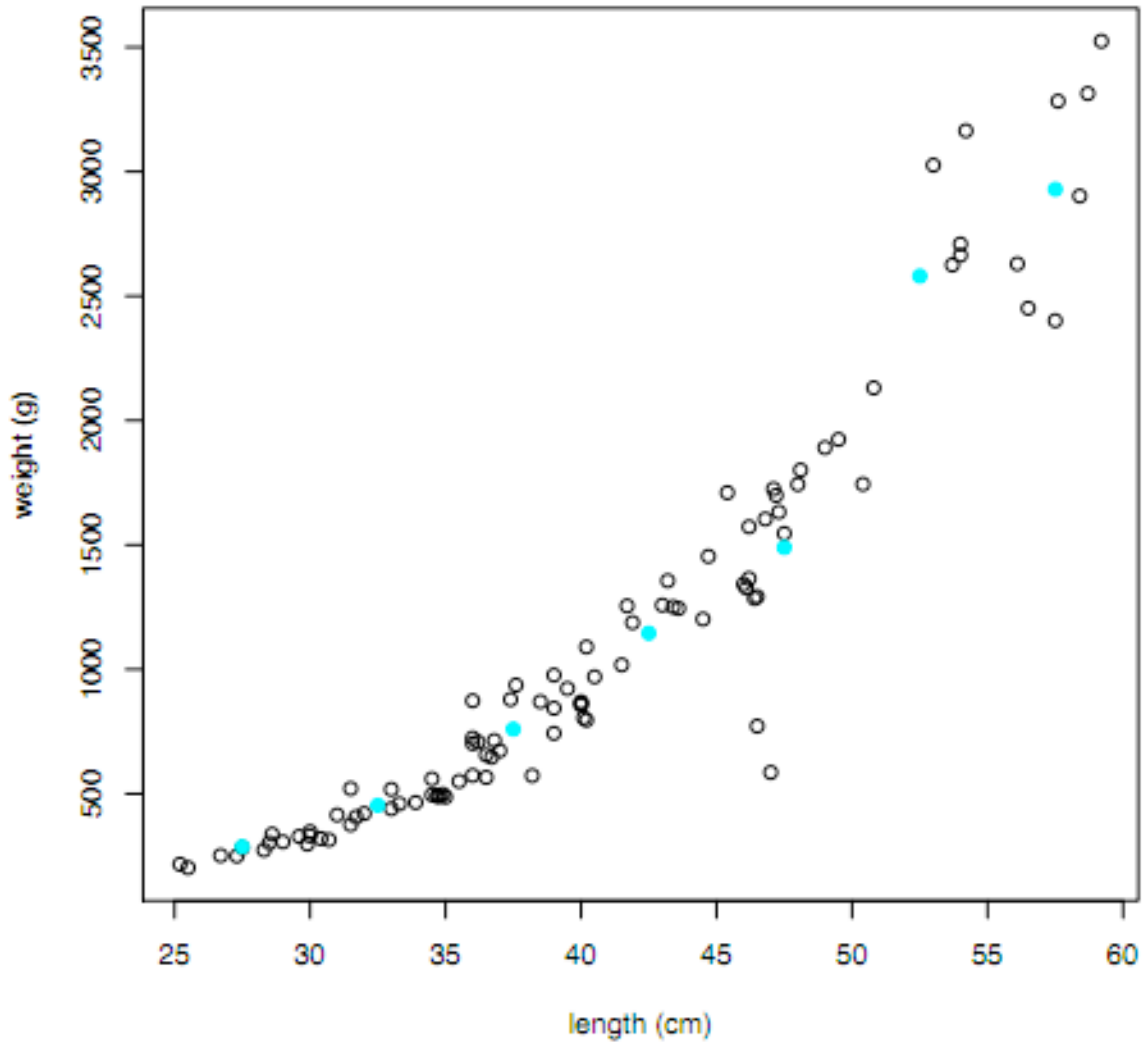

length
weight
mercury
500
1500
2500
3500
25
30
35
40
45
50
55
60
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

