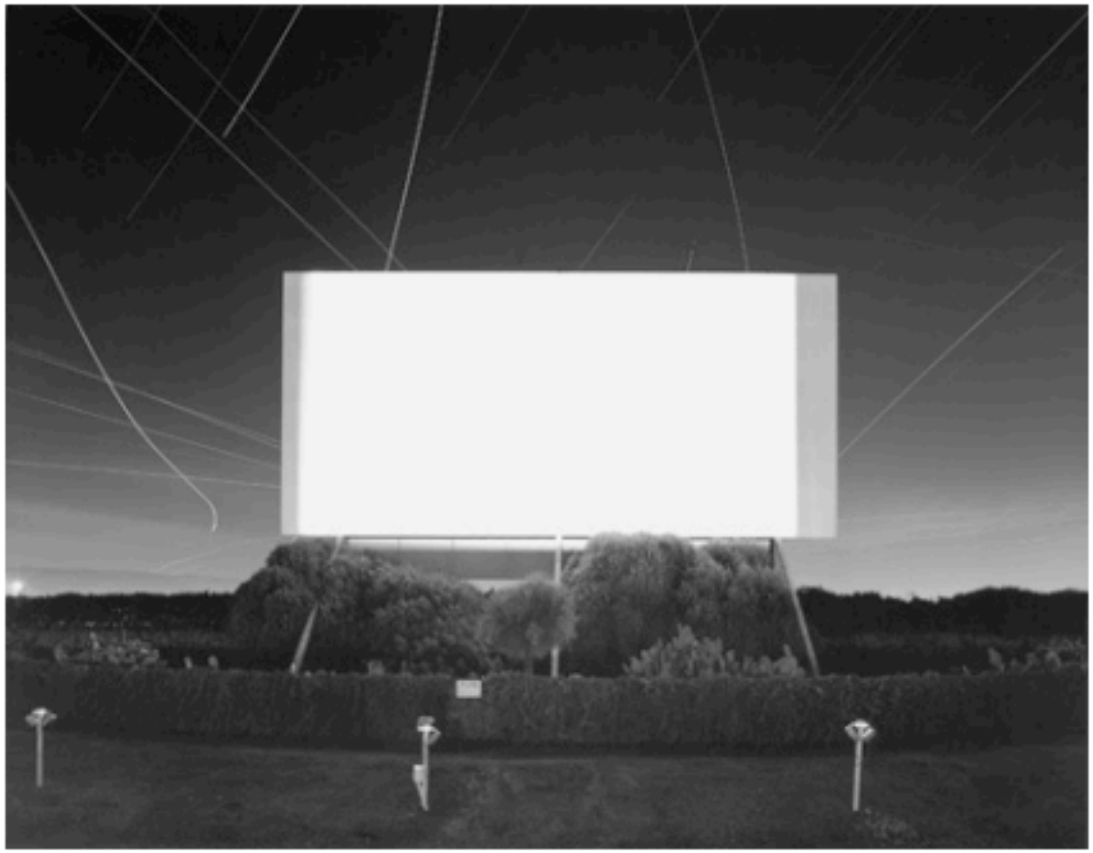


Lecture 14: Regression

Sprawling places

Decision trees are powerful tools for data analysis -- Consider one use of your CDC data -- A researcher at JHU wanted to see if there was a connection between urban sprawl obesity

The original CDC BRFSS survey was augmented by a researcher at JHU with a variable that indicates how sprawling a city is; the researcher then applied a technique called logistic regression to study the effect of sprawling places

While that is beyond our reach at a technical level, we can still get a sense of the relationships between obesity and these variables using a tree mode

The variables

The response variable is 1/0 whether someone is obese or not; and obesity was defined via the respondents BMI -- 25 or over is declared obese and in the sample about 20% of the respondents were obese

The potential explanatory variables are

sprawl we've seen this in our lab

gender 1/0 for female or not

afam 1/0 for African American or not

hisp 1/0 for Hispanic or not

age in years

income 1: < 10K; 2: 10-15K; 3: 15-20K; 4: 20-25K; 5: 25-25K; 6: 25-50K; 7:50-75K; 8: 75K+)

education 1: never attended school or only kindergarten; 2: grades 1-8; 3: grades 9-11; 4: grade 12 or GED; 5: college 1-3 years; 6: 4 or more years of college

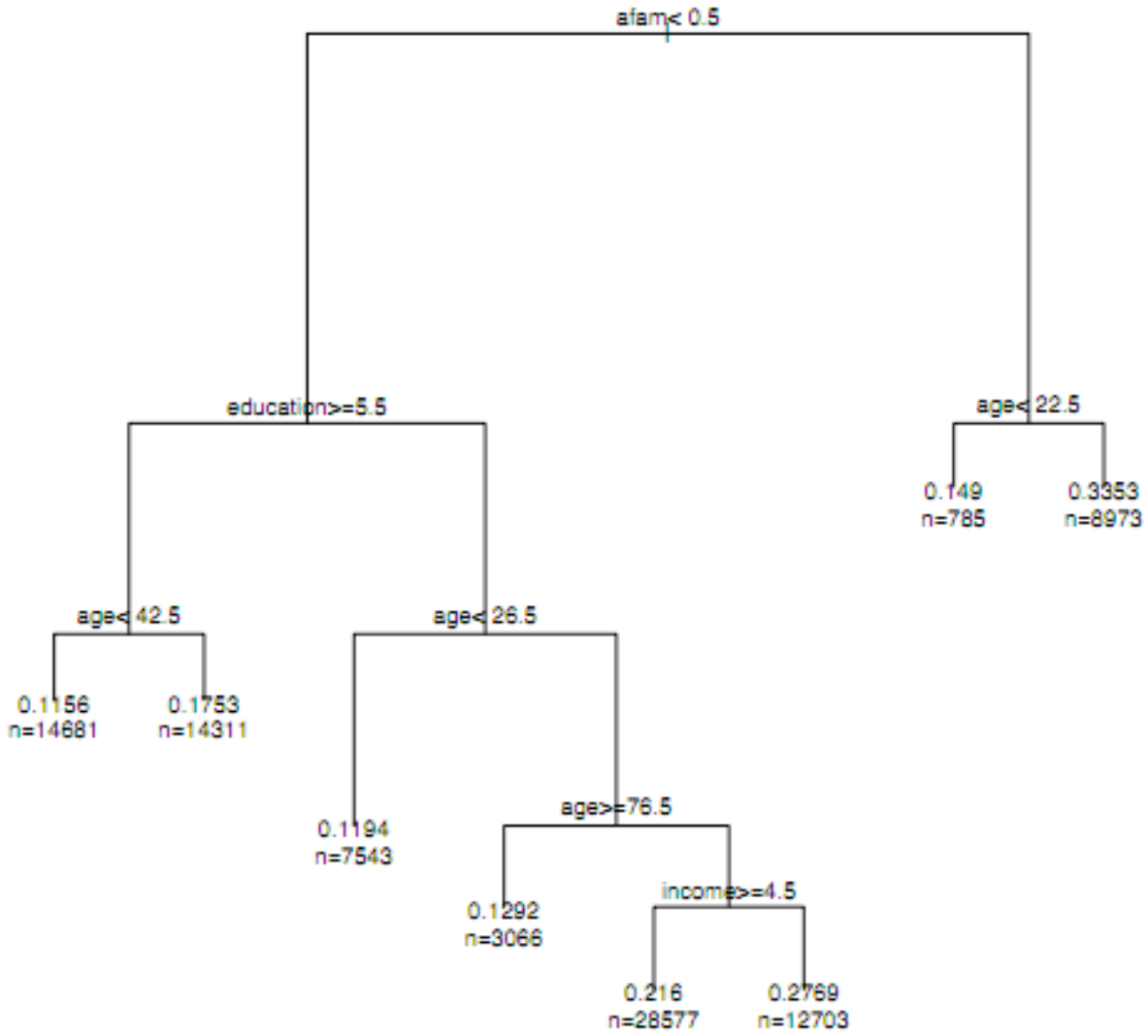
Some caveats

The trees we have been growing are elegant data analytic tools; but **they are not the final word in modeling -- not by a long shot**

They are relatively coarse in that they carve up your variables in to boxes; "smoother" behavior is hard to capture with these boxes -- there are ways to smooth them up a bit, but that's best left to an advanced class

We have also only used these models for 0/1 data, where the value at the leaves is a majority vote; if we have continuous data, we would want to minimize something other than the misclassification error

That something else is the subject of our next topic...


afam< 0.5
education>=5.5
age< 22.5
0.149
n=785
0.3353
n=8973
age< 42.5
age< 26.5
0.1156
n=14681
0.1753
n=14311
0.1194
n=7543
age>=76.5
0.1292
n=3066
income>=4.5
0.216
n=28577
0.2769
n=12703

Linear models

We are going to step out of the trees for the moment and consider a much smaller data set; this will let us audition a few more classical tools for visualizing the relationships between a small number of variables

We will then begin the basic framework for the so-called linear model (regression analysis); we'll start out slow today, just motivating the basic "loss function" involved and next time we'll introduce some statistics

A new example: Mercury contamination

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

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

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

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

Mercury contamination

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

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

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

The New York Times
nytimes.com

March 16, 2005

New Rules Set for Emission of Mercury

By MATTHEW L. WALD

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

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

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

Summaries

Lumber (n=73)

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

Waccamaw River (n=98)

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

Mercury contamination

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

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

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

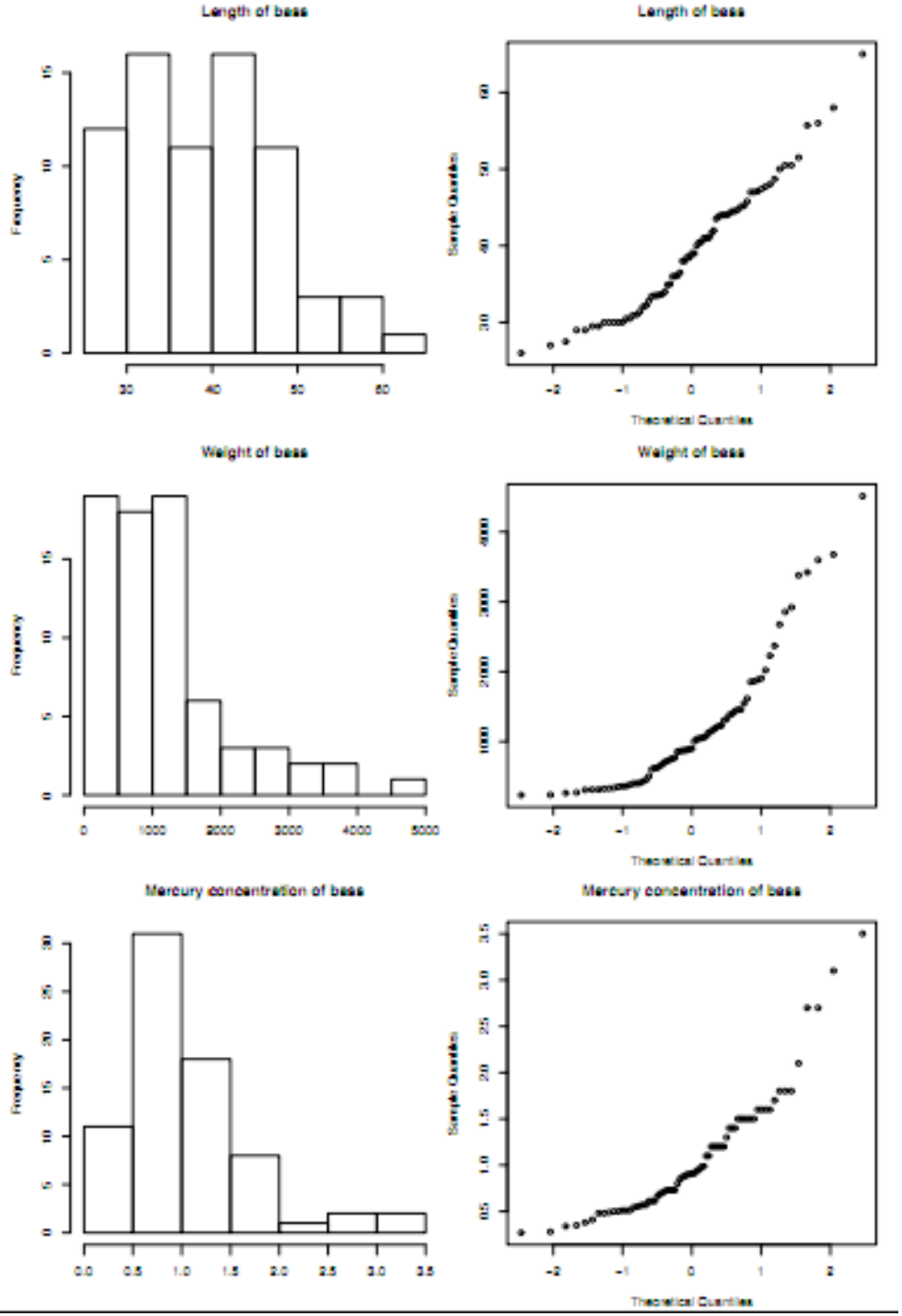
`river = 1, stn=7,...,15`

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

Mercury contamination

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

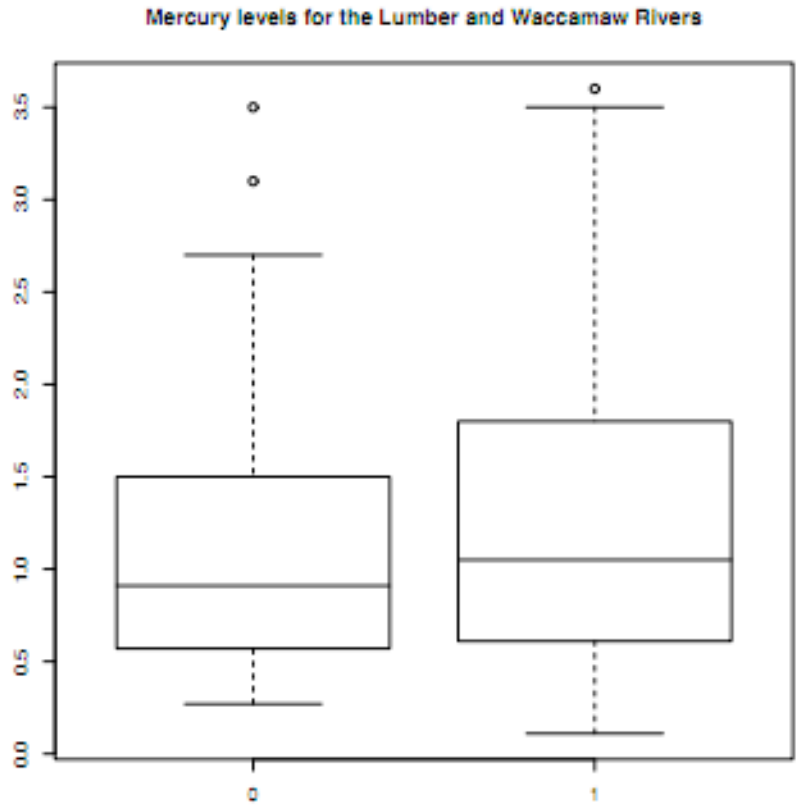
At the right we have our old friends the histogram and the boxplot



