Lecture 13

Last time

You were introduced to simple linear regression -- We approached the topic largely historically, tracing the initial technique back to Francis Galton in the late 1880s and his studies on inheritance

As we noted, the regression framework came primarily from Galton and we see hints of his early work even in more modern extensions to the topic

Some history

Galton collected data from 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 "transmuted" the heights of the girls and women in his data set, multiplying these heights by 1.08 and then forms a table of the heights of children versus the heights of their mid-parents (the average height of the father and transmuted mother)

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).

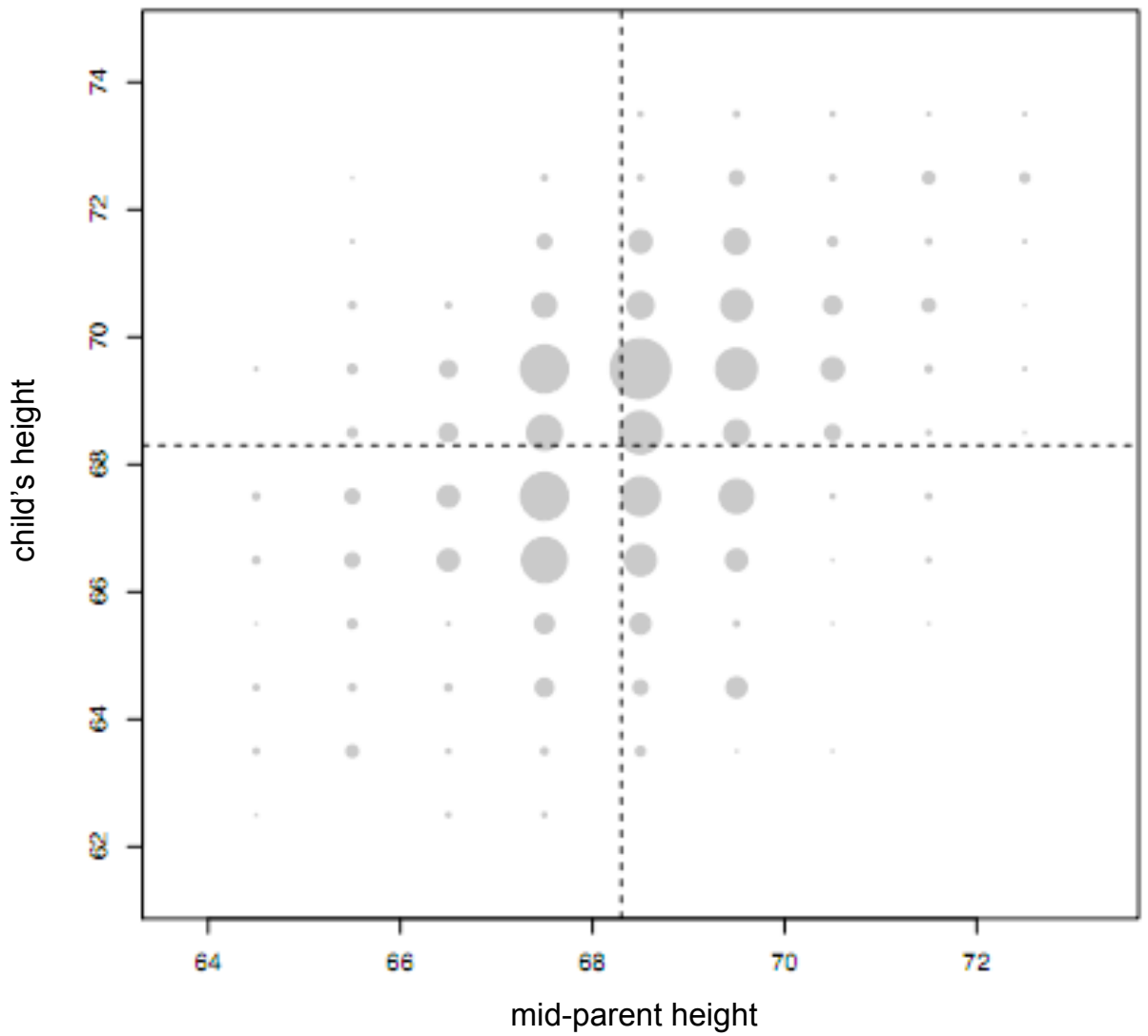
Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	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.

Heights

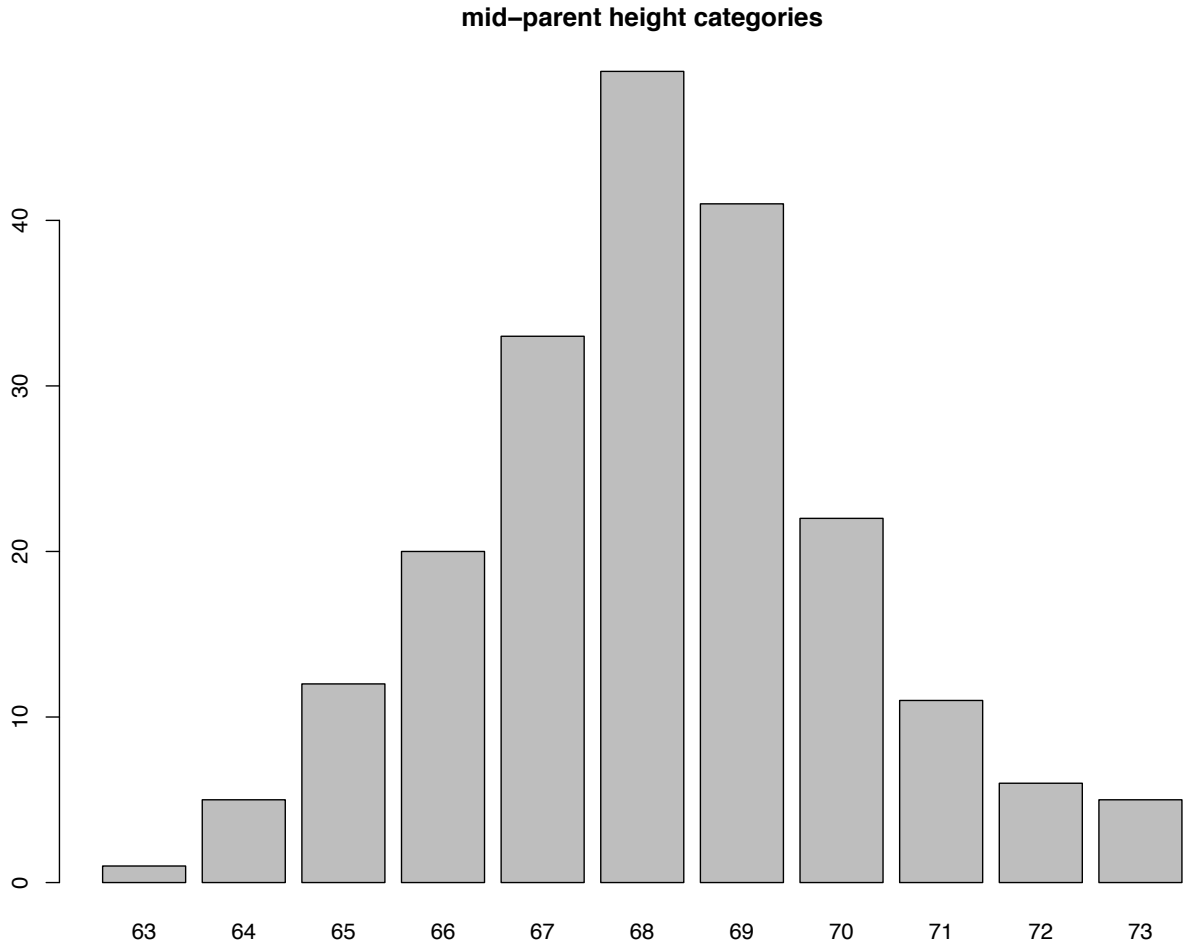
On the next slide we present another view of this table -- Here the different cells in the table are represented by circles that are sized according to the counts in each cell

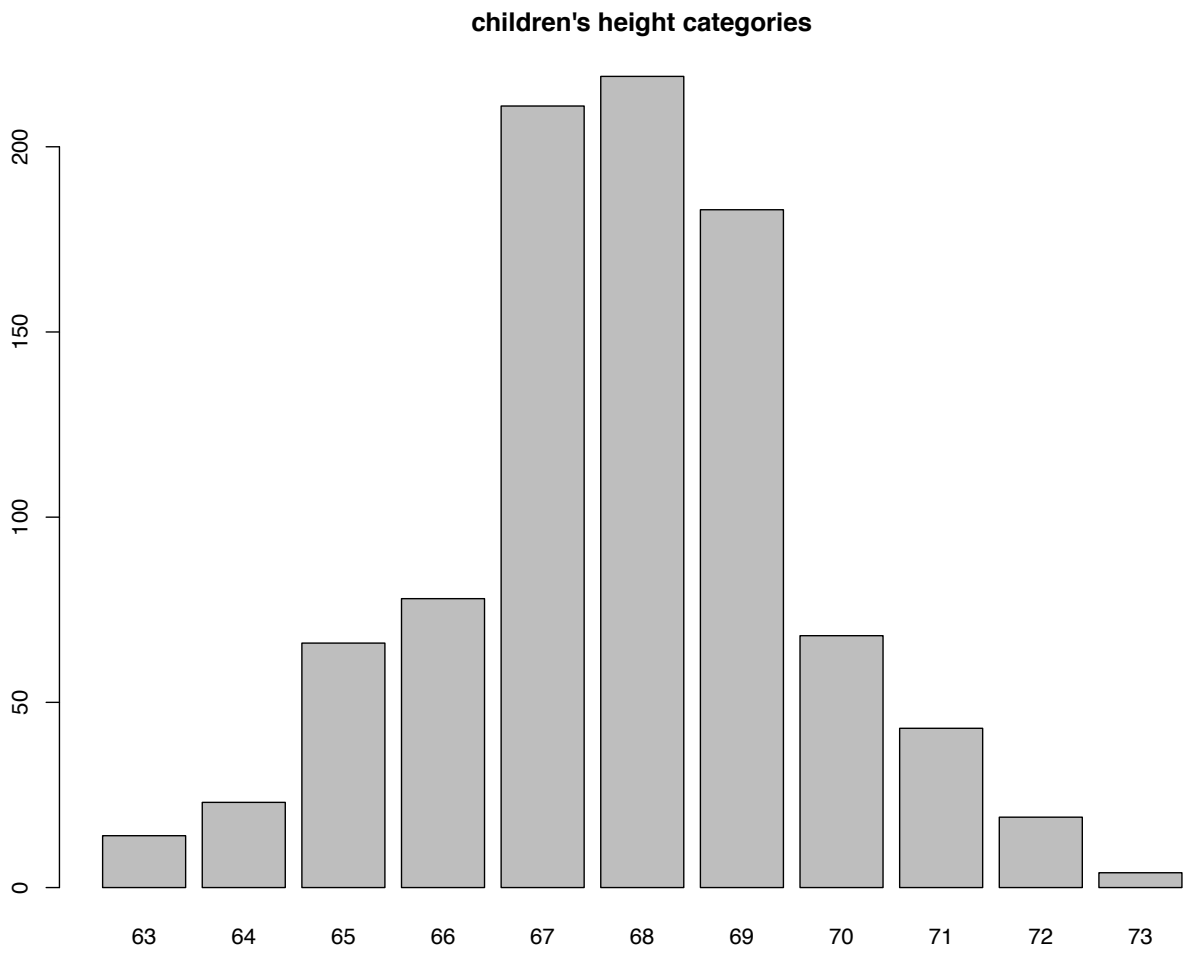
The dashed lines mark the means of the mid-parents' and (transmuted) children's heights (both about 68.3 inches) -- What does this display and the table suggest about the "data," the paired values of child and mid-parent heights?


child's height
mid-parent height
62
64
66
68
70
72
74
64
66
68
70
72

Heights

In addition to the elliptical look of the data distribution, the “marginal” height distributions (say, the distribution of mid-parent heights in the study considered on their own) also look normal...

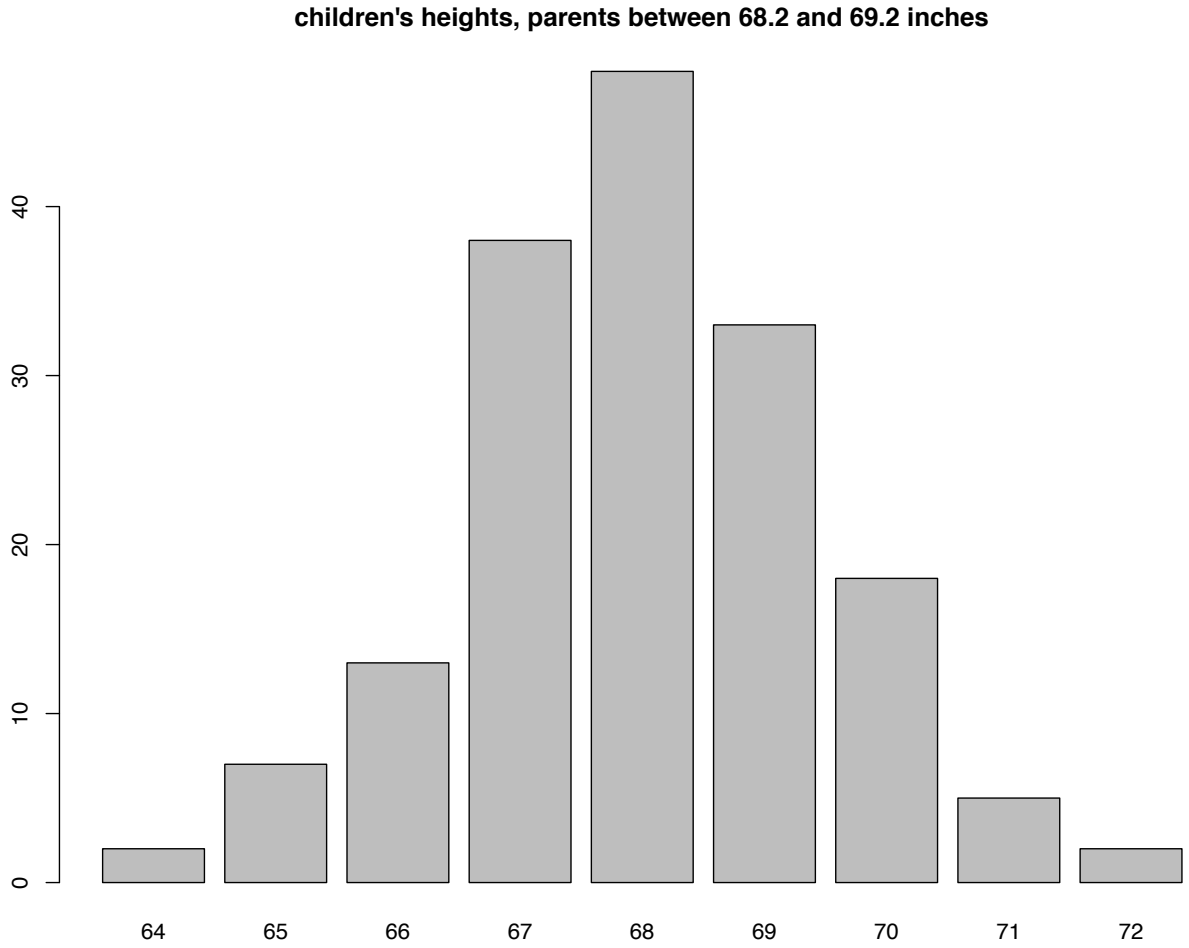

mid−parent height categories
0
10
20
30
40
63
64
65
66
67
68
69
70
71
72
73


children's height categories
0
50
100
150
200
63
64
65
66
67
68
69
70
71
72
73

Heights

He went on to further notice that in each of his separate categories of mid-parent heights, the (transmuted) children's heights also had normal distributions -- That is, if you look at just one of the columns of his table, you see again a normal distribution

Here's one example...


children's heights, parents between 68.2 and 69.2 inches
0
10
20
30
40
64
65
66
67
68
69
70
71
72

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**

FIG .7.

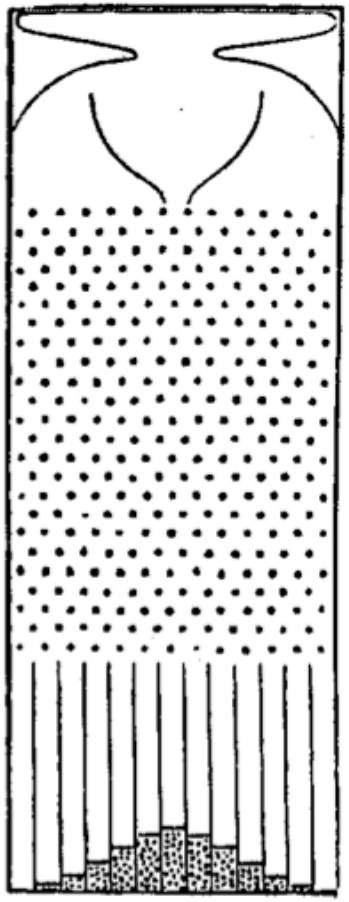


FIG .8.

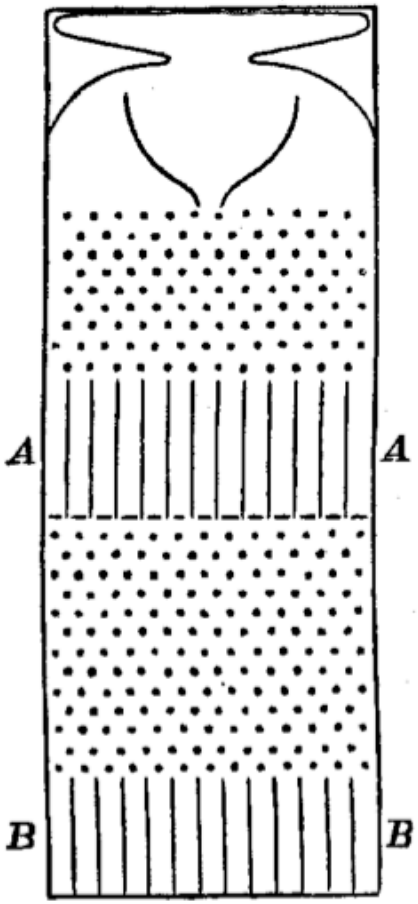
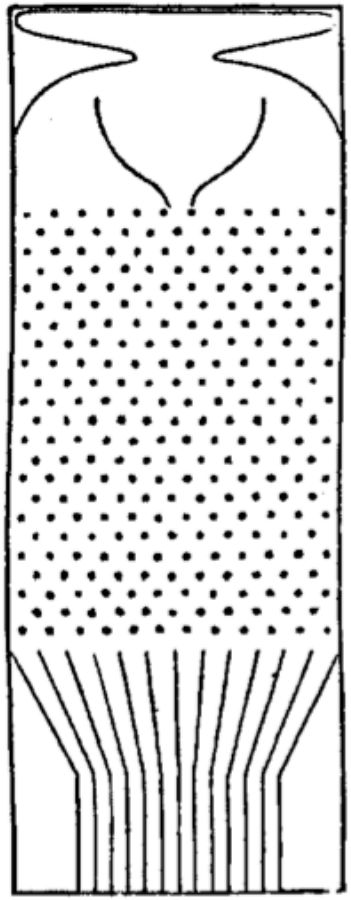


FIG .9.



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

In effect, he established that a normal mixture of normals is itself normal -- But with this idea in hand, **we see his tables of human measurements in a different light**...

FIG .7.

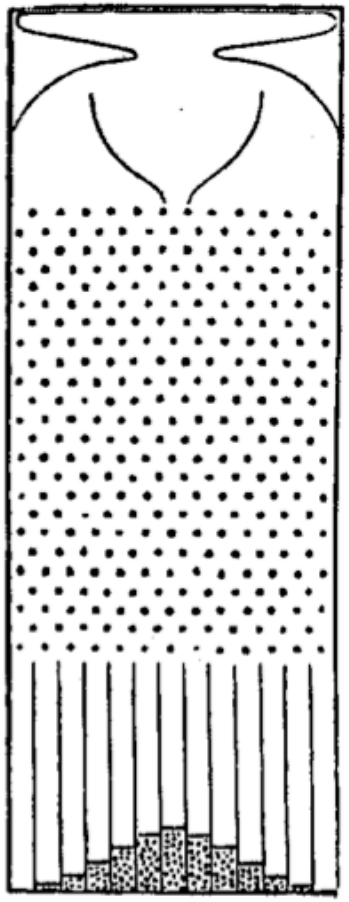


FIG .8.

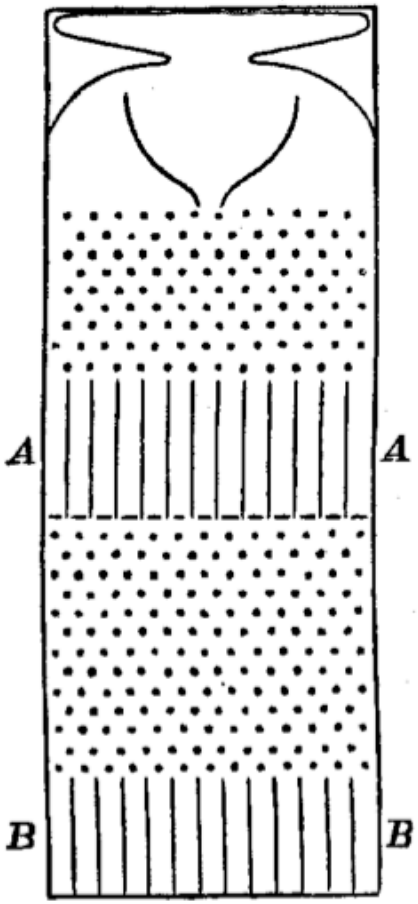


FIG .9.

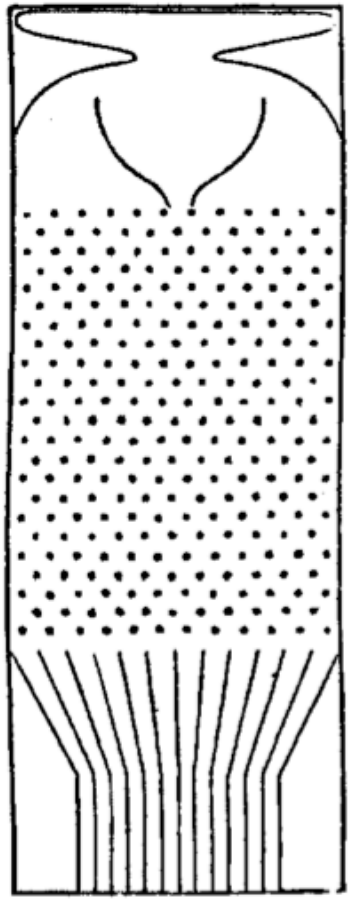


TABLE I.