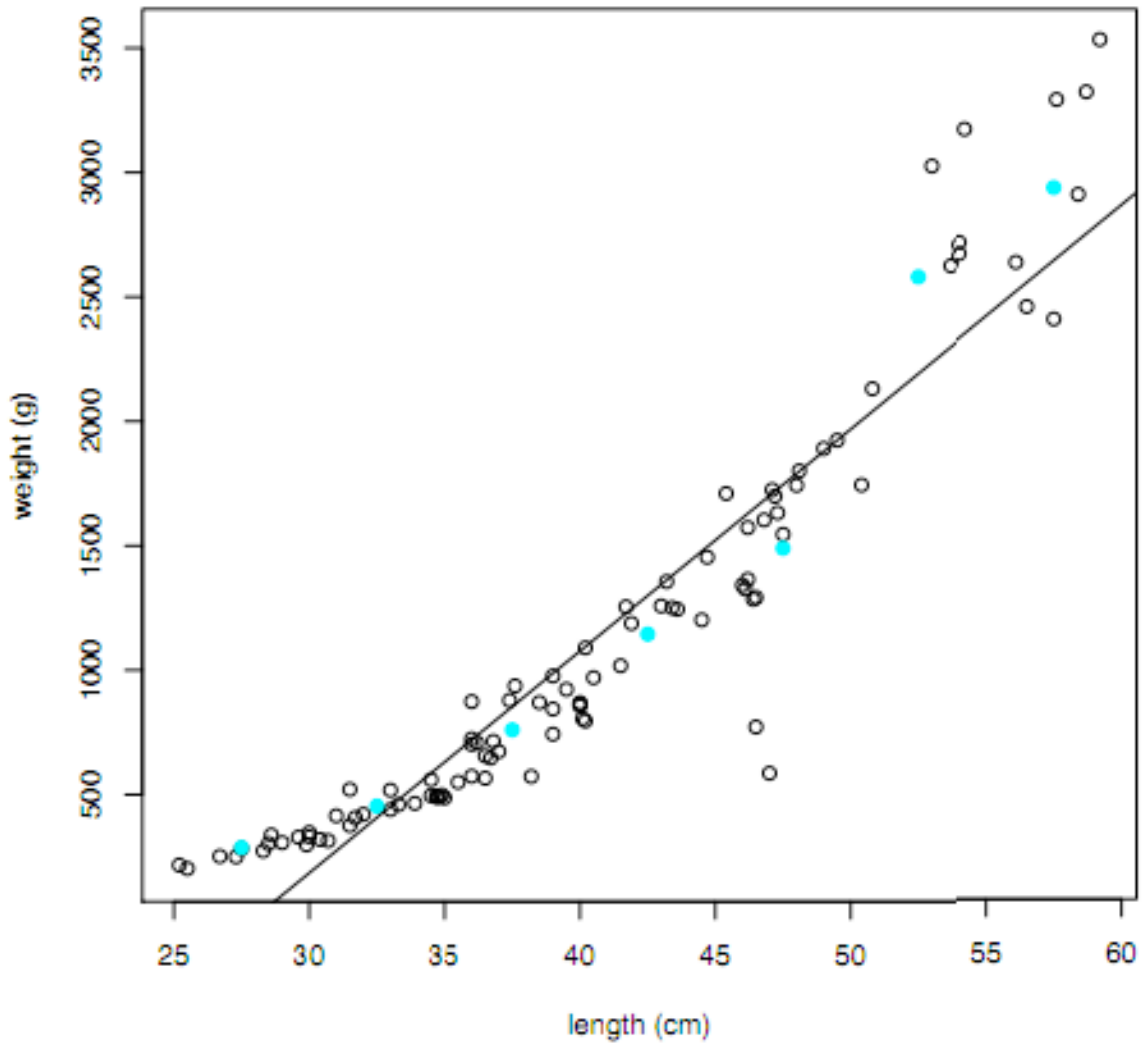
Weight

Either recalling Galton's mental image of data or our somewhat more practical application of least squares, **what do you think of this relationship?**

How would we model it mathematically? Will a line work?




weight (g)
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length (cm)


weight (g)
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length (cm)

Assessing the fit

When examining a regression model, there are several things to consider

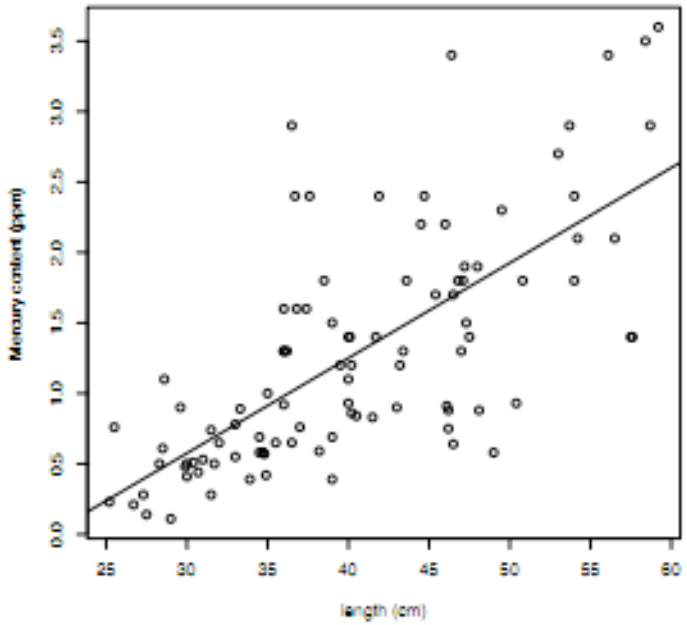
1. Is the relationship between outcomes and predictors linear?
2. Is the error variance constant?
3. Do the errors look roughly normally distributed?

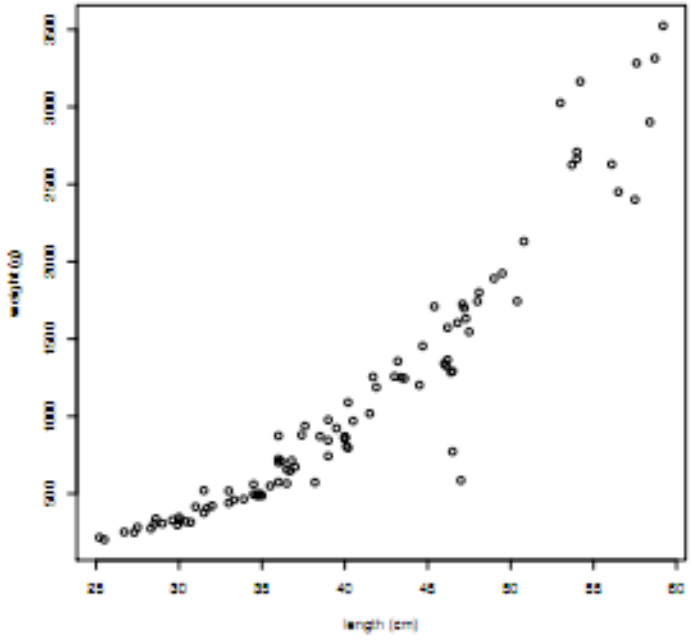
Regression analysis

For a simple linear regression with just one predictor variable, we can make scatterplots to assess the relationship between inputs and outputs easily

Assuming a linear relationship, then the errors from our least squares procedure should look like a sample from the normal distribution

This suggests other plots...