Relationships between variables

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

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

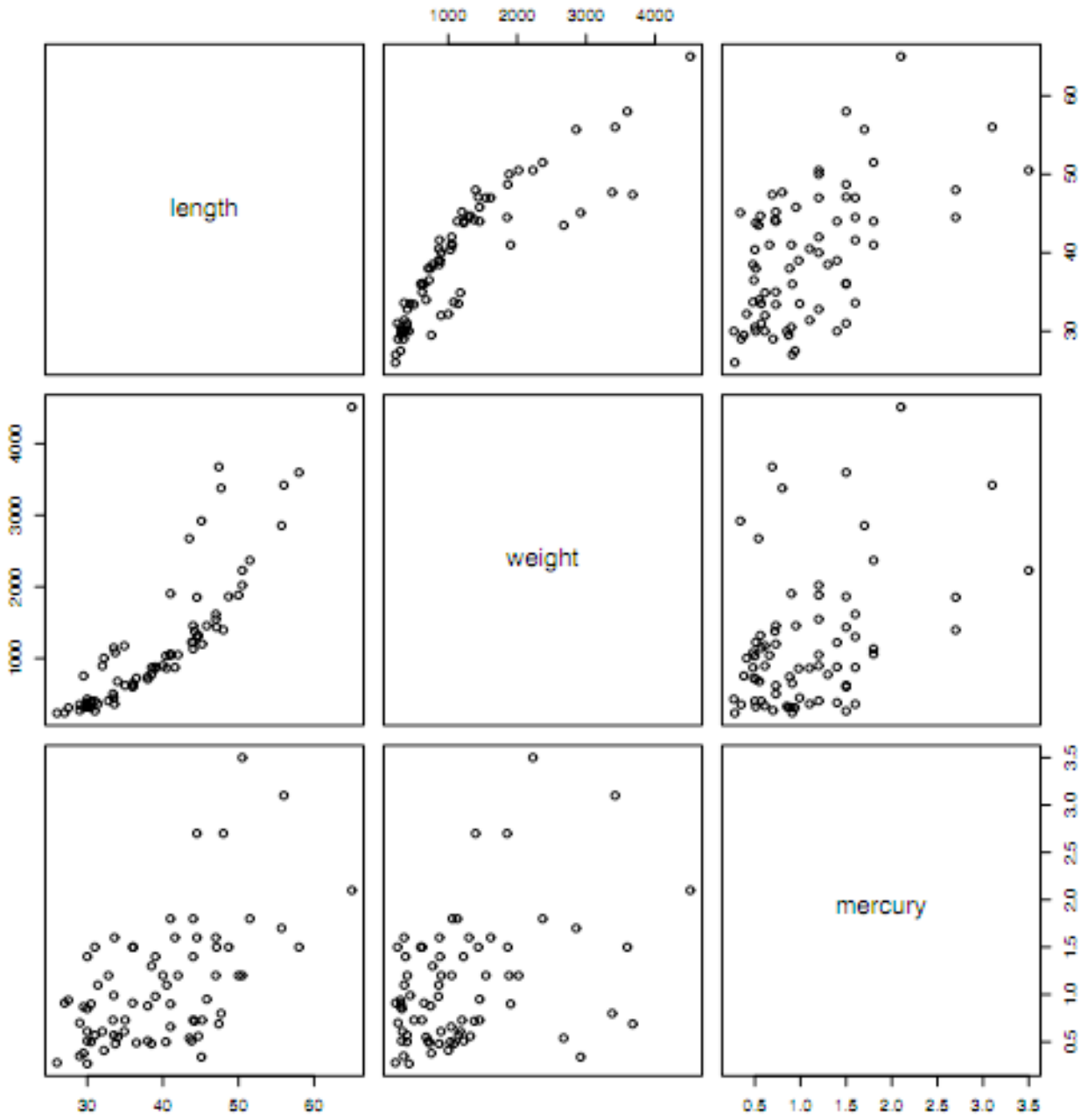


Mercury contamination

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

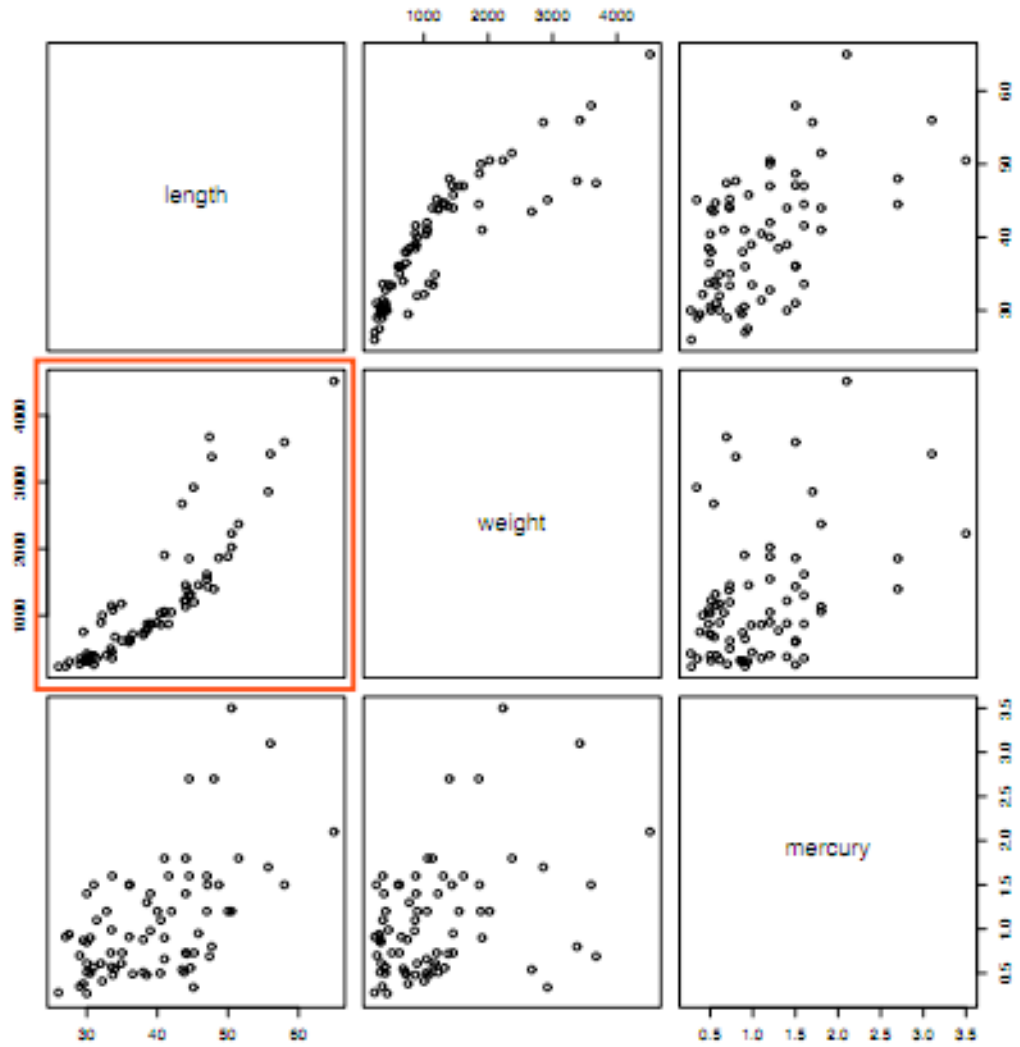
For this, we look at scatterplots; the scatterplot matrix allows us to look at several pairs of variables at one time

Lumber River



Lumber River

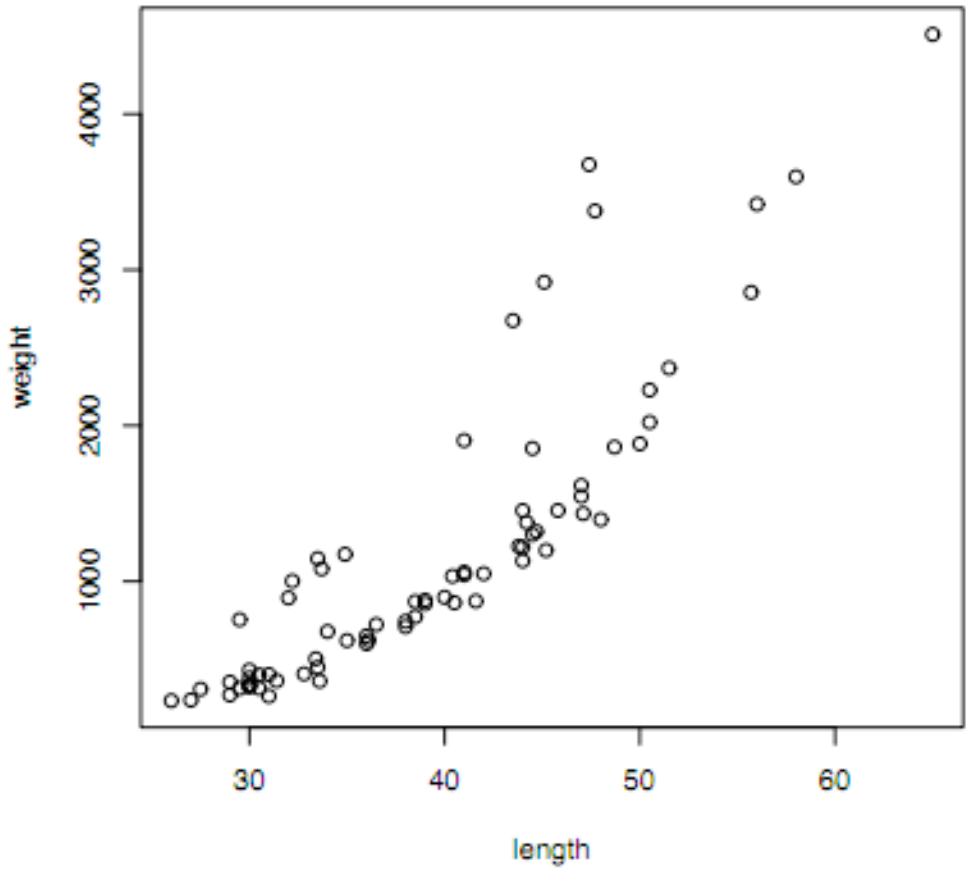
The plot in row 2 and column 1 is a scatter plot of weight on the y-axis and length on the x-axis



A scatterplot

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

What do we notice?



A scatterplot

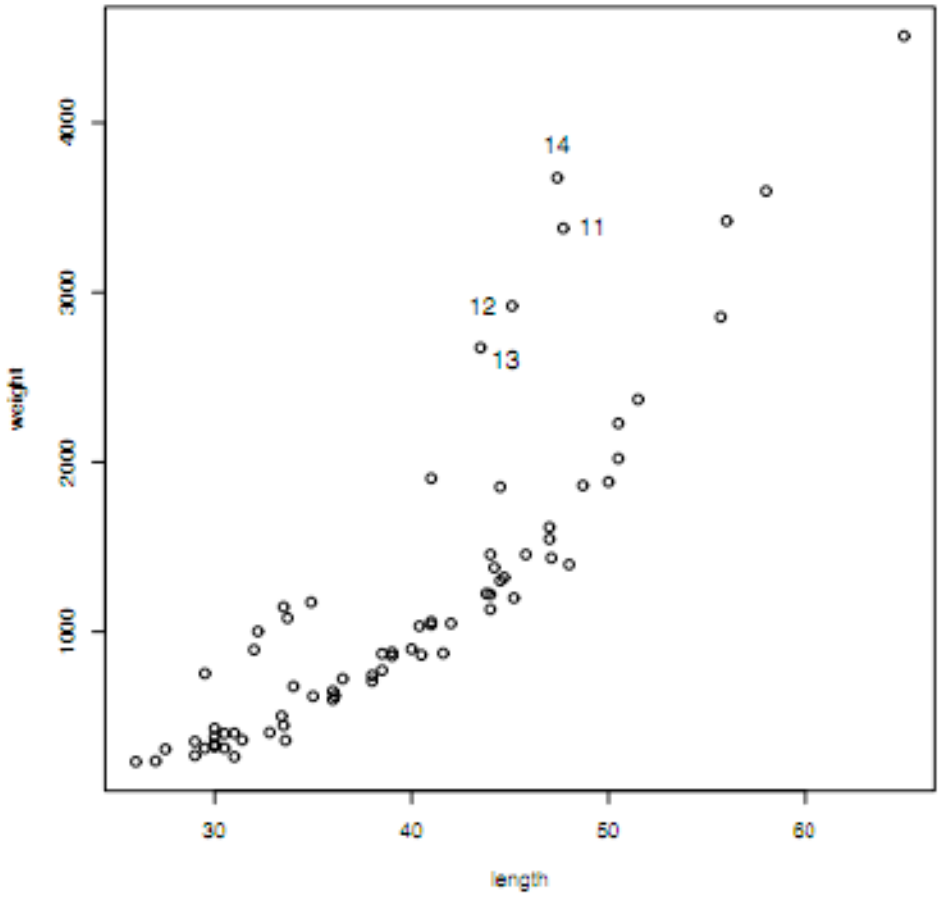
We can explore things a bit more using the `identify` command

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

A scatterplot

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

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



A scatterplot

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

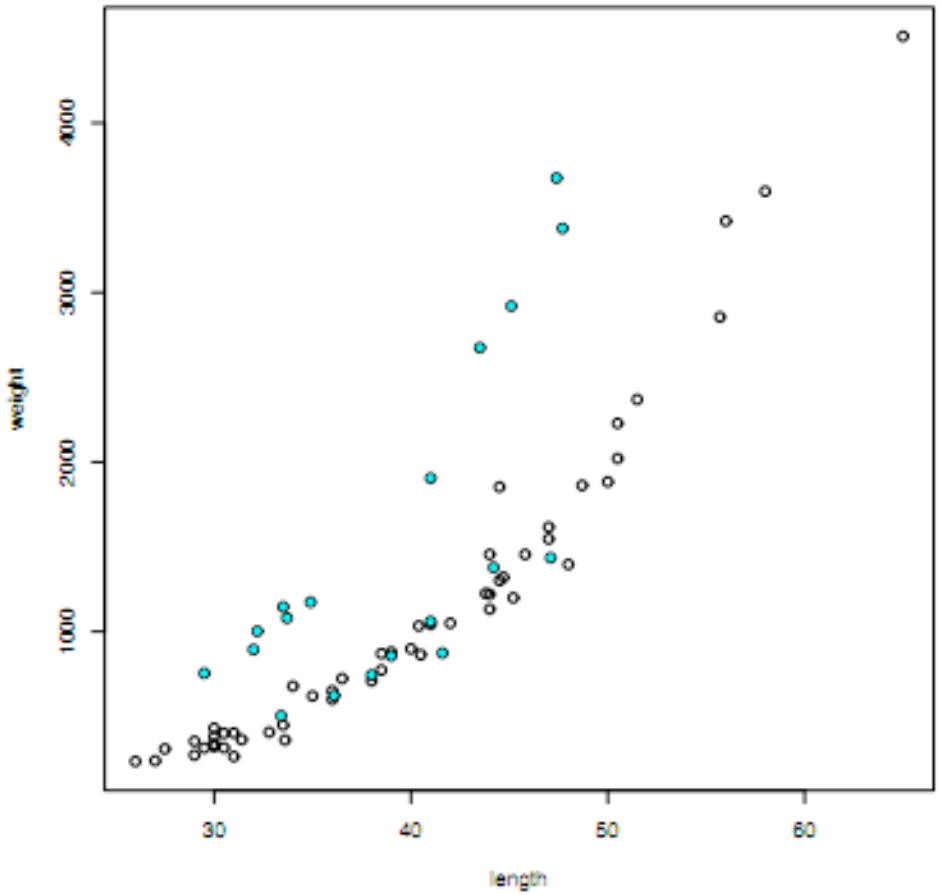
```
> plot(fish$length, fish$weight)
> uu = identify(fish$length,fish$weight)
> uu
 [1] 11 12 13 14 15 16 17 18 19 20 21
```