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**

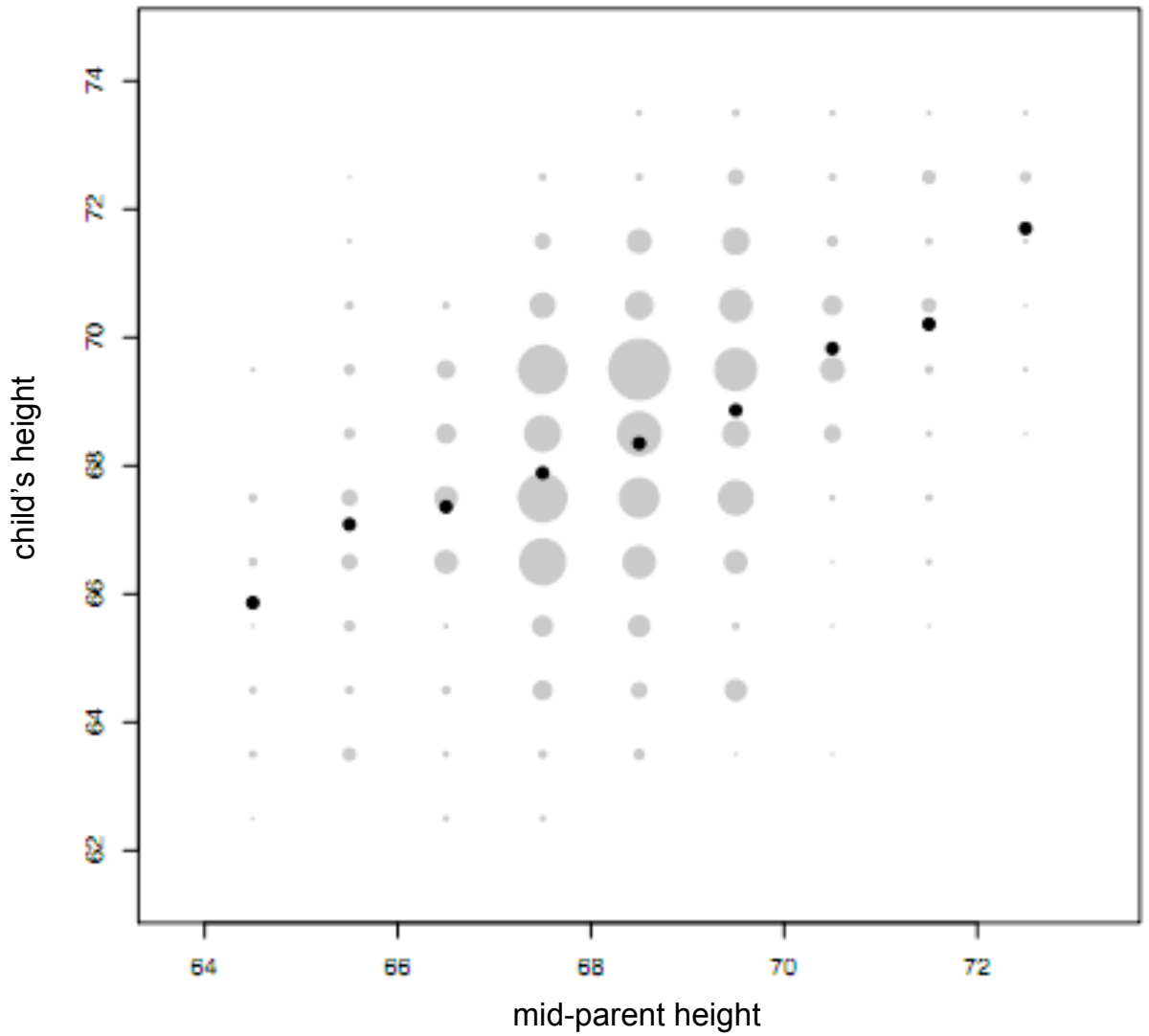
By 1877, Galton was starting to examine these ideas mathematically -- He essentially **discovered the important properties of the bivariate normal distribution** (the bivariate normal had been derived by theorists unknown to Galton, but they did not develop the idea of regression, nor did they attempt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

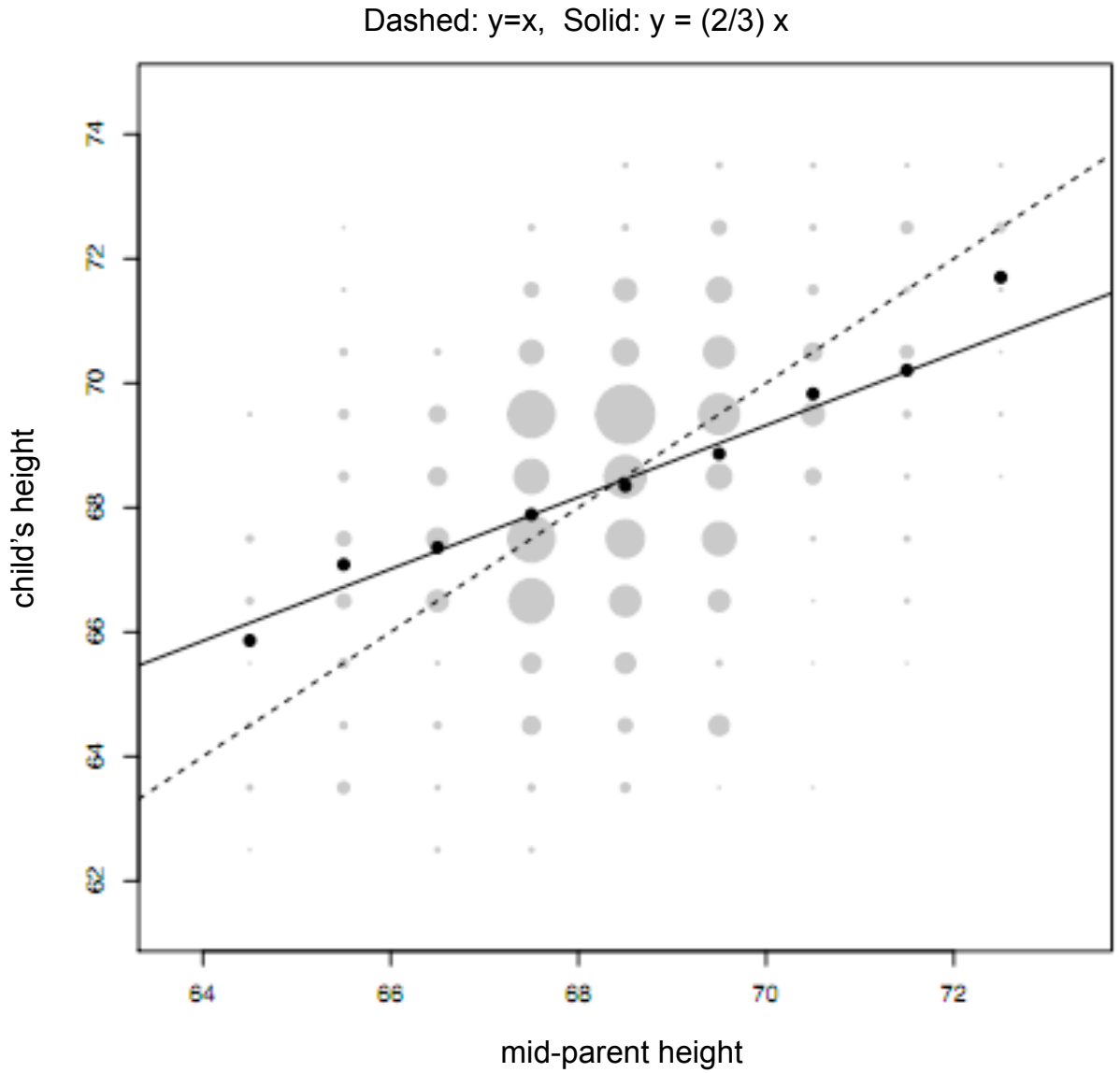
Galton and regression

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)

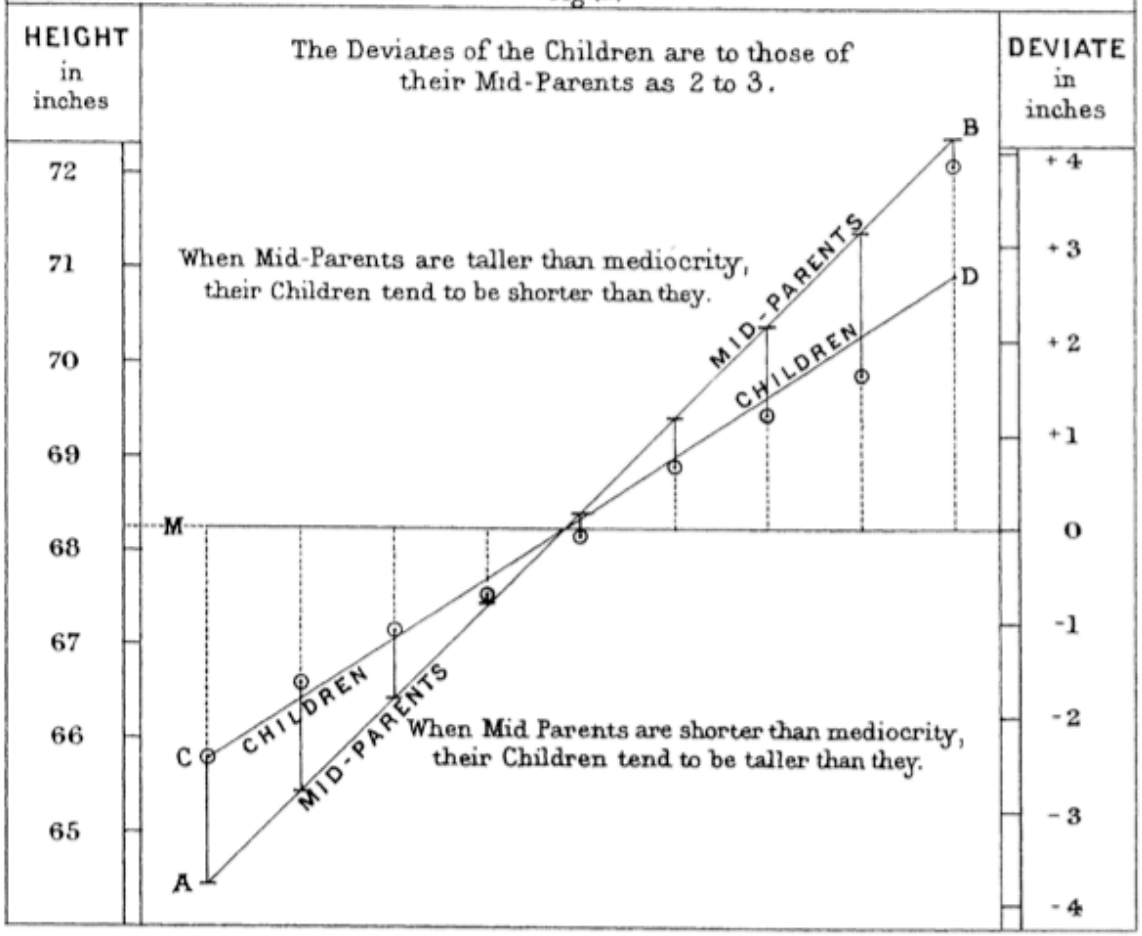
He found that the slope was about 2/3 -- If children were on average as tall as their parents, he'd expect a slope of 1, leading him to coin the phrase "regression toward mediocrity"


child's height
74
72
70
68
66
64
62
64
66
68
70
72
mid-parent height


Dashed: y=x, Solid: y = (2/3) x
child's height
mid-parent height
62
64
66
68
70
72
74
64
66
68
70
72

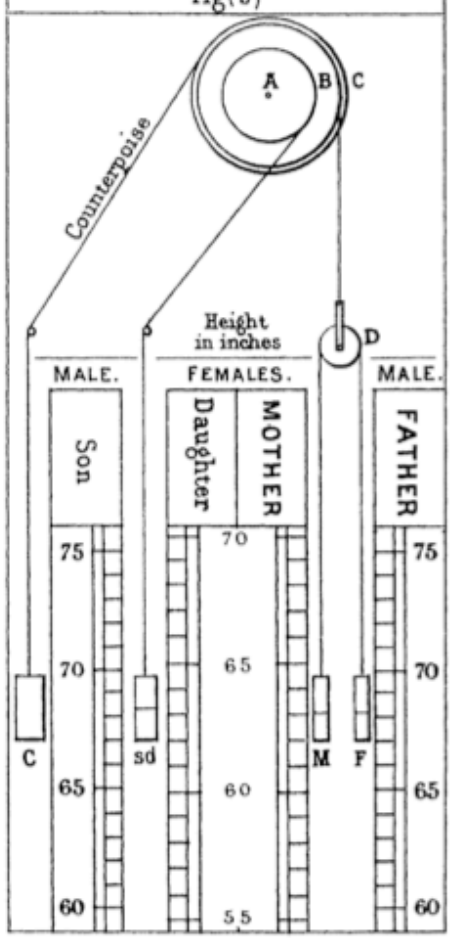
RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)



FORECASTER OF STATURE

Fig (b)



J P & W R Emslie, lith

Journ. Anthropolog. Inst., Vol. XV, Pl. IX

Galton and regression

What Galton found through essentially geometric means was the following relationship

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

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

The r here is the so-called correlation coefficient -- If we are given data in n pairs, say, mid-parent heights and children's heights, $(X_1, Y_1), \ldots, (X_n, Y_n)$, then it is computed as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\widehat{\sigma}_X}\right)\left(\frac{Y_i - \overline{Y}}{\widehat{\sigma}_Y}\right)$$

where $\overline{X}$ and $\widehat{\sigma}_X$ are the mean and sample standard deviation of $X_1, \ldots, X_n$ and $\overline{Y}$ and $\widehat{\sigma}_Y$ are the mean and sample standard deviation of $Y_1, \ldots, Y_n$

We will see what this is estimating shortly -- For now, notice that it seems to be measuring the degree of agreement (the co-relation) between X and Y

FIG. 10.

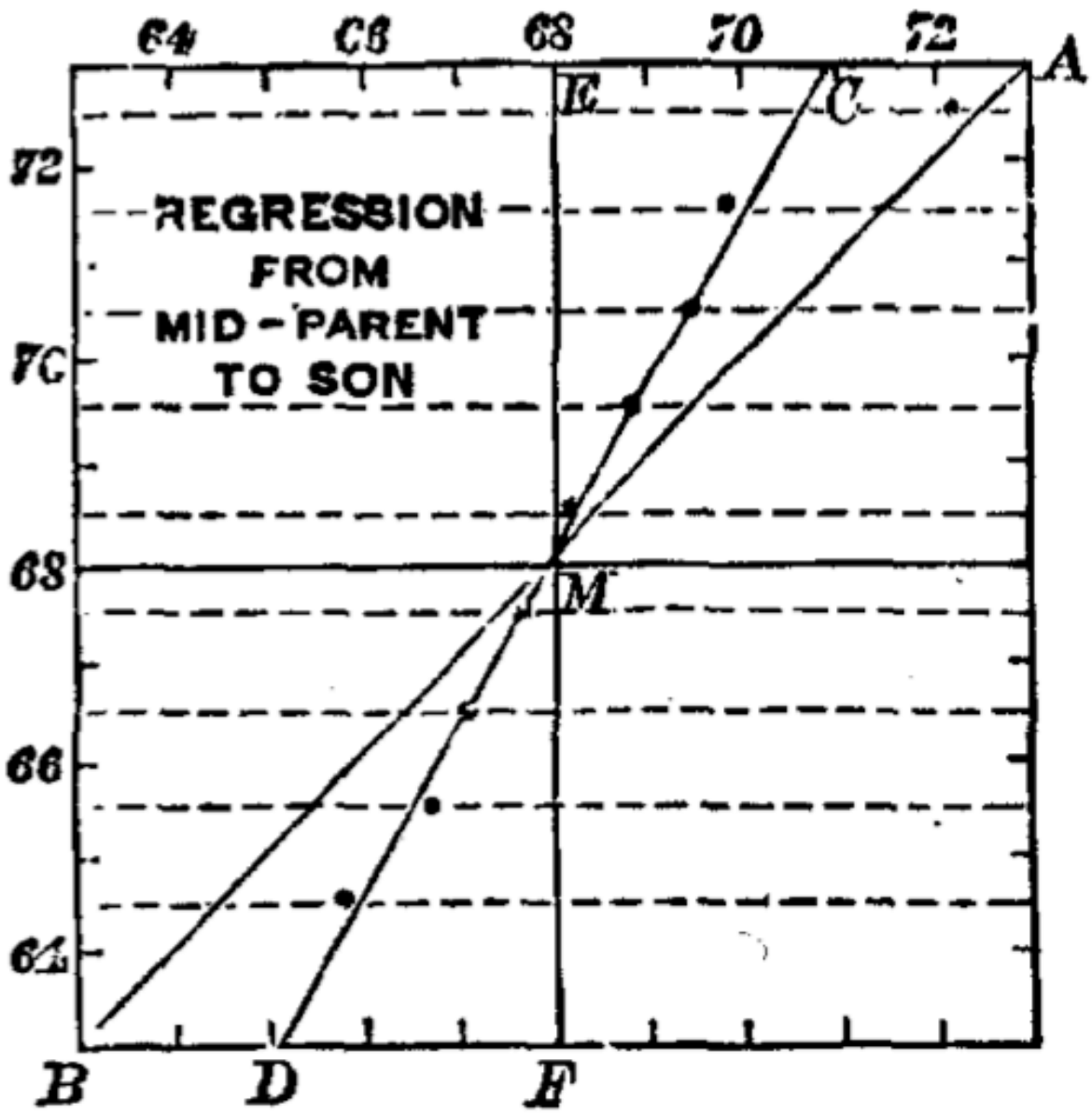
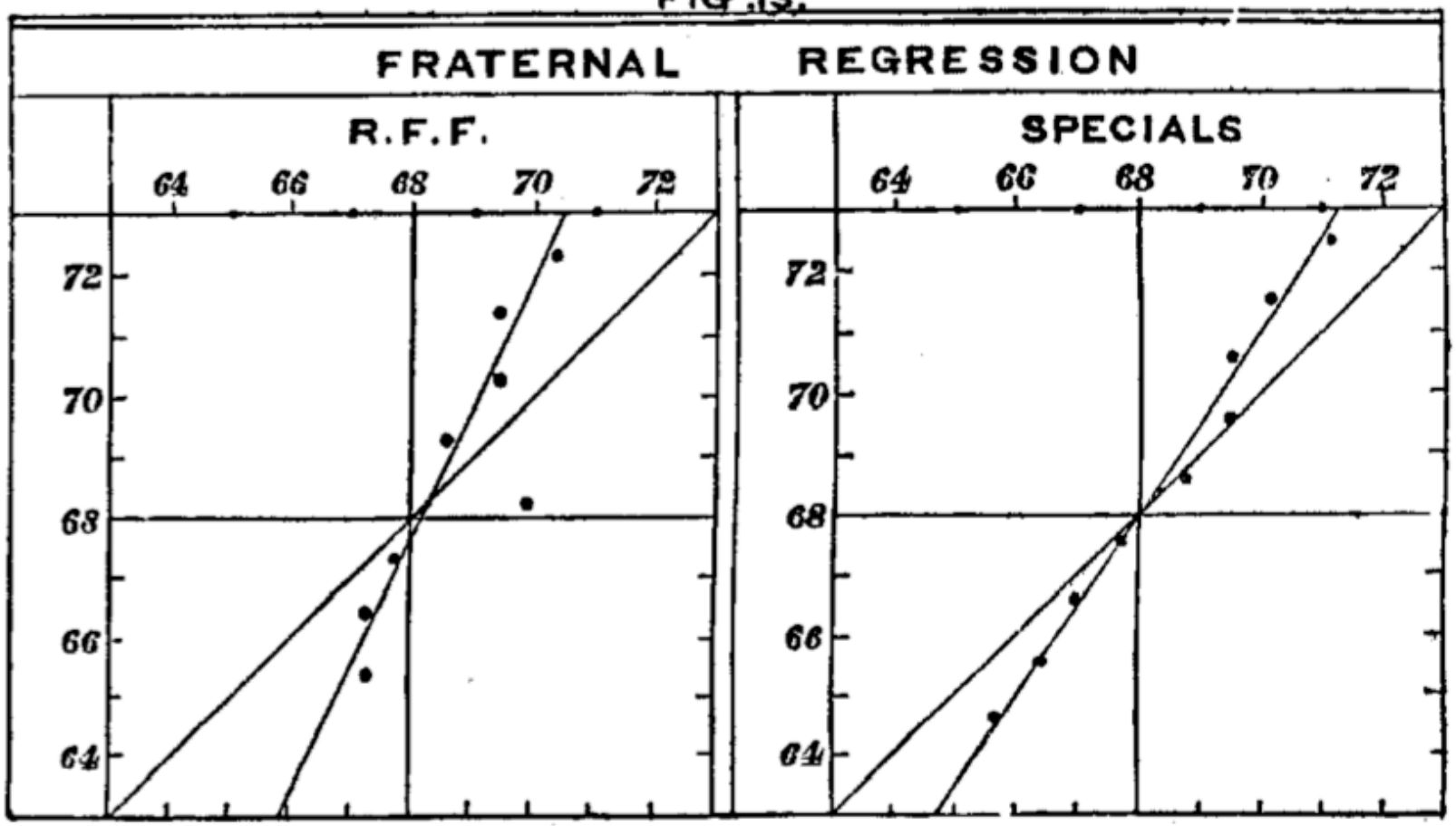

64
66
68
70
72
A
E
C
72
REGRESSION
FROM
MID-PARENT
TO SON
70
68
M
66
64
B
D
F

FIG. 13.


FRATERNAL REGRESSION
R.F.F.
SPECIALS
64
66
68
70
72
72
70
68
66
64

Galton and regression

Galton also noticed, however, that **a similar kind of regression happened in reverse** -- That is, that if you transposed the table, you'd find a slope of 1/3 relating the average mid-parent's height to that of their children

He surmised that the regression effect was more a fact about **the bivariate normal distribution** than anything else -- This is a lesson that many researchers have failed to appreciate even now

Here's Galton -- Notice that he's not content to just invent regression, but he also exhibits one of the first (if not the first) bivariate kernel density estimate!

I found it hard at first to catch the full significance of the entries in the table, which had curious relations that were very interesting to investigate. They came out distinctly when I "smoothed" the entries by writing at each intersection of a horizontal column with a vertical one, the sum of the entries in the four adjacent squares, and using these to work upon. I then noticed (see [fig. 6.6]) that lines drawn through entries of the same value formed a series of concentric and similar ellipses. Their common centre lay at the intersection of the vertical and horizontal lines, that corresponded to 68.25 inches. Their axes were similarly inclined. The points where each ellipse in succession was touched by a horizontal tangent, lay in a straight line inclined to the vertical in the ratio of 2/3; those where they were touched by a vertical tangent lay in a straight line inclined to the horizontal in the ratio of 1/3. These ratios confirm the values of average regression already obtained by a different method, of 2/3 from mid-parent to offspring, and of 1/3 from offspring to mid-parent, because it will be obvious on studying [fig. 6.6] that the point where each horizontal line in succession is touched by an ellipse, the greatest value in that line must appear at the point of contact. The same is true in respect to the vertical lines. These and other relations were evidently a subject for mathematical analysis and verification. (Galton 1885c, 254–255)

Galton and regression

To complete this story, Galton enlisted the help of a mathematician, Hamilton Dickson -- The problem he wanted solved was the following

Suppose x and y are expressed as deviations from the mean and that x is normal with mean zero and standard deviation Q_x

Also suppose that conditional on a fixed value of x, y is also normal with mean $\beta_{y|x}$ and standard deviation $Q_{y|x}$

What is the joint density of x and y and, in particular, are the contours of equal probability elliptical?

What is the conditional distribution of x given y, and in particular, what is the relation between the two regression coefficients?

DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1·08)

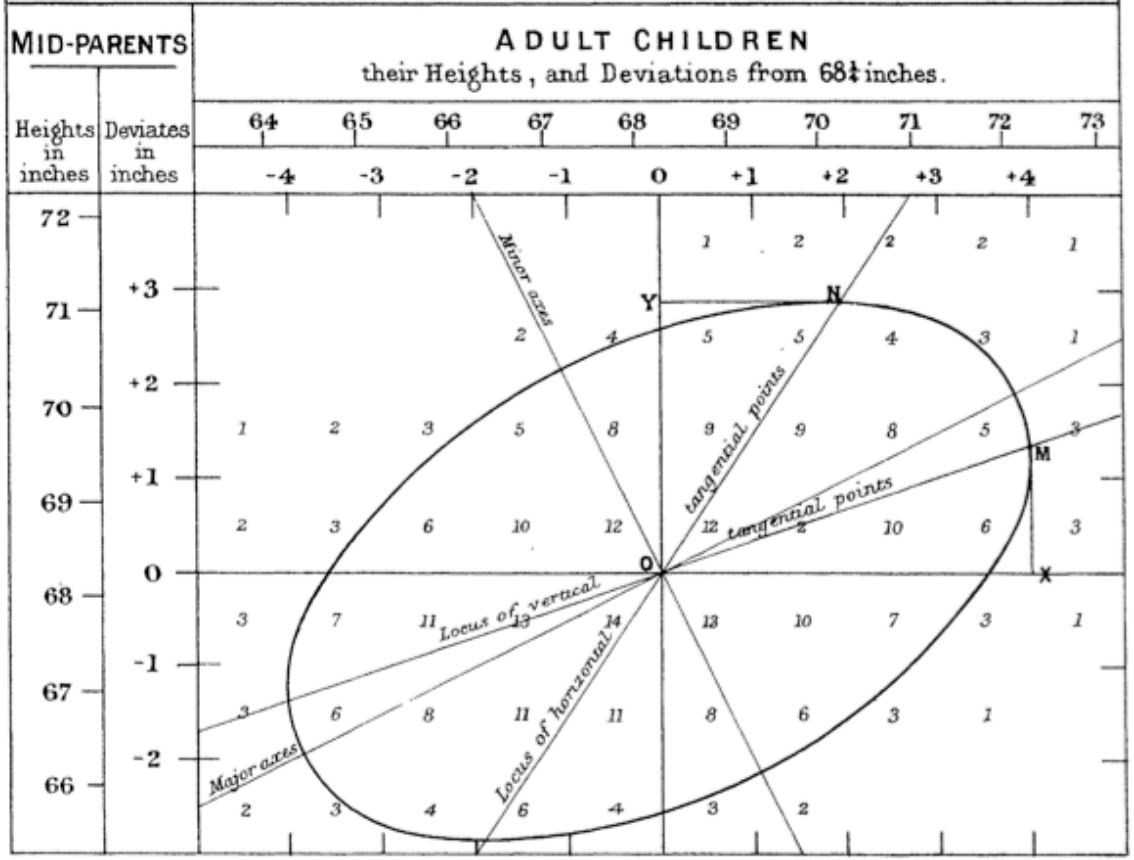
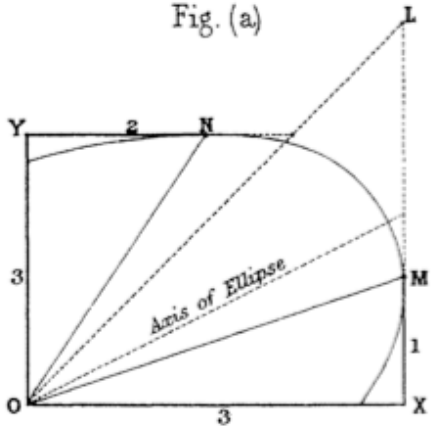


Fig. (a)



J.P. & W.R. Emslie, lith.