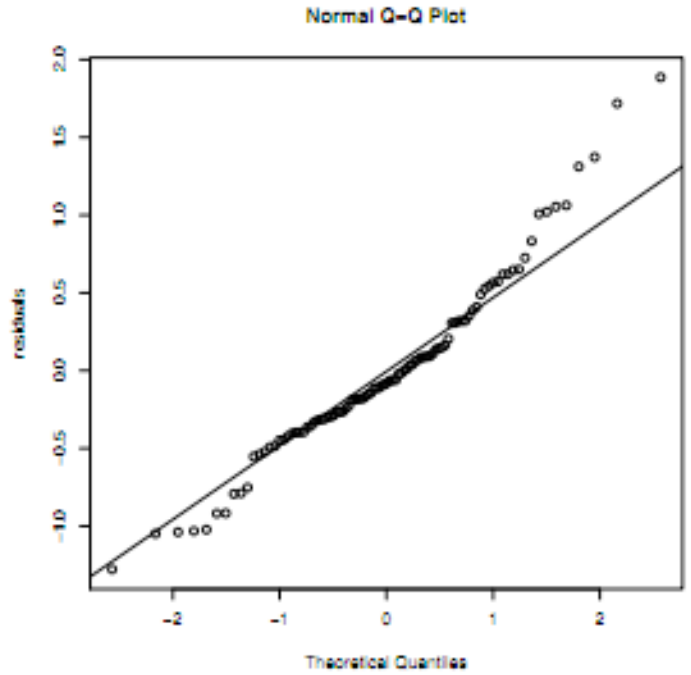


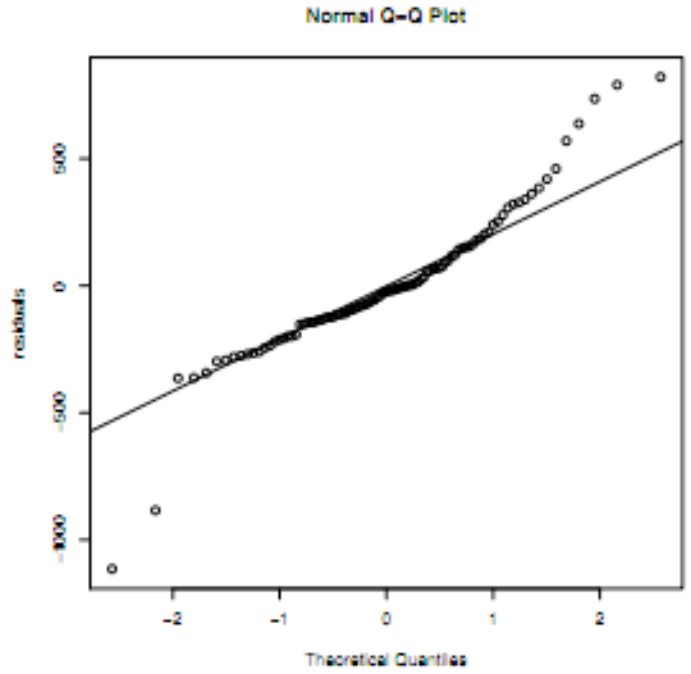


Residual analysis

Again, we want to inspect the residuals for “bends” which would indicate departures from normality

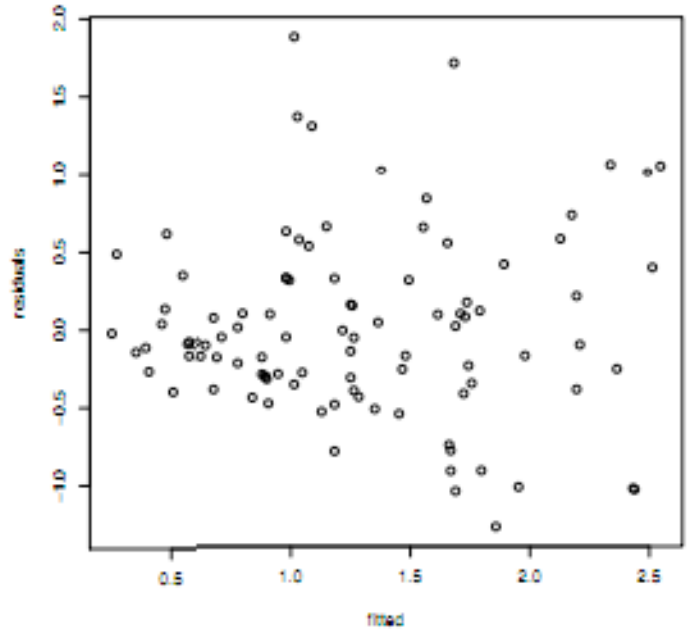
We should also examine the plots for large (positive or negative) values that might indicate outlying points





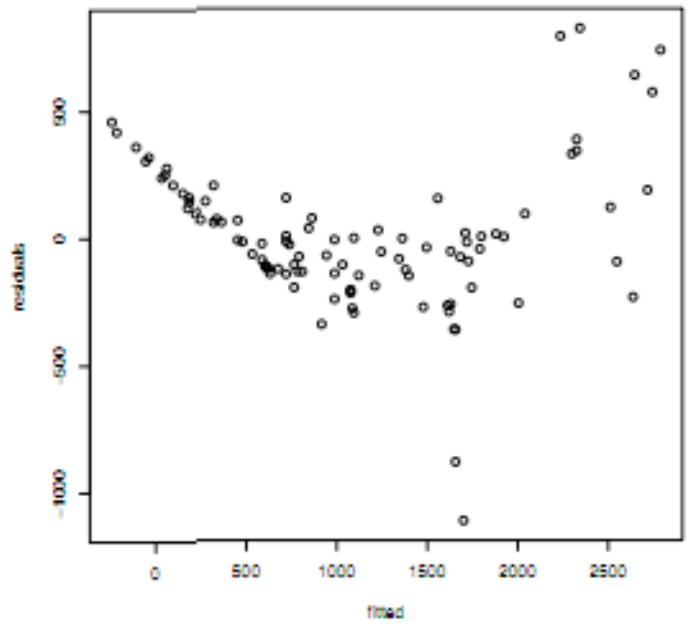
Residual analysis

Another informative plot compares the residuals to the fitted values (the points on the line); ideally **we should see no pattern here** - remember our model assumes the errors are independent normal observations



For the Mercury regression we see a slight indication of changing variability -- **for short fish we see less variation in the residuals than the long fish**

For the weight regression, we see that the fitted model **consistently overestimates (positive errors) the weights of small and large fish**, giving the plot a U-shape and suggesting a problem with the model



Polynomials

The relationship between weight and length is not linear and our basic inferential model breaks down when this assumption is violated (we no longer have just random errors from our model, but also considerable bias from the structural components we've left out)

Often, we consider fitting low-degree polynomials instead of just a line; on the next few slides, we go from a linear fit to a cubic -- the R command `poly()` returns a polynomial with the indicated degree

Fitting a line

Below we provide the code to fit a line using the poly() function; this allows us to go from degree 1 to 2 to 3 easily; what do you see?

```
> fit = lm(weight~poly(length,1),data=waccamaw)
> summary(fit)

Call:
lm(formula = weight ~ poly(length, 1), data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-1114.82  -141.65   -22.94   136.13   821.17

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1111.22      28.86   38.50   <2e-16 ***
poly(length, 1)  7625.26     285.70   26.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 285.7 on 96 degrees of freedom
Multiple R-squared: 0.8812,   Adjusted R-squared:  0.88
F-statistic: 712.3 on 1 and 96 DF,  p-value: < 2.2e-16
```