Mercury levels in water

```
> fish[uu,]

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

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



Seeing relationships between variables

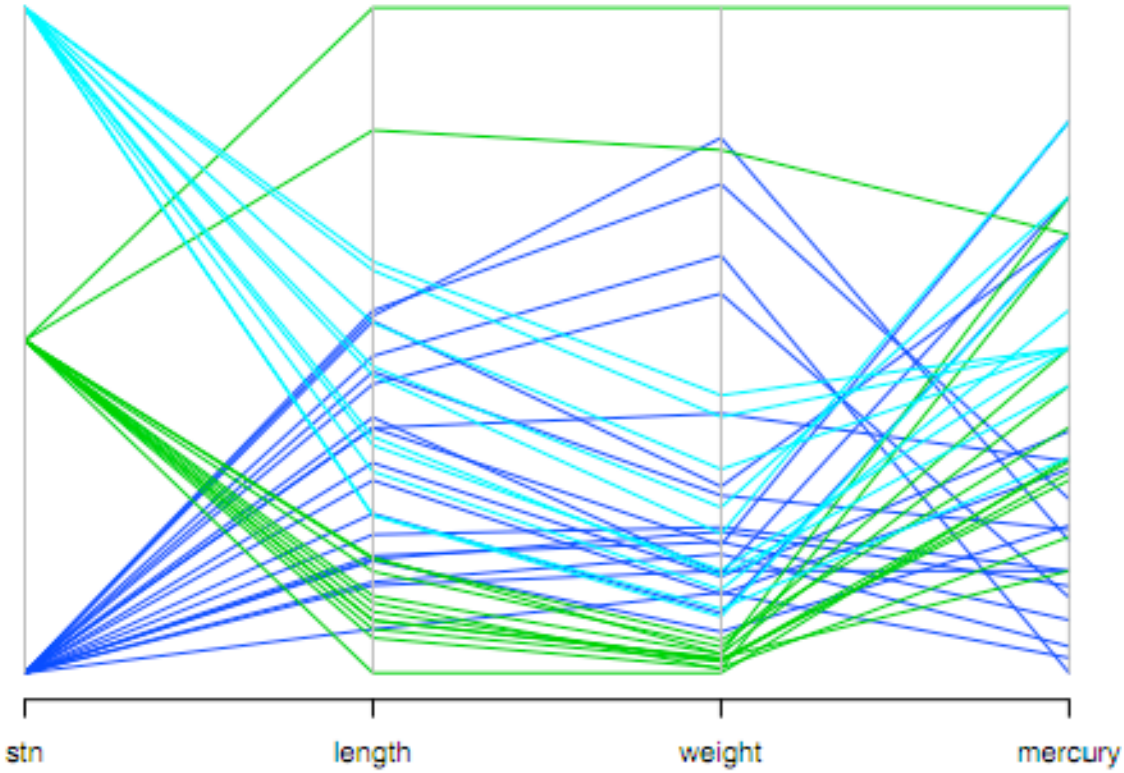
There are a variety of tools in R for helping you "see" the relationships between 2 or more variables

The scatterplot matrix is one device; here is another, perhaps less intuitive plot, which involves so-called parallel coordinates

Parallel coordinates

Here we show data for just the Lumber river and stations 1, 2 and 3

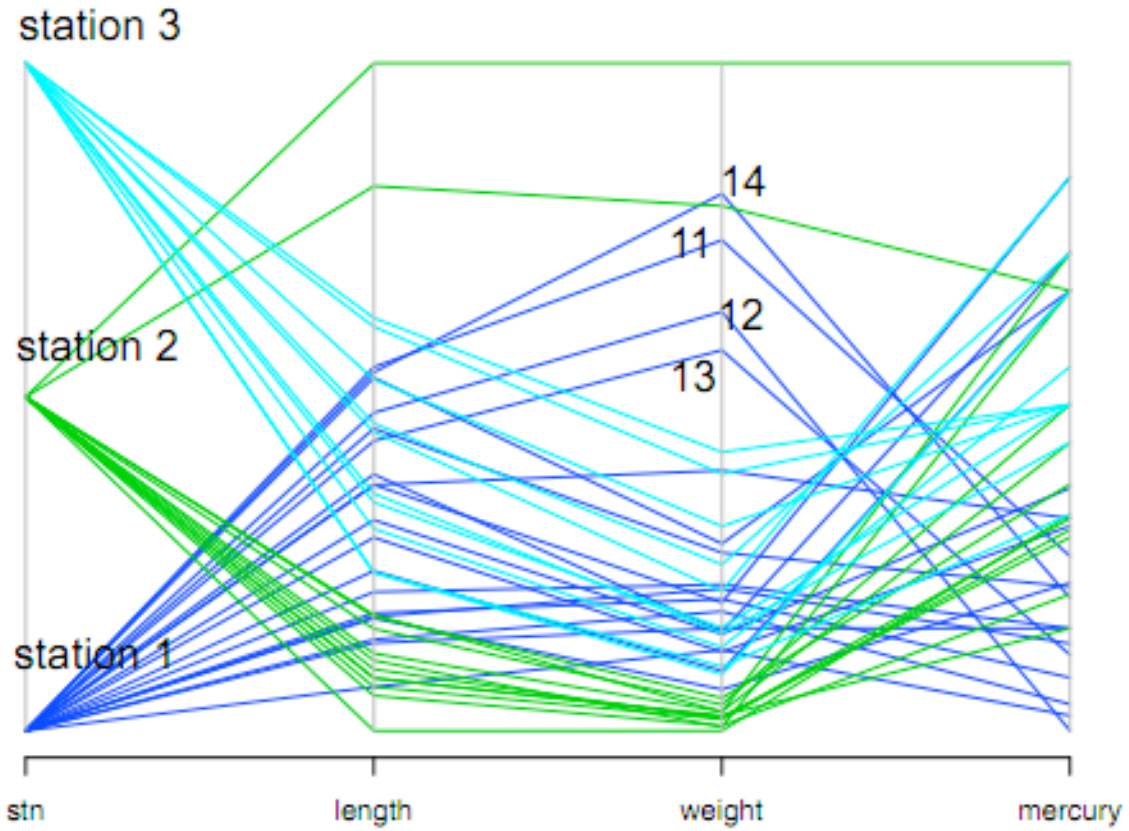
The vertical axes represent variables in the data set, each line connects the values for a single data point



Parallel coordinates

Different colors represent different stations

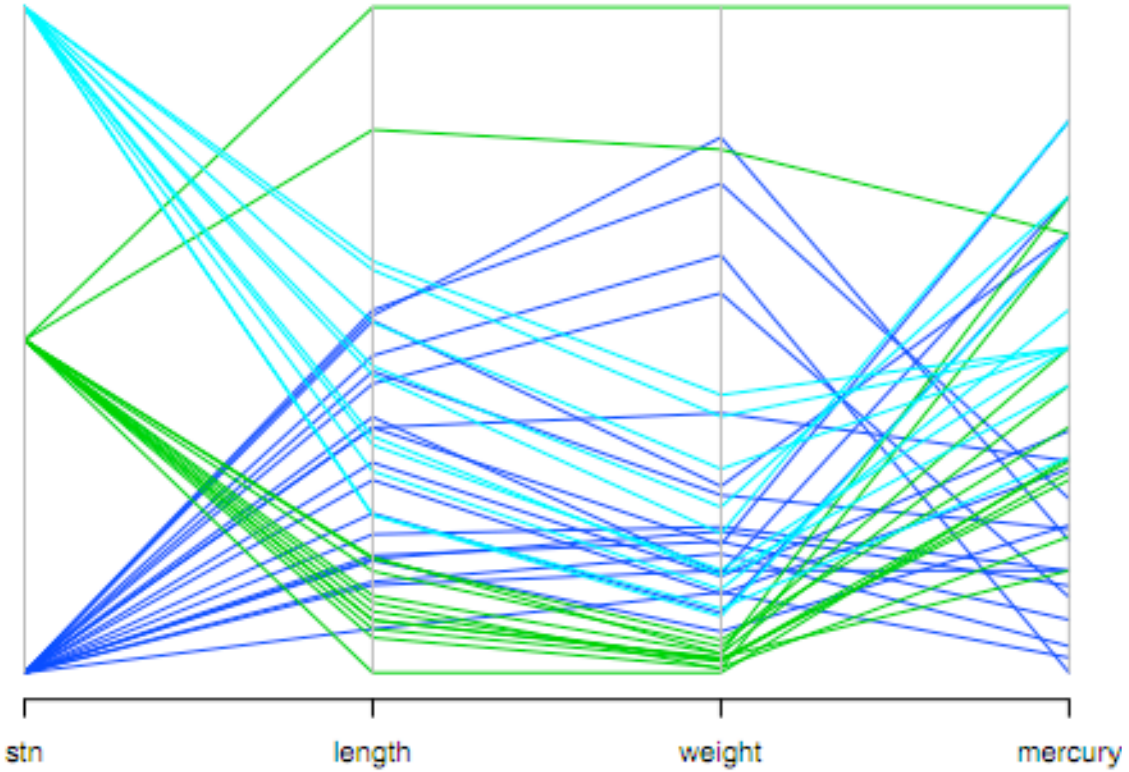
Notice the four “outliers” in weight among the dark blue lines that we highlighted in the scatterplot



Parallel coordinates

You can also see “clusters” here and differences between the stations

Look at the split in the green lines; how do the green lines relate to the cyan?

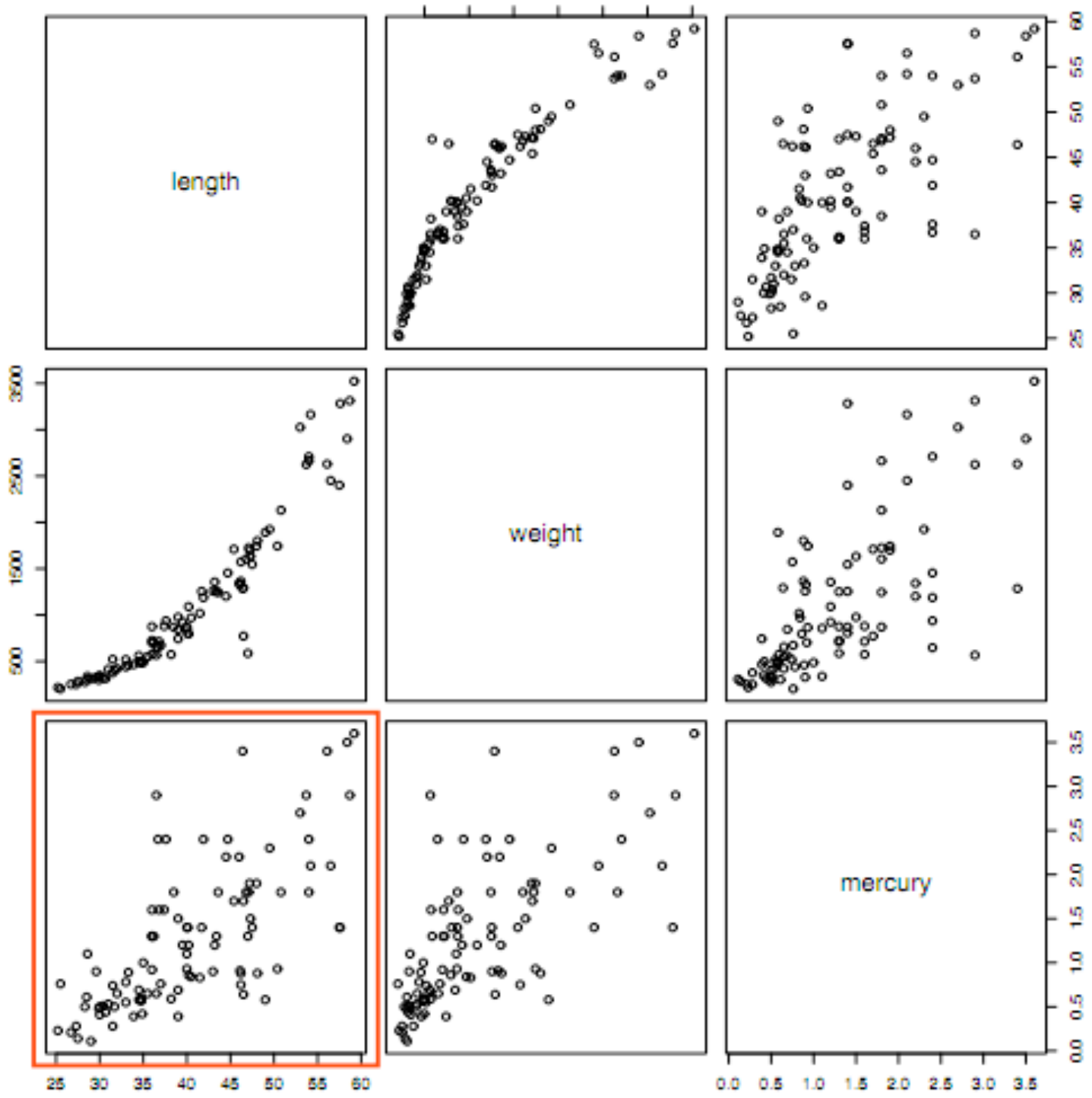


Mercury contamination

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

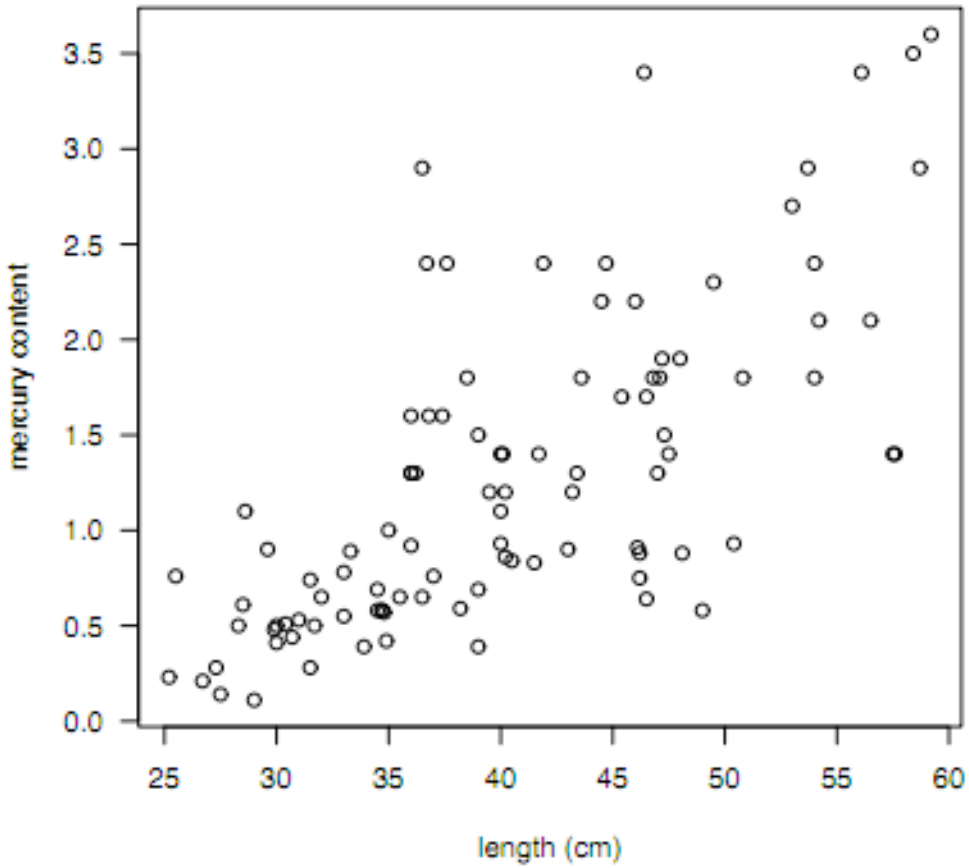
We will first use data on fish taken from the Waccamaw


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

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?



Modeling

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

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

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

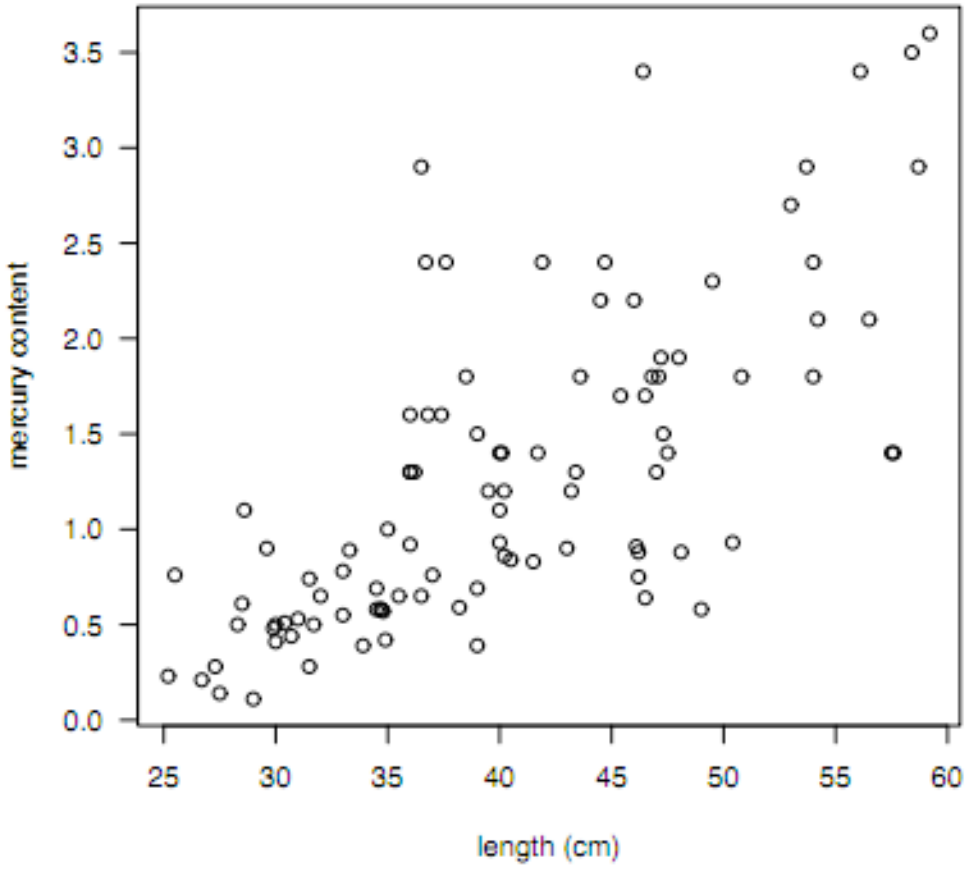
A linear model

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

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

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

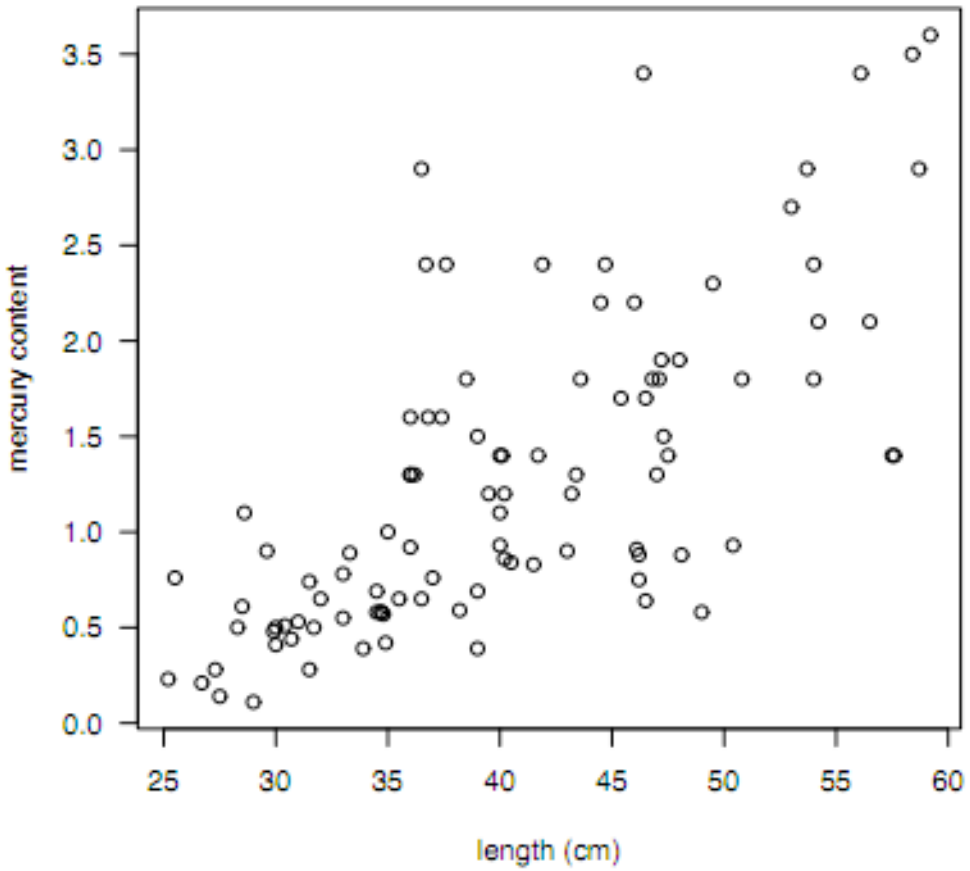
where β_0 and β_1 are parameters, the slope and intercept



A linear model

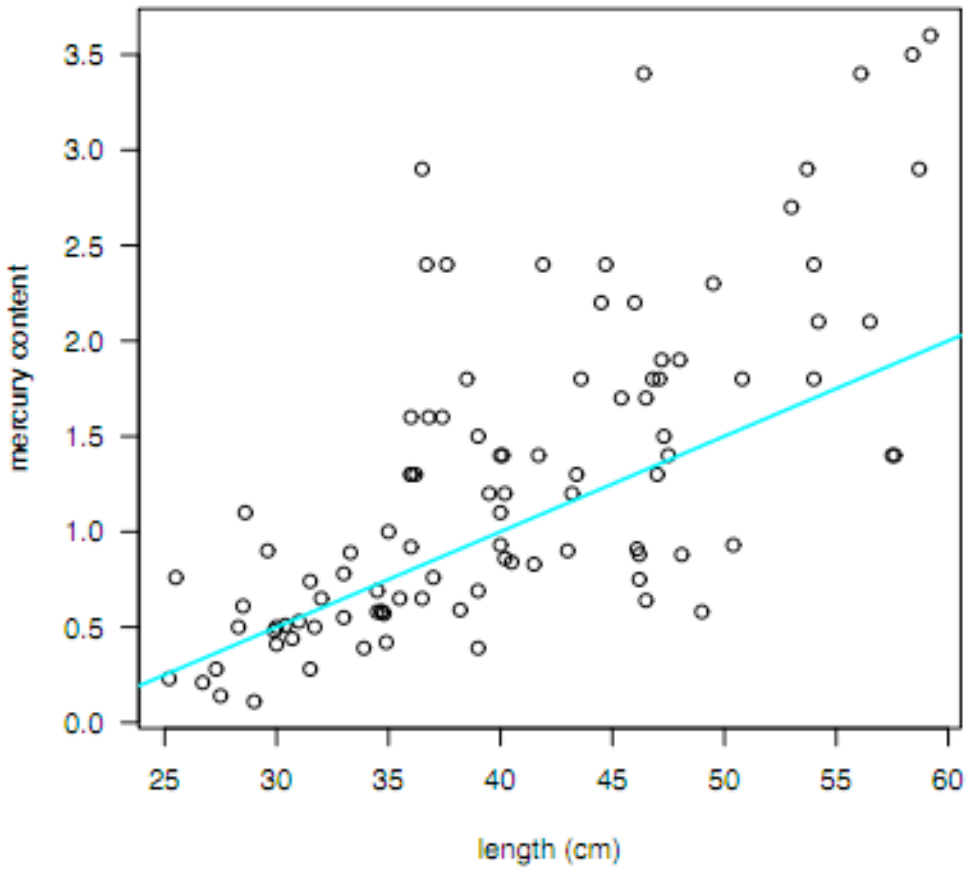
In terms of our data, we might posit a model of the form

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



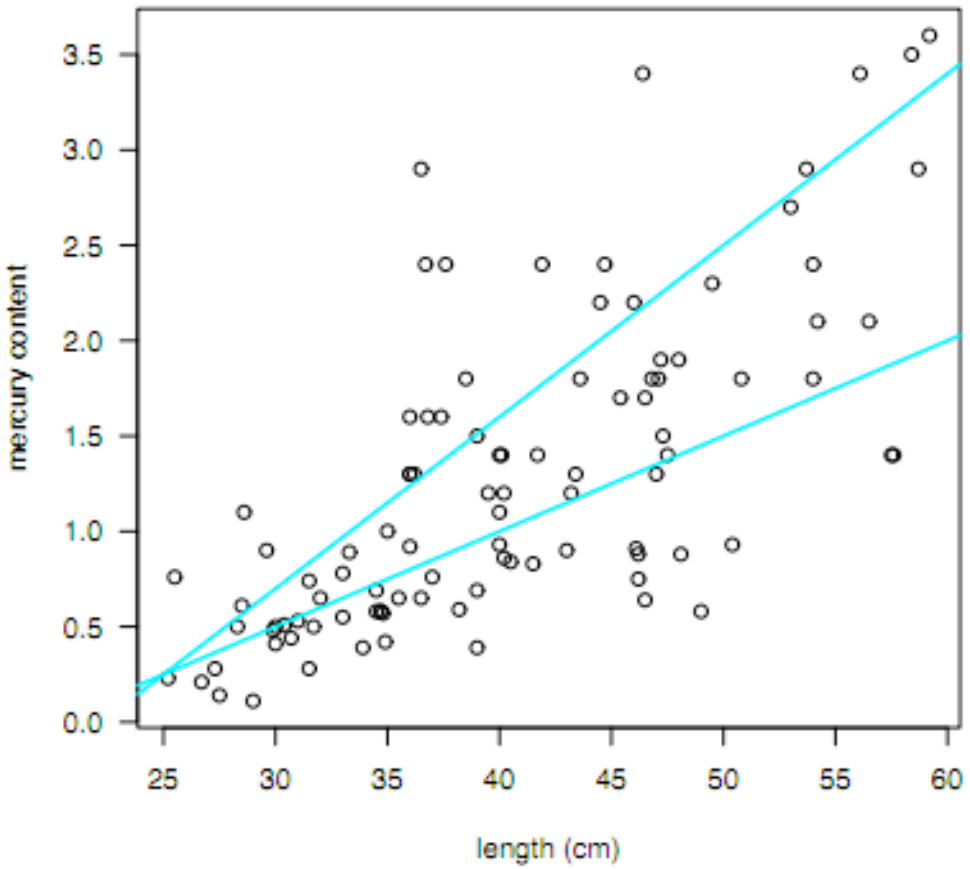
Linear models

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



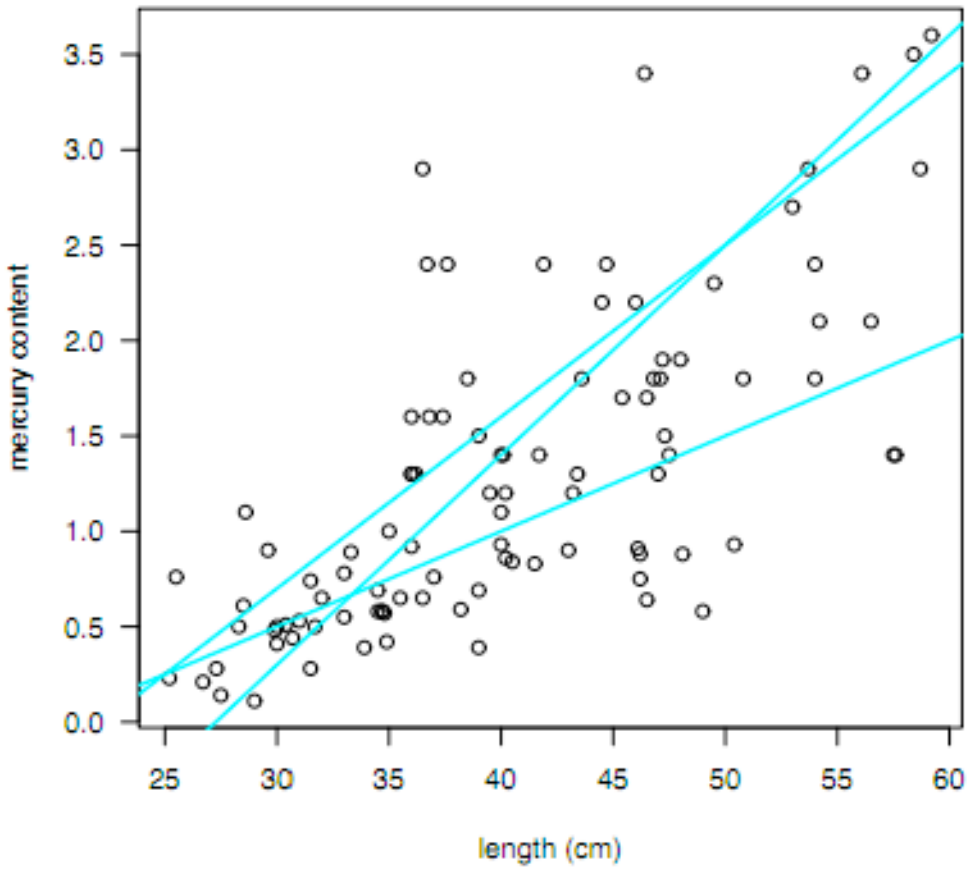
Linear models

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



Linear models

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



Least squares

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

x = fish length and y = mercury content

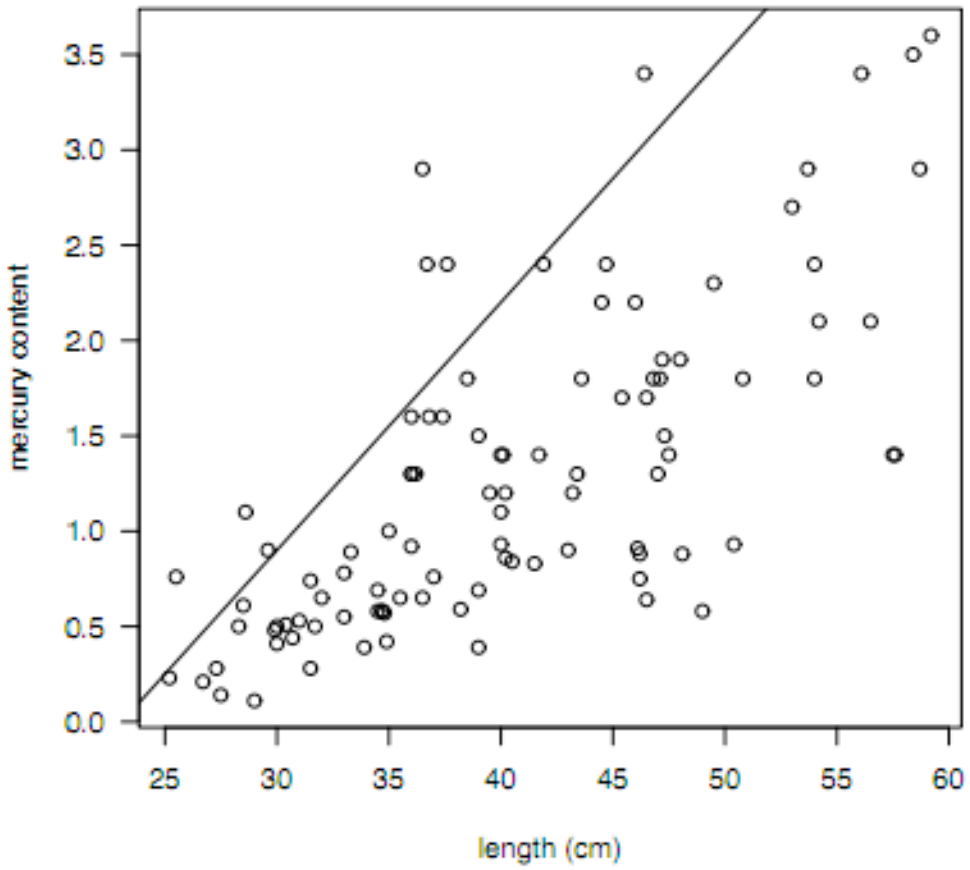
We then label our data set $(x_1, y_1), \ldots, (x_{98}, y_{98})$

Least squares

Specify a choice for the slope and intercept

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

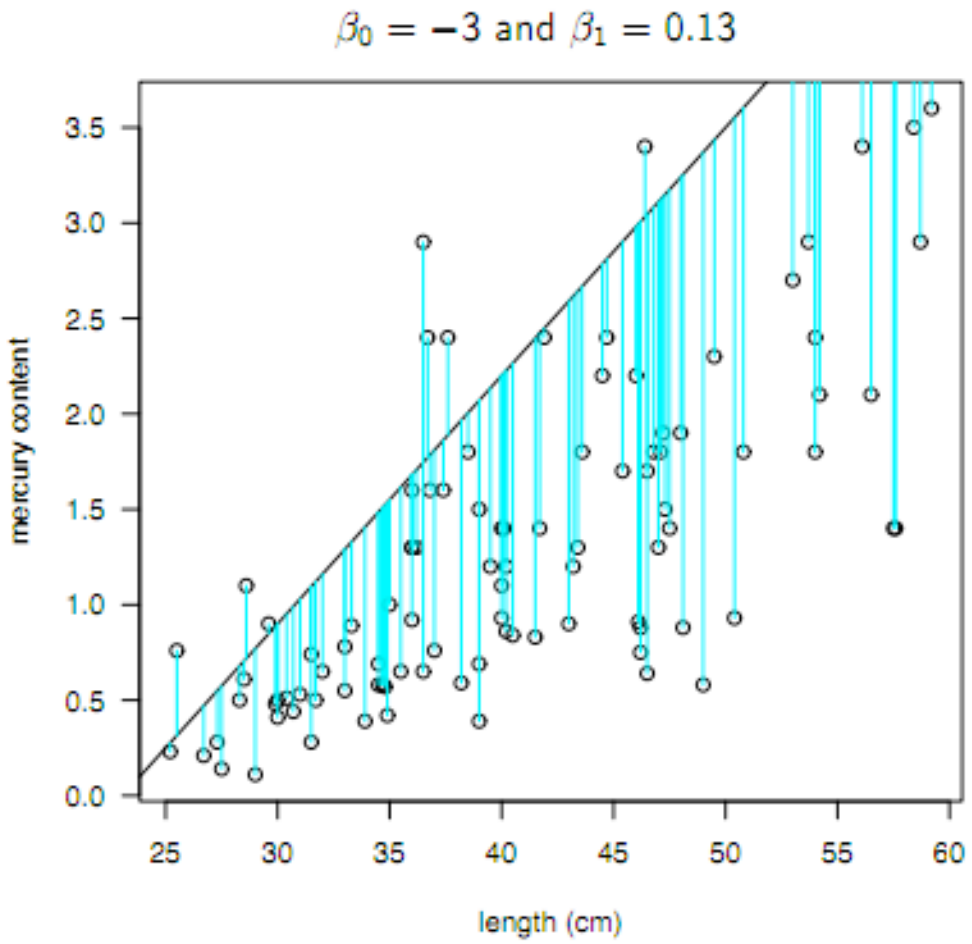
$\beta_0 = -3$ and $\beta_1 = 0.13$



Least squares

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

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



Least squares

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

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

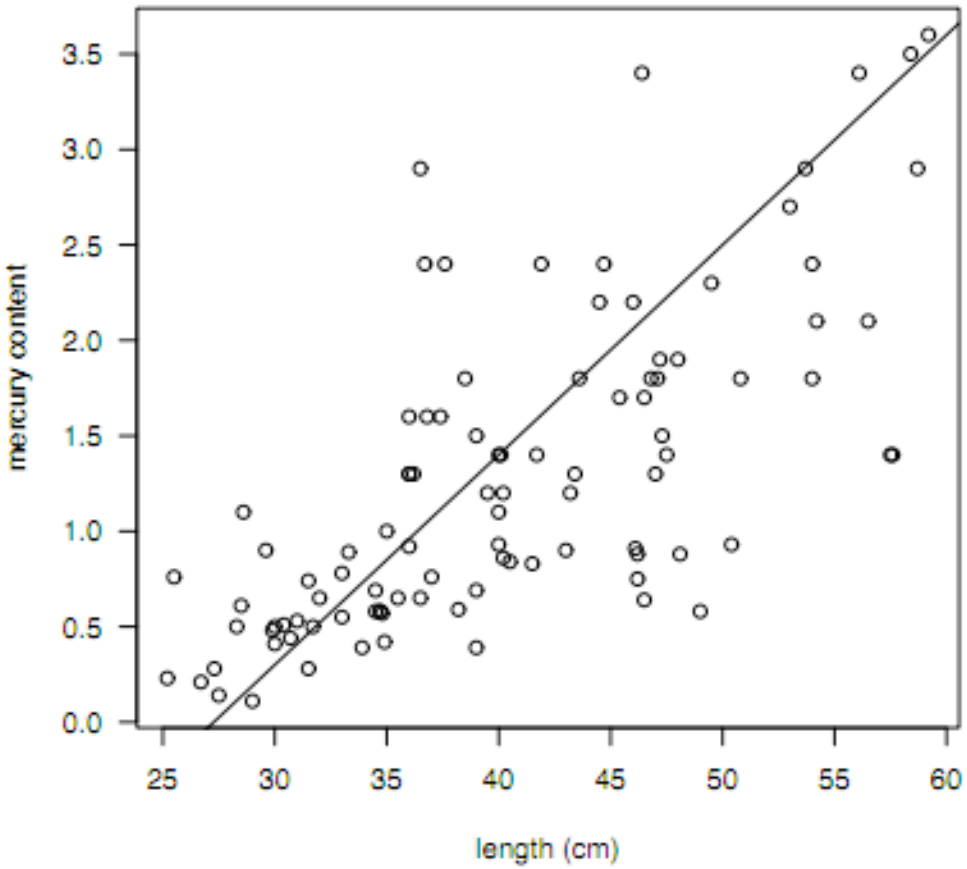
In this case, the squared error is 154.9

Least squares

Let's try another line

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

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

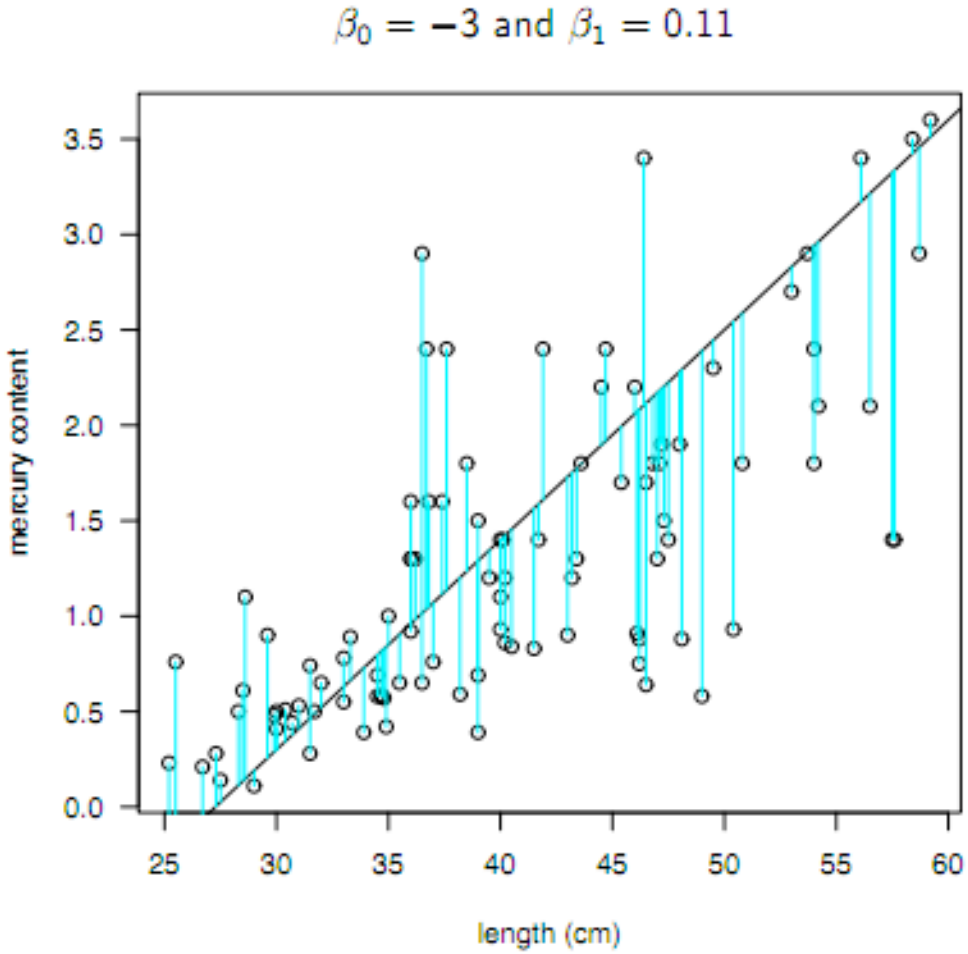


Least squares

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

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

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



Least squares

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

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

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

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

Least squares

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

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

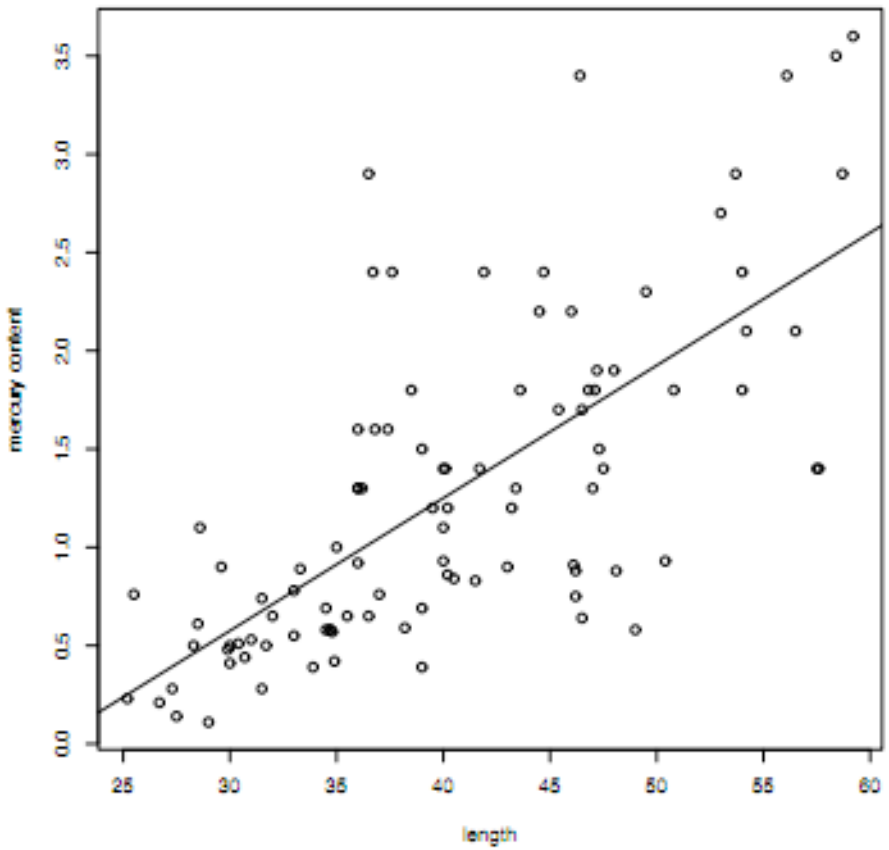
and the associated **sum of squares** is 33.4 (our simple trial and error approach was pretty far off!)

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

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

are called **residuals**

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



Least squares

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

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

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

Some interpretation

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

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

A useful comparison

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

What can you say about the value of b that minimizes

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