Journ. Anthropolog. Inst., Vol. XV, Pl. X.

The bivariate normal

In his response, Dickson derived the bivariate normal distribution and the associated marginals and conditionals -- Suppose Y_1 is normal with mean μ_1 and variance σ_1^2 and that Y_2 is normal with mean μ_2 and variance σ_2^2

The pair Y_1 and Y_2 are said to have a bivariate normal distribution if their density is given by

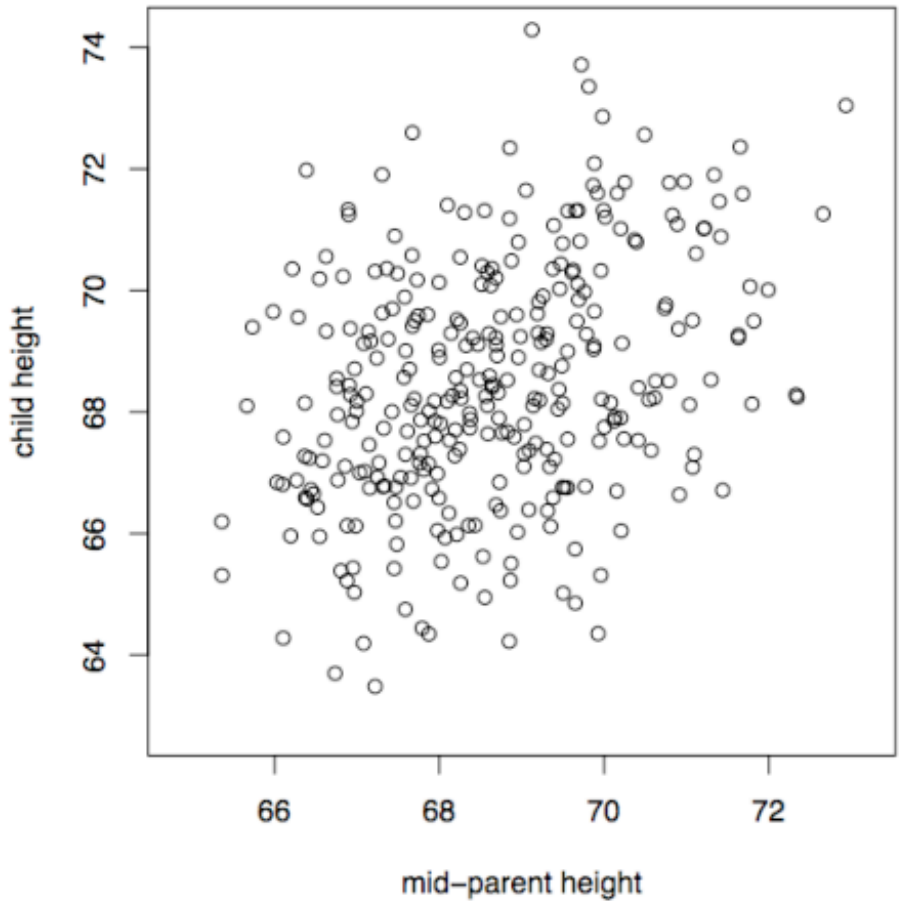
$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}\right\}\right]$$

where ρ is the so-called correlation between Y_1 and Y_2 -- This is again a parametric family with 5 parameters, μ_1, μ_2, σ_1, σ_2, and ρ

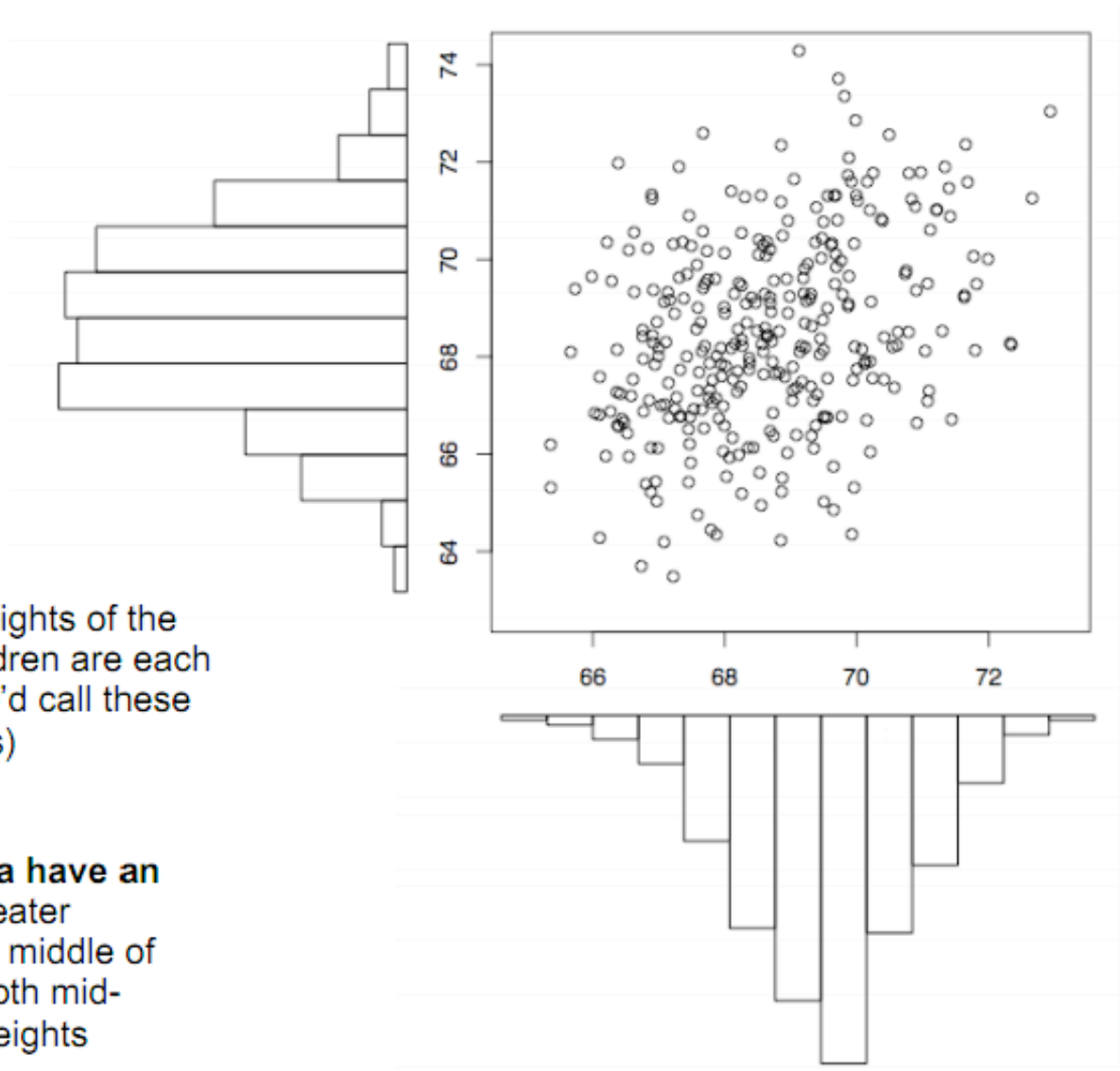
The bivariate normal

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

Correlation

The covariance between two variables Y_1 and Y_2 is defined to be

$$\sigma_{12} = \text{cov}(Y_1, Y_2) = \text{E}\left[(Y_1 - \mu_1)(Y_2 - \mu_2)\right]$$

where covariance between a variable and itself is just its variance (these concepts were part of your probability course -- I'm hoping!)

We can then define the correlation between Y_1 and Y_2 to be

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

To specify a bivariate normal distribution, we need either ρ or σ_{12} -- We can reexpress the distribution a few slides back in terms of σ_{12} instead

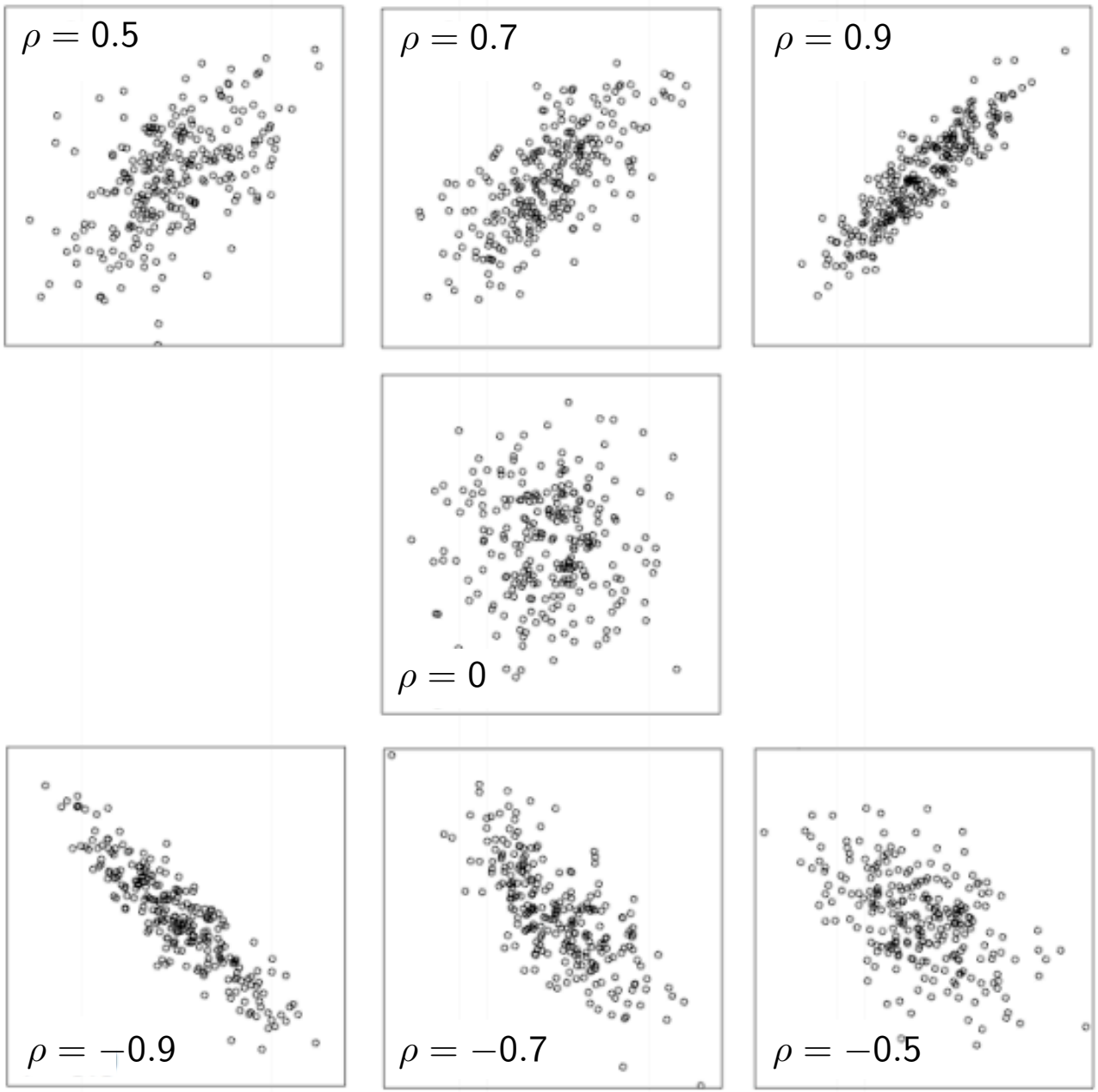
Correlation

As a standardized quantity, however, the correlation has a couple of nice properties

1. It is bounded between -1 and 1 -- With either -1 or 1 indicating a “perfect” linear 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

The sample correlation coefficient r is an estimate of ρ


$\rho = 0.5$
$\rho = 0.7$
$\rho = 0.9$
$\rho = 0$
$\rho = -0.9$
$\rho = -0.7$
$\rho = -0.5$

The bivariate normal

After a little algebra, the conditional distribution of Y_2 given that $Y_1 = y_1$ is univariate normal with mean

$$E(Y_2|Y_1 = y_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1)$$

and variance $(1 - \rho^2)\sigma_2^2$

Galton and regression

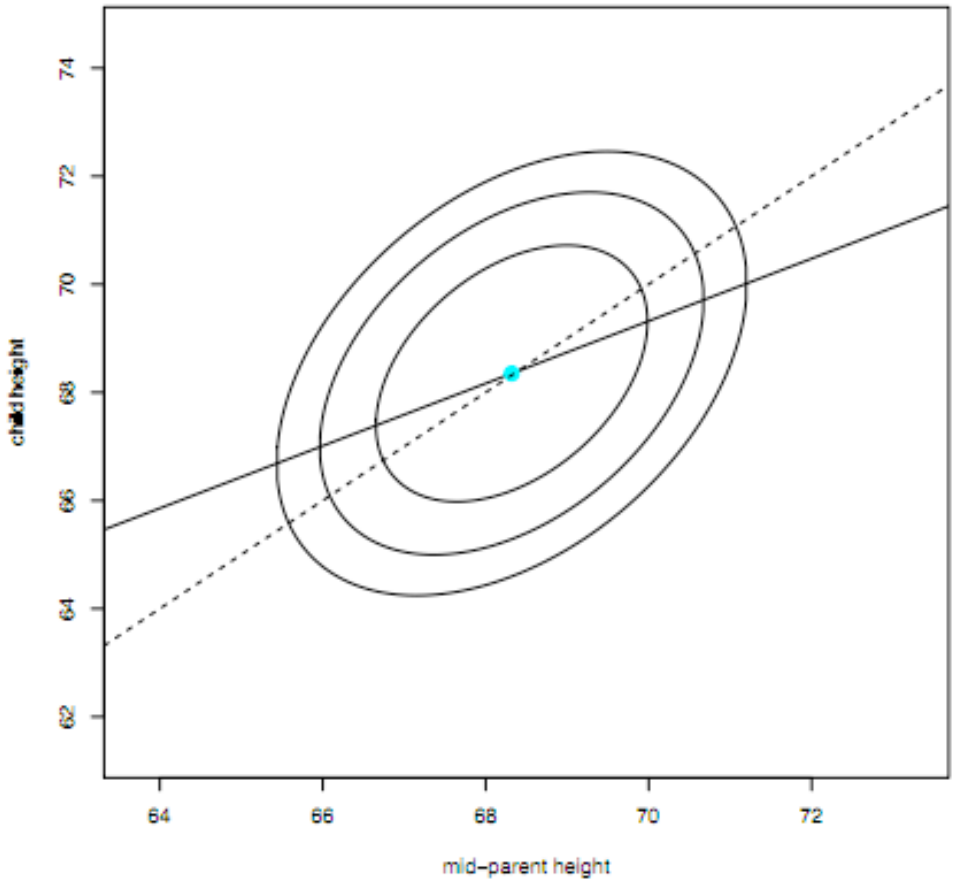
If you look at this formula, you'll see that the conditional means trace out a line as a function of y_1, which is exactly the solid line drawn in Galton's images -- Here, for example, we let Y_1 be mid-parents heights and Y_2 be children's heights we

Therefore, Galton's regression work is really about conditioning using the bivariate normal distribution -- He was driven to these results empirically and then directed a mathematician to establish them formally

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)

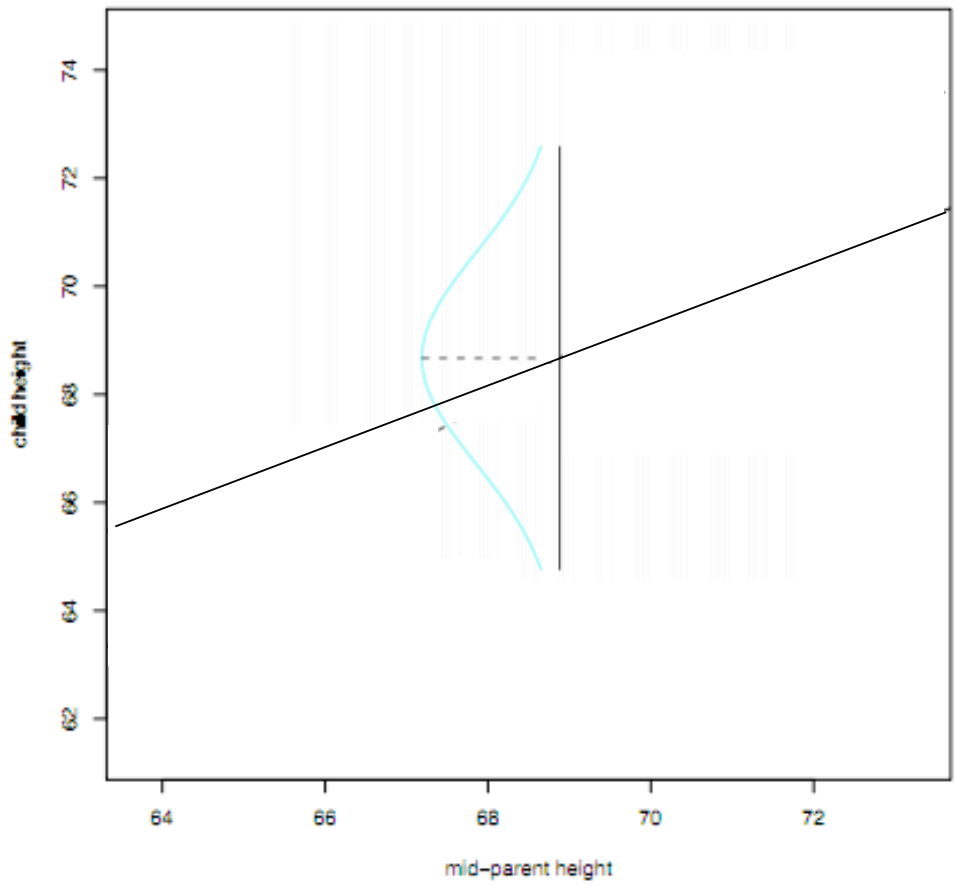
The solid line is the regression line of child height on mid-parent height and the dashed line is y=x



Putting it all together

The regression line sweeps out the conditional means of child's heights given mid-parent heights (restrict your attention to children from mid-parents with a given height), then you again have a normal distribution with mean on the line

These, again, are just properties of the bivariate normal, but were motivations for Galton (and provide some useful insight)



Conditioning

The implication is that if we wanted to estimate the mean height of children born to mid-parents of a given height (say, 62”), we would simply form the average of all the children’s heights we have in that category -- The conditional distribution is normal and we know how to estimate the mean of a normal!

This is what Galton did with his plot of points that eventually followed a line -- What we see from the bivariate normal is that the conditional means should fall on a line, the regression line

Multivariate normal

To refresh your memory on the subject, if $Z_1, \ldots, Z_m$ are independent standard normal random variables (means zero and standard deviations 1), then, taken as a vector, the joint distribution of $Z = (Z_1, \ldots, Z_m)^t$ is

$$\begin{aligned} f(z) &= f(z_1, \ldots, z_m) \\ &= \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-z_j^2/2} \\ &= \frac{1}{(2\pi)^{m/2}} e^{-\sum_j z_j^2/2} \\ &= \frac{1}{(2\pi)^{m/2}} e^{-z^t z/2} \end{aligned}$$

Multivariate normal

In general, if the vector $Y = (Y_1, \ldots, Y_m)^t$ has a m-dimensional multivariate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_m)^t$ and an m-by-m variance-covariance matrix Σ (symmetric and positive definite), its density is given by

$$f(y|\mu, \Sigma) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} e^{-(y-\mu)^t \Sigma^{-1}(y-\mu)/2}$$

As with the univariate normal, we can generate any member of this family by taking a linear transformation of independent normals

Multivariate normal

That is, if Z has an m-dimensional multivariate normal distribution with mean 0 (a vector of 0's) and variance-covariance matrix $I_{m\times m}$, the m-by-m identity matrix, then $Y = \mu + \Sigma^{1/2}Z$ has a multivariate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_m)^t$ and variance-covariance matrix Σ

Similarly, if Y has a multivariate normal distribution with mean μ and variance-covariance matrix Σ , then $\Sigma^{-1/2}(Y - \mu)$ consists of m independent standard normals

Recall that since Σ is positive definite, we can find its "square root" -- That is, a symmetric matrix $\Sigma^{1/2}$ such that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$ and $\Sigma^{-1/2}\Sigma^{1/2} = \Sigma^{1/2}\Sigma^{-1/2} = I$ where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$

A more recent set of examples

Galton is interesting because of the completeness of his treatment -- Working out the mechanics of the bivariate normal and establishing the regression line in terms of conditional distributions

We will now take up our example from last time, approaching regression from a more computational direction, starting from least squares

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

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`

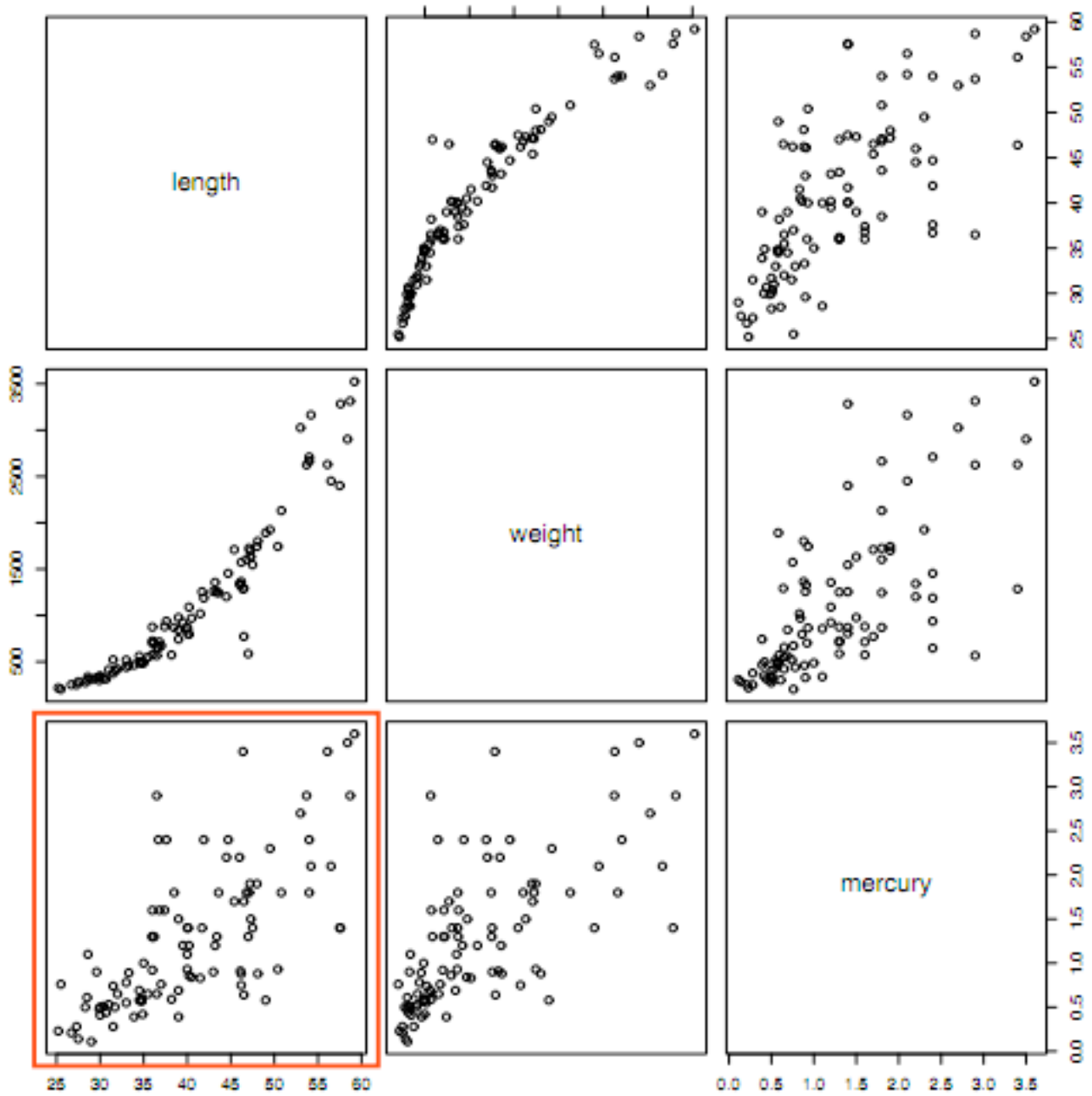
	river	stn	length	weight	mercury
1	0	0	47.0	1616	1.60
2	0	0	48.7	1862	1.50
3	0	0	55.7	2855	1.70
4	0	0	45.2	1199	0.73
5	0	0	44.7	1320	0.56
6	0	0	43.8	1225	0.51
7	0	0	38.5	870	0.48
8	0	0	45.8	1455	0.95
9	0	0	44.0	1220	1.40
10	0	0	40.4	1033	0.50
11	0	1	47.7	3378	0.80
12	0	1	45.1	2920	0.34
13	0	1	43.5	2674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

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
25 30 35 40 45 50 55 60
500 1500 2500 3500
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?