Fitting a quadratic

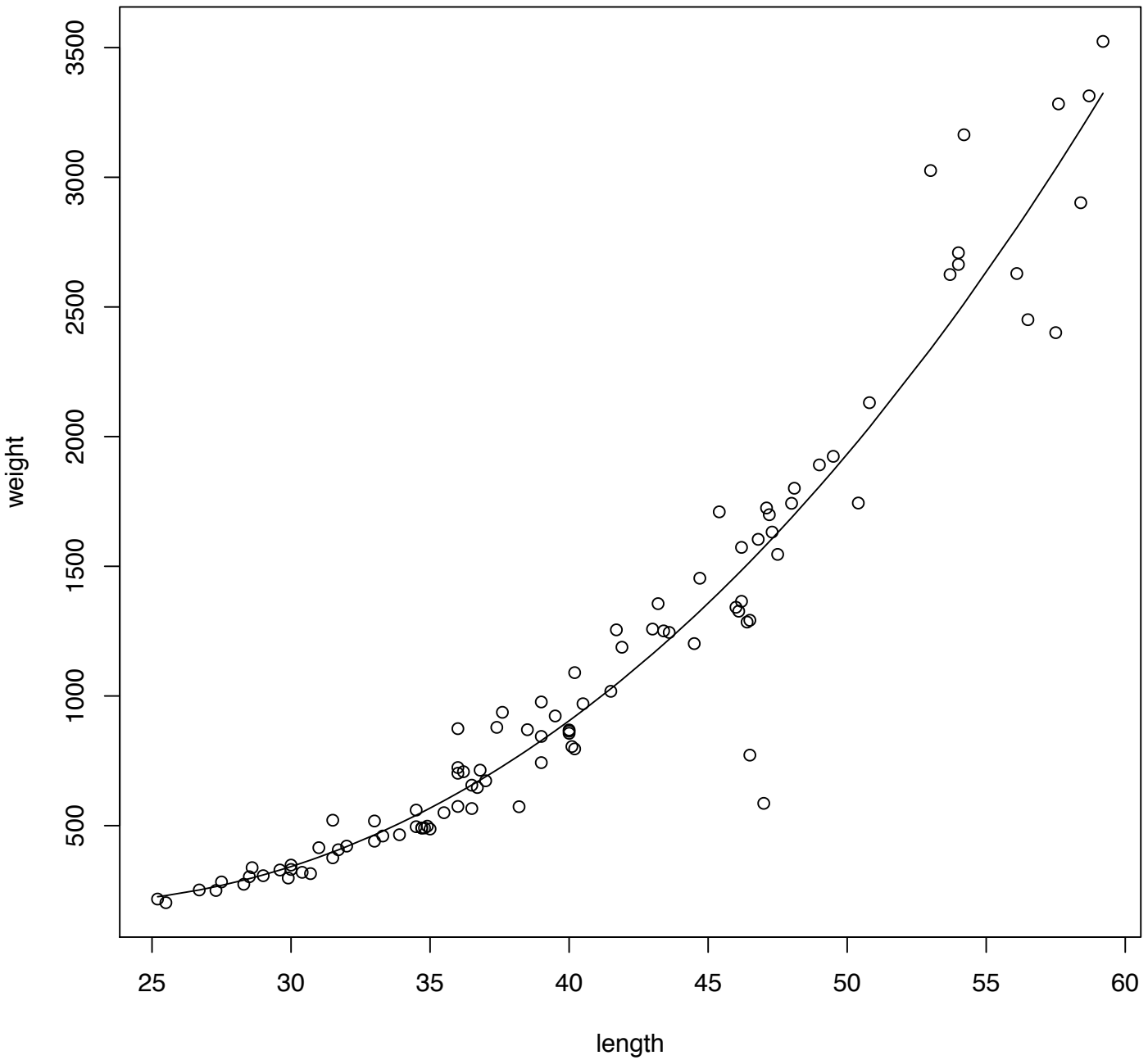
Below we provide the code to fit a quadratic (including both length and $length^2$ in the model) relationship between weight and length; what do you see?

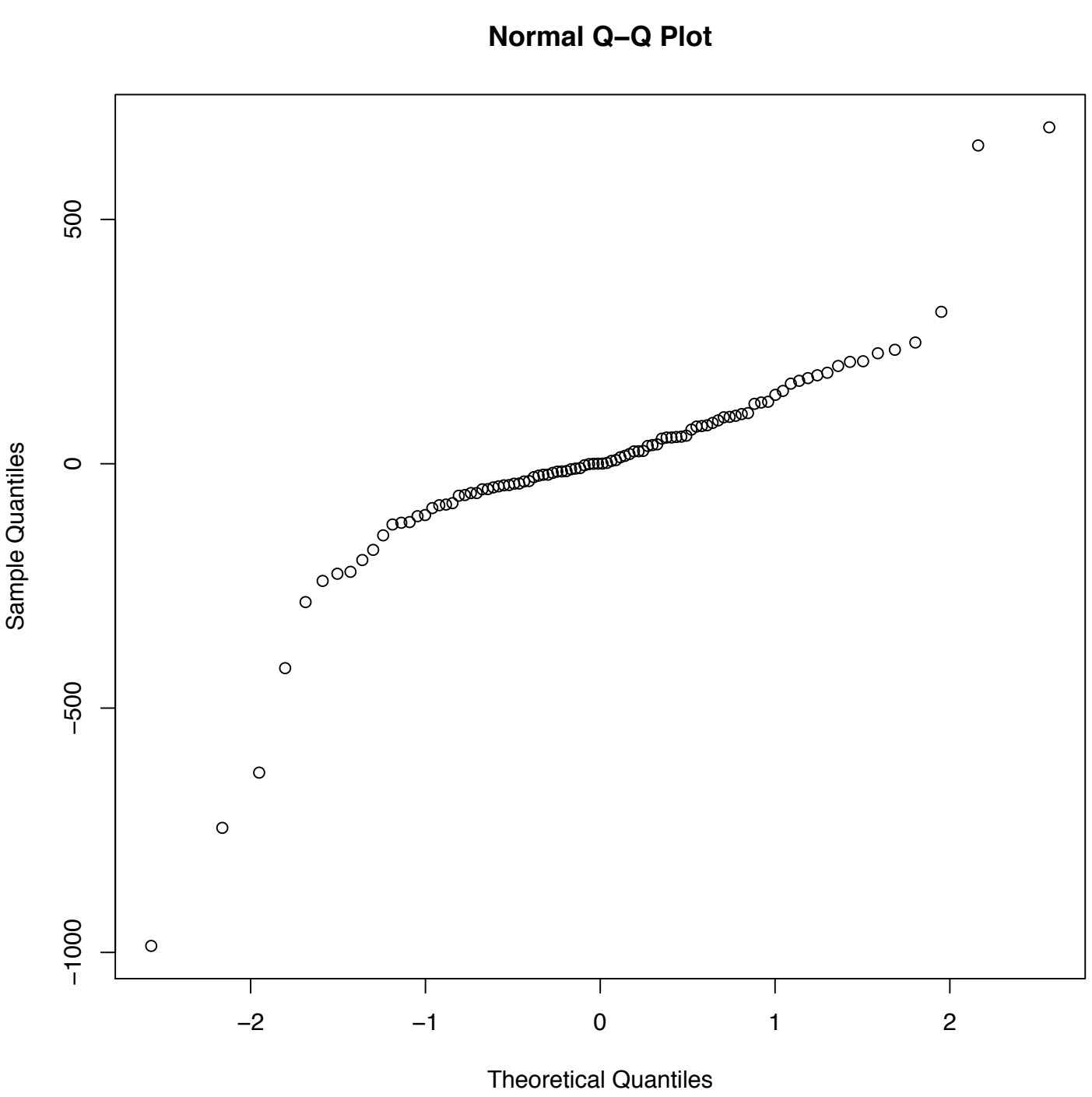
```
> names(waccamaw)
[1] "river"   "station" "length"  "weight"  "mercury"

> fit = lm(weight-poly(length,2),data=waccamaw)

> summary(fit)

Call:
lm(formula = weight - poly(length, 2), data = waccamaw)

Residuals:
     Min       1Q   Median       3Q      Max
-992.013  -49.733    3.498   87.098  684.114

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       1111.22      21.27   52.24  < 2e-16 ***
poly(length, 2)1  7625.26     210.58   36.21  < 2e-16 ***
poly(length, 2)2  1903.54     210.58    9.04 1.87e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 210.6 on 95 degrees of freedom
Multiple R-squared: 0.9362,   Adjusted R-squared: 0.9348
F-statistic: 696.5 on 2 and 95 DF,  p-value: < 2.2e-16
```