Flashback: The sample mean and standard deviation

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

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

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

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

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

Residual standard deviation

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

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

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

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

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

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

More interpretation: Residuals

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

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

More interpretation: Residuals

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

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

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

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

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

More interpretation: Residuals

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

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

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

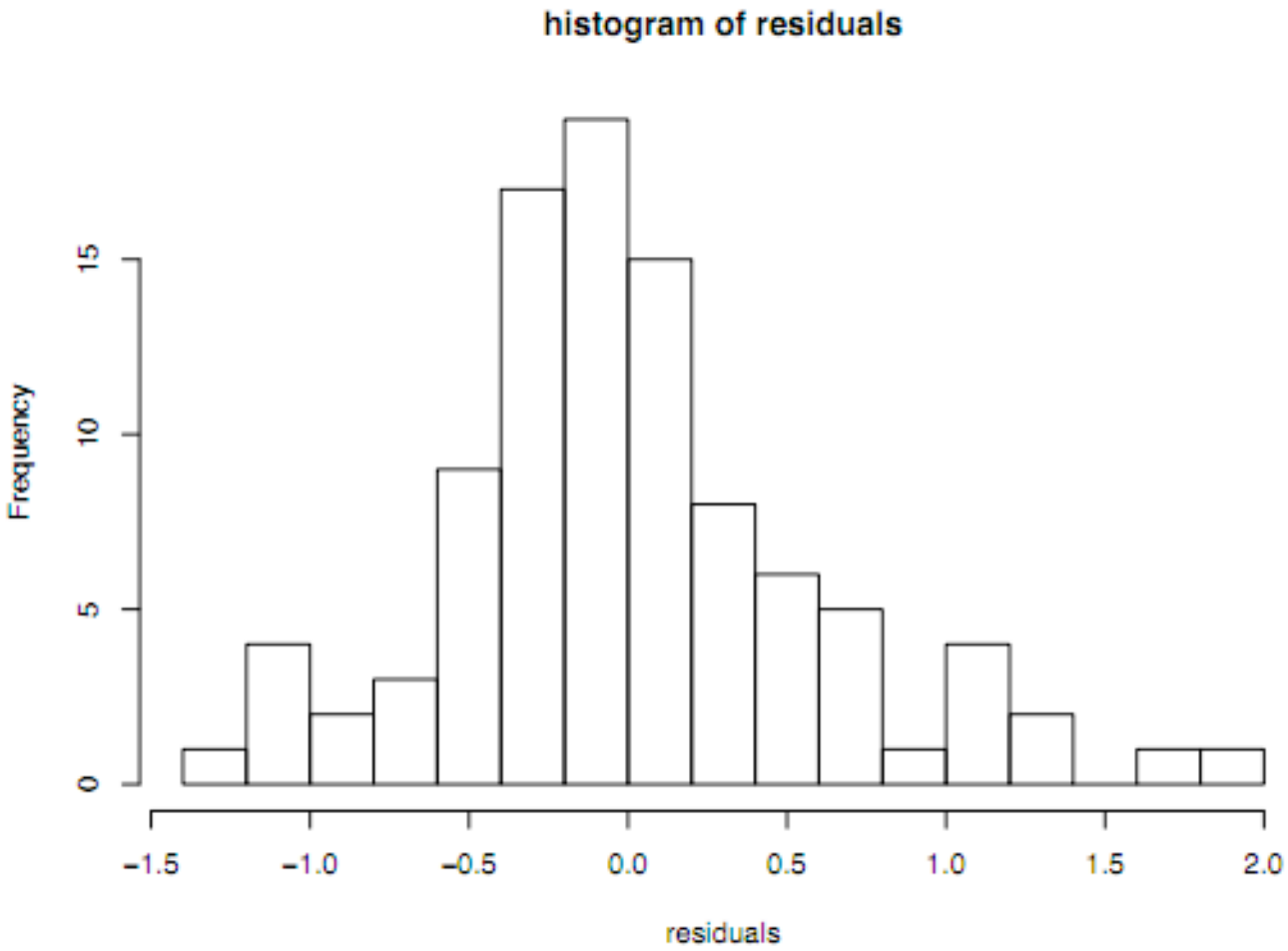
An example

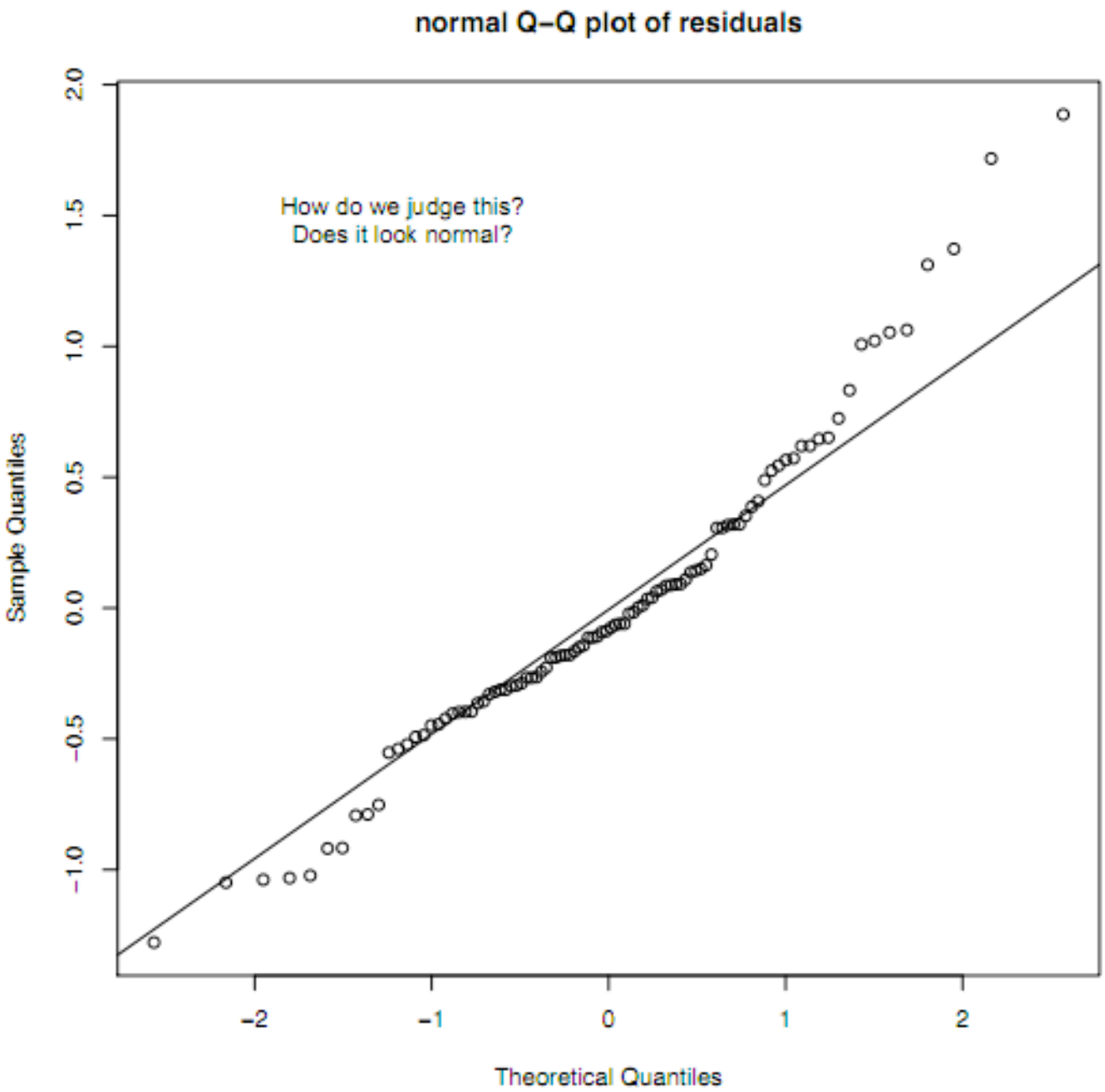
On the next two pages, we plot the residuals from the least squares fit to the fish data; the regression relating fish length and mercury content

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

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

Does this view of the residuals match your expectations?


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


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

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**

Question

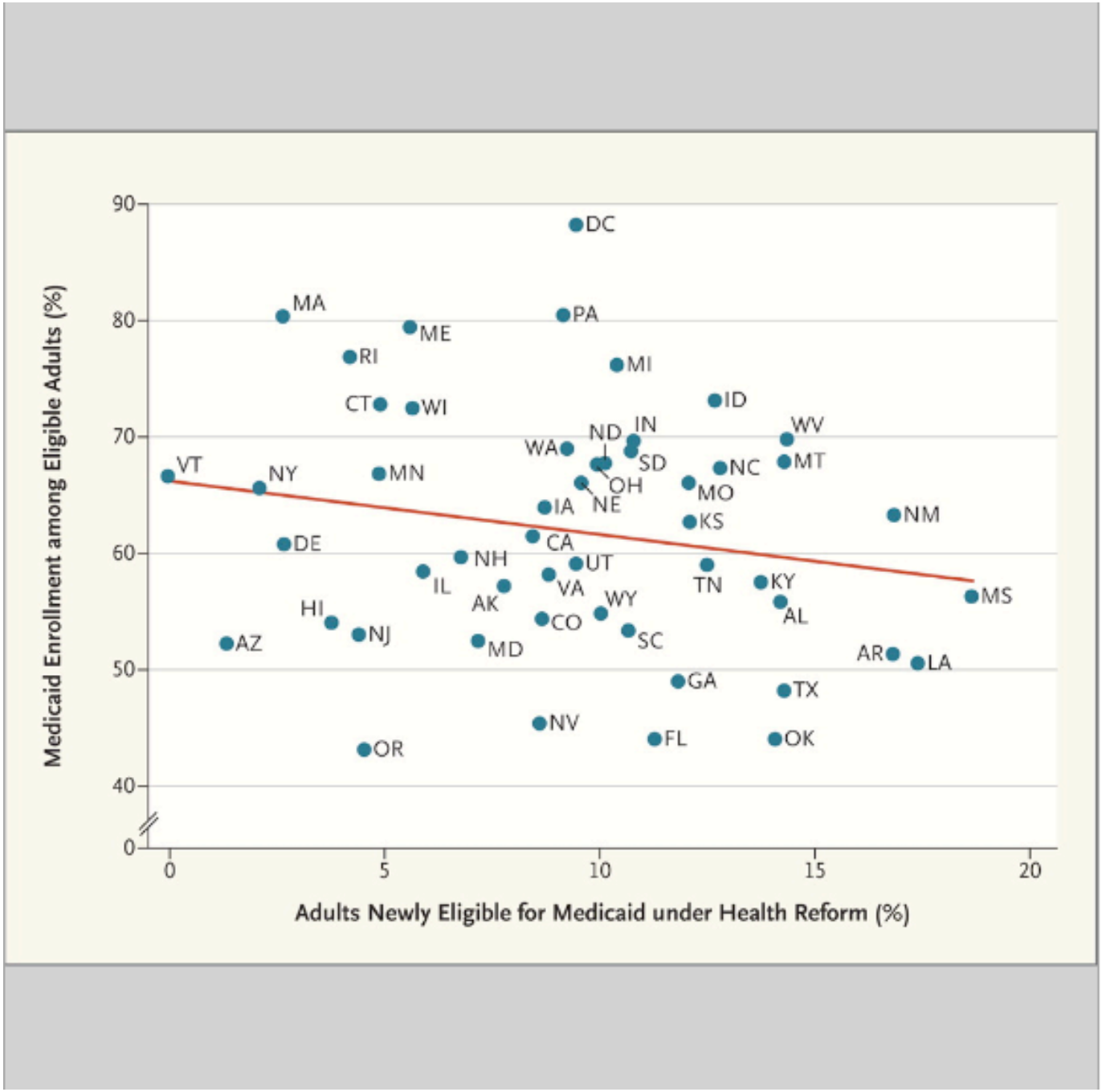
What are $\hat{\beta}_0$ and $\hat{\beta}_1$? What are we estimating?

Regression today

Regression has is a powerful tool in many quantitative disciplines -- In many cases, a regression model acts as a kind of social probe, providing researchers with a glimpse into the workings of some larger phenomenon

OK, that's generous. **It's also a highly abused tool**, one for which the elegant mathematics breaks down rather quickly once you hit modern practice -- Researchers often choose between many competing models, often through exhaustive searchers; data go missing and some form of imputation is often required; the underlying functional form is rarely linear and must also be estimated...

But here's what regression looks like in various fields...



Medicaid Enrollment among Currently Eligible Adults (2007 through 2009) and Percentage of Adults Who Will Become Eligible in 2014 under Health Care Reform, by State.

The population sample was restricted to eligible adults with no other form of health insurance; noncitizens were excluded from the analysis. Results are based on an analysis of data from the Current Population Survey of 2007 through 2009. The red line shows the regression equation: Enrollment=0.660.46×Newly Eligible (P=0.17).

Table 3. Regression Equations Predicting Sales, Number of Customers, Market Share, and Relative Profitability with Racial and Gender Diversity and Other Characteristics of Establishments

	Model 1	Model 2	Model 3	Model 4
Independent Variables	Sales	Customers	Market Share	Profitability
Constant	4.998***	61545.4	3.403***	3.363***
Racial diversity	.093***	433.86***	.007**	.006*
Gender diversity	.028**	195.642**	.001	.005**
Proprietorship	−.821	−370.78	−.232*	−.161
Partnership	.663	−6454.6	−.017	.256
Public corporation	−.109	7376.29	.214*	.202
Private corporation	−1.484**	−8748.7*	.008	.019
Company size	.00001*	.352*	.000**	.000**
Establishment size	.00001**	.119	.000	.000
Organization age	.013**	44.813	.001	.001
Agriculture	−1.942	4188.66	−.206	−.033
Mining	.739	−28856*	−.168	−.264
Construction	−.967	−875.7	−.152	−.036
Transportation/communications	−.052	1498.75	.119	.226
Wholesale trade	.008	−16383**	.136	−.064
Retail trade	−1.183**	7209.83*	.08	−.087
F. I. R. E.	−.683	−8335.4	−.212	.085
Business services	−1.49*	1552.28	.204*	.112
Personal services	−1.566*	1480.03**	.423**	−.001
Entertainment	−4.708**	−1504.5	−.191	−.076
Professional services	−.615	−13539**	.138	−.023
North	2.196***	23143.9***	−.06	−.039
Midwest	2.616***	14968.3***	−.023	−.073
South	1.82***	21152.8***	−.055	.059
R^2	.165***	.155***	.075**	.064**
N	506	506	469	484

Notes: Coefficients are unstandardized. For the dummy (binary) variable coefficients, significance levels refer to the difference between the omitted dummy variable category and the coefficient for the given category.