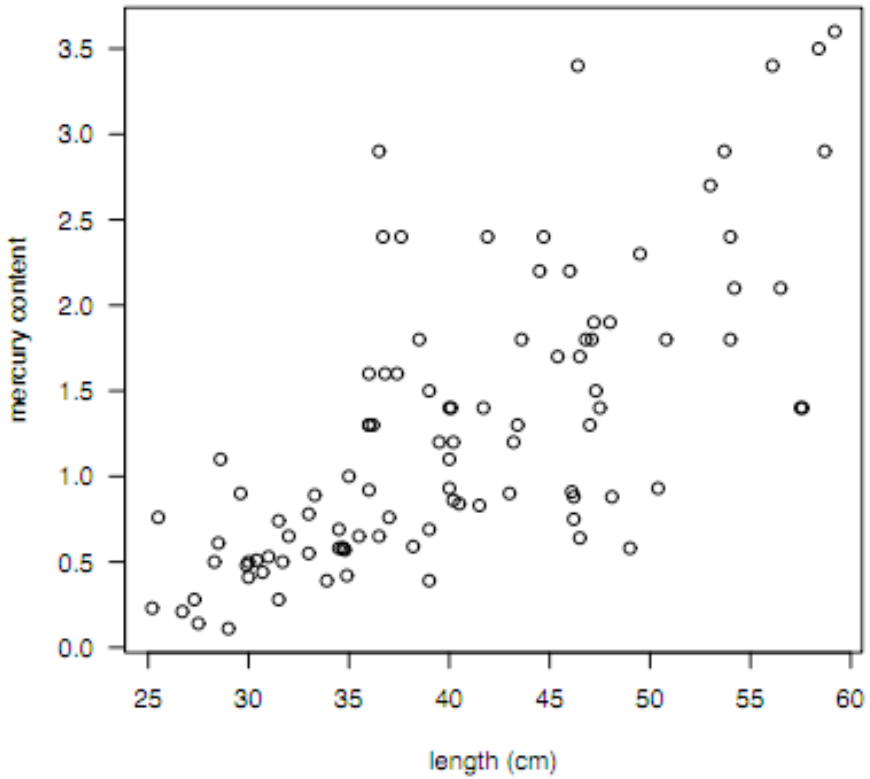
A linear model

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

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

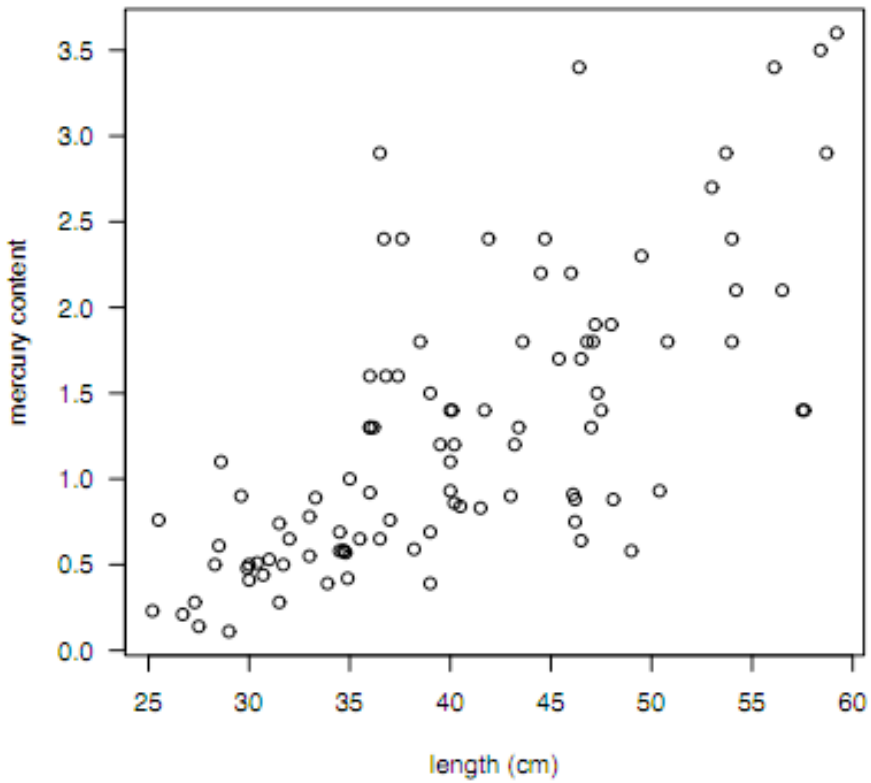
where β_0 and β_1 are parameters, the slope and intercept



A linear model

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

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



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

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

Least squares

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

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

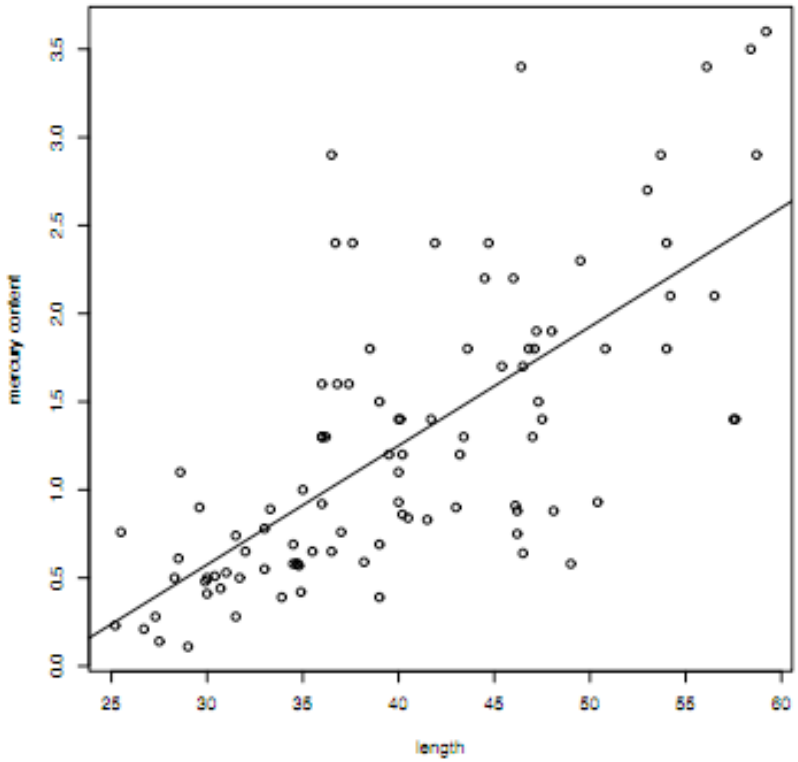
and the associated **sum of squares** is 33.4 (our simple trial and error approach was pretty far off!)

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**

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

Downstairs in the expression for $\hat{\beta}_1$ 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?

Least squares

We can also express this answer in terms of the correlation coefficient r that we introduced earlier -- That is

$$\widehat{\beta}_1 = \mathsf{r}\,\frac{\widehat{\sigma}_{\mathsf{y}}}{\widehat{\sigma}_{\mathsf{x}}}$$

which should look a lot like Galton's solution in the height example

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**

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)

More interpretation:

We can add another interpretation to the coefficient of determination -- That is, after a little algebra, we can show that R^2 is just the correlation between our original response values $y_1, \ldots, y_{98}$ and their predictions from the model $\widehat{y}_1, \ldots, \widehat{y}_{98}$

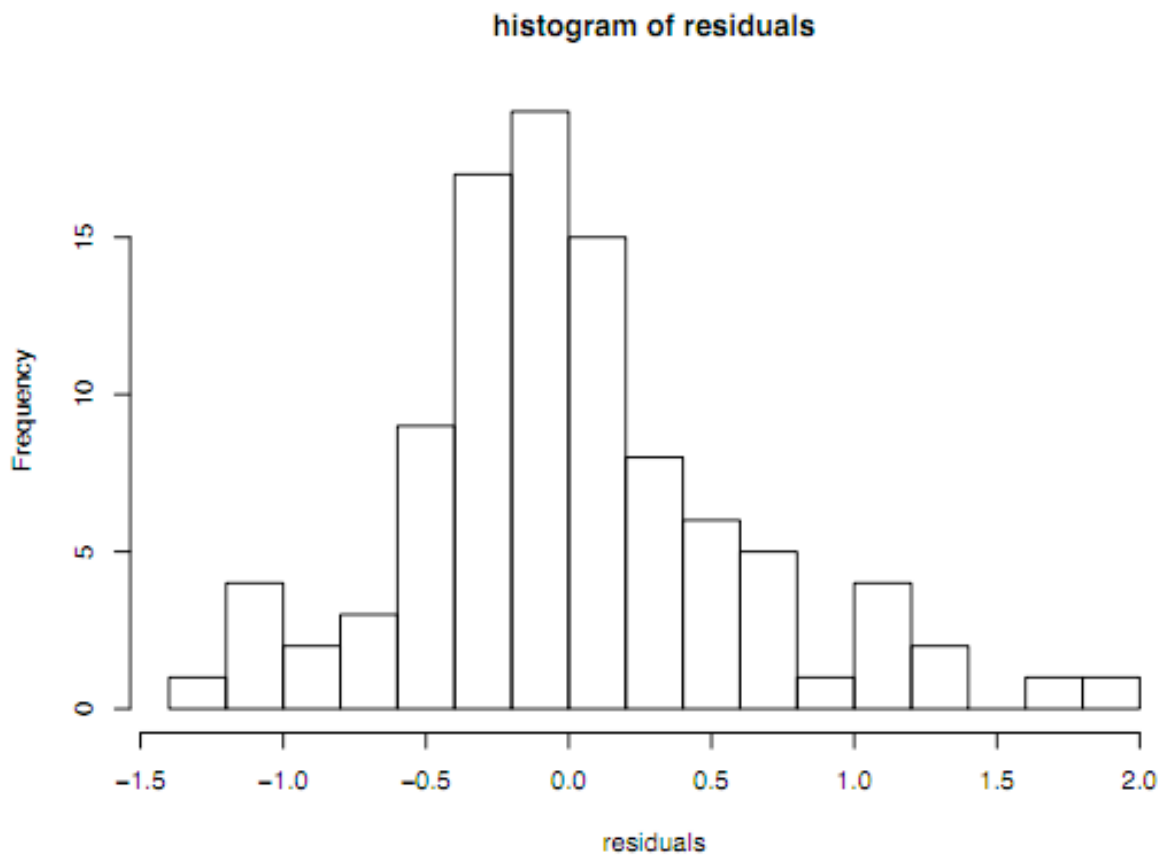
An example

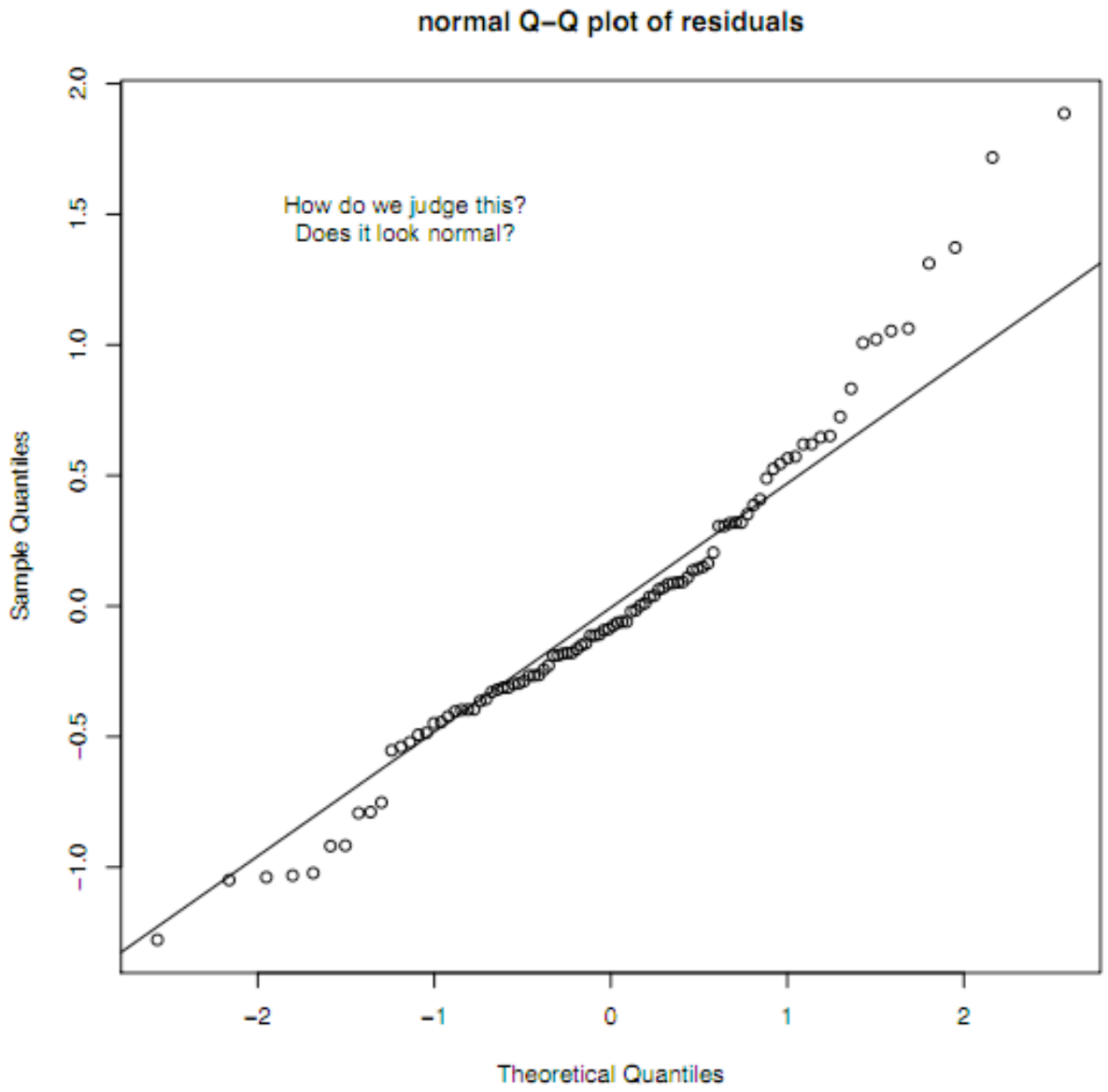
On the next two pages, we plot the residuals from the least squares fit to the fish data; the regression relating fish length and mercury content

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

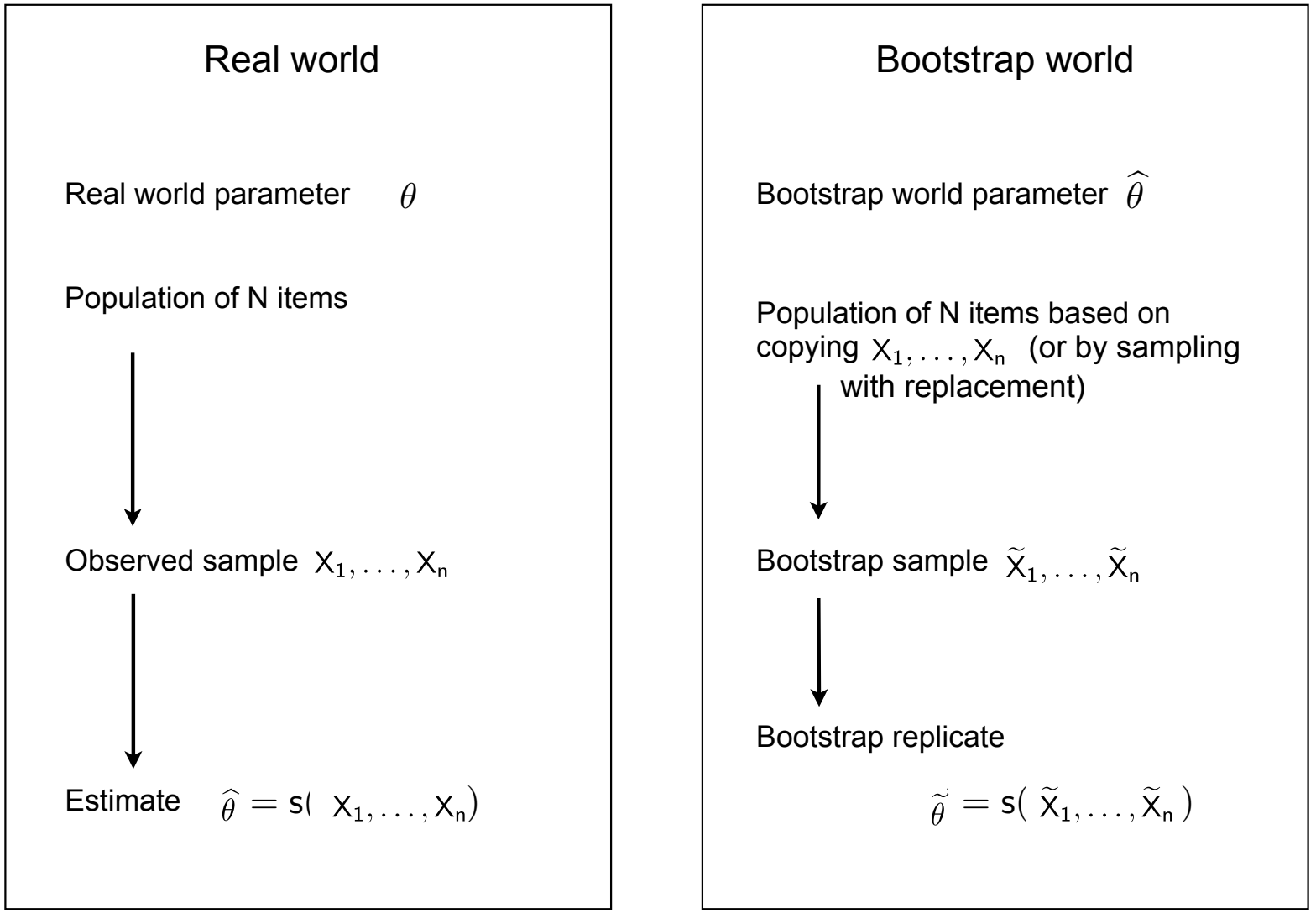

normal Q–Q plot of residuals
How do we judge this?
Does it look normal?
Sample Quantiles
Theoretical Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

Inference

Last time you also considered how we might make inferences from our sample of 98 fish to the population of fish living in the river -- Because these were a random sample, we could appeal to the (non-parametric) bootstrap

Again, we assume that if we had access to all the fish in the river, we could form a least squares fit which would result in our population level coefficients for slope and intercept -- Because we have just a sample, we want to assess plausible values for the population quantities

Here's the bootstrap for confidence intervals again...


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

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 $\widehat{\beta}_1^*$

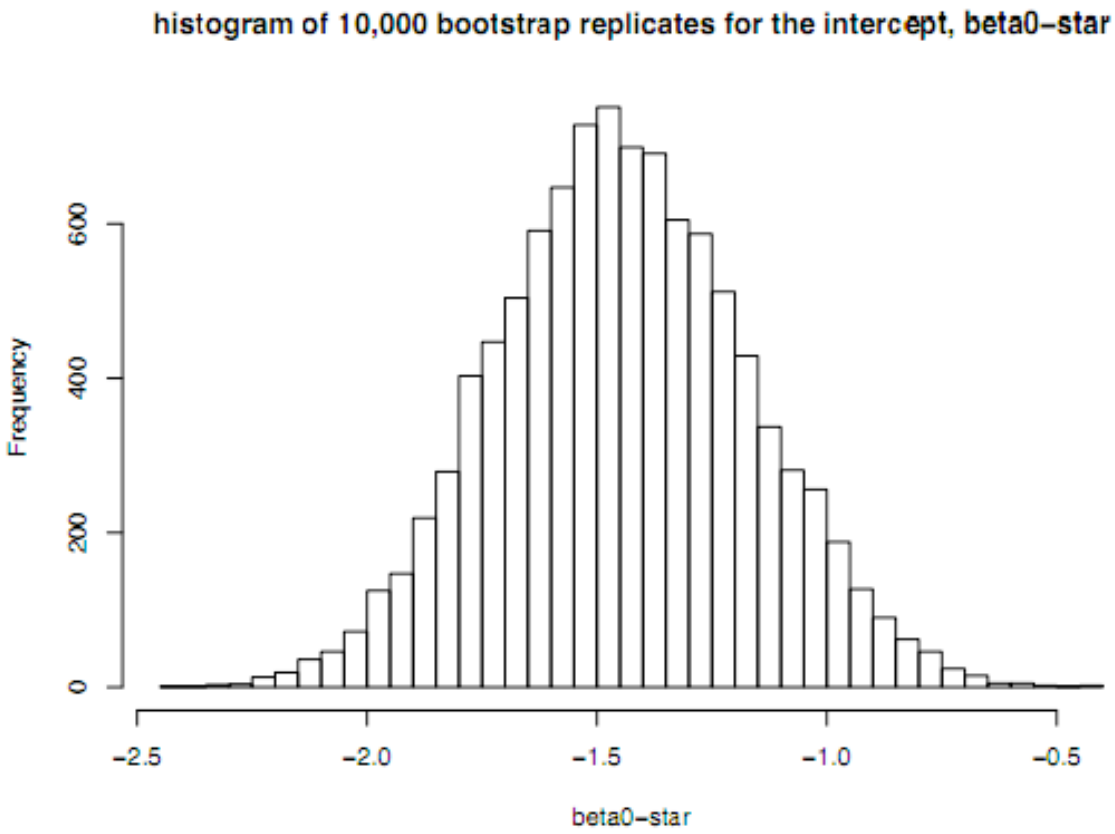
4. Repeat steps 1-3 a large number of times, say 10,000, to obtain a set of bootstrap replicates

