Statistics

'There are lies, damned lies, and statistics.'

(Mark Twain)

Statistics

- ‘Statistics is the art of never having to say you’re wrong.’

- ‘ . . . mysterious, sometimes bizarre, manipulations performed upon the collected data of an experiment in order to obscure the fact that the results have no generalizable meaning for humanity. Commonly, computers are used, lending an additional aura of unreality to the proceedings.’

A Definition of Statistics

Statistics

is a collection of

procedures and principles

for gaining and

processing information

in order to make decisions

when faced with uncertainty.

Does aspirin prevent heart attacks?

In 1988 the Steering Committee of the Physicians' Health Study Research Group in the US published results of a 5-year study to determine the effects upon heart attacks of taking an aspirin every other day. The study had involved 22,071 male physicians aged 40 to 84. The results were

Condition	**Heart attack**	**No heart attack**	**Attacks per 1000**
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

What can make of this data? Is it evidence for the hypothesis that aspirin prevents heart attacks?

MLE and decision-making

You and a friend have agreed to meet sometime just after 12 noon. You have arrived at noon, have waited 5 minutes and your friend has not shown up. You believe that either your friend will arrive at X minutes past 12, where you believe X is exponentially distributed with an unknown parameter λ, $\lambda > 0$, or that she has completely forgotten and will not show up at all. We can associate the later event with the parameter value $\lambda = 0$. Then

$$\begin{aligned}\mathbb{P}(\text{data} \mid \lambda) &= \mathbb{P}(\text{you wait at least 5 minutes} \mid \lambda) \\ &= \int_5^\infty \lambda e^{-\lambda t}\, dt \\ &= e^{-5\lambda}\,.\end{aligned}$$

Thus the maximum likelihood estimator for λ is $\hat{\lambda} = 0$.

If you base your decision as to whether or not you should wait a bit longer only upon the maximum likelihood estimator of λ, then you will estimate that your friend will never arrive and decide not to wait. This argument holds even if you have only waited 1 second.

Example 6.1

It has been suggested that dying people may be able to postpone their death until after an important occasion. In a study of 1919 people with Jewish surnames it was found that 922 occurred in the week before Passover and 997 in the week after. Is there any evidence in this data to reject the hypothesis that a person is as likely to die in the week before as in the week after Passover?

Example 6.2

In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions of genetic theory.

type	observed count	prediction frequency	expected count
smooth yellow	315	$9/16$	312.75
smooth green	108	$3/16$	104.25
wrinkled yellow	102	$3/16$	104.25
wrinkled green	31	$1/16$	34.75

Is there any evidence in this data to reject the hypothesis that theory is correct?

Example 9.1

In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions of genetic theory.

type i	observed count o_i	prediction frequency	expected count e_i
smooth yellow	315	9/16	312.75
smooth green	108	3/16	104.25
wrinkled yellow	102	3/16	104.25
wrinkled green	31	1/16	34.75

Is there any evidence in this data to reject the hypothesis that theory is correct?

Here the Pearson chi-squared statistic is

$$\begin{aligned}\sum_{i=1}^{4}\frac{(o_i-e_i)^2}{e_i} &= \frac{(315-312.75)^2}{312.75}+\frac{(108-104.25)^2}{104.25}\\ &\quad+\frac{(102-104.25)^2}{104.25}+\frac{(31-34.75)^2}{34.75}\\ &= 0.618.\end{aligned}$$

Here $|\Theta_1| = 3$ and $|\Theta_0| = 0$. So under H_0 the test statistic is approximately χ_3^2, for which the 10% and 95% points are 0.584 and 7.81. Thus we certainly do not reject the theoretical model. Indeed, we would expect the observed counts to show even greater disparity from the theoretical model about 90% of the time.

Example 9.2

Here we have observed (and expected) counts for the study about aspirin and heart attacks described in Example 1.2.

We wish to test the hypothesis that the probability of heart attack or no heart attack is the same in the two rows.