* $p < .1$; ** $p < .05$; *** $p < .01$.

Table 2. Percent Approving Same-Sex Marriage Ban; OLS Estimates, U.S. Counties

Independent Variable	1	2	3	4
Percent Women Not Working in Labor Force	.152**	.170**	.075*	.090*
Occupational Sex Segregation	16.310***	18.048***	8.216***	9.559***
Percent Households Married with Children	.219*	.322***	.156**	.217***
Percent Same-Sex Households	−2.038	−2.237	−5.428***	−5.956***
Percent Unmarried Opposite-Sex Households	−1.682***	−1.720***	−1.314***	−1.210***
Residential Instability	−.033	−.024	−.083**	−.066*
Percent Homes Not Owner Occupied	−.076*	−.084*	−.079***	−.109***
Crime Rate		.099***		.022*
Percent Production or Construction Occupations	.194***	.234***	.067**	.063*
Percent Professional Occupations	−.178*	−.100	−.118*	−.119*
Percent Self-Employed	−.100	−.093	−.198***	−.234***
Median Family Income ($1,000s)	−.138***	−.045	−.166***	−.148***
Percent Receiving Public Assistance	−.054	.037	.125	.207
Mean Years of Education	−1.328*	−2.611***	−2.040***	−2.606***
Percent Enrolled in College	−.060	.020	−.185**	−.139
Population Density (logged)	−.111	−.332	.001	.040
Percent Urban	.029**	.011	.015**	.011
Republican Voting (percent Bush 2000)	.350***	.381***	.287***	.317***
Percent Evangelical	.180***	.179***	.031***	.020
Percent Catholic	.063***	.064***	−.025**	−.035**
Median Age	−.155	−.022	−.314***	−.247**
Percent African American	.092***	.085***	−.006	.009
Percent Latino	.043*	.024	−.108***	−.118***
LGBT Organizations	−.114	.974	−2.096***	−1.576**
Civil Rights Organizations	−.007	−.004	.073	.113
Antidiscrimination Legislation	−3.283***	−2.797***	−2.656***	−1.918*
Alabama			−6.679***	−6.132***
Arizona			−23.814***	−23.878***
Arkansas			−9.454***	−9.085***
California			−2.079*	−1.867*
Colorado			−13.112***	−12.478***
Florida			−7.906***	−8.263***
Georgia			−2.787***	−3.307***
Idaho			−17.638***	−17.591***
Kansas			−5.857***	−6.051***
Kentucky			−7.623***	−9.111***
Louisiana			−4.771***	−5.492***
Michigan			−15.107***	−15.211***
Mississippi			−2.865***	−1.619
Missouri			−5.502***	−5.583***
Montana			−8.701***	−8.893***
Nebraska			−5.916***	−5.754***
Nevada			−3.886***	−3.964***
North Dakota			−4.316***	−4.057***
Ohio			−13.621***	−13.853***
Oklahoma			−8.002***	−7.835***
Oregon			−10.999***	−10.574***
South Carolina			−3.390***	−3.947***
South Dakota			−30.373***	−29.254***
Tennessee			−1.166	−.920
Utah			−18.009***	−18.697***
Virginia			−17.326***	−17.351***
Wisconsin			−12.236***	−11.792***
Texas (basis of comparison)				
Number of Observations	2,231	1,602	2,231	1,602
R-Square	.713	.735	.909	.915

* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 2. OLS Estimates of Effect of Selected Measures of Residential Segregation on Log of Total Foreclosures

	Dissimilarity Index		Isolation Index	
Variables	*B*	*SE*	*B*	*SE*
Index of Segregation				
African Americans	3.718**	.725	2.122**	.619
Hispanics	−.773	.596	.080	.656
Asians	−2.080*	.920	−2.161	1.636
Control Variables				
Housing Starts Ratio	2.980**	.960	3.067**	1.077
Wharton Land Use Index	.250**	.082	.272**	.096
Change in Housing Price Index	.082**	.024	.092**	.029
CRA-Covered Lending Share	−1.295	.912	−.810	1.061
Subprime Loan Share	3.022*	1.353	4.310**	1.581
MSA Credit Score Index	−.015*	.007	−.016*	.007
Log of Population	1.008**	.089	1.013**	.093
Percent with College Degree	−1.341	1.315	−.997	1.459
Log Median Household Income	.253	.509	.340	.515
Percent with Second Mortgage	.751	3.687	.225	4.350
Percent Workforce Unionized	−.025**	.011	−.022*	.011
Unemployment Rate	−.010	.064	.012	.071
Change in Unemployment Rate	.245**	.052	.213**	.063
Age of Housing Stock	.004	.012	.014	.013
Region				
Midwest	.434*	.200	.631**	.200
South	.042	.257	.081	.296
West	.463	.384	.679	.436
Coastal MSA	−.053	.123	.070	.133
Borders Rio Grande	−1.030**	.370	−1.054**	.380
Constant	1.960	7.557	.979	8.150
R^2	.91		.90	
Joint *F*-Test for Region	3.35*		7.97**	
Joint *F*-Test for Segregation	10.48**		6.28**	

Note: $N = 99$. Robust standard errors. Model also includes percent black, percent Hispanic, and percent Asian.

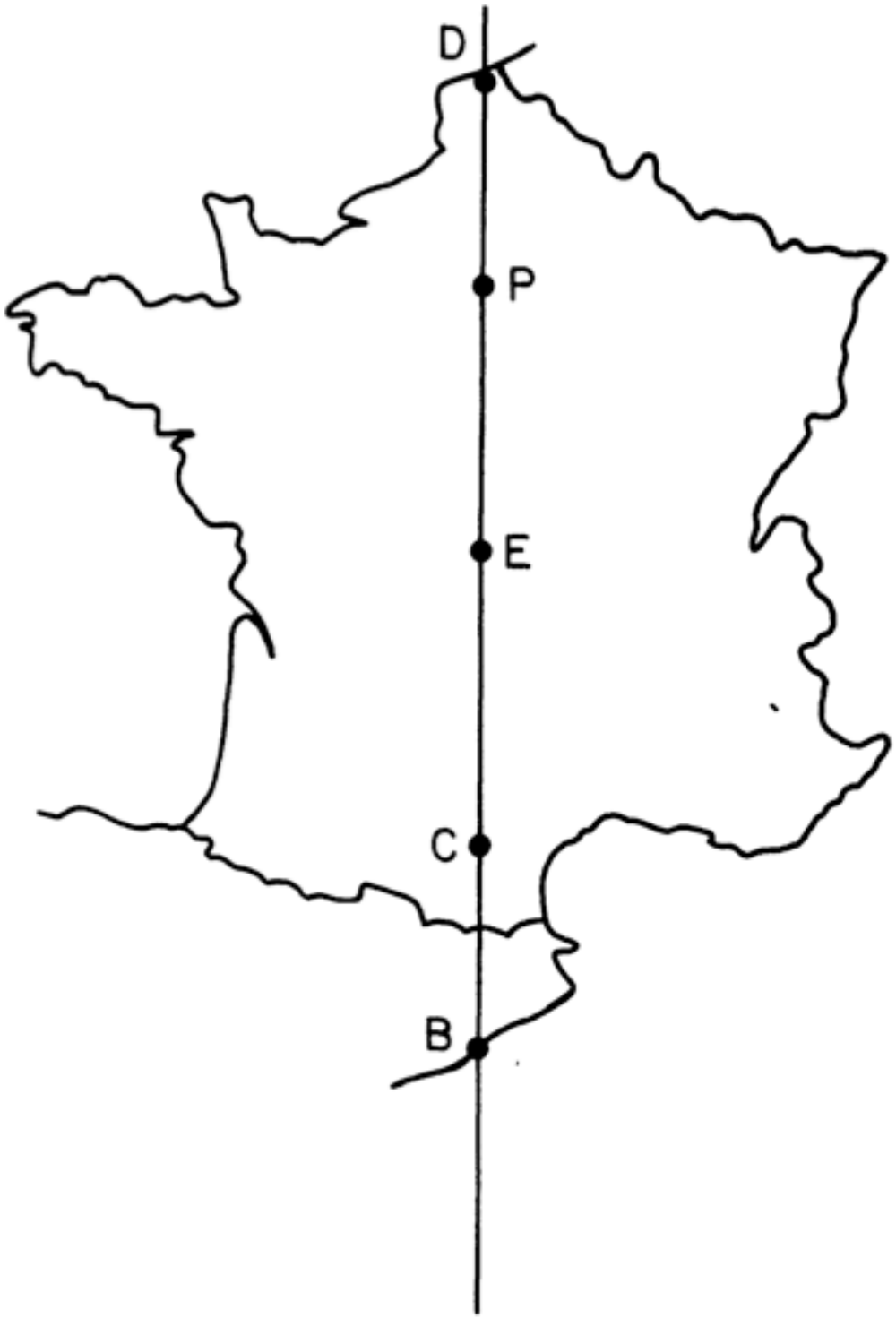
*$p < .05$; **$p < .01$ (two-tailed tests).

The contested origins of least squares

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

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

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


D
P
E
C
B

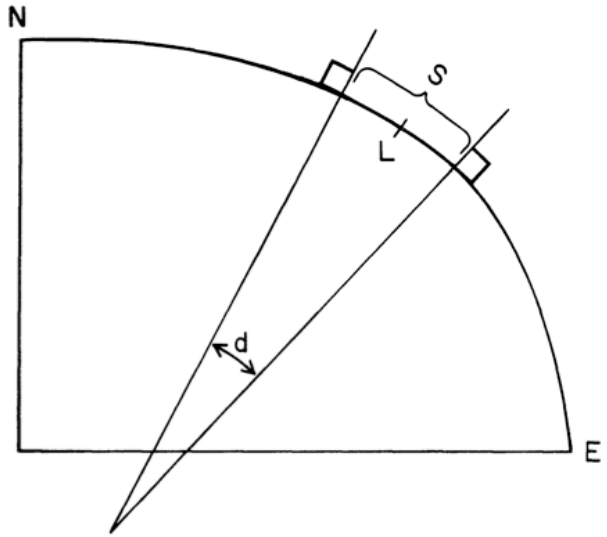

N
s
L
d
E

TABLE 1.

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

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

Least squares

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

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