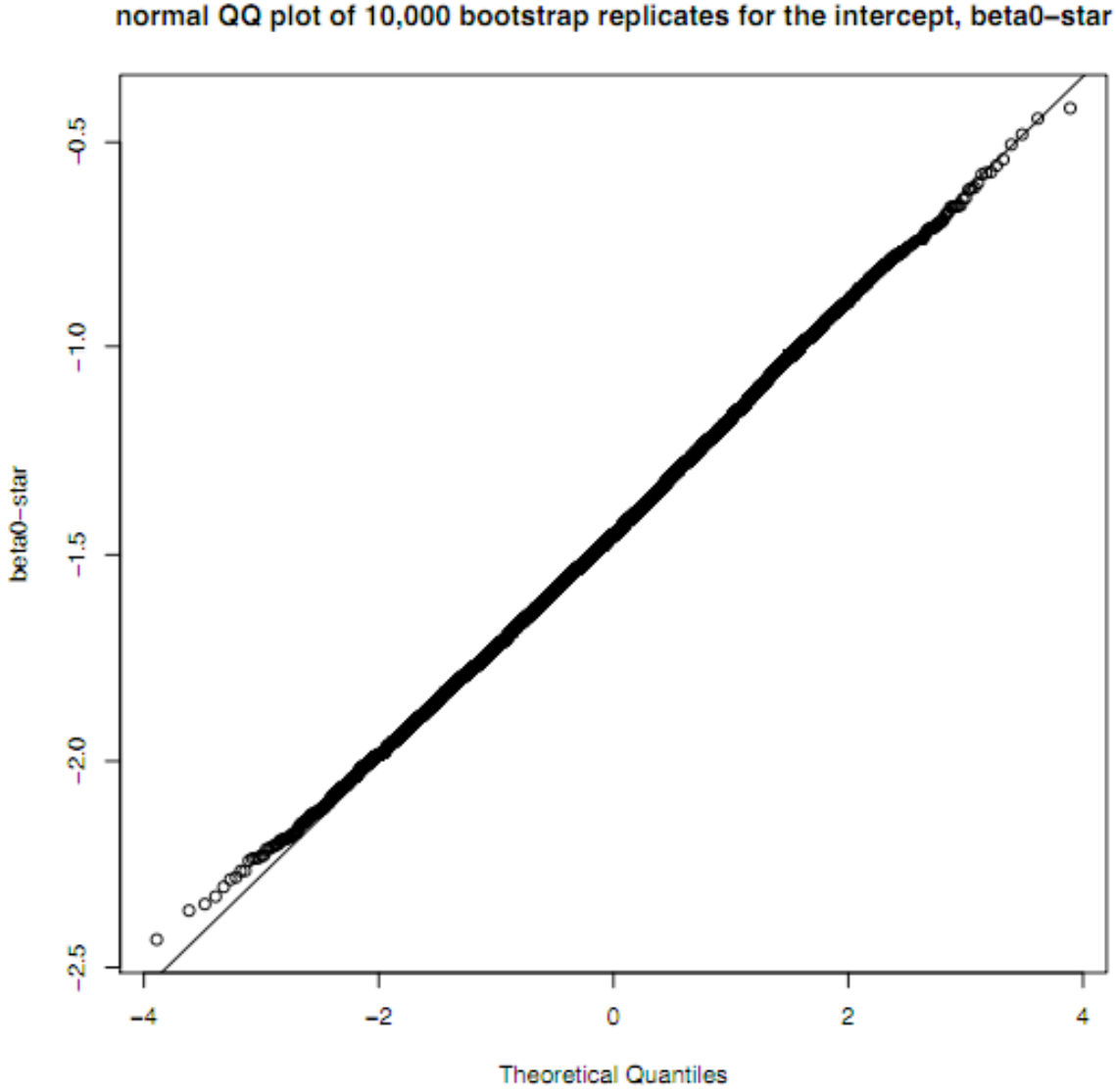
The bootstrap

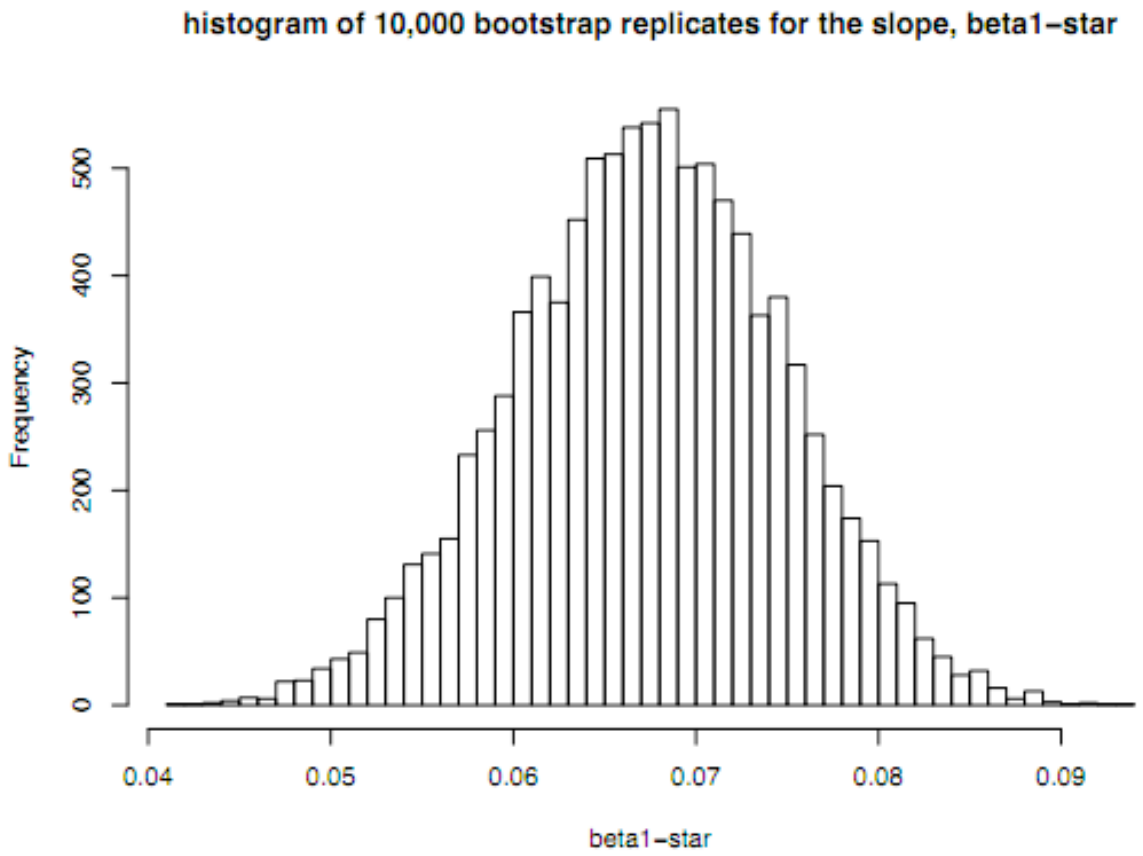
The bootstrap distribution for $\widehat{\beta}_0^*, \widehat{\beta}_1^*$ again is an estimate of their sampling distribution and we can use it to estimate the standard error of each, together with confidence intervals

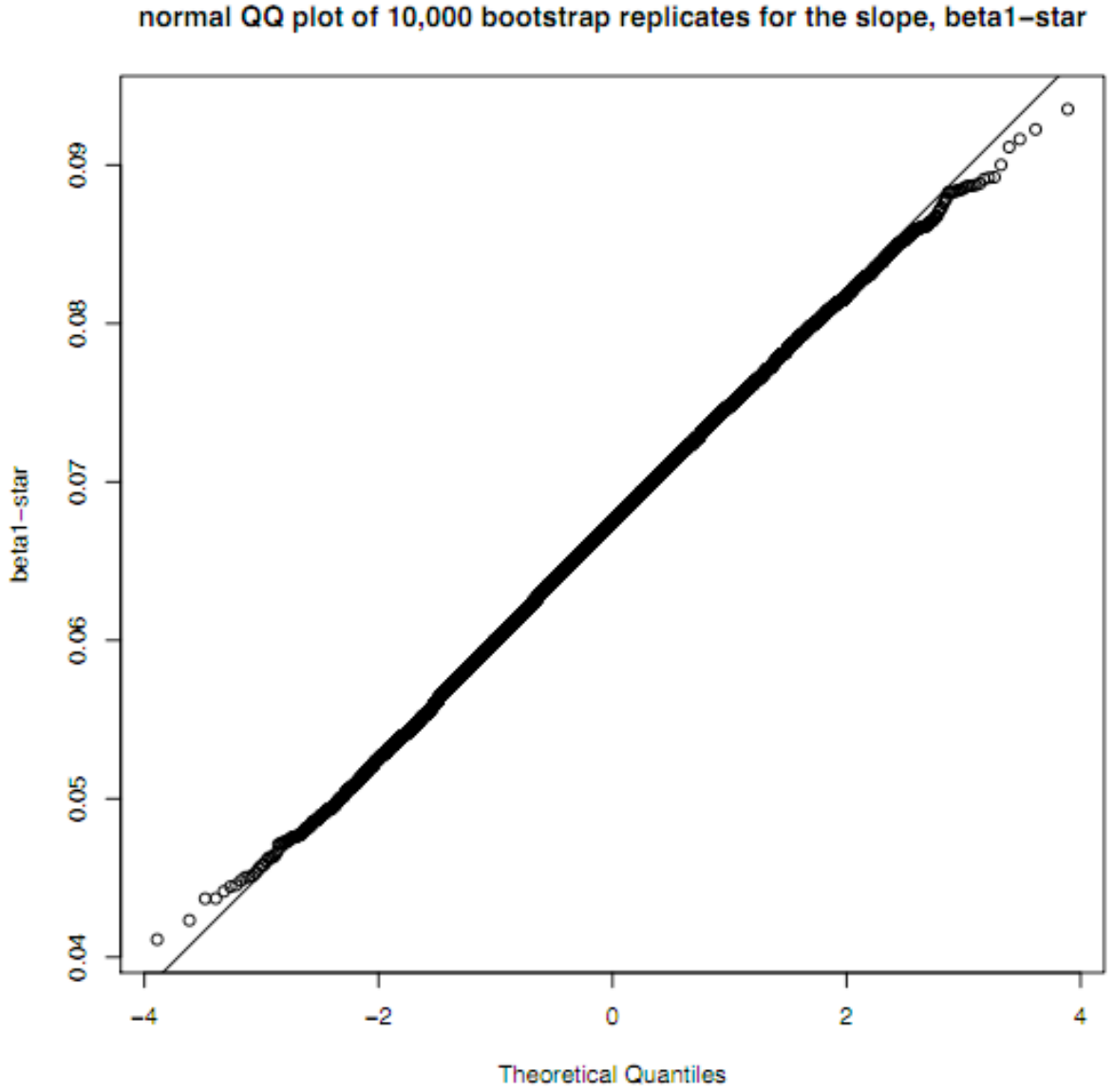
First, $\widehat{\beta}_0 = -1.45$ and the standard deviation of the bootstrap replicates is 0.28; a confidence interval using the 0.025 and 0.975 quantiles agrees with $-1.45 \pm 1.96 * 0.28 = [-2.00, -0.90]$

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

A population model

Our concern for the residuals a few slides back comes from the model that is underneath (at least implicitly) a regression model -- Recall that **Galton came to a linear form for the conditional means** when our data (both input and response) follow a **bivariate normal distribution**

In many (most?) applications, we rarely make such strong assumptions -- Instead we might simply start with a model of the form

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

where we will make assumptions on the error term

One approach is to assume that the errors (conditional on the input data) are all independent, identically distributed with mean 0 and common variance σ^2 -- We can make this even more restrictive by assuming our errors have a normal distribution with mean 0 and variance σ^2

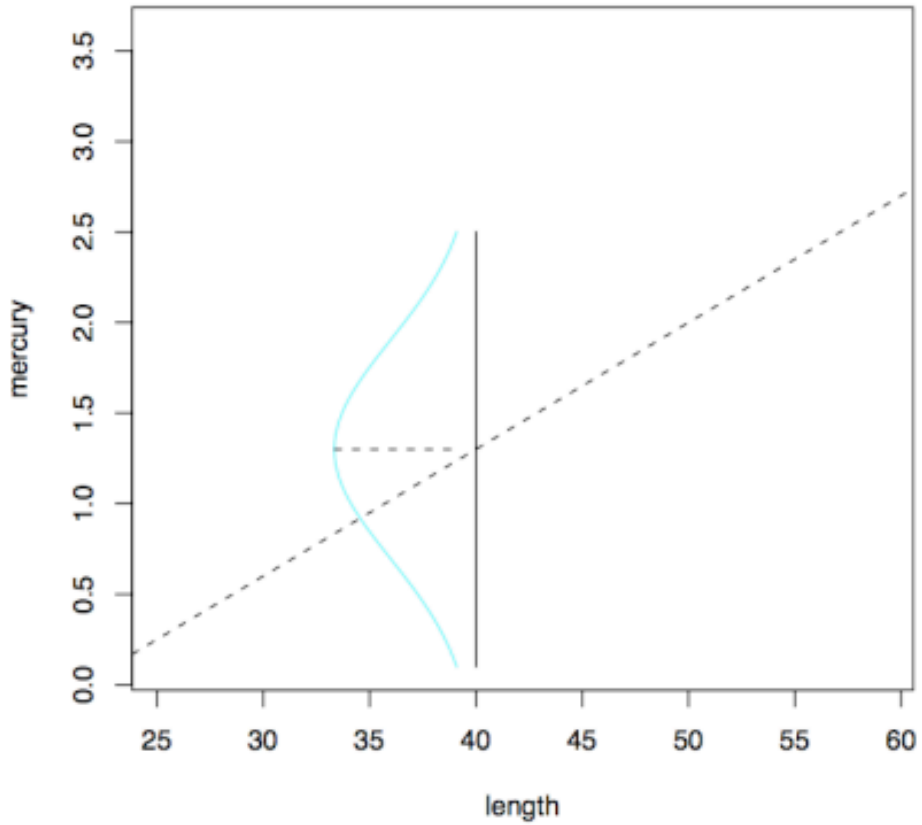
A population model

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!



A population model

Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$, pairs consisting of an input (independent variable, predictor) and an output (dependent variable, response) and we have a model linking them of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where β_0, β_1 are coefficients to be estimated and $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent normal random variables with mean 0 and common variance σ^2

Regression

Put another way, we see that the conditional distribution of Y given X=x is normal with mean $\beta_0 + \beta_1 x$ and variance σ^2

This is a bit less restrictive than Galton's approach which required assuming things about the distribution of X, the inputs -- Here we will do our analysis holding the input data fixed

Estimation

Given a parametric family, we can now write down a likelihood --

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-(Y_i - \beta_0 - \beta_1 X_i)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2/2\sigma^2}\end{aligned}$$

and setting $\tau = \sigma^2$, the log-likelihood

$$l(\beta_0, \beta_1, \tau) = -\frac{n}{2}\log 2\pi\tau - \frac{1}{2\tau}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Estimation

In this way, the MLE is intimately linked to least squares -- We see that our estimates for the regression coefficients are the same

In this way, we have tied several ideas together -- Galton's original conditioning ideas and the least squares solution

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 $\hat{\beta}_0, \hat{\beta}_1$; 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 $\widehat{\beta}_0, \widehat{\beta}_1$

Approximately normal sampling distributions for large n

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

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

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

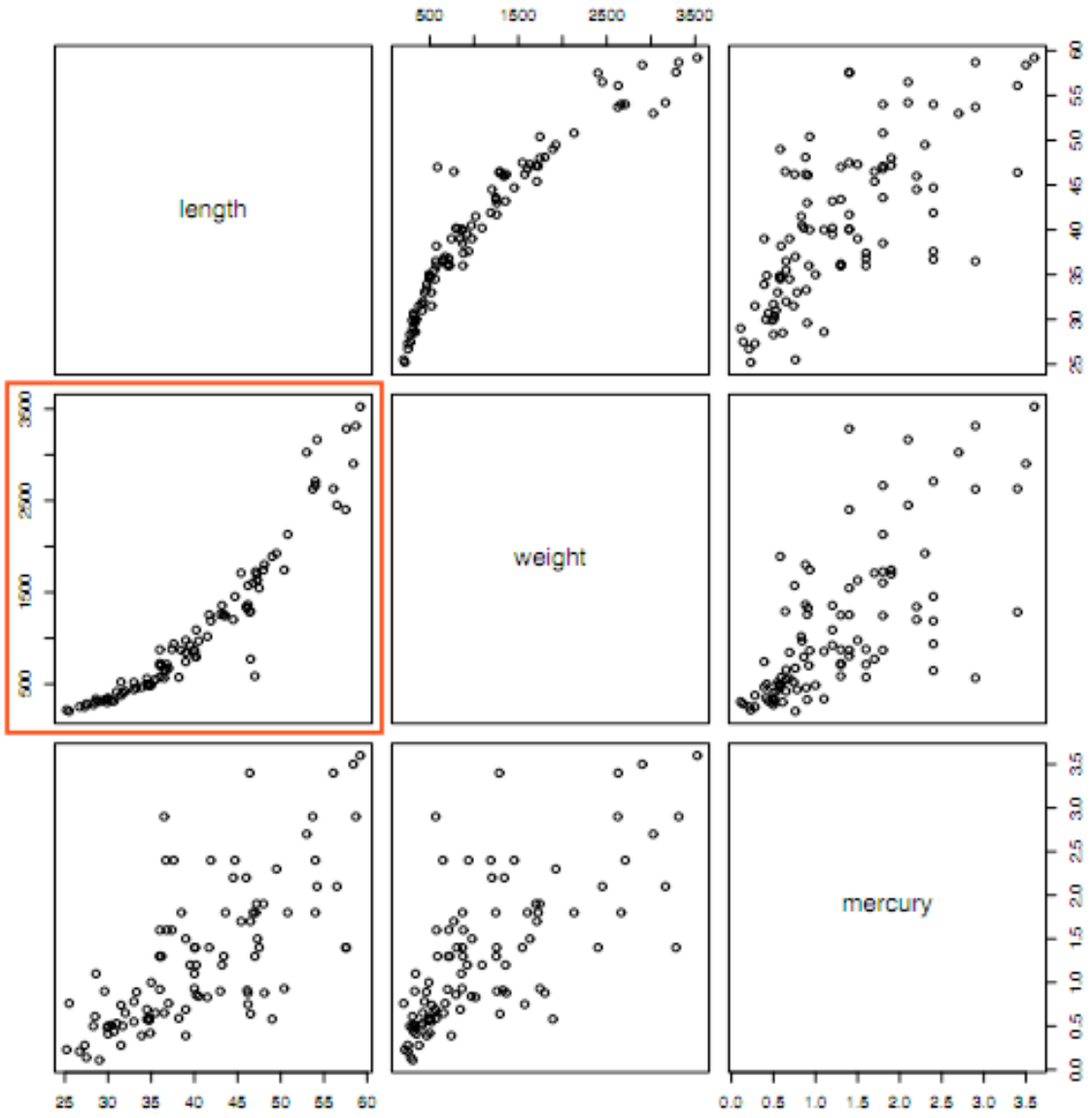
The classics

We will dig a bit deeper into the MLE and the classical approach to estimating a regression equation and assessing the uncertainty next time, for now, let’s continue with another example

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

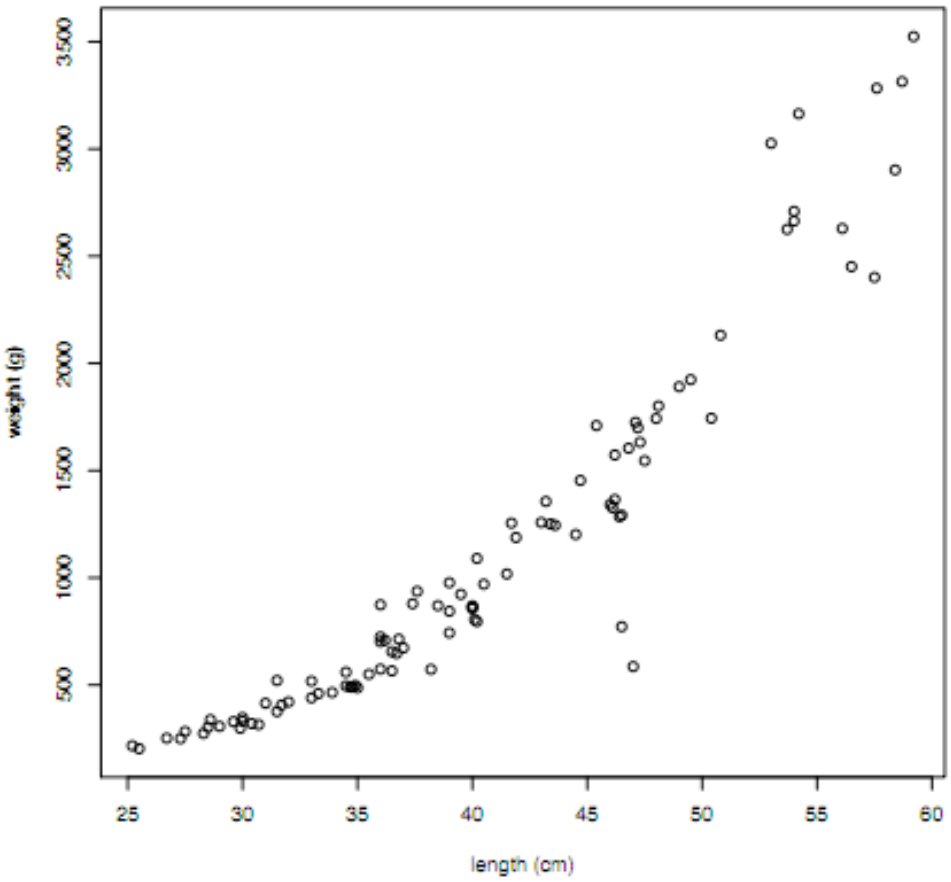
Let's have a look at the relationship between fish length and fish weight; recall our scatterplot matrix...

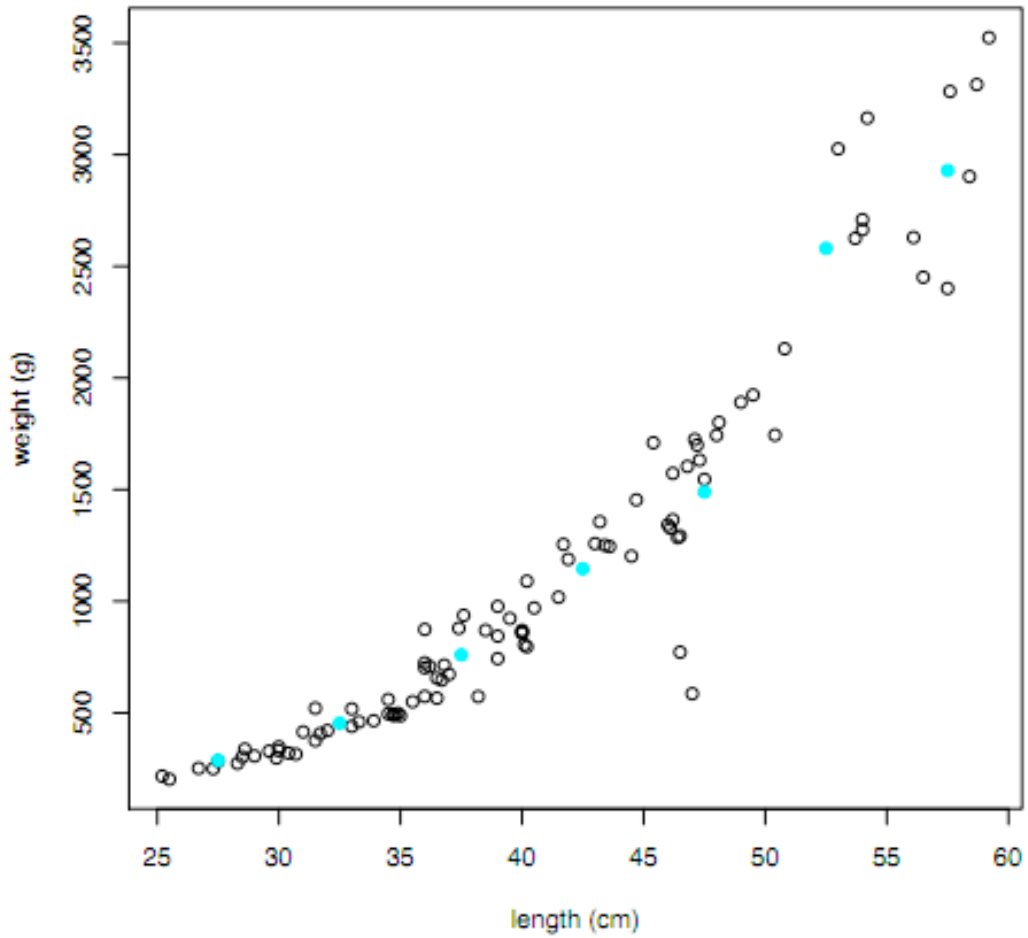

length
weight
mercury
500
1500
2500
3500
25
30
35
40
45
50
55
60
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

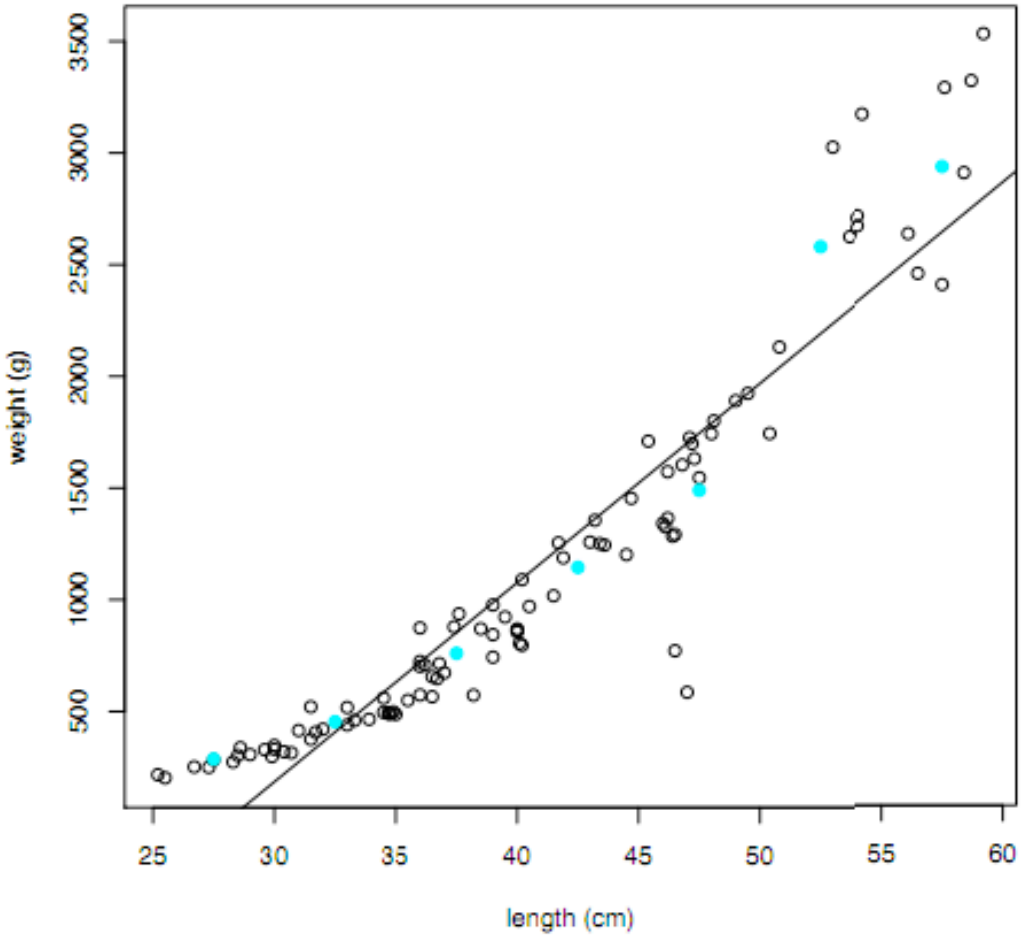
Weight

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

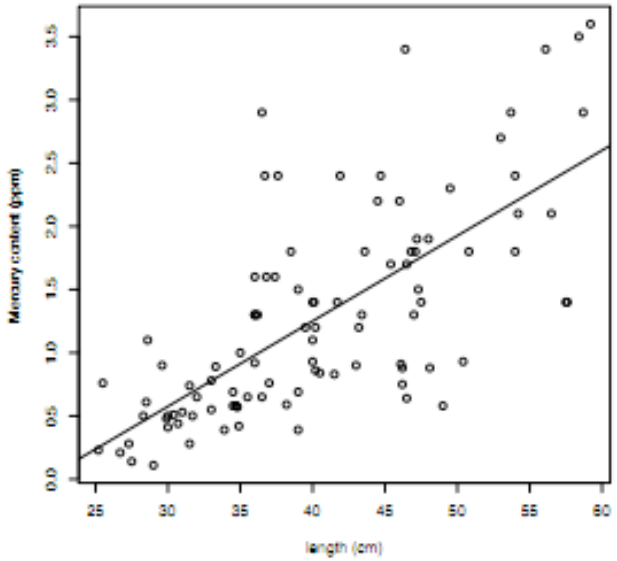
1. Is the relationship between outcomes and predictors linear?
2. Is the error variance constant?
3. Do the errors look roughly normally distributed?