Fitting a cubic

Below we provide the code to fit a cubic (including both length, $length^2$ and $length^3$ in the model) relationship between weight and length; w you see?

```
> fit = lm(weight-poly(length,3),data=waccamaw)
> summary(fit)

Call:
lm(formula = weight - poly(length, 3), data = waccamaw)

Residuals:
      Min        1Q    Median        3Q       Max
-986.6574  -52.1110    0.3027   87.4199  688.4783

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1111.22      21.38  51.978  < 2e-16 ***
poly(length, 3)1   7625.26     211.64  36.030  < 2e-16 ***
poly(length, 3)2   1903.54     211.64   8.994 2.53e-14 ***
poly(length, 3)3     48.03     211.64   0.227    0.821
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 211.6 on 94 degrees of freedom
Multiple R-squared: 0.9362,   Adjusted R-squared: 0.9342
F-statistic: 459.7 on 3 and 94 DF,  p-value: < 2.2e-16
```


weight
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
500
0
−500
−1000
−2
−1
0
1
2

Where to?

There are numerous open questions

1. How do you know you've included the most informative variables?
2. How do you decide on their functional form?
3. What do you do about missing data?
4. How do you handle changing variances?
5. What if my data are not normally distributed but are discrete counts or even binary?

A new data set

For the last couple of summers, I have done work related to climate change and the differential impacts that the accompanying extreme climate events have globally

Broadly, those least responsible for climate change are most impacted -- Several organizations have tried to quantify this statement, creating indices of various kinds to order countries according to their vulnerability to climatic events

Call and response

To get started on this topic, I sent out a handful of emails to try to acquire some of the base vulnerability data from researchers publishing indices -- Here's a typical note, with identifying references stripped

This is "the call"...

hi

i'm a professor in the statistics department at ucla... i'm writing to see if i might be able to get a copy of the data you assembled for your XXXXXX article on vulnerability to extreme climate events. i was specifically looking for the data referred to for your first regression in your XXXXXX section.

is that possible?

thanks!

M.

Mark Hansen | www.stat.ucla.edu/~cocteau

Call and response

In a graduate class I teach, we discussed how much of computational science is not particularly reproducible -- Two of the emails I received tell the story nicely...

Mark,

Sure. It has been a little while, and I am not 100% sure which files I used, but I am 99% sure that it was the sheet "late with population" in the attached excel spreadsheet, which ought to match what is in the STATA file. If this doesn't seem right, let me know, and I can spend a few more minutes hunting for the right data.

Cheers,

XXXXX

Dear Mark

This will take some digging through old files and backup disks as this was some time ago and my filing system is probably not what it should be. However, I'll try and find some time to do the necessary excavation. The data we used were all publicly available, and you should be able to reproduce the analysis based on the description of the methodology in the paper, if that's your interest. In any case feel free to give me a nudge in a week or so. When do you need to have the data, given your teaching schedule?

It would be good to have a professional statistician cast an eye over this, and the data we used were up to 2000, and could do with being updated. In any case I'm sure you'll find much to criticise! As someone who finds themselves having to deal with vulnerability indices I'm very sceptical of them, even the ones I've produced myself.

All the best

XXXXXX

Call and response

What's beautiful here is that you find two very different ways of working -- In one, the researcher is able to produce data almost immediately that (while not exactly right) got us really close to the published results

In the other case, there's a certain amount of hunting that has to take place -- I don't mean to fault this researcher in any way, I simply wanted to make the case that we should strive to be more like the first researcher (and better)

It's also worth noting the reception that statisticians get if we're not careful -- Writing to researchers outside of statistics often elicits a kind of fear that we're going to check up on them or criticize their work in some way

My (unwanted and highly biased) advice to you is to be the kind of statistician that tries to help!

PNAS PNAS PNAS

Estimating least-developed countries' vulnerability to climate-related extreme events over the next 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwabena Asante[d], Marc Metzger[e,f], Jose Rafael[g], Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]Climate Systems Analysis Group, University of Cape Town, Rondebosch 7701, South Africa; [c]Institute of Environmental Science and Technology, Autonomous University of Barcelona, 08193 Bellaterra, Spain; [d]Climatus LLC, Mountain View, CA 94041; [e]Centre for the Study of Environmental Change and Sustainability, University of Edinburgh, EH8 9XP, Scotland; [f]Alterra, Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Department of Geography, University of Eduardo Mondlane, Maputo, Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences, 1040 Vienna, Austria; and [i]Department of Oceanography, University of Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approved December 4, 2009 (received for review September 10, 2009)

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may

sensitivity to those stressors, which in turn is determined by a complex set of social, economic, and institutional factors collectively described as determining its adaptive capacity (5, 6). As the UNFCCC secretariat suggested in its needs assessment, "one of the key limitations in estimating the costs of adaptation is the uncertainty about adaptive capacity. Adaptive capacity is essentially the ability to adapt to stresses such as climate change. It does not predict what adaptations will happen, but gives an indication of differing capacities of societies to adapt *on their own* to climate change or other stresses" (1, p. 97).