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

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

Label the four data points in the previous table

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

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

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

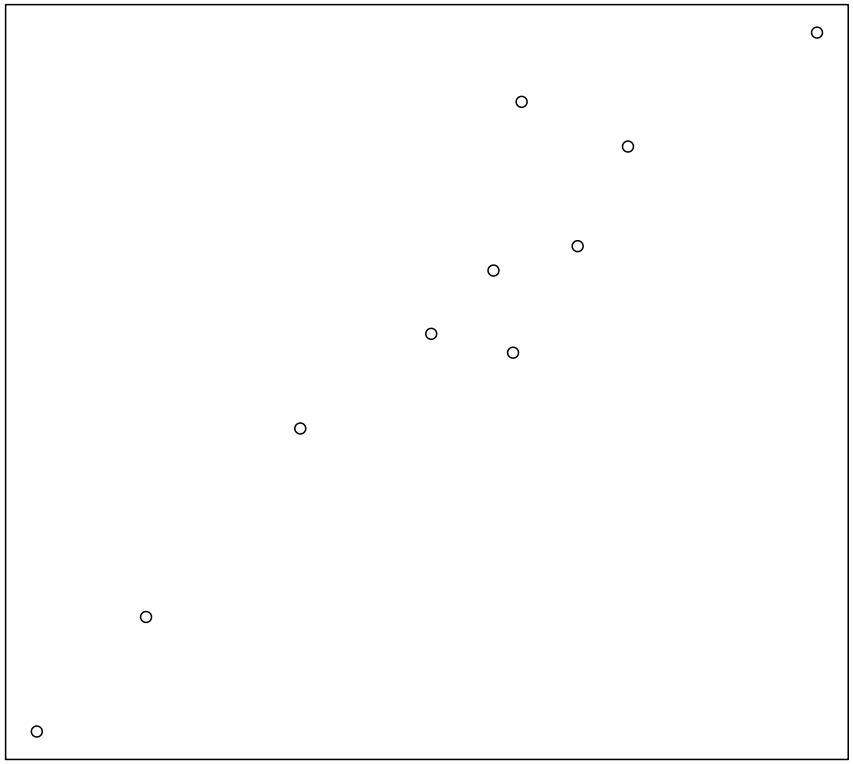
Least squares

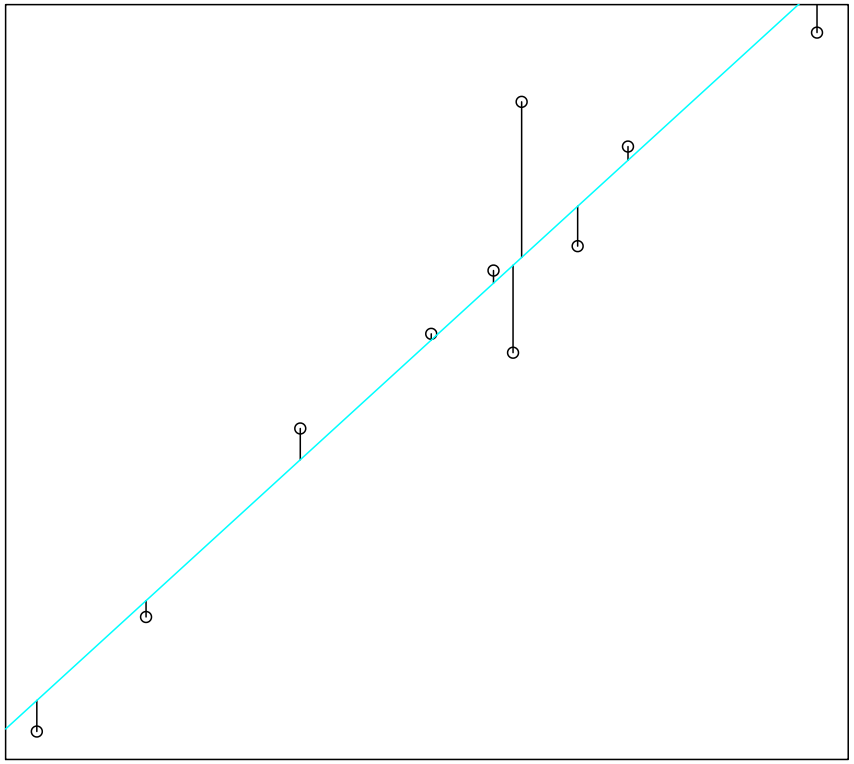
Given a set of predictor-response pairs $(x_1, y_1), \ldots, (x_n, y_n)$, we can write the ordinary least squares (OLS) criterion (as opposed to a weighted version that we'll get to) as

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

Least squares

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





Gauss and least squares

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

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

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

Galton and regression

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

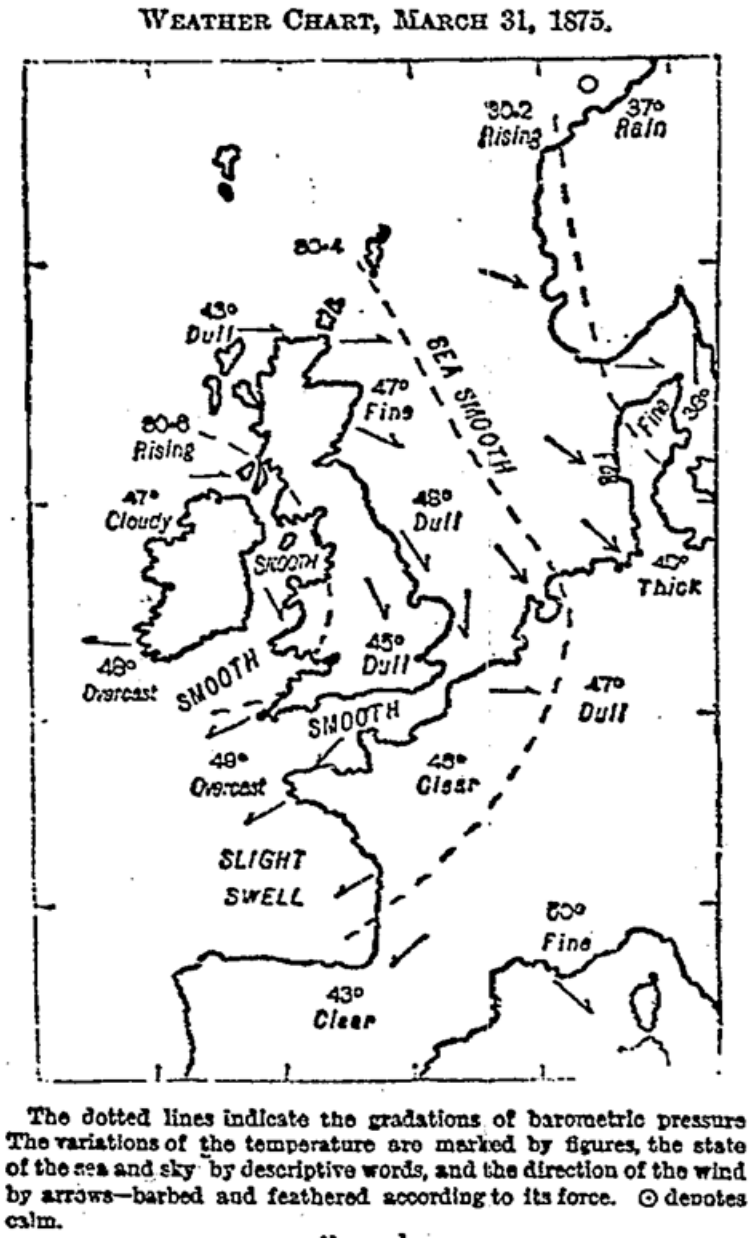
Stigler writes “Few conceptual advances in statistics can be as unequivocally associated with a single individual. Least squares, the central limit theorem, the chi-squared test -- all of these were realized as the culmination of many years of exploration by many people. Regression too came as the culmination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

Galton and regression

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

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

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



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

Galton and regression

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

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

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

Galton and regression

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

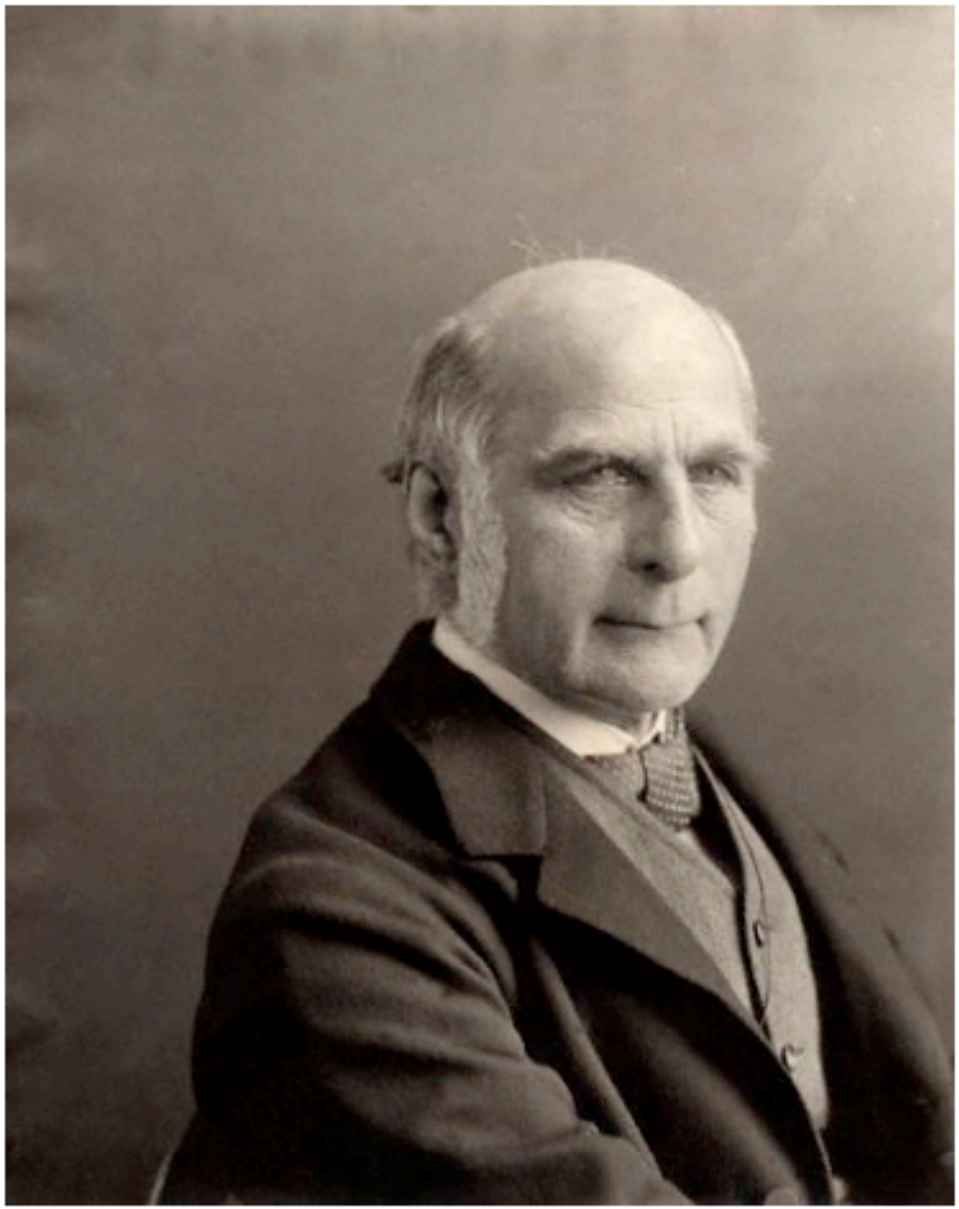
Galton writes “Order in Apparent Chaos -- I know of scarcely anything so apt to impress the imagination as the wonderful cosmic order expressed by the Law of Frequency of Error. The law would have been **personified by the Greeks and deified**, if they had known of it.”

Galton and regression

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

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

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



http://www.npg.org.uk

Some history

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

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

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

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

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

[With Plates IX and X.]

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

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

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

2 FATHER ...

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

2. Occupation. Clerk in Holy Orders. Residences.

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

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

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

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

8. General appearance. Slender.

9. Bodily strength and energy, if much above or below the average. During 22 years from Ordination, have preached 3600 times, yearly average 164. Unable to preach (from temporary indisposition) only 4 Sundays during these 22 years.

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

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

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

1. Date of birth. August 7th 1838. Birthplace. South Glamorganshire.
2. Occupation. Clerk in Holy Orders. Residence.
3. Age at marriage. The place for this entry is at 4 in next page. 23.
4. do. of wife. The place for this entry is at 3 in next page. 23.
5. Mode of life so far as affecting growth or health.
6. Was early life laborious? why and how? No.
7. Adult height. 5 ft. 6 in. Colour of hair when adult. Dark brown. Colour of eyes. Blue.
8. General appearance. Slender.
9. Bodily strength and energy, if much above or below the average. During 5½ years from Ordination, have preached three times, quite [illegible] unable to preach (from temporary indisposition) only on Sundays during last 22 years.
10. Keenness or imperfection of sight or other senses. Have always possessed good sight, both for near & distant objects. No failure as yet. (age 46).
11. Mental powers and energy, if much above or below the average. Rapid reader.
12. Character and temperament. Cool, cautious, methodical.