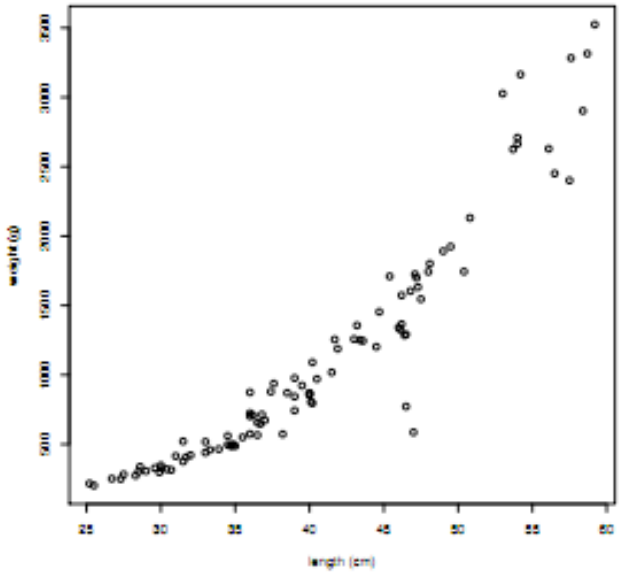
Regression analysis

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...

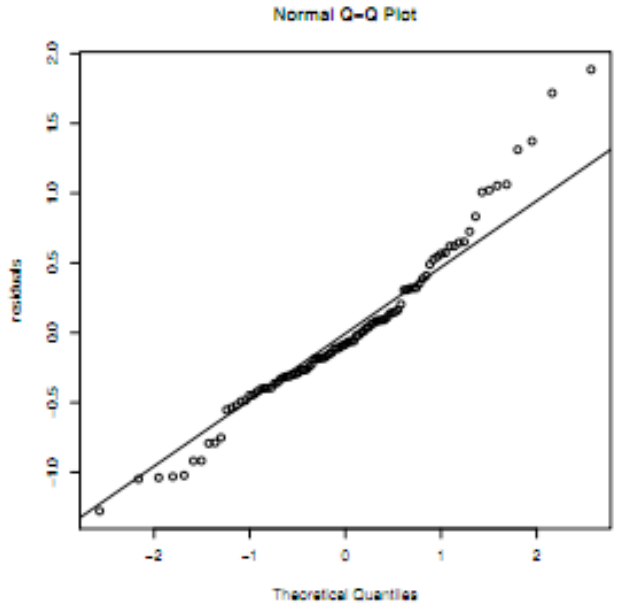


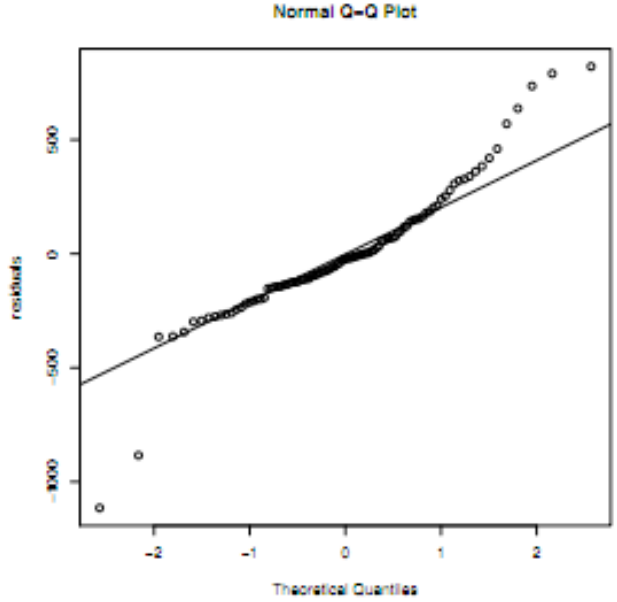


Residual analysis

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

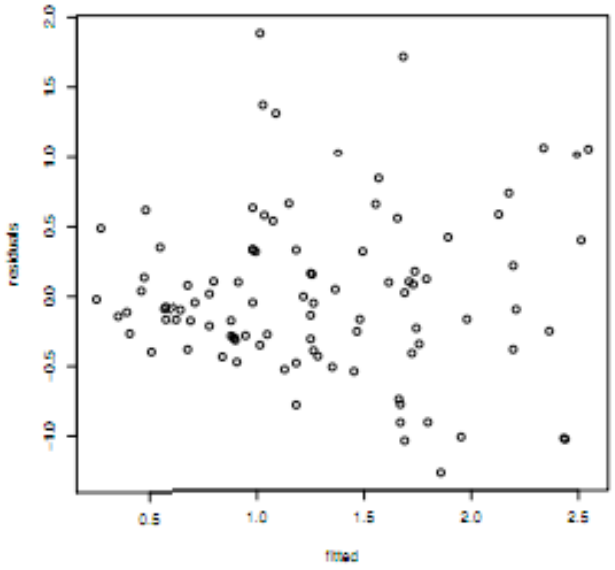
We should also examine the plots for large (positive or negative) values that might indicate outlying points





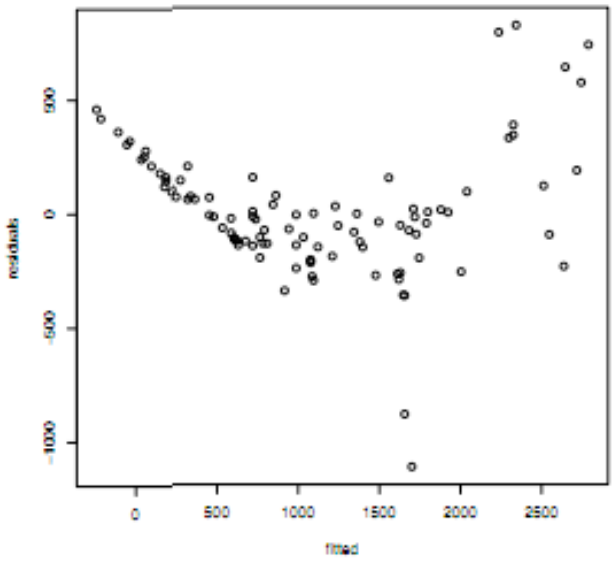
Residual analysis

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

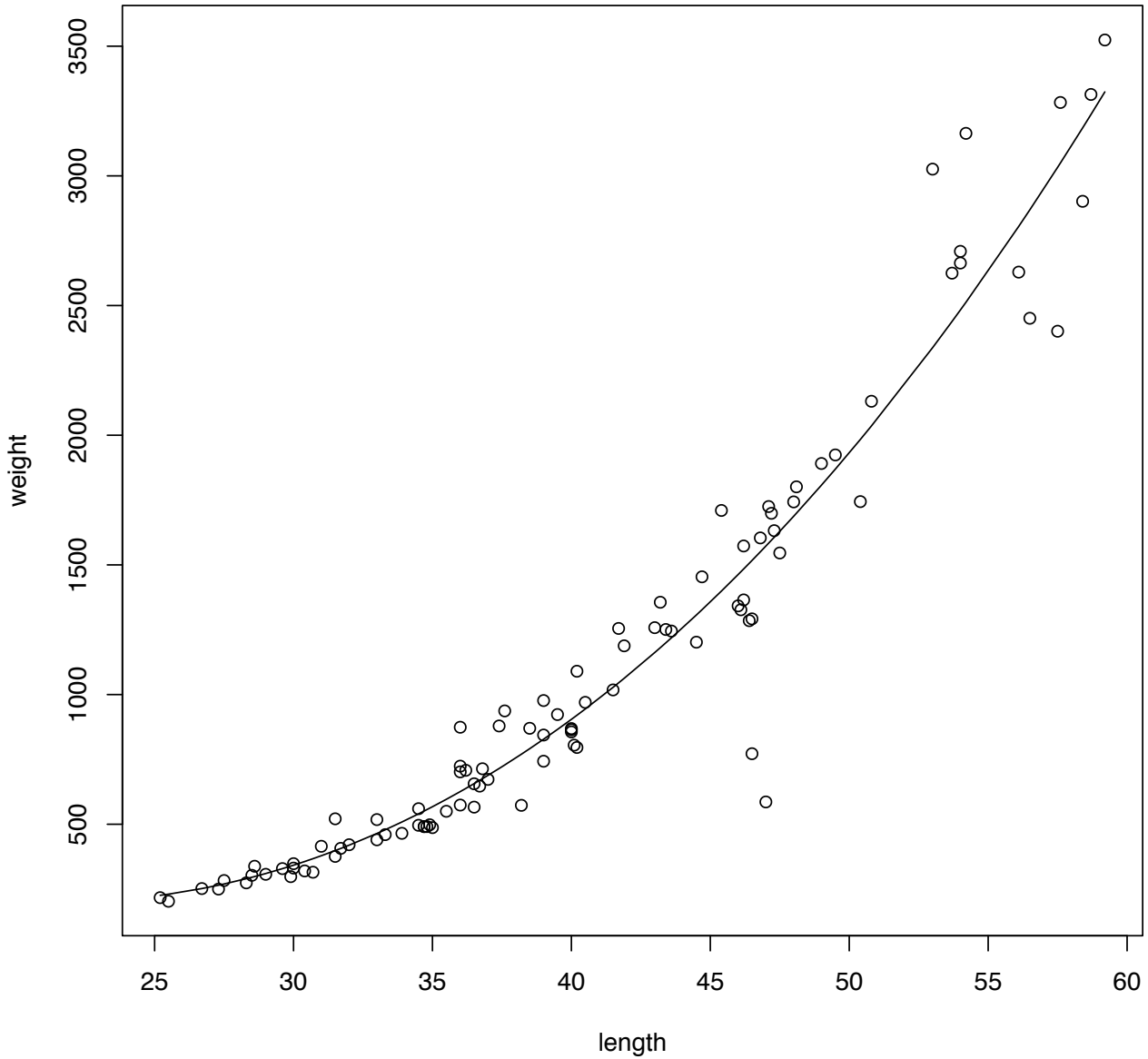
Below we provide the code to fit a quadratic (including both length and $length^2$ in the model) relationship between weight and length; what do you see?

```
> 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?

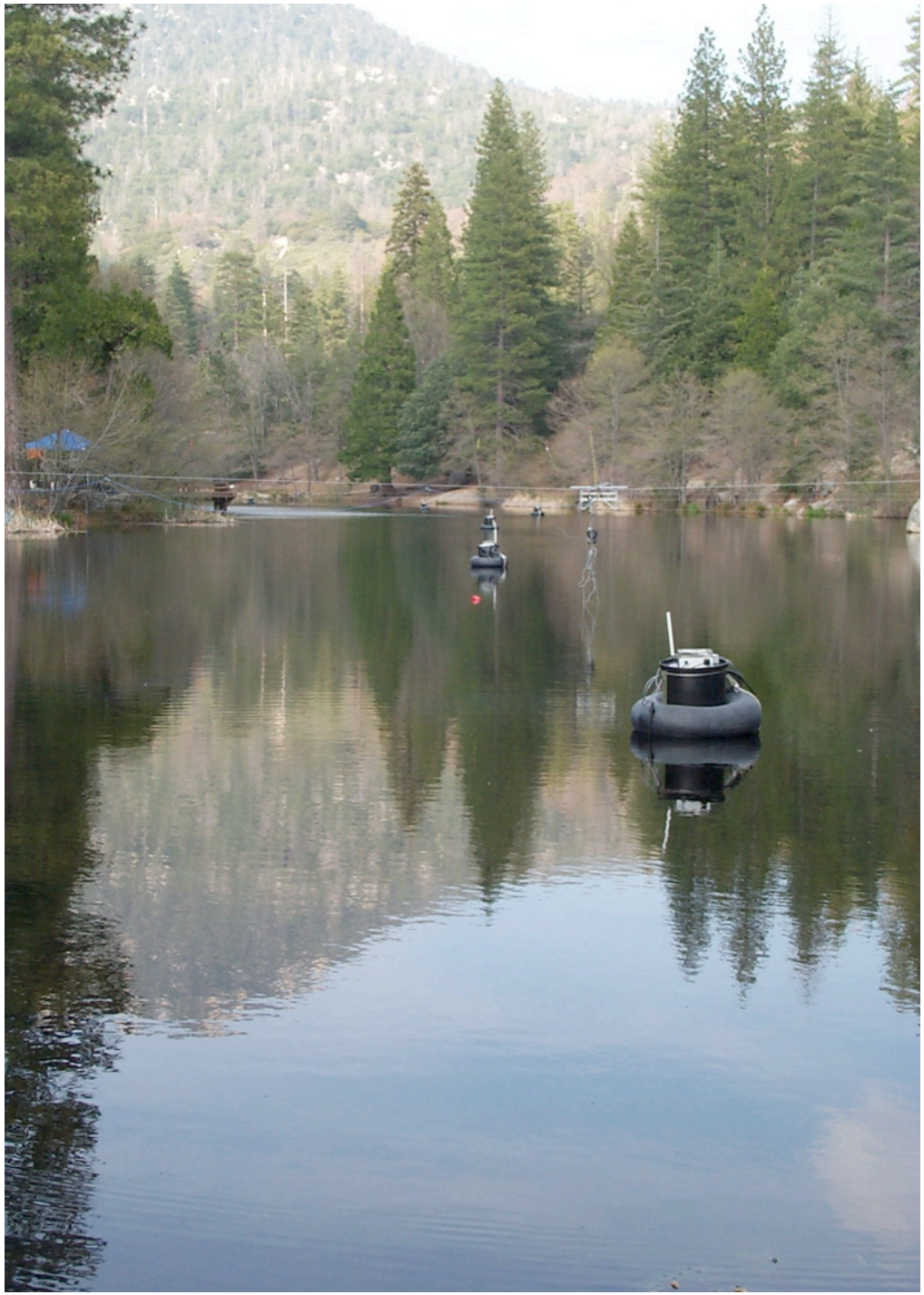
```
> fit = lm(weight-poly(length,3),data=waccamaw)
> summary(fit)

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

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?

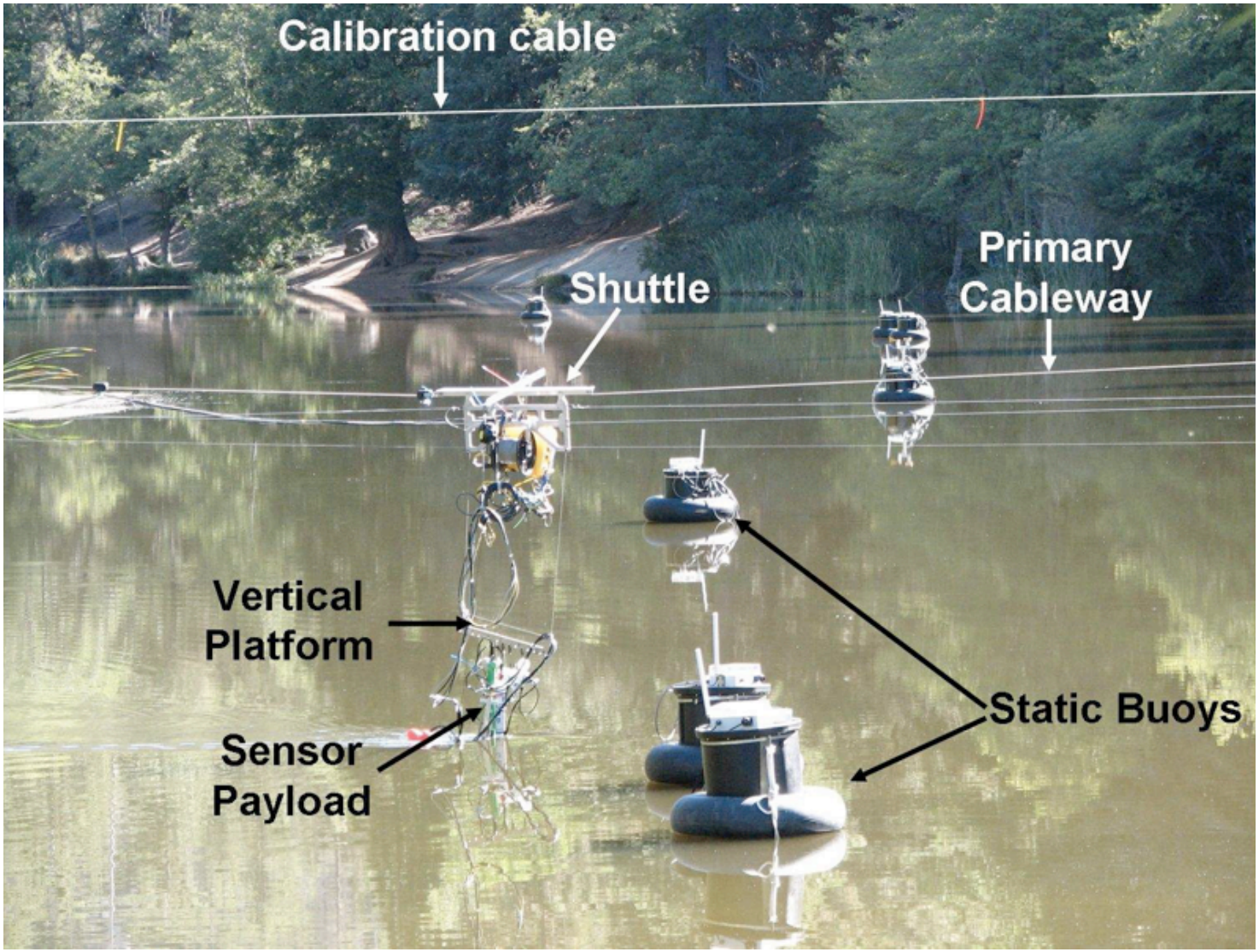


Dynamics of an urban lake

This is Lake Fulmor, located in the San Jacinto Mountains; it is adjacent to James Reserve, part of the UC Natural Reserve System

James Reserve runs several testbeds for embedded sensing systems; sensors have been deployed throughout the area to study microclimates, plant phenology, climate change, you name it

The data we will study come from a week-long deployment that attempted to assess the dynamics of various biological and chemical processes in the lake


Calibration cable
Shuttle
Primary
Cableway
Vertical
Platform
Sensor
Payload
Static Buoys

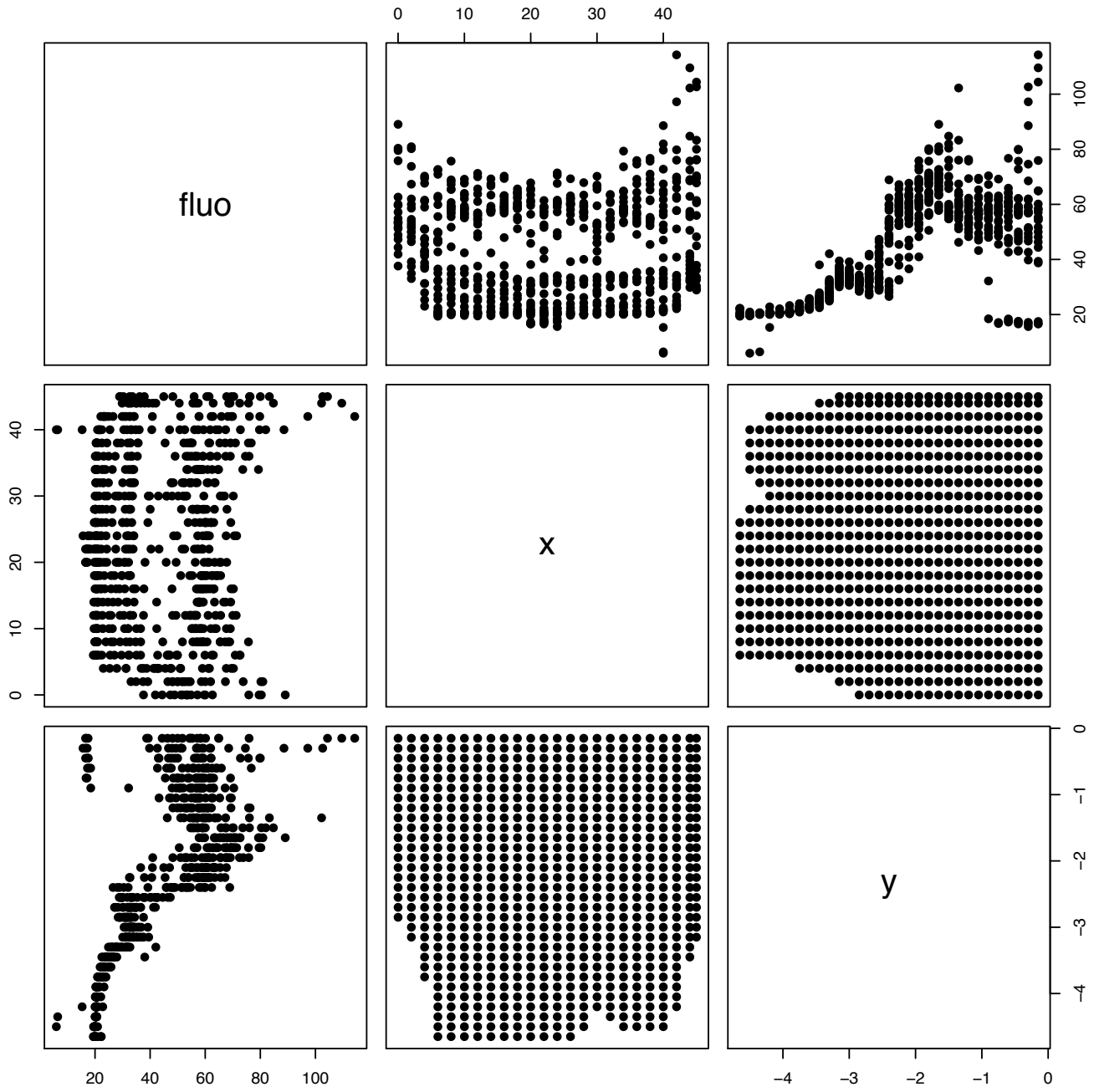
Dynamics of an urban lake

The measurements were collected by a robotic sensing system, producing a vertical profile of about 10 different kinds of measurements

The robot itself consisted of two parts: A shuttle that rode along a tether stretched across the lake, and a sensor pack that could be lowered into the water at any depth

Over the course of an hour, the robot visited 685 different locations; we will focus on data from the fluorometer, a device that responds to chlorophyll levels... here are some plots

```
# load the data
> lake = read.csv(url("http://www.stat.ucla.edu/~cocteau/lake.csv"),head=T)
> names(lake)
[1] "fluo" "x"    "y"

# our friend the scatterplot matrix
> pairs(lake,pch=16)

# something new
> library(lattice)
> cloud(fluo~x+y,data=lake)
> wireframe(fluo~x+y,data=lake,drape=T)
```