Human losses to extreme weather events can serve as a reliable indicator for this vulnerability, and with it the need for financial assistance, for two reasons. First, measures to reduce vulnerability to extreme weather events account for a particularly large share of estimated adaptation financial needs (1). Second, in the context of efforts to achieve a wide range of development goals, it is only

Vulnerability

The underlying question here is interesting and relevant (they usually are, for what it's worth) -- Here we are interested in understanding how climate change (and the accompanying increase in extreme weather events) will affect different parts of the world

Specifically, the researchers produce a model that relates variables capturing some notion of vulnerability to the impacts that weather-related natural disasters have had, country by country

Estimating least-developed to climate-related extreme 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwa
Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]
South Africa; [c]Institute of Environmental Science and Technology, Autonomous
View, CA 94041; [d]Centre for the Study of Environmental Change and Sustainab
University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Depa
Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences
Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approv

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may begin to offset rising climate exposure in the second quarter of the century, and that it is in the period between now and then that vulnerability will rise most quickly. This implies an urgency to the need for international assistance to finance adaptation.

vulnerability | adaptive capacity | development | natural disasters | natural hazards

Results

The first stage of our analysis was to estimate statistical models of losses from climate-related disasters, based on a set of climatic and socio-economic variables that will likely change over time, which appear in Table 1. The dependent variables are logged values of the number of people per million of national population killed or affected, respectively, by droughts, floods, or storms over the period 1990–2007. The variable number of disasters is the logged value of numbers reported by each country over the same period, and accounts for climate exposure; estimated coefficient values greater than 1 in both models indicate that average losses per disaster are higher in more disaster-prone countries. We expected that larger countries are likely to experience disasters over a smaller proportion of their territory or population, and also benefit from potential economies of scale in their disaster management infrastructure, both resulting in lower average per capita losses; the negative coefficient estimates for the variable national population in both models are consistent with this expectation. The variable HDI represents the Human Development Index, a United Nations (UN) indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth. Recent studies of disaster losses —not limited to climate-related events—have shown that countries with medium HDI values experience the highest average losses, whereas countries with high HDI values experience the lowest (14, 15). We therefore included the logged HDI values in quadratic form. Negative coefficient estimates for both HDI and HDI^2 in both models are thus consistent with these expectations, given that logged HDI values are always negative, and the square of the logged values are in turn positive. Finally, we considered several additional socio-economic variables not directly captured by HDI, and found only two that improved model fit. For the model of the number of people killed, the positive coefficient estimate for female fertility indicates that countries with higher birth rates experience greater average numbers of deaths. We do not take this to mean that there is a direct connection between fertility and natural hazard deaths, but rather that higher birth rates are associated with lower female empowerment, and lower female empowerment is associated with higher disaster vulnerability, as has been shown previously (16, 17). For the model of the number of people affected, the negative coefficient estimate for the proportion urban population is consistent with urban residents being less likely to require post-disaster assistance than rural residents, also observed previously (18, 19). Both models yield an R^2 statistic slightly greater than 0.5, indicating that variance in the independent variables explains just over half of the variance in the numbers killed and affected. This is consistent with results from past analyses based on similar data and methods (8–10).

Vulnerability

In the end, a great deal of attention is paid to a regression table (below), the form of which we should be fairly familiar with

In each row they present the regression of the logarithm of the number of people killed by weather-related natural disasters from 1990 to 2007 as a function of several predictors, one of which is slightly special...

Table 1. Ordinary least-squares regression results

Independent variables	Killed	Affected
Number of disasters	1.36* (0.15)	1.88* (0.19)
National population	−0.56* (0.09)	−0.79* (0.11)
HDI	−5.97* (1.95)	−13.55* (2.16)
HDI^2	−6.26* (1.52)	−9.82* (1.86)
Female fertility	1.45* (0.43)	
Proportion urban population		−0.41 (0.37)
Constant	−3.86* (0.49)	5.33* (1.71)
Number of observations	150	154
R^2	0.52	0.55

The dependent variable in the Killed model is the logged value of the number of people reported by CRED as killed by the three types of disasters considered (droughts, floods, and storms) divided by population. The dependent variable in the *Affected* model is the same for the number of people reported affected, but not killed, by the same disasters. All independent variables are logged values. Because HDI occupies the range of 0–1, all logged HDI values were negative, whereas the squares of these values were positive. *Values significant (*two-tailed student's t* test) at the 99% confidence level. Values in parentheses are SEs.

HDI

The HDI or Human Development Index, a United Nations (UN) is an “indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth”

From the table on the previous page, we see that the variable and its square are both included in the final model and are given the following interpretation

“Of particular importance to rapidly developing countries is the observed nonlinear relationship between HDI and disaster losses. Fig. 1 illustrates the magnitude of this effect in both models, compared with the background variance, and taking into account the effects of the other variables. The estimated regression curve in Fig. 1A suggests that the risk of being affected by a climate disaster is highest in countries with HDI values of ~0.5, whereas the curve in Fig. 1B suggests that the highest risk level is for countries with HDI values somewhat higher, ~0.6. This suggests that for countries with HDI values of less than 0.5, the transition to higher levels of development could potentially, in the absence of targeted intervention, exacerbate vulnerability.”