13. Favorite pursuits and interests. Artistic aptitudes. From earliest childhood a decided inclination to the mechanical. Fond of music. [illegible]
14. Minor ailments to which there was special liability { in youth. Colds in the head. Few headaches. { in middle age. Very rarely suffer now from these ailments.
15. Graver illnesses { in youth. None, excepting measles & whooping cough, but mild. { in middle age. None, excepting lumbago occasionally.
16. Cause and date of death, and age at death. Still living.
17. General remarks.

1. Date of birth. Dec. 18, 1838. Birthplace. Lindley, Huddersfield.
Residences. Lindley, [illegible]
2. Occupation.
3. Age at marriage } 23 Total No. of { sons 1 { daus. 1 No. of sons deceased. None. ages
4. ditto of husband } 23 No. of daus. deceased. None. ages
5. Mode of life so far as affecting growth or health. Educated at private schools.
6. Was early life laborious? why and how? No.
7. Adult height. 5 ft. [illegible] in. Colour of hair when adult. Dark brown. Colour of eyes. Brown.
8. General appearance. Slender frame. Dark complexion.
9. Bodily strength and energy, if much above or below the average.
10. Keenness or imperfection of sight or other senses. Sight & other senses good.
11. Mental powers and energy, if much above or below the average.
12. Character and temperament. [illegible]
13. Favorite pursuits and interests. Artistic aptitudes.
14. Minor ailments to which there was special liability { in youth. Bilious attacks, with vomiting & headache. { in middle age. Indigestion.
15. Graver illnesses { in youth. { in middle age. Bronchitis.
16. Cause and date of death, and age at death. Still living.
17. General remarks.

TABLE I.

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

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

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

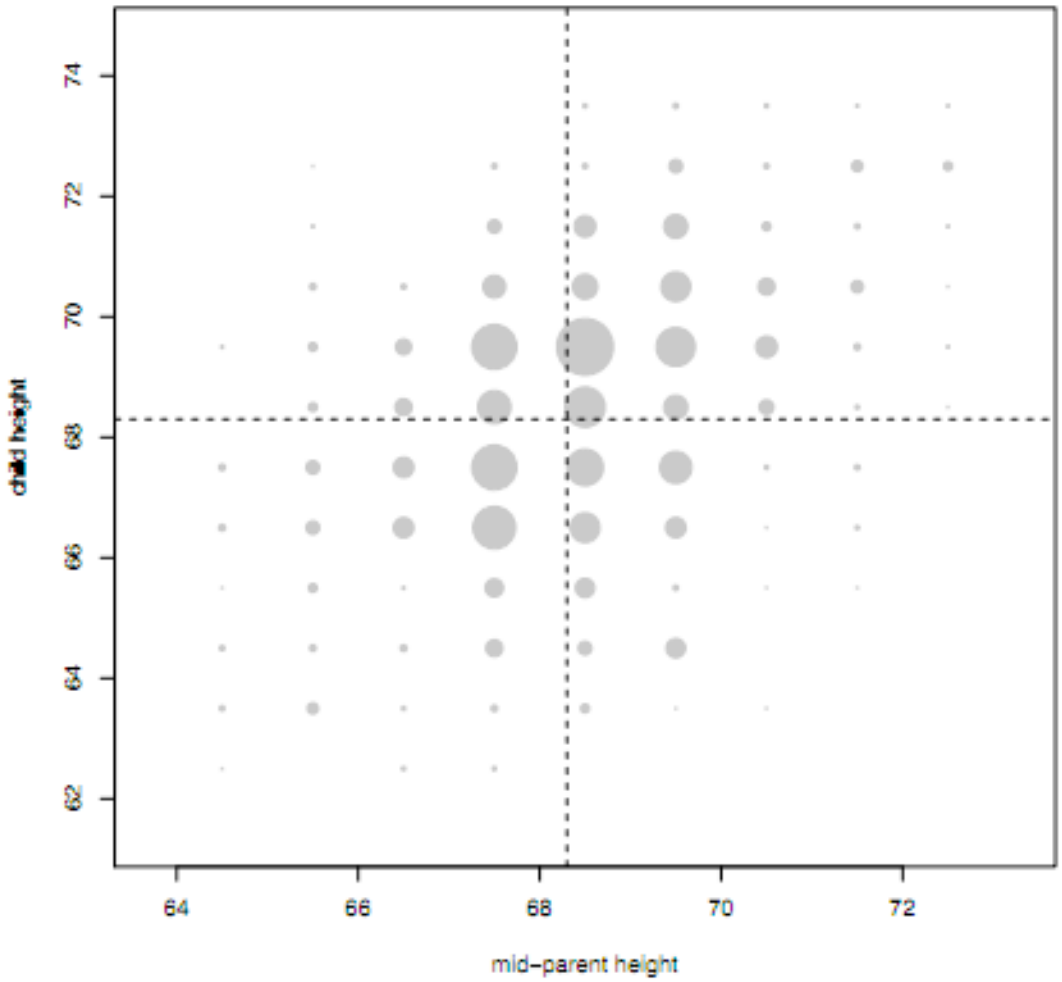
NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a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

Some history

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

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

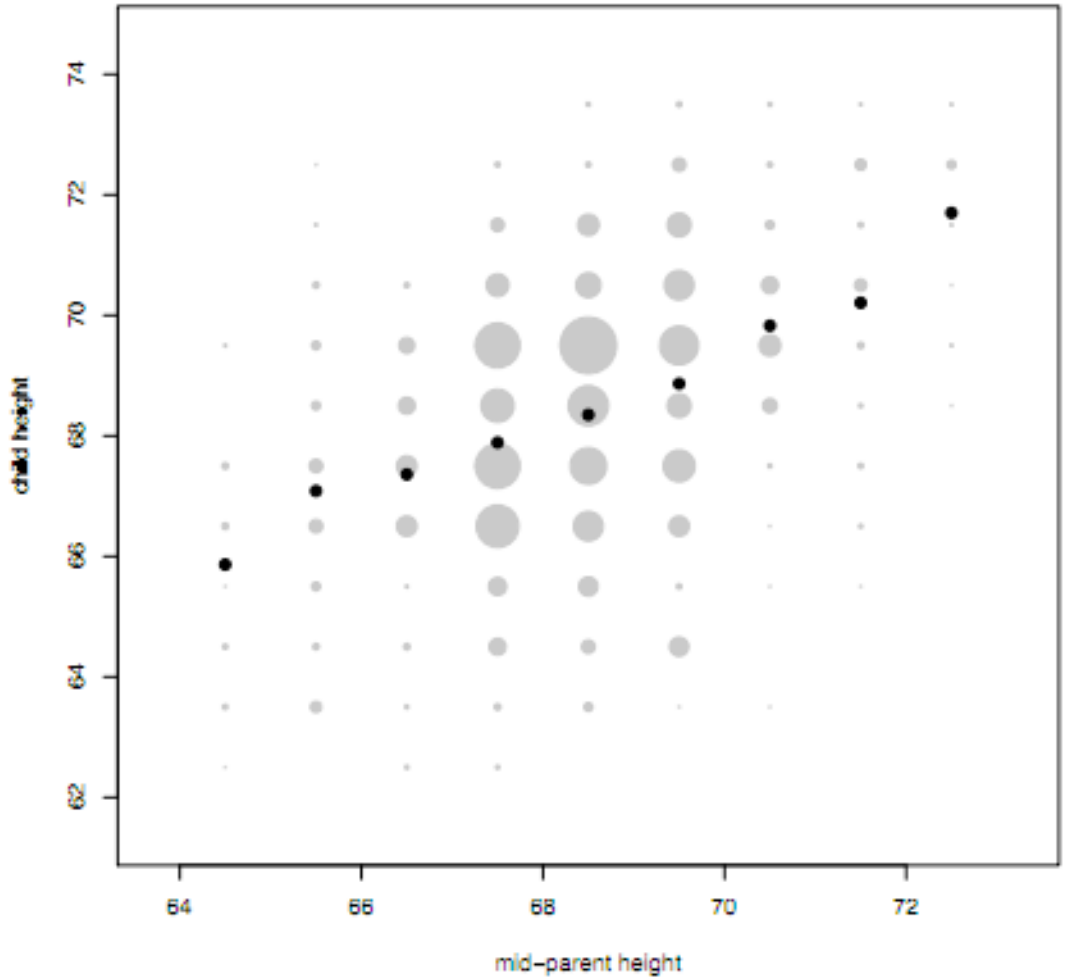
What do we notice about this pattern?



Some history

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

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



Some history

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

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

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

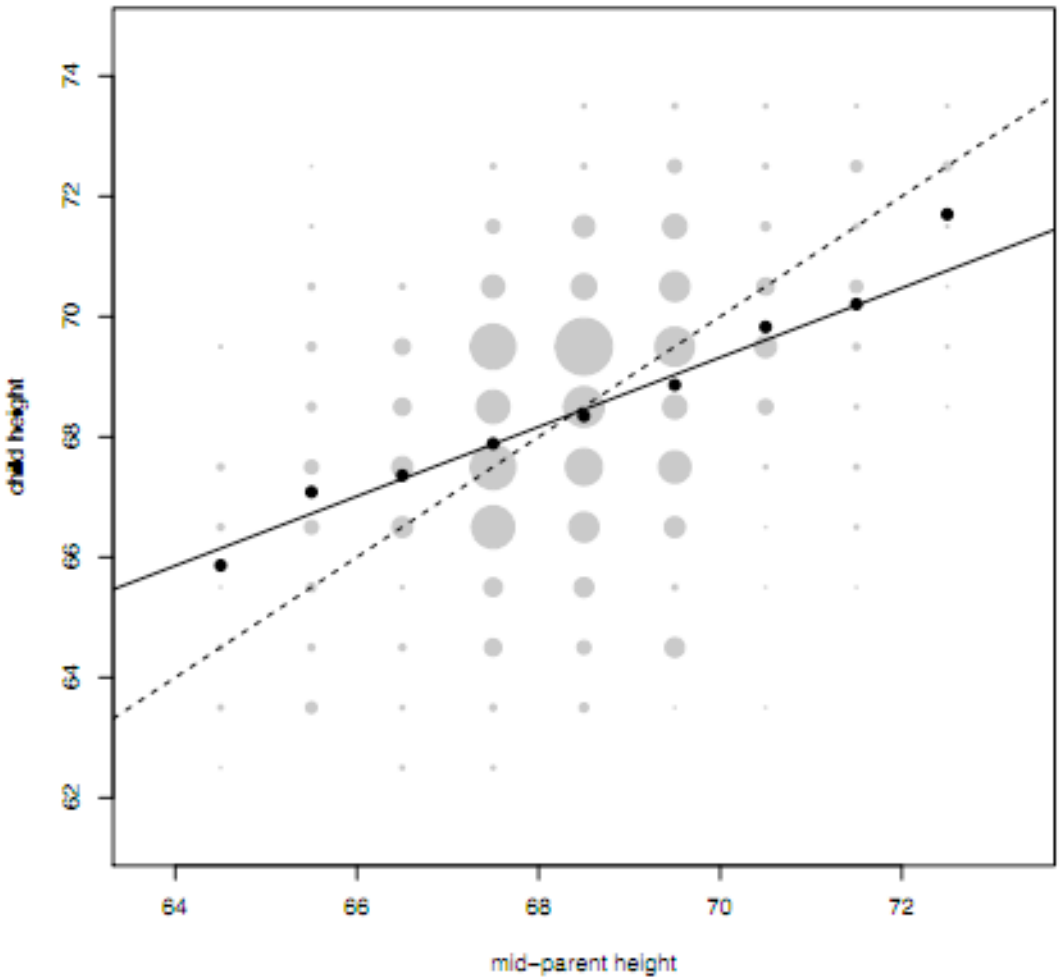


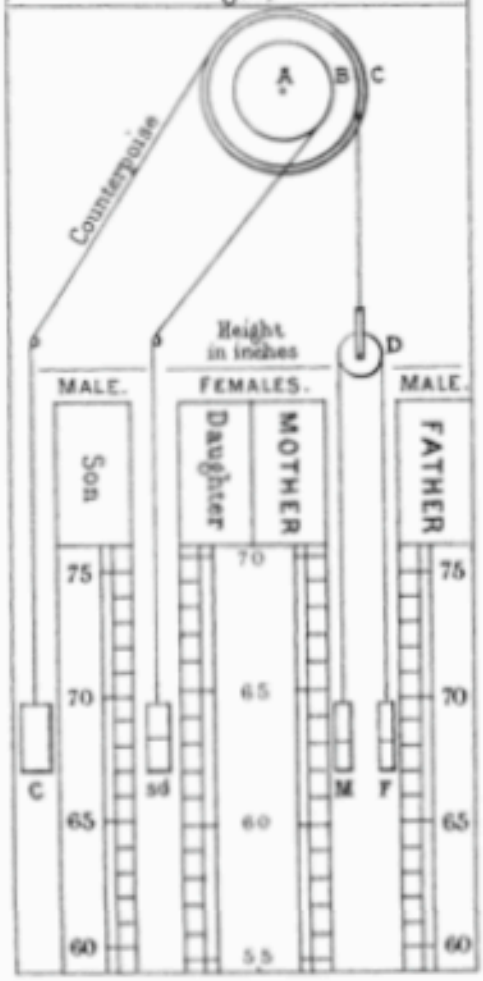
Plate IX.

RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)

FORECASTER OF STATURE

Fig. (b)



J F & W E Swaine, lith

Galton and regression

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

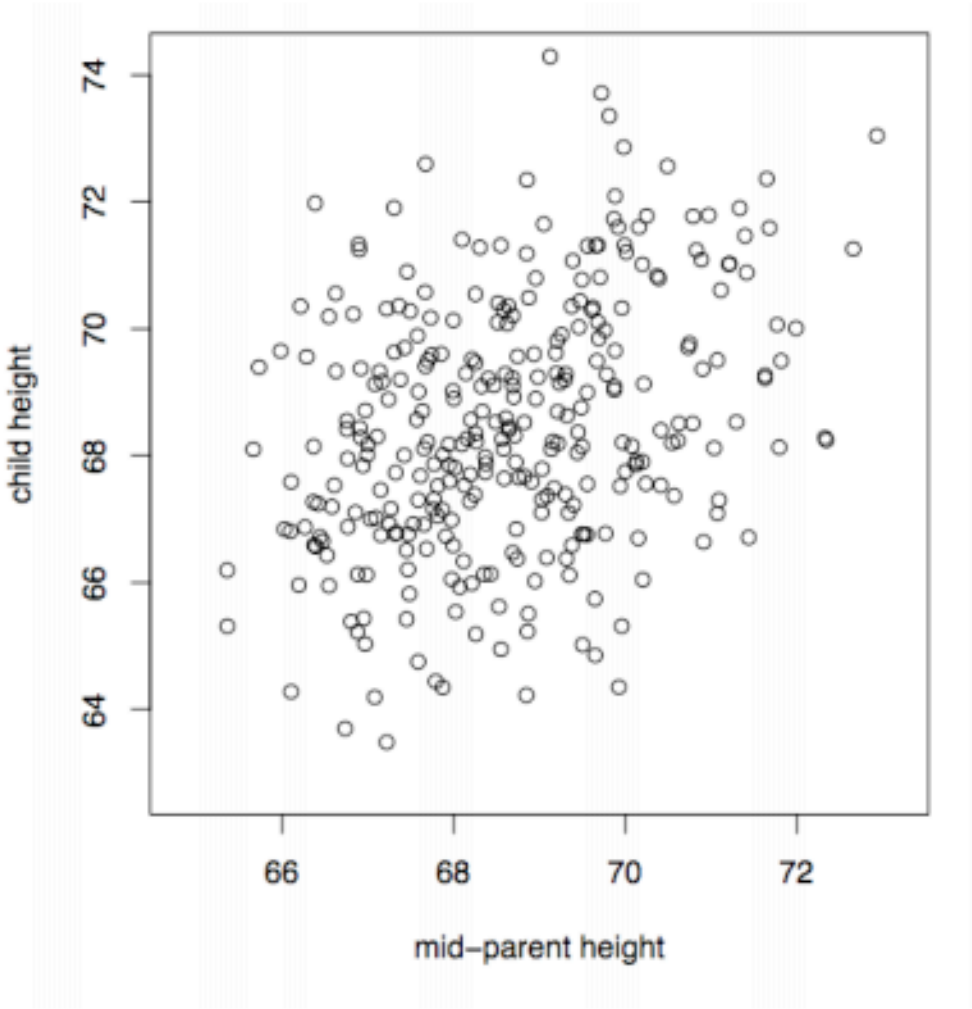
He then considered the heights of the children associated with different columns in his table, plotting median values against mid-parental height and finding a straight line (which he fit by eye)

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

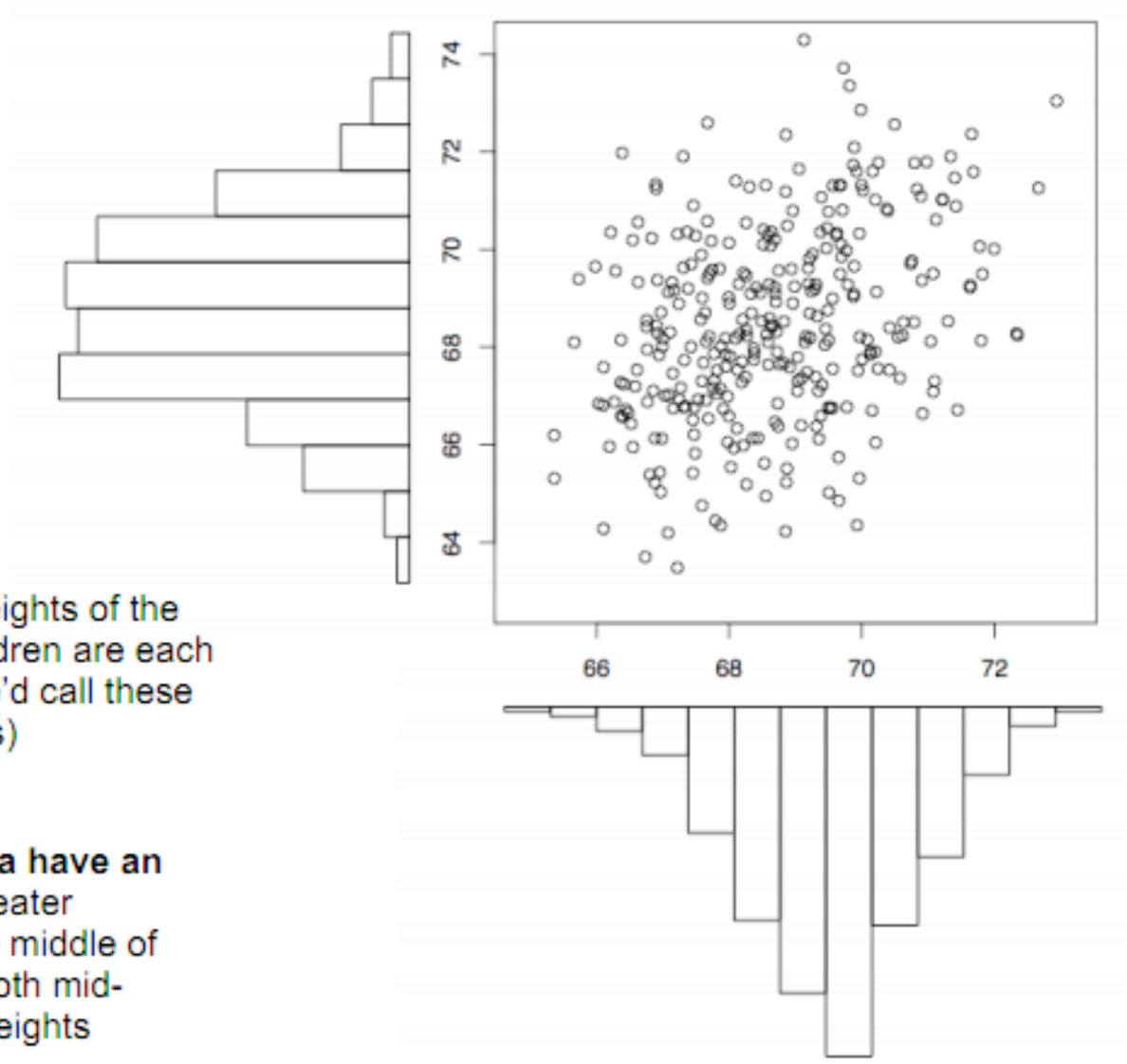
The bivariate normal

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

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



The bivariate normal



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

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

Galton and regression

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

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

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

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

Galton and regression

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

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

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

NATURAL INHERITANCE

BY

FRANCIS GALTON, F.R.S.

AUTHOR OF

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

FIG. 7.

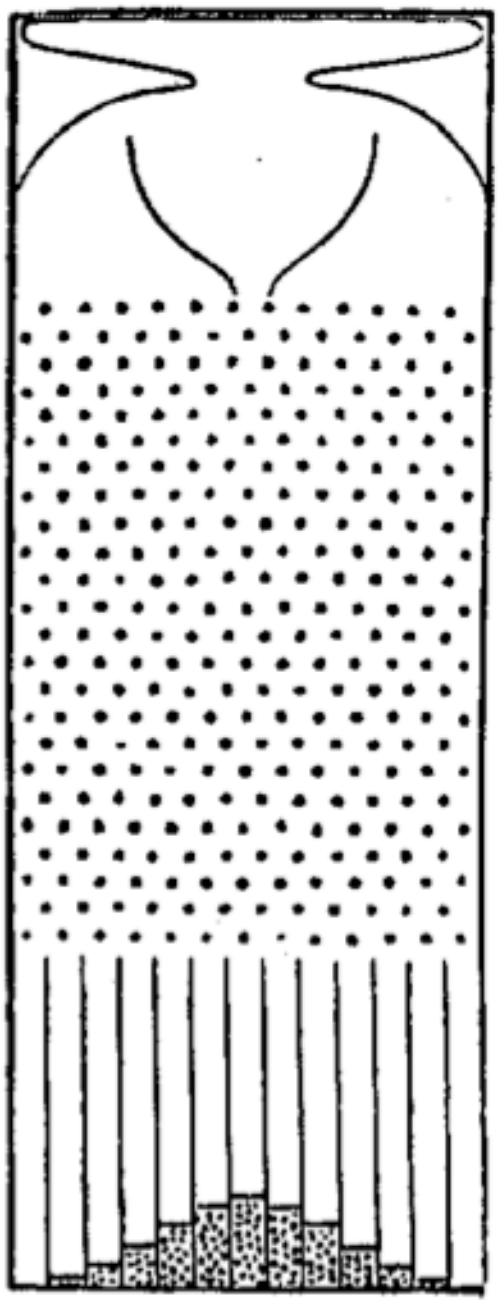


FIG. 8.

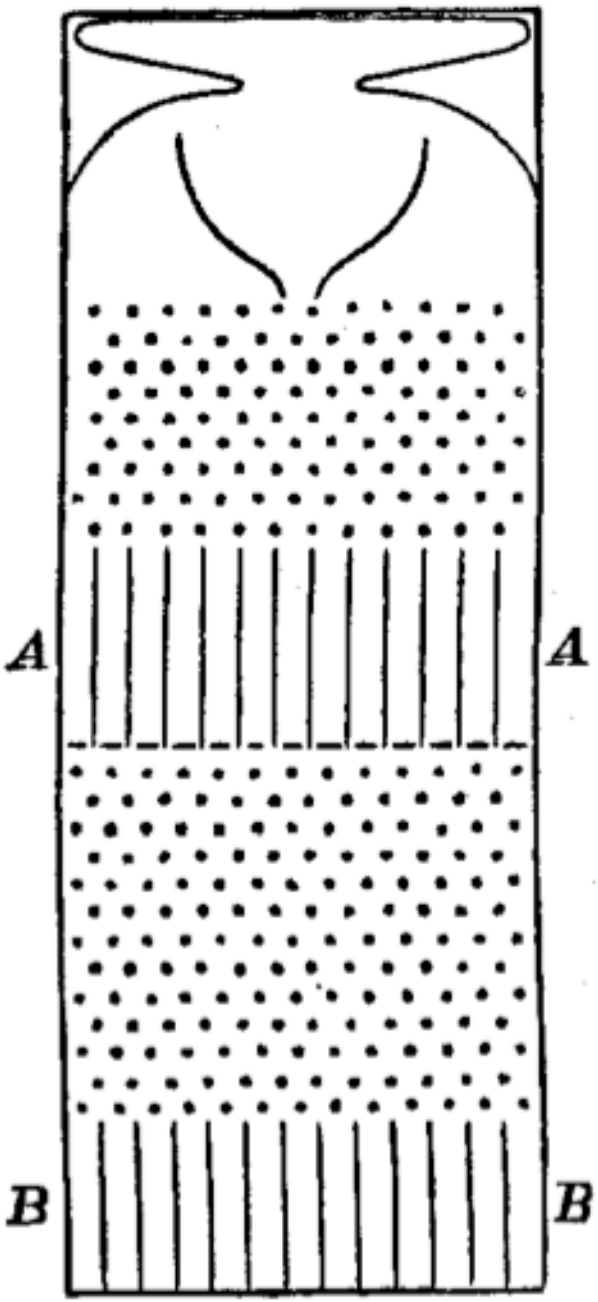
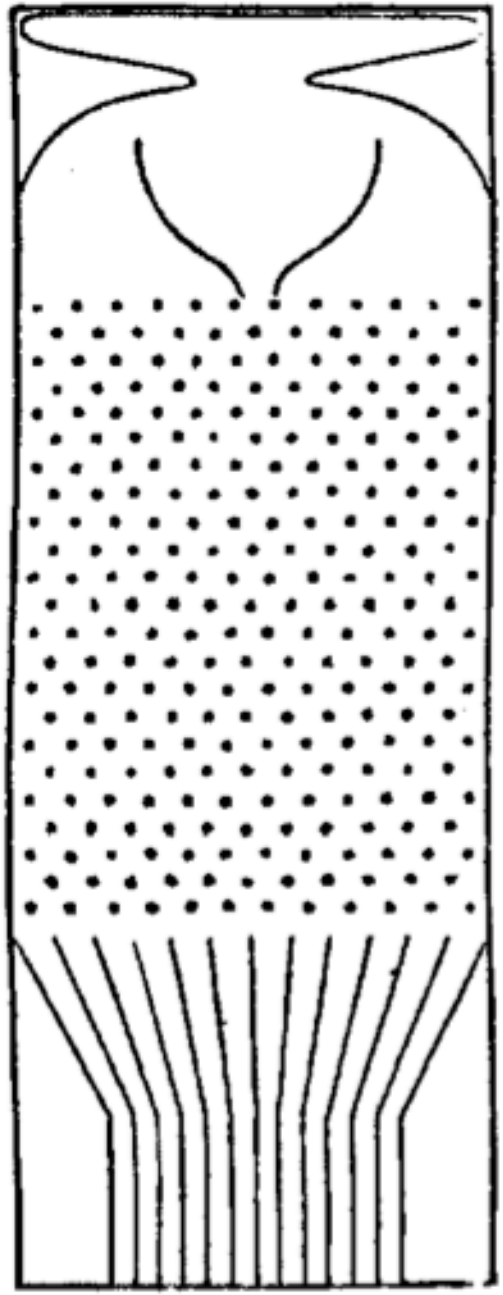


FIG. 9.



Galton and regression

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

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

TABLE I.

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

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

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

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

Table 13 (Special Data).

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

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

Galton and regression

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

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

Some history