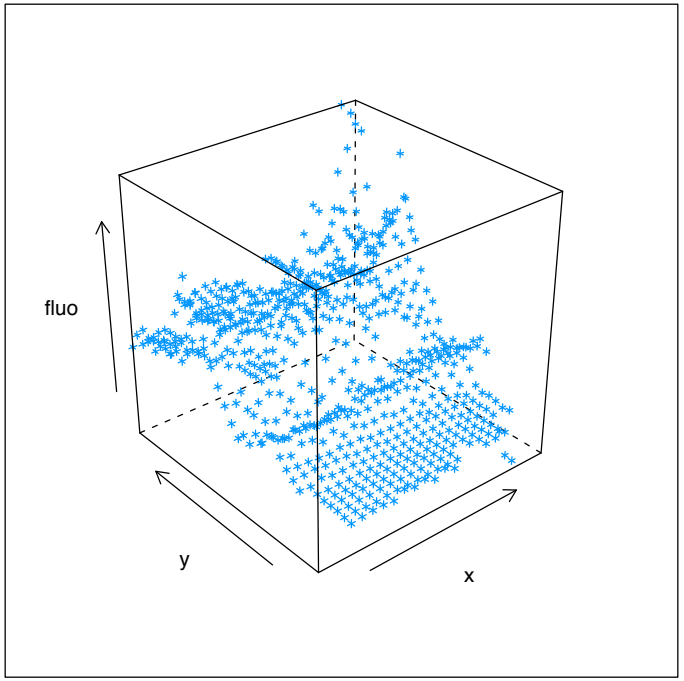

fluo
x
y
0
10
20
30
40
20
40
60
80
100
−4
−3
−2
−1
0

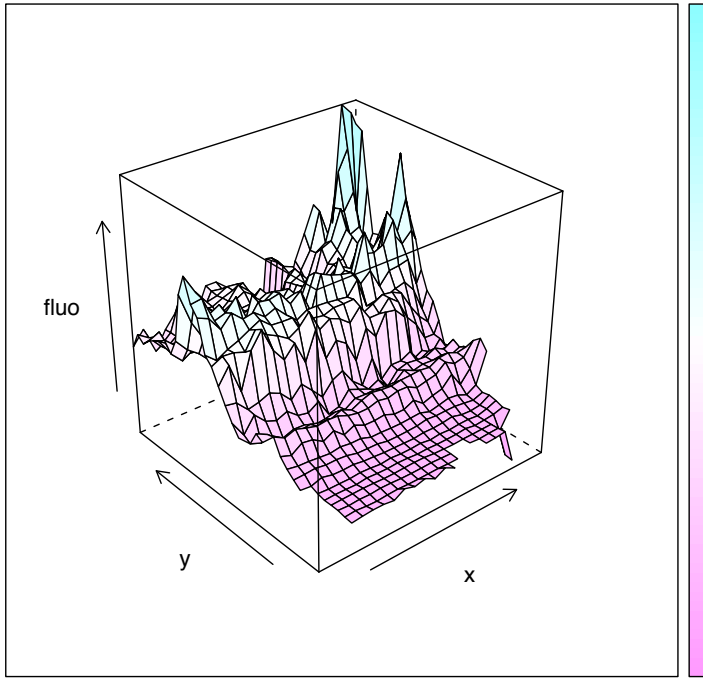
Dynamics of an urban lake

The variable x denotes distance along the lake, and y represents depth below the surface; we see from the pairs plot that the locations were part of a (mostly) regular grid

The plots on the right show that the chlorophyll levels start low near the surface, peak at about 2m deep, and 100 then trail off as you approach the bottom of the lake

It turns out that this pattern is chlorophyll is expected, in that it closely matches the thermal profile in the lake

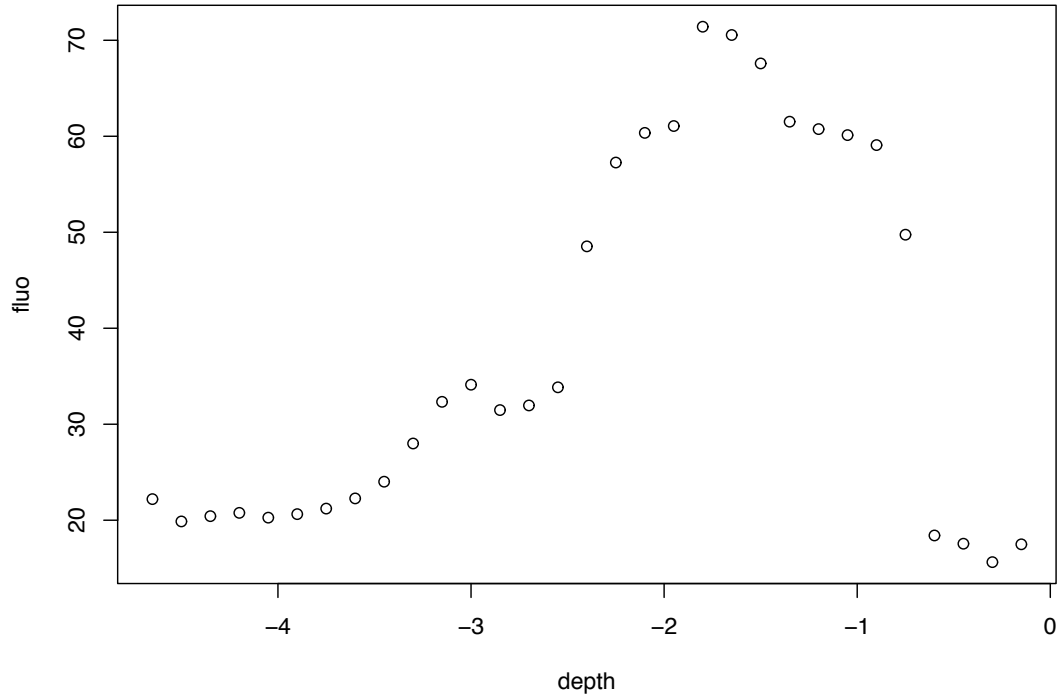




A single slice

We'll start by considering simple smoothing; therefore, we'll select a single slice through the data, the chlorophyll profile corresponding to x=24m across the lake

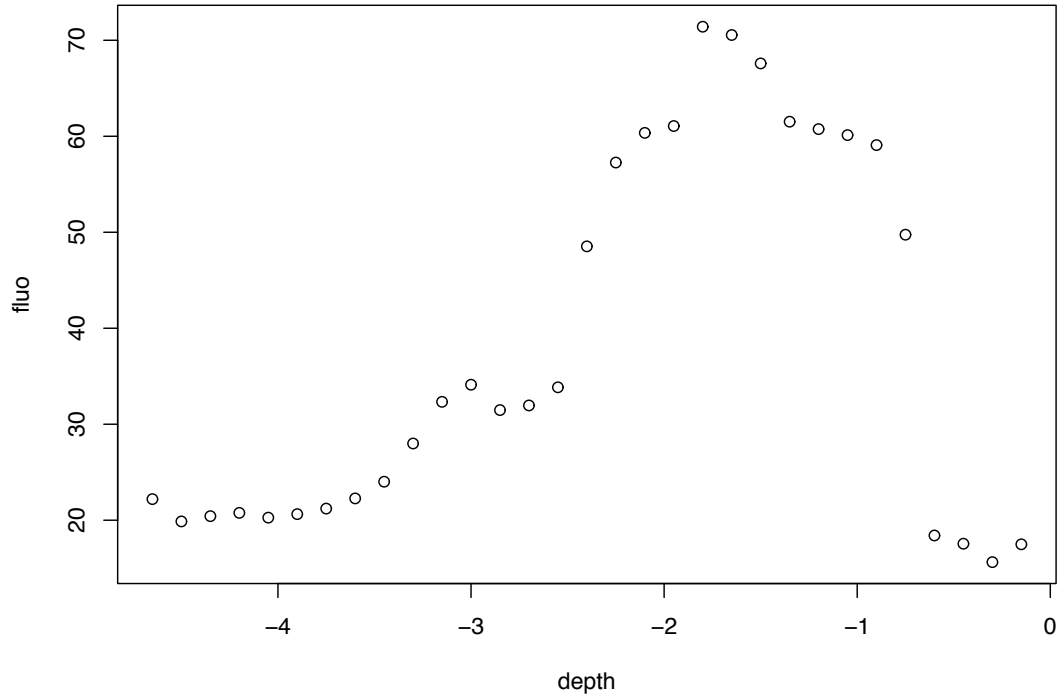
The robot was lowered to 31 different depths, starting at 15cm below the surface to 465cm in steps of 15cm



Smoothing

How do we model data like these? Intuitively the goal seems clear in that we'd like to run a smooth curve of some kind through the "center" of these data

It seems unlikely that a linear or quadratic or even a cubic polynomial will be flexible enough -- What principle should we leverage here to come up with this curve?



Parametric...

The regression models we have seen so far are defined by a set of parameters; the relationship between the predictors and the response is governed by, say, a polynomial function of the input data

Over the last few lectures, we saw that many of the common analysis methodologies in this context focus on the presence or absence of terms in the model, their sign and their relative magnitude

... vs. Nonparametric

In the statistics literature, smoothing is often referred to as “nonparametric” -- In short, the object of our attention shifts from parameters to features of the smooth curve

We are interested in looking at the shape of the curve, in counting modes or bumps, in examining regions of sharp change -- You might say that our analysis becomes a lot more visual

Rather than “seeing” the model through a table of parameters, we instead look at the curve itself in a lot more detail; with this shift, our underlying model also changes

Smoothing (Nonparametric regression)

Let's assume that our underlying model is of the form

$$y = f(x) + \epsilon$$

for x in some region $\mathcal{X}$, where again we assume normal errors with mean zero and variance σ^2; as usual we have observations $(x_1, y_1), \ldots, (x_n, y_n)$

In the case of the lake slice, we might take $\mathcal{X} = [-4.65, 0]$, depths ranging from the surface to the bottom of the lake; if we were to model the entire set of data, then $\mathcal{X}$ could be the 2-dimensional region that stretches across the lake in one dimension and from the surface to the lake bottom in the other (the upper boundary being straight, the lower boundary jagged)

If we don't make any assumptions on $f(x)$, there's nothing to anchor an estimation problem on; that is, if we cannot relate $f(x_1)$ to $f(x_2)$ for two points $x_1, x_2 \in \mathcal{X}$, then the best we can do is estimate $\widehat{f}(x_i) = y_i$

Smoothing (Nonparametric regression)

For the normal linear model, we assumed a particular parametric form for f ; that is, we said that the predictors and the response were related via a linear equation

In the case of smoothing, our implicit assumption has to do with the functional characteristics of f ; one common assumption of this type is that f is, well, smooth

We can make this formal by saying that f has two continuous derivatives or something like that; smoothness means that if x_1 and x_2 are near each other $f(x_1)$ and $f(x_2)$ should be close

Smoothing via local polynomials

It also provides some guidance about how we might create a good smoother: Recall from Taylor's Theorem that any (sufficiently) smooth function can be written locally as a polynomial

That is, given some point $x_0 \in \mathcal{X}$, we have the approximation

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \\ &\quad \frac{1}{2} f''(x_0)(x - x_0)^2 + \frac{1}{6} f'''(x_0)(x - x_0)^3 + \cdots \end{aligned}$$

This, then will be the basis of a class of smoothers called local polynomials

An aside

Note that Taylor's theorem could be used to justify higher and higher degree polynomials -- In a theoretical sense, the more terms we add the better the approximation to the underlying function

Keep in mind, however, that there's a difference between theoretical approximation and what we can do with data -- We'll see a little later what will happen when we try ramping up the degree of an approximating polynomial

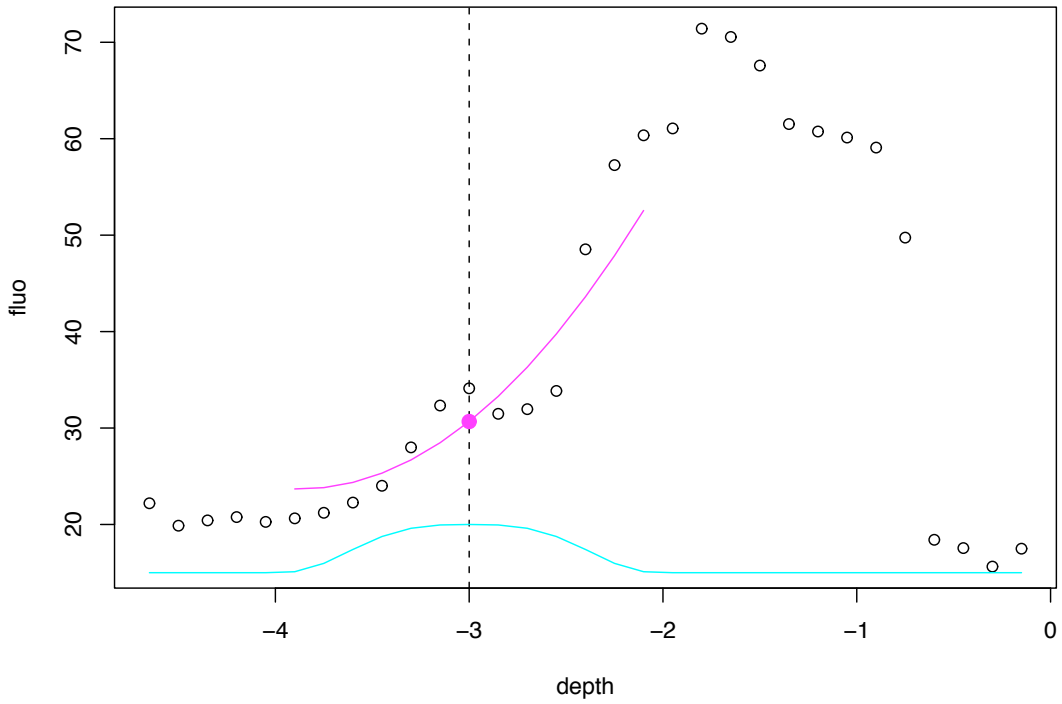
And it isn't pretty!

An example

Rather than fit these data by working with higher and higher degree polynomials, we might instead fit the polynomials locally

Here, for example, we are interested in estimating at the value of $f(x)$ at $x = -3$, or 3 meters below the surface

At the right, we perform a "local" quadratic fit that weights data according to their distance from -3

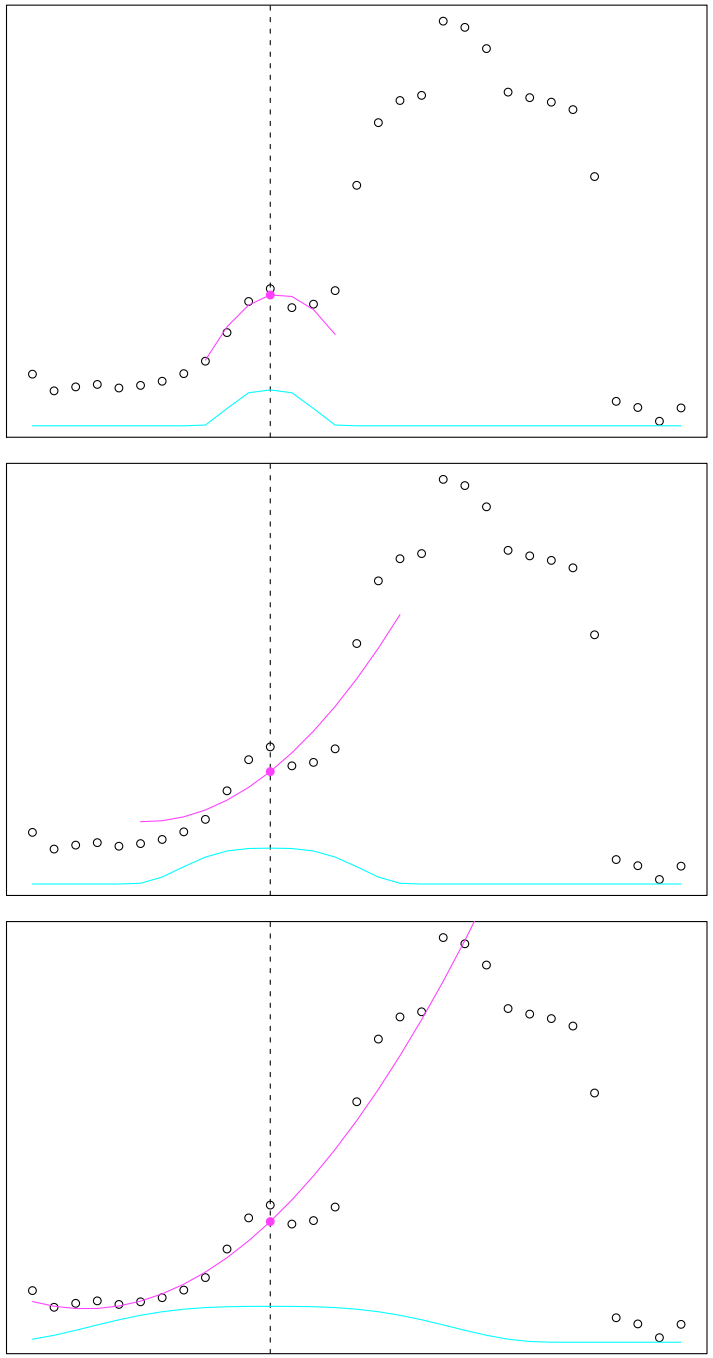


Local polynomials

To accomplish this “local” fit, we have to introduce the notion of a weighted regression; simply, instead of solving the OLS criterion, we are given weights for each data point

In the case of a local polynomial, these weights are biggest near the point we’d like to predict and then tend to zero as you move farther away; at the right we plot a few sets of weights and the associated local fits

In general, we have defined these weights using a kernel function; the wider the kernel’s “bandwidth,” the more points are included in the regression locally



Weighted regressions

To create these fits, we need the notion of a weighted regression; let's again consider the case of fitting a quadratic in the neighborhood of $x_0 = -3$ meters; for each $i = 1, \ldots, n$, let $w_i \geq 0$ be a weight (say, the cyan colored lines from the previous slide)

Then, we seek to solve not the OLS criterion but instead the weighted criterion

$$\sum_{i=1}^{n} w_i \left[y_i - \beta_0 - \beta_1 (x - x_0) - \beta_2 (x - x_0)^2 \right]^2$$

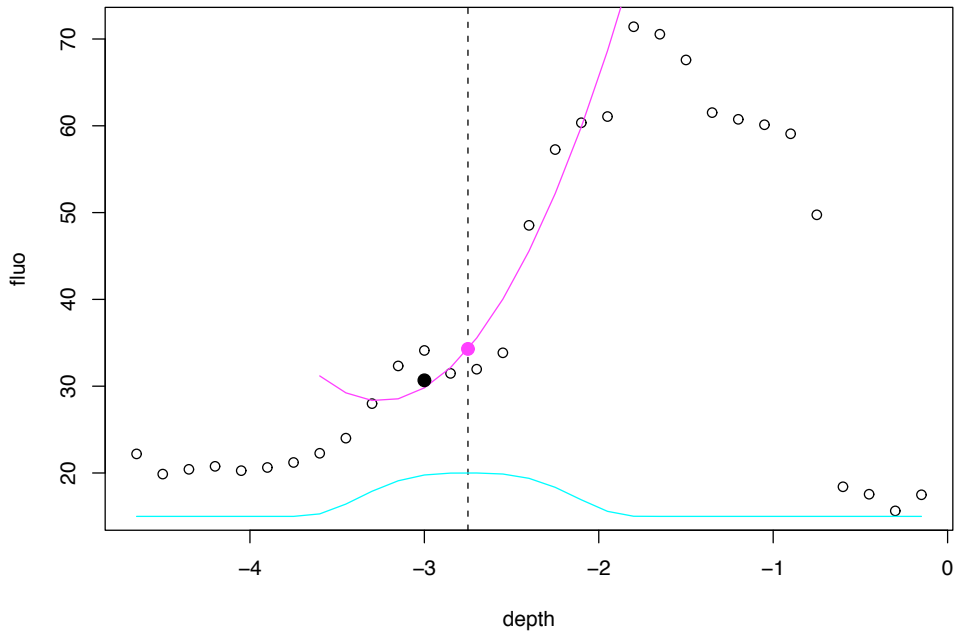
which gives $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$; we then take our estimate of $f(x_0)$ to be $\widehat{\beta}_0$

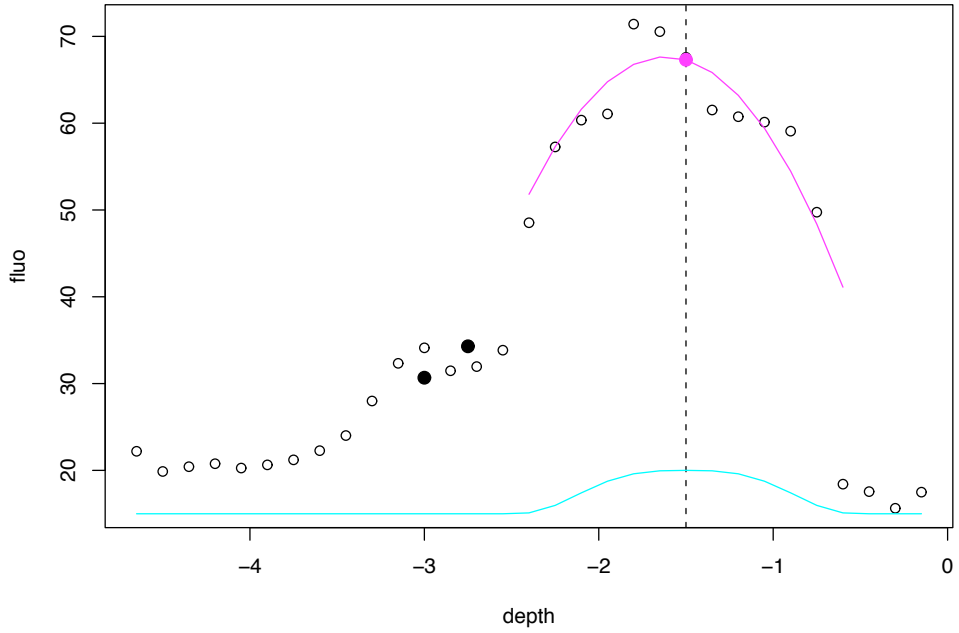
An example

We can continue the process and make predictions for any depth; here's -2.75 and -1.5, keeping the previous predictions marked in black

Again, keep in mind the basic character of the fit; a polynomial is being fit locally using the weights in cyan

Notice also that we're making predictions at any point; this procedure works whether or not we have an observation at x_0