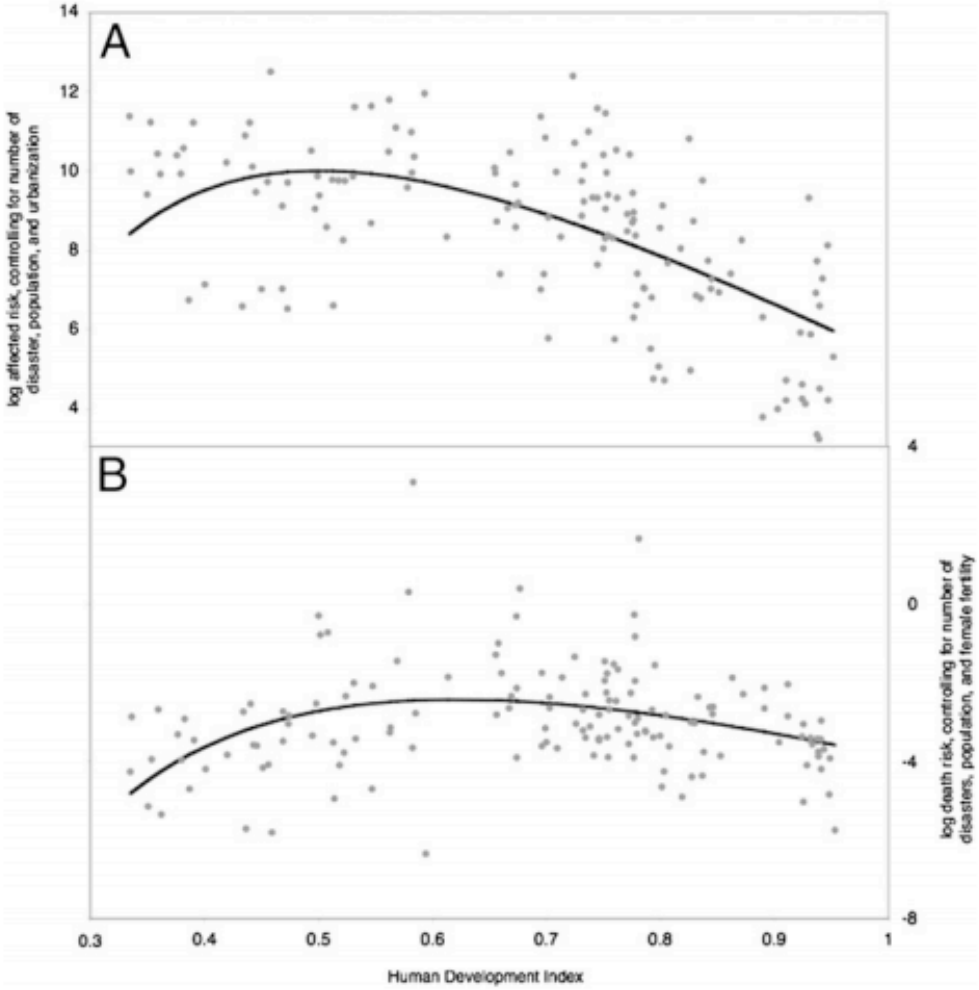


Fig. 1. Relationship between risk and HDI for (*A*) the number of people affected, i.e., needing emergency or recovery assistance, by a flood, drought, or cyclone, per million of population, and (*B*) the number of people killed. Each dot represents a country in the CRED database during the period 1990–2007, with its position on the vertical scale being the logarithm of the annual value per million population, after subtracting the predicted influence of other risk factors. Regression line in each figure shows predicted values including the influence of HDI.

Looking at the data

The data we were given consist of measurements associated with 144 different countries -- For each we have the following variables

`country_name` the name of the country

`ln_events` the natural logarithm of the number of droughts, floods and storms occurring in the country from 1990-2007

`ln_pop` the natural logarithm of the country's population

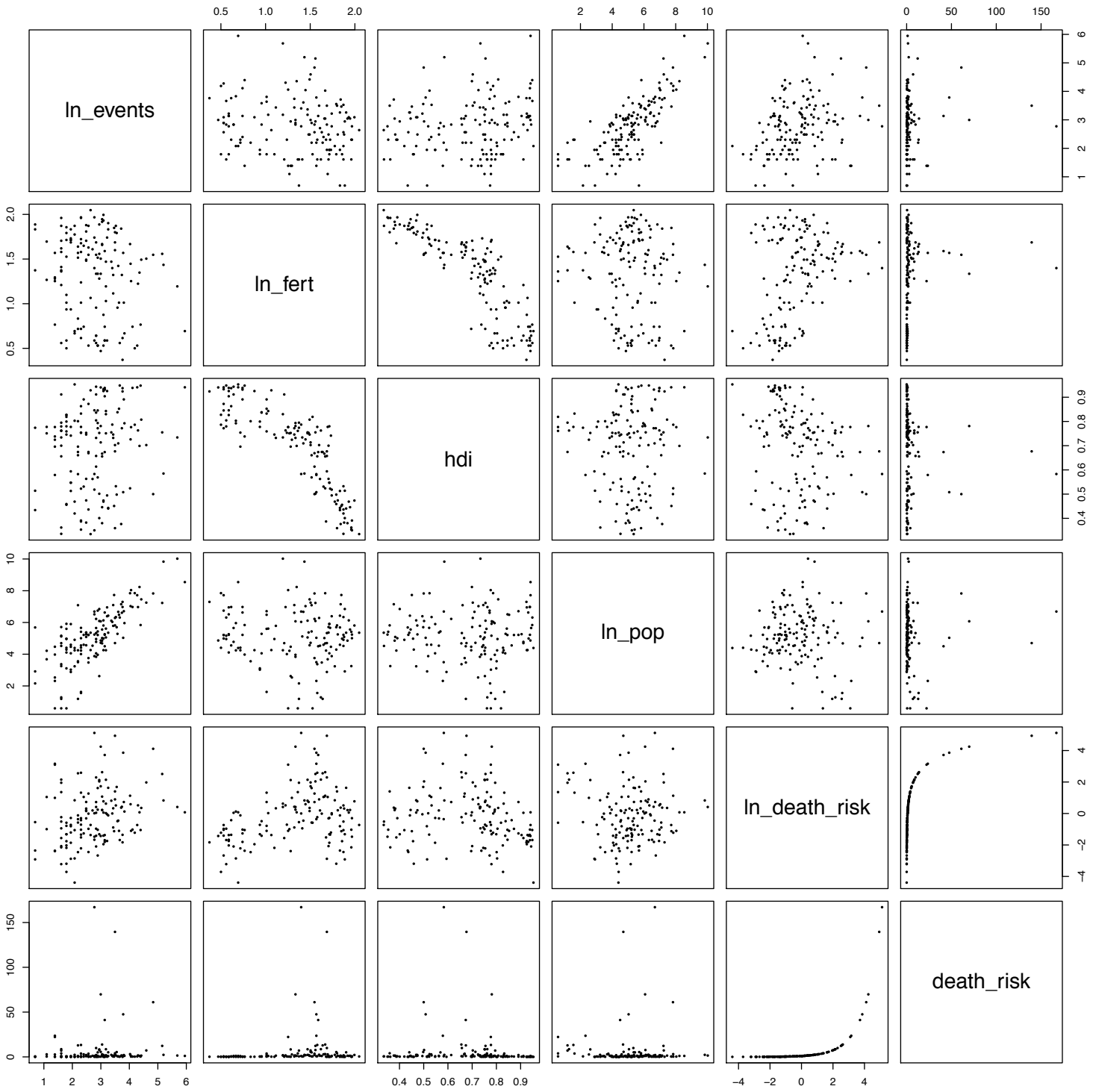
`ln_fert` the natural logarithm of an estimate of the country's female fertility