Regression to the mean is not a fact about genetics really, it's a fact about statistics and **about regression analysis in general**

Next time, we will discuss this effect in a little more detail and introduce the notion of correlation (along with bootstrap confidence intervals, oh the joy!)

I wanted to close with a few of Galton's experiments with averages, with means; but experiments that are decidedly non-mathematical...

Lantern slide of composite photograph. Comprises faces of private soldiers collected by Chatham. Leonard Darwin. son of Charles Darwin. C. L. Darwin was in the Royal Engineers and was later President of the Eugenics Society.

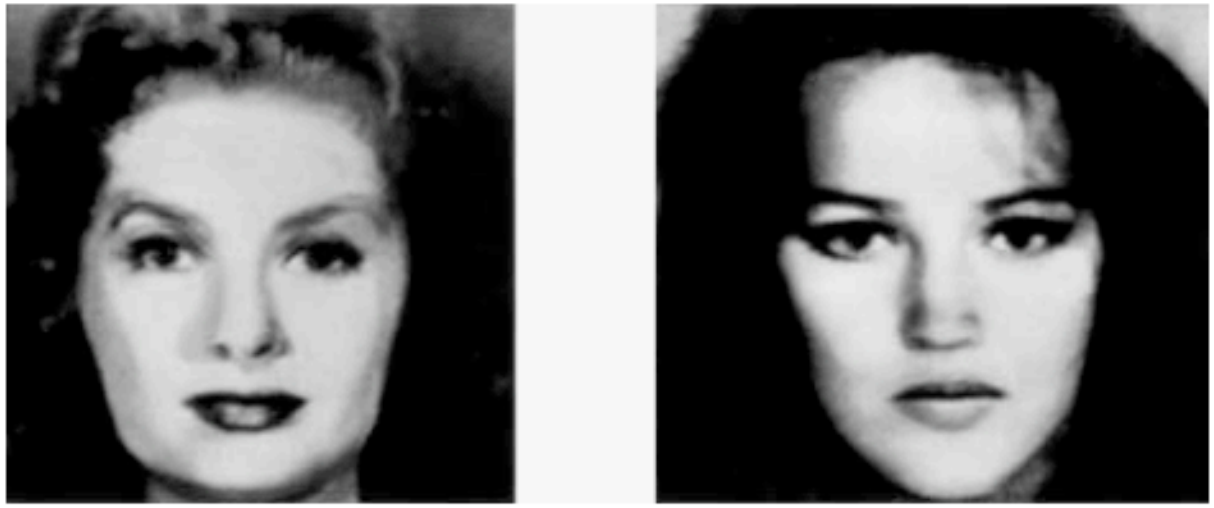
It might be expected that when many different portraits are fused into a single one, the result would be a mere smudge. Such, however, is by no means the case... There are then so many traits in common, to combine and to reinforce one another that they prevail to the exclusion of the rest. All that is common remains, all that is individual tends to disappear.

Composite pictures are... the equivalents of those large statistical tables whose totals divided by the number of cases and entered in the bottom line are the averages. They are real generalizations because they include the whole of the material under consideration.

Portraits of Napoleon prepared for making composites exhibited by Francis Galton at the Universal Exhibition, Paris, in 1889. Comprises 5 small photos of different likeness of Napoleon, as he appeared on various coins \medals. The 6th photo is a composite of the other 5.

Lantern slide of family composite photograph. Comprises photos of a father, mother, 2 sons & 2 daughters.

Lantern slide of composite photograph. Comprises 2 horses' heads - Raconteur & St Marnock. 1898



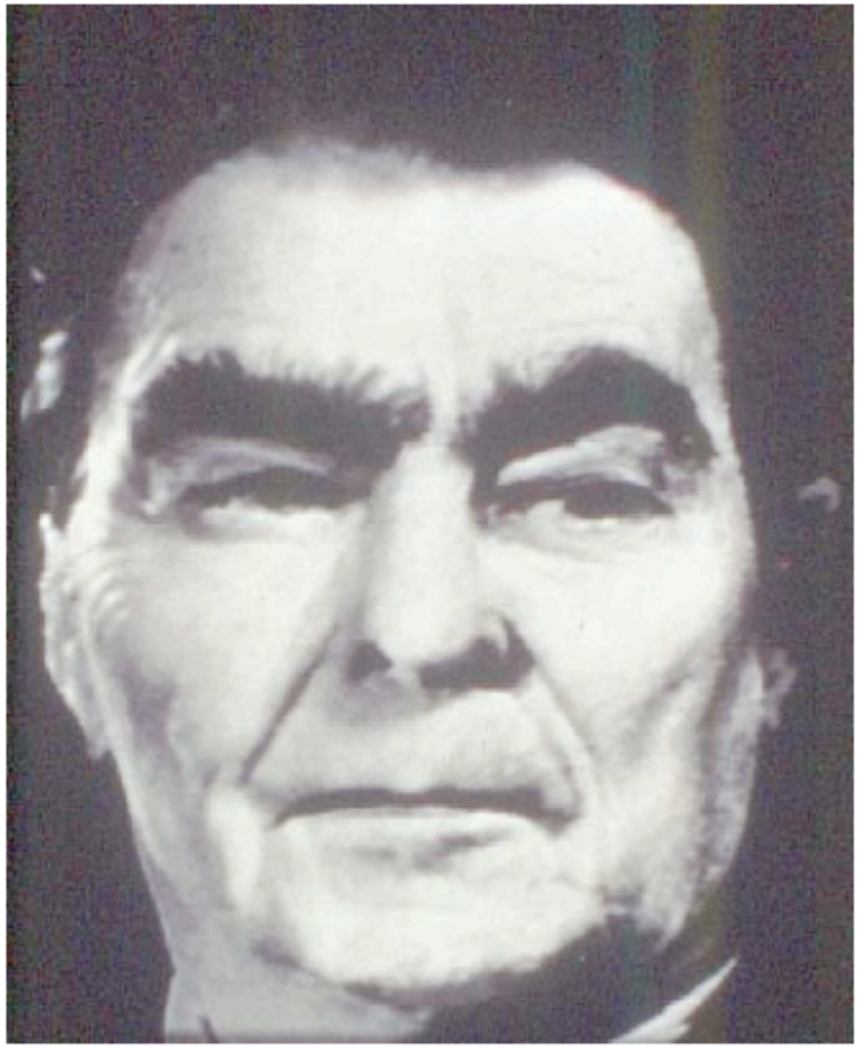
Nancy Burson *First and Second Beauty Composite* 1984

One of the first artists to use digitial technologies, Burson's early focus upon faces led to hauntingly altered portraits. *First and Second Beauty Composite* (1984) combines silver screen stars Bette Davis, Monroe, Loren, Hepburn and Grace Kelly into an amalgamated lush beauty. The "Second" is a more angular and androgynous mix of Jane Fonda, Brooke Shields, Meryl Streep, Diane Keaton and Jacqueline Bisset.

To link us back to Galton and the criminal fixation, Burson later combined three assassins, Lee Harvey Oswald, Sirhan Sirhan and James Earl Ray.

Nancy Burson, *Androgyny* (6 Men + 6 Women), 1982

As a visual experiment, Burson combined the faces of six men and six women, attempting to see which gender would dominate. She found that if you cover the mouth, the face appears more feminine.



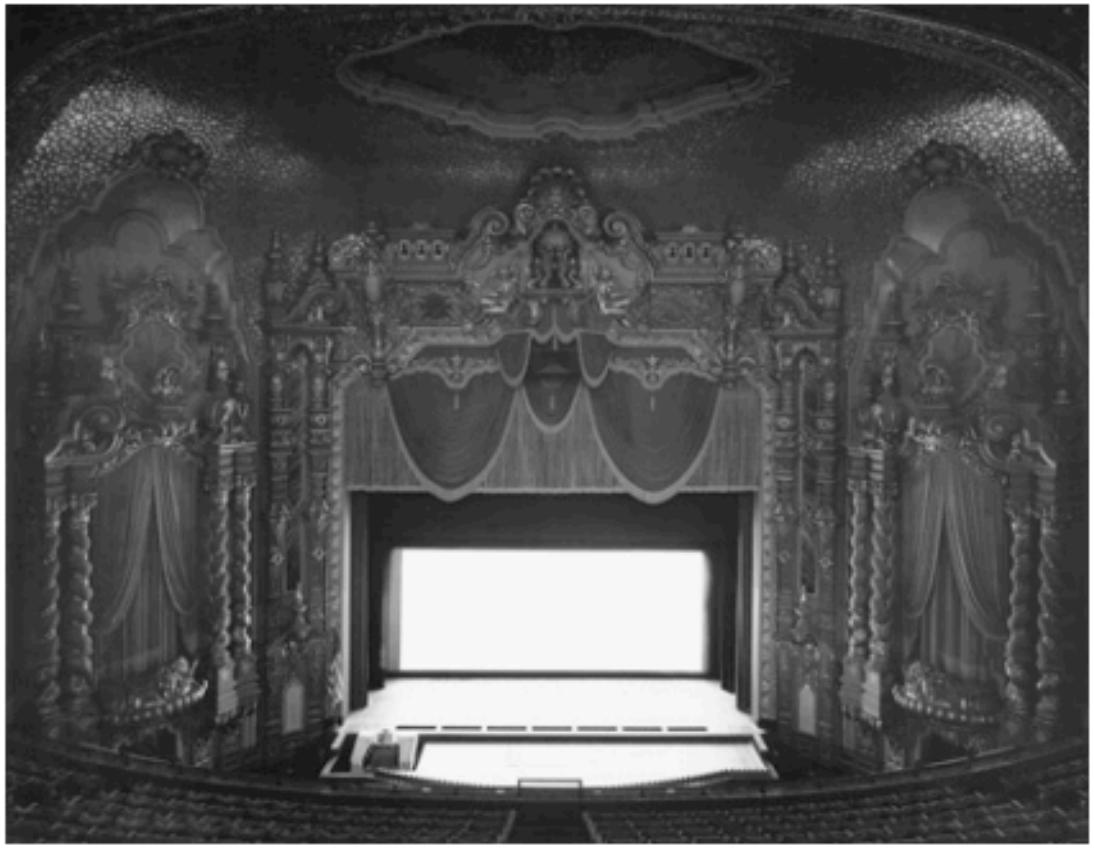
Nancy Burson *Warhead I* 1982

The composition -- 55 percent Reagan, 45 percent Brezhnev, and 1 percent Deng, Thatcher and Mitterand -- is still chilling, 20 years after its creation.

I am a habitual self-interlocutor. One evening while taking photographs at the American Museum of Natural History, I had a near-hallucinatory vision. My internal question-and-answer session leading up to this vision went something like this: "Suppose you shoot a whole movie in a single frame?" The answer: "You get a shining screen." Immediately I began experimenting in order to realize this vision. One afternoon I walked into a cheap cinema in the East Village with a large-format camera. As soon as the movie started, I fixed the shutter at a wide-open aperture. When the movie finished two hours later, I clicked the shutter closed. That evening I developed the film, and my vision exploded before my eyes.

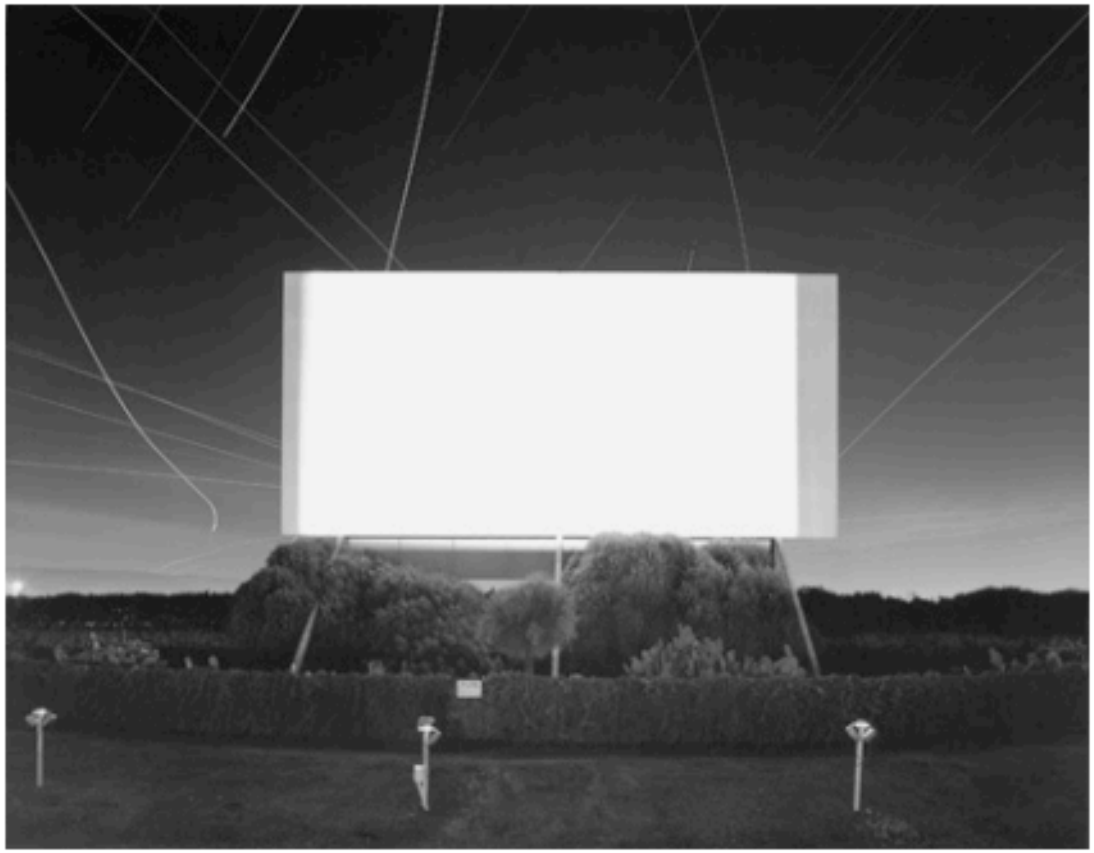
-Hiroshi Sugimoto

Radio City Music Hall, New York, 1978, private collection.



Ohio Theater, Ohio, 1980, private collection.

valon Theater, Catalina Island, 1993, private collection.



Union City Drive-In, Union City, 1993, private collection.

In the Class of 1967 & Class of 1988 suite of prints, a custom averaging process is applied to graduating yearbook photos from my family history. The Class of 1988 is an amalgamation of all of the young men and women in my graduating high school class. The Class of 1967 is composed of all the members from my mother's graduating class from the same hometown, Fort Worth, Texas.

Jason Salavon

Extending an interest in the everyday to the special occasion, each of these works utilize 100 unique commemorative photographs culled from the net. In a recent development, the final compositions are arrived at using both the mean and median. This gives the data (and resulting meta-portrait) a "crunchier" quality, splitting the difference between a specific norm and an ideal one.

Jason Salavon, *100 Special Moments*