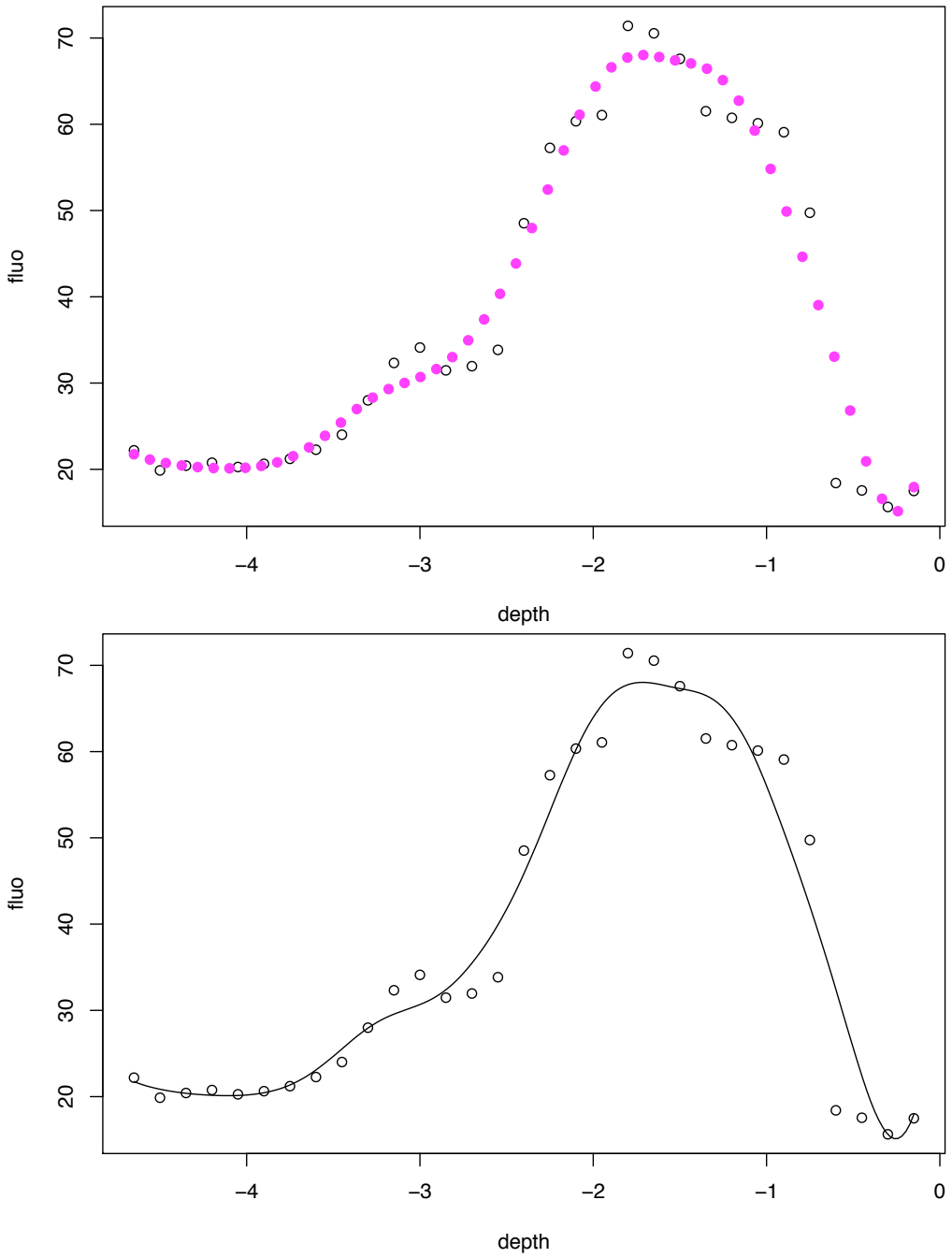


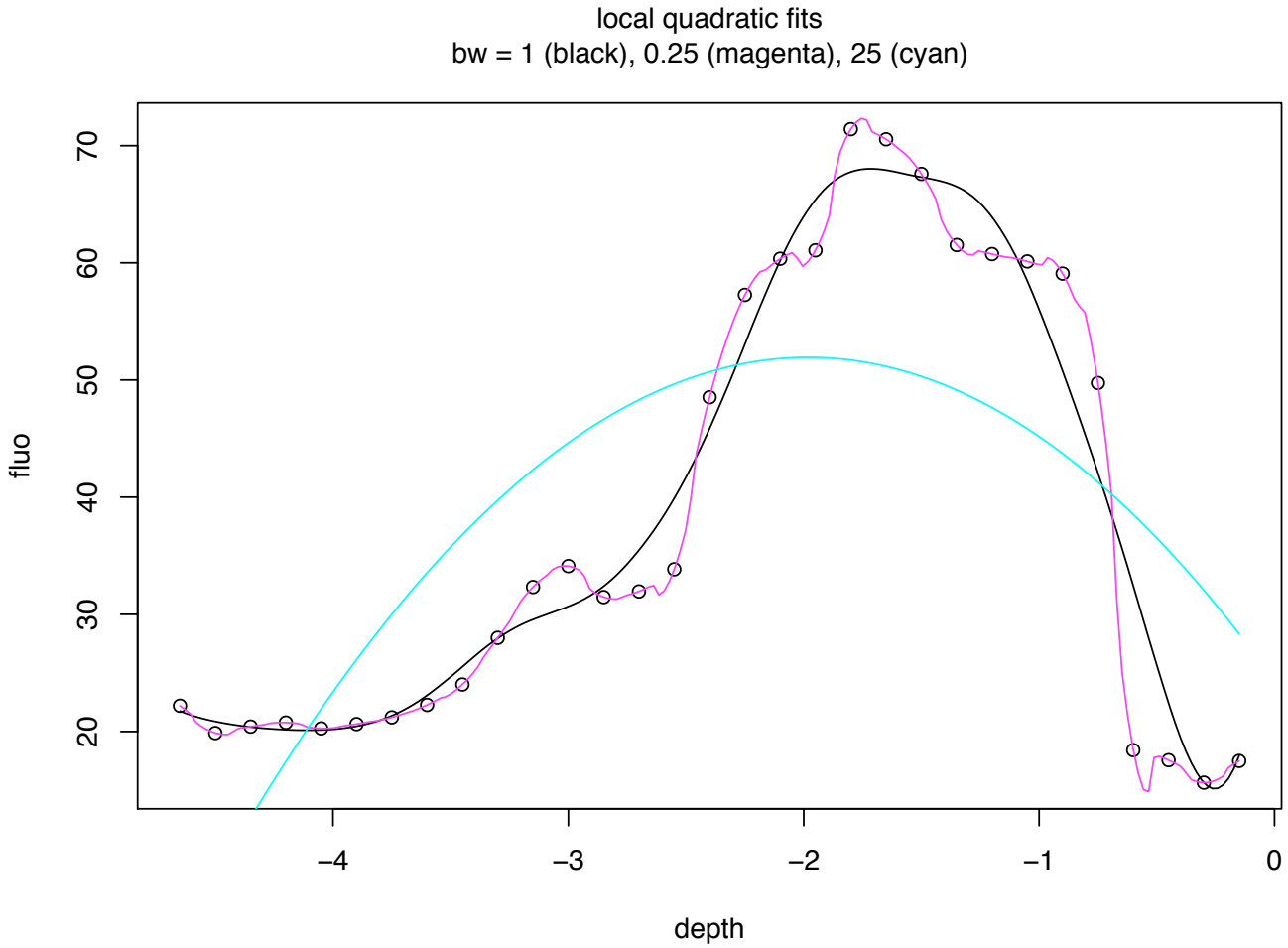
Local polynomials

And voila! If we continue this process at a number of points we create a smooth curve (why does the curve have to be smooth?)

The fit at the right uses a “bandwidth” of 1; as we vary the bandwidth, we get smoother or wigglier fits

If the bandwidth is chosen very large, then we are essentially fitting a quadratic polynomial; if it is very small, we “interpolate the data”




local quadratic fits
bw = 1 (black), 0.25 (magenta), 25 (cyan)
fluo
20
30
40
50
60
70
−4
−3
−2
−1
0
depth

```
# first, load in a function to compute the weights; it is called tric()

source(url("http://www.stat.ucla.edu/~cocteau/tric.R"))

# now, focus on the point 24 meters across the lake
# recall that the condition before the comma selects rows;
# here we take just those rows that correspond to x=24

slice = lake[lake$x==24,]
plot(slice$y,slice$fluo,xlab="depth",ylab="fluo")

# now, let's fit a local polynmoial at x0=-3; the weight
# function takes arguments of your data, the point x0 and
# the bandwidth (which here is 1)

x0=-3
weights = tric(slice$y,x0,1)
plot(slice$y,weights)

# now do the local fit

plot(slice$y,slice$fluo,xlab="depth",ylab="fluo")

fit = lm(fluo~I(y)+I(y^2),weight=tric(y,x0,1),data=slice)
points(x0,predict(fit,newdata=data.frame(y=x0)),pch=20,col=6)

# or do a series of them

x0 = seq(-4.65,-0.15,len=50)

for(i in 1:50){

   fit = lm(fluo~I(y)+I(y^2),weight=tric(y,x0[i],1),data=slice)
   points(x0[i],predict(fit,newdata=data.frame(y=x0[i])),pch=20,col=6)
}
```

Binning?

In some sense, this is a continuous version of the **binning** that Galton performed when he was relating the heights of children to mid-parents -- Galton, you'll recall, looked at averages of children's heights across (mid-)parents who had heights in separate groups

The mechanics behind binning involves (essentially) **fitting models piecewise**, carving the data into pieces and fitting a separate model in each -- We'll see more piecewise fits later...

In Galton's case, these binned averages followed (essentially) a straight line, providing evidence that his model was reasonable -- Interestingly, Galton appealed to a "nonparametric" tool when examining the goodness of his model

Defining locality

Right, the idea is pretty clear (although there are lots and lots of questions about practicalities); but first, let's figure out how best to define "local"

What properties do we want from the weights? Well, we need them to be centered at x_0 ; it's probably good if they are symmetric around that point also; they should decrease the farther you are from

A common choice is the so-called tri-cube function (gosh, I wonder why it's called that?)

$$w(u) = \begin{cases} (1 - |u|^3)^3 & |u| < 1 \\ 0 & \text{else} \end{cases}$$

and with it, given a bandwidth h , you define weights

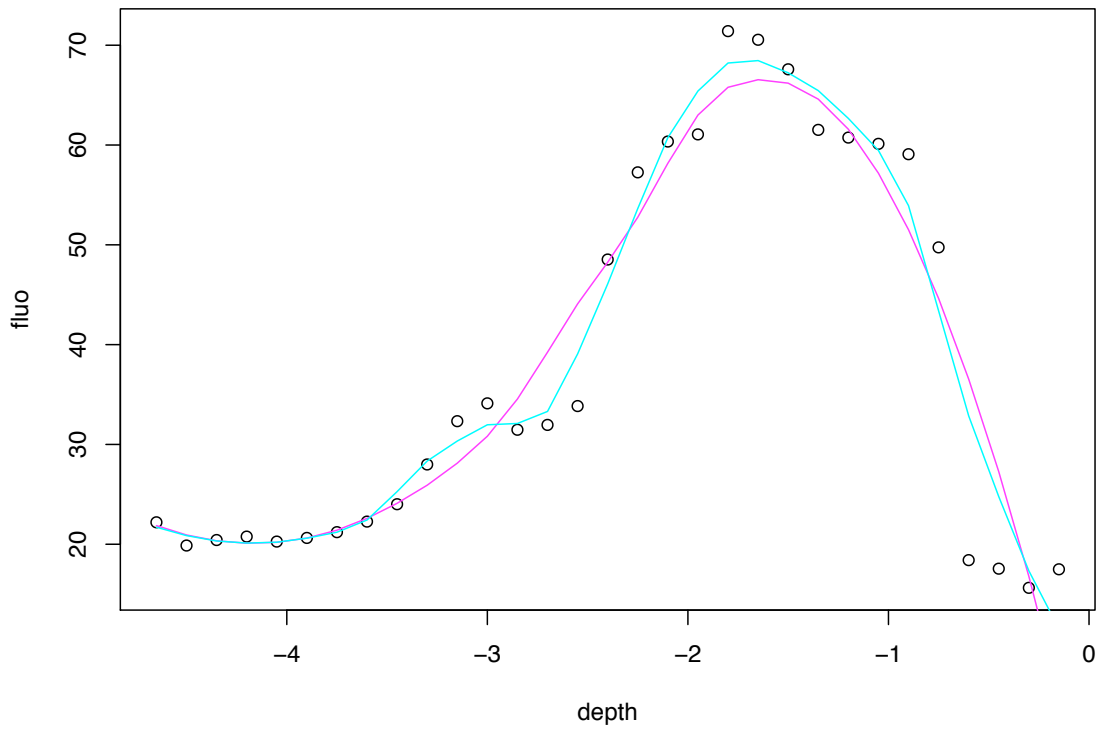
$$w_i = w\left(\frac{x_i - x_0}{h}\right)$$

Local polynomials

`loess()` exposes two ways to control the "amount" of smoothing

1. The first is a parameter called `span`; it is a number between 0 and 1 and determines the fraction of the data that should be included in the fit (using the tri-cube function, notice that data points beyond a certain distance are left out or assigned zero weight)

2. The second parameter is called `enp.target`; this translates the amount of smoothing into an equivalent number of parameters or degrees of freedom (this is essentially the same trick we played with ridge regression when thinking about the amount of shrinkage taking place)

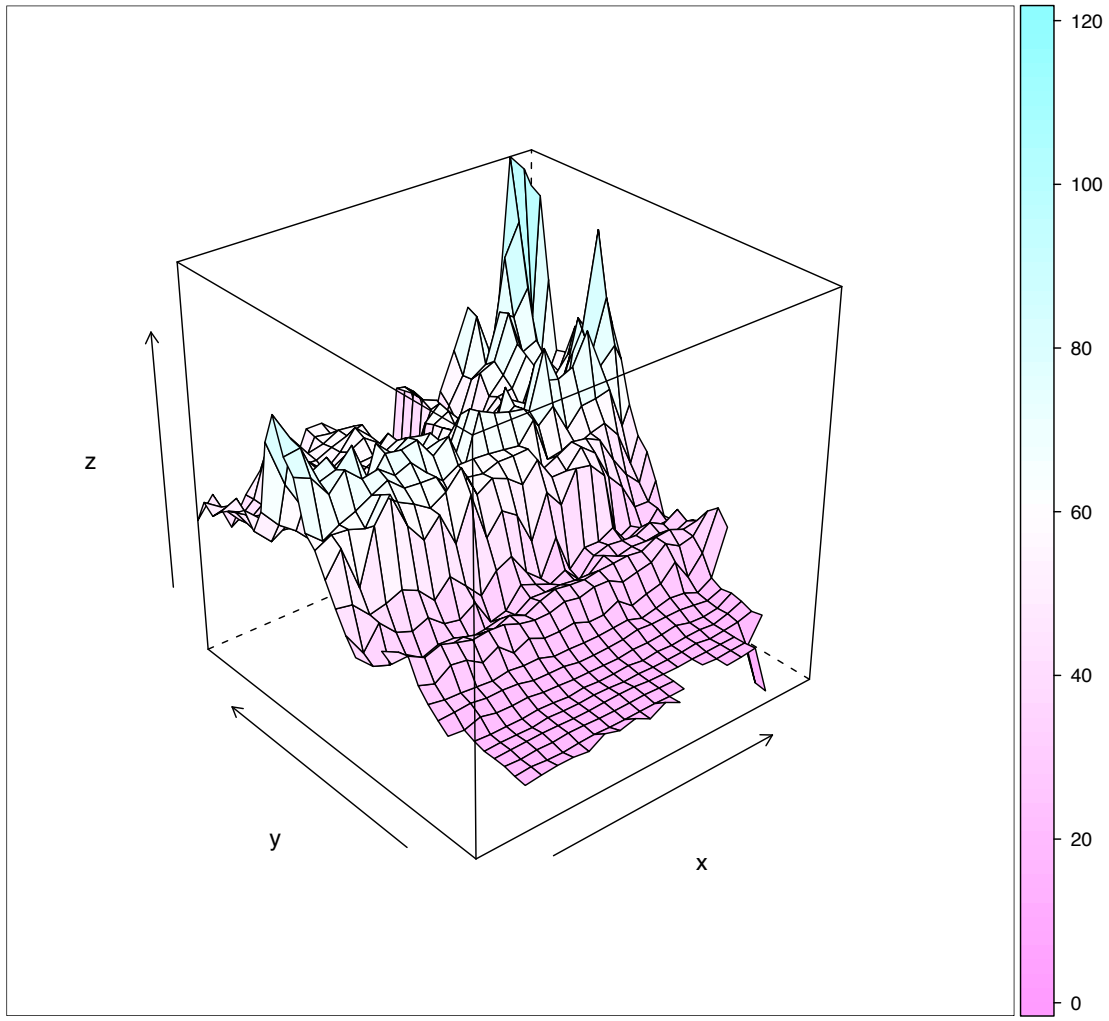



```
slice = lake[lake$x==24,]
plot(slice$y,slice$fluo,xlab="depth",ylab="fluo")

# fit with about 5 degrees of freedom

fit = loess(fluo~y,data=slice,enp.target=5)
lines(slice$y,predict(fit),col=6)

# fit with about 10 degrees of freedom

fit = loess(fluo~y,data=slice,enp.target=10)
lines(slice$y,predict(fit),col=5)
```

Multivariate modeling

Here we have the chlorophyll values in both depth and distance along the transect

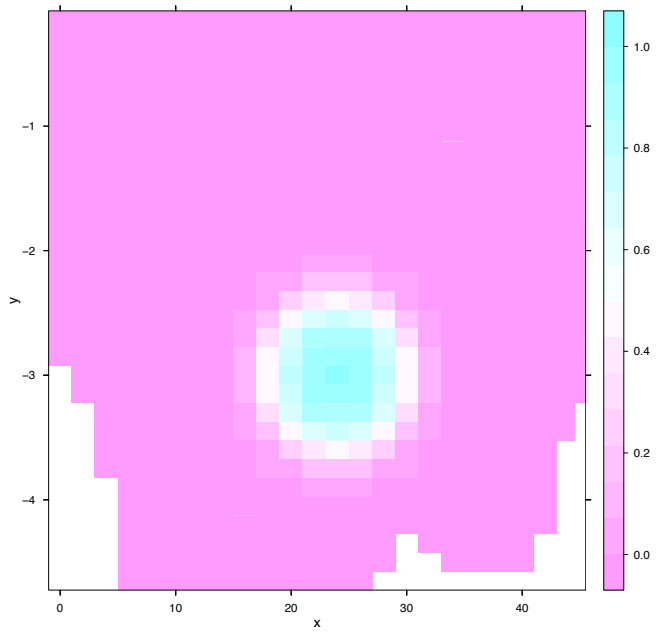
Applying the local polynomial ideas here, we select a location (x,y) at which we want to make a prediction and fit a quadratic polynomial (now in both x and y)...

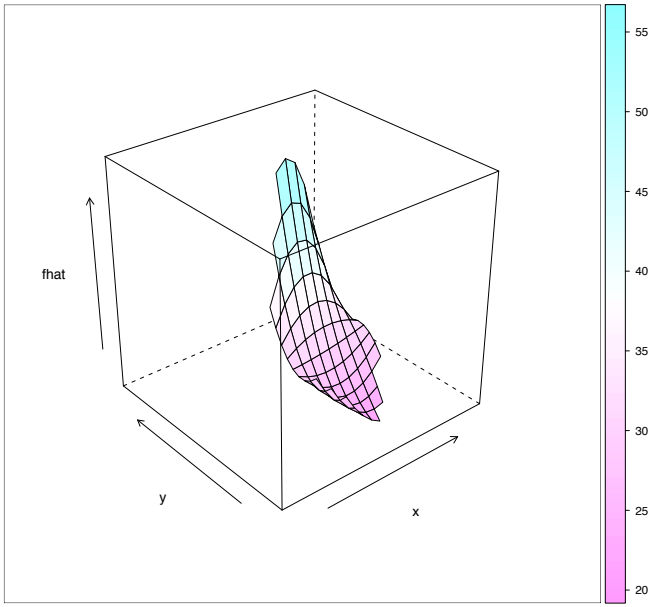


Multivariate modeling

Here we choose the point (2.4,-3) and plot the tri-cube kernel centered at this point

This defines a region for us and the resulting polynomial fit in the region is given in the lower panel

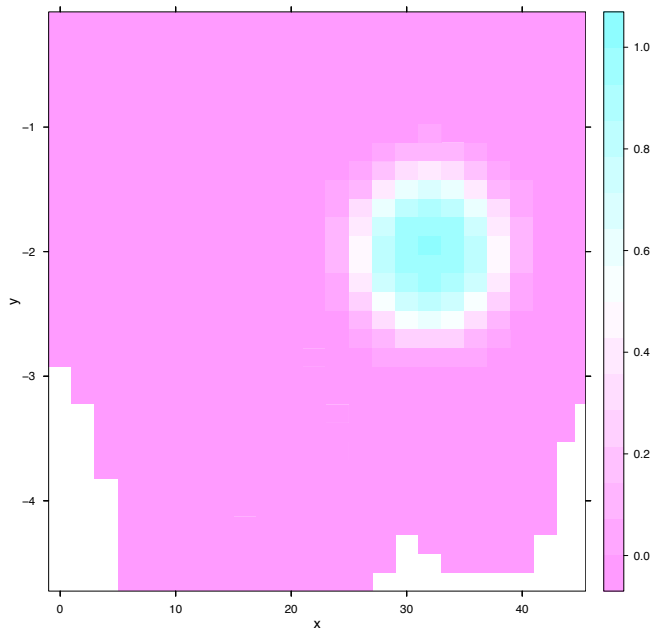


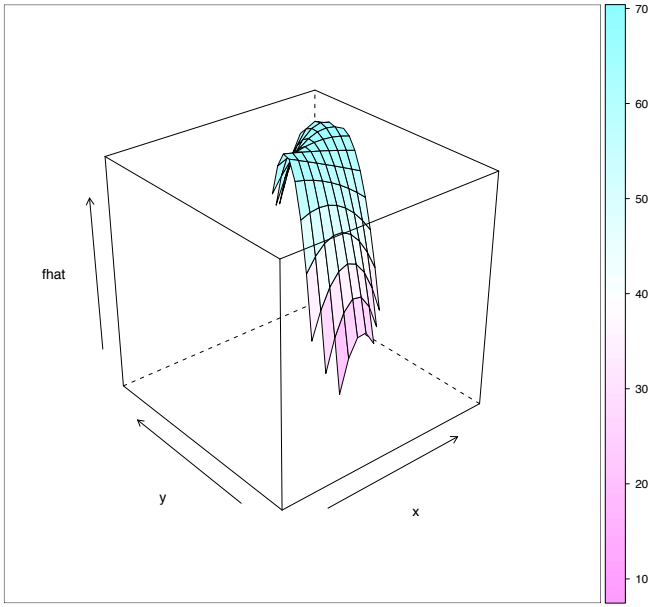


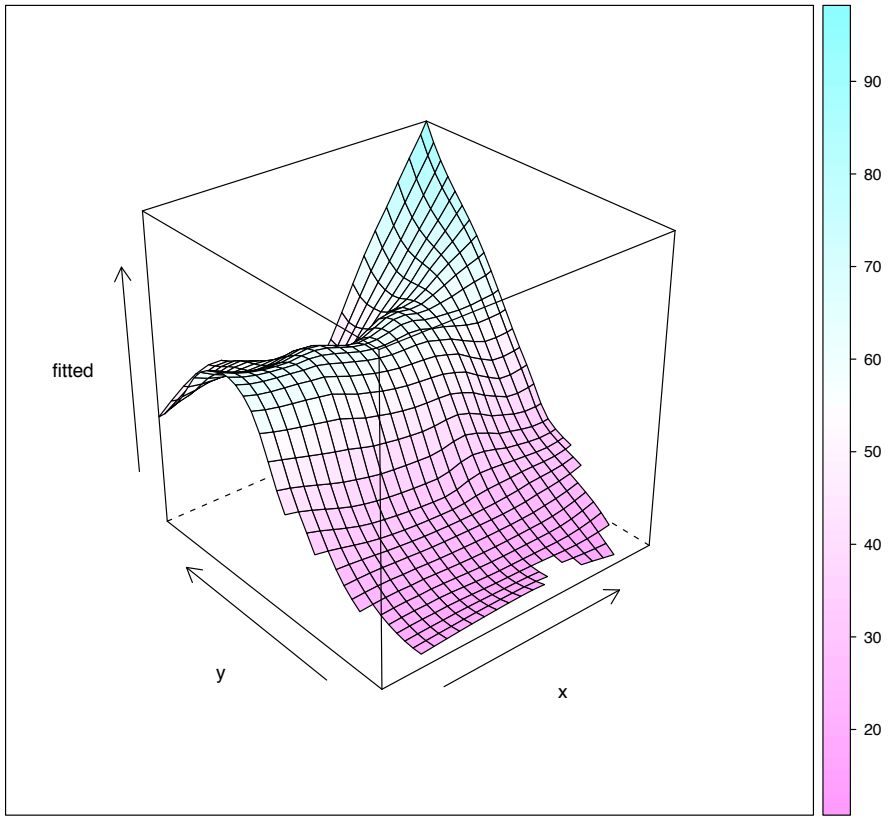
Multivariate modeling

Here we choose the point (3.2,-2) and plot the tri-cube kernel centered at this point

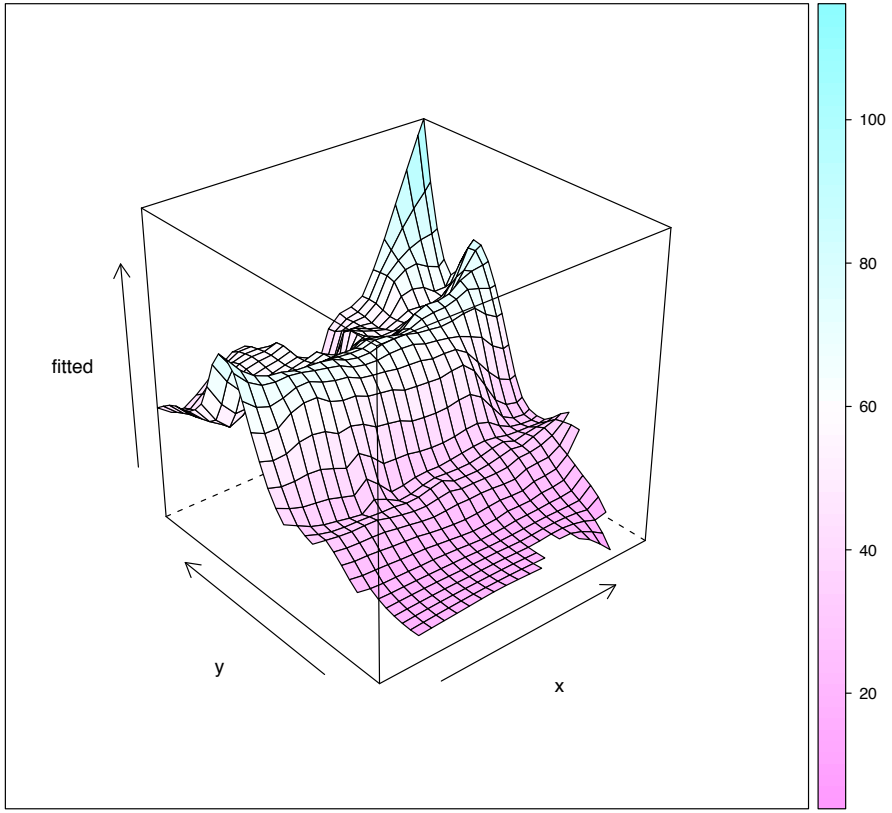
This defines a region for us and the resulting polynomial fit in the region is given in the lower panel








```
# load up graphics library to make lovely 3d plots

library(lattice)

# fit a local polynomial in two dimensions using the full
# lake data set

fit = loess(fluo~x+y,data=lake,enp.target=25)

# create a new data frame to plot with and plot the fit

newlake = data.frame(x=lake$x,y=lake$y,fitted=predict(fit))
wireframe(fitted~x+y,data=newlake,drape=T)
```




```
# load up graphics library to make lovely 3d plots

library(lattice)

# fit a local polynomial in two dimensions using the full
# lake data set; this time with 100 degrees of freedom

fit = loess(fluo~x+y,data=lake,enp.target=100)

# create a new data frame to plot with and plot the fit

newlake = data.frame(x=lake$x,y=lake$y,fitted=predict(fit))
wireframe(fitted~x+y,data=newlake,drape=T)
```