`hdi` the Human Development Index for the country

`death_risk` the proportion of people out of 1M in population killed in droughts, floods and storms

There are four predictor variables (if you count HDI and its square as one) which, while not big by any stretch of the imagination, is complex enough to keep us from "seeing" the whole data set

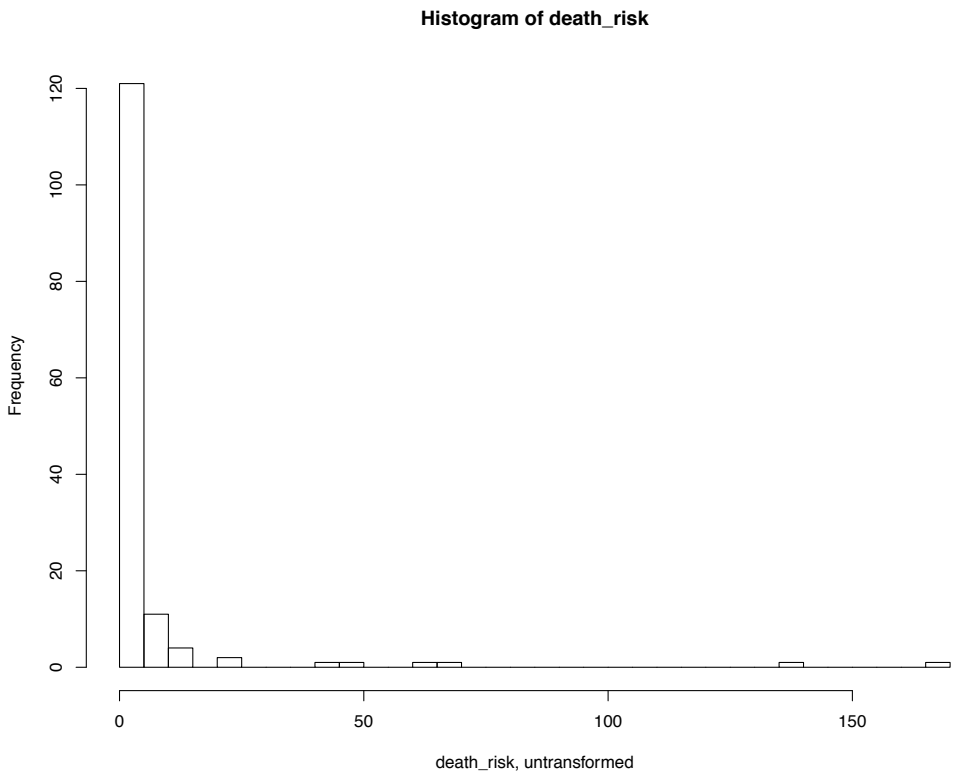
Instead, we might opt for partial views...

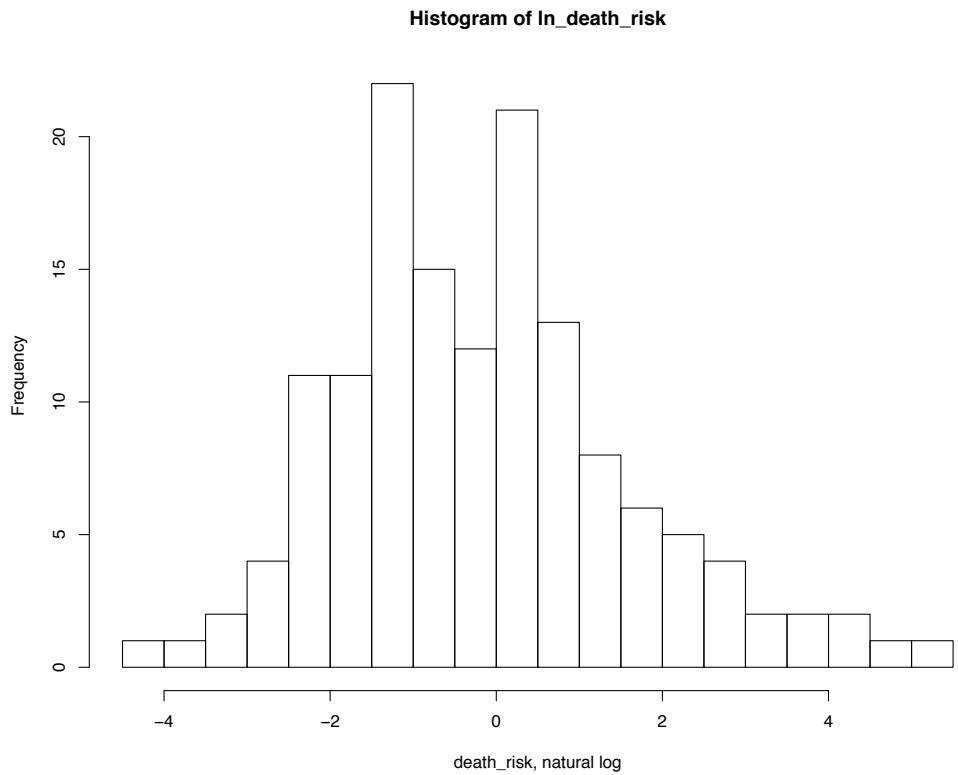

ln_events
ln_fert
hdi
ln_pop
ln_death_risk
death_risk

A first model

Rather than repeat the author's final analysis, we might usefully question the rationale for taking each step -- Let's start with the response variable and the decision to take a logarithm

If we look at the response itself, we see that it's quite skewed...


Histogram of death_risk
Frequency
0
20
40
60
80
100
120
0
50
100
150
death_risk, untransformed


Histogram of ln_death_risk
Frequency
0
5
10
15
20
−4
−2
0
2
4
death_risk, natural log