	Heart attack	**No heart attack**	**Total**
	o_{i1} (e_{i1})	o_{i2} (e_{i2})	
Aspirin	104 (146.52)	10,933 (10890.5)	11,037
Placebo	189 (146.48)	10,845 (10887.5)	11,034
Total	293	21,778	22,071

E.g., $e_{11} = \left(\frac{293}{22071}\right) 11037 = 146.52.$

The χ^2 statistic is

$$\sum_{i=1}^{2} \sum_{i=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(104 - 146.52)^2}{146.52} + \frac{(189 - 146.48)^2}{46.48}$$

$$+ \frac{(10933 - 10890.5)^2}{10890.5} + \frac{(10845 - 10887.5)^2}{10887.5}$$

$$= 25.01\,.$$

The 95% point of χ_1^2 is 3.84. Since $25.01 >> 3.84$, we reject the hypothesis that heart attack rate is independent of whether the subject did or did not take aspirin.

Example 9.3

A researcher pretended to drop pencils in a lift and observed whether the other occupant helped to pick them up.

	Helped	**Did not help**	**Total**
Men	370 (337.171)	950 (982.829)	1,320
Women	300 (332.829)	1,003 (970.171)	1,303
Total	670	1,953	2,623

E.g. $e_{11} = \hat{p}_1 \hat{q}_1 n = \left(\frac{1320}{2623}\right)\left(\frac{670}{2623}\right) 2623 = 337.171.$

$$\sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 8.642.$$

This is significant compared to χ^2_1 whose 5% point is 3.84.

Example 10.1 (Simpson's paradox)

These are some Cambridge admissions statistics for 1996.

	Women			**Men**		
	applied	accepted	**%**	applied	accepted	**%**
Computer Science	26	7	**27**	228	58	**25**
Economics	240	63	**26**	512	112	**22**
Engineering	164	52	**32**	972	252	**26**
Medicine	416	99	**24**	578	140	**24**
Veterinary medicine	338	53	**16**	180	22	**12**
Total	1184	274	**23**	2470	584	**24**

In all five subjects women have an equal or better success rate in applications than do men. However, taken overall, 24% of men are successful but only 23% of women are successful.

Sexual activity and the lifespan

In 'Sexual activity and the lifespan of male fruitflies', *Nature*, 1981, Partridge and Farquhar report experiments which examined the cost of increased reproduction in terms of reduced longevity for male fruitflies. They kept numbers of male flies under different conditions. 25 males in one group were each kept with 1 receptive virgin female. 25 males in another group were each kept with 1 female who had recently mated. Such females will refuse to remate for several days. These served as a control for any effect of competition with the male for food or space. The groups were treated identically in number of anaesthetizations (using CO2) and provision of fresh food.

To verify 'compliance' two days per week throughout the life of each experimental male, the females that had been supplied as virgins to that male were kept and examined for fertile eggs. The insemination rate declined from approximately 1 per day at age one week to about 0.6 per day at age eight weeks.

Fruitfly data

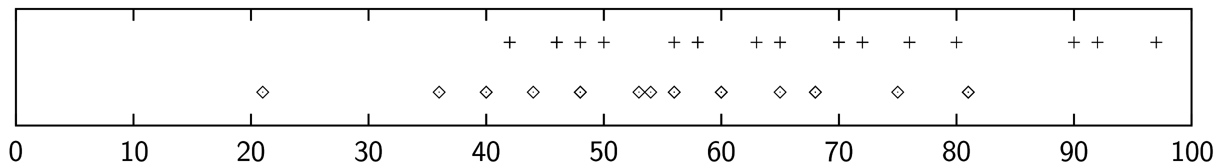
Here are summary statistics

Groups of 25 males kept with	mean life (days)	s.e.
1 uninterested female	64.80	15.6525
1 interested female	56.76	14.9284

It is interesting to look at the data, and doing so helps us check that lifespan is normally distributed about a mean. The longevities for control and test groups were

42 42 46 46 46 48 50 56 58 58 63 65 65 70 70 70 70 72 72 76 76 80 90 92 97

21 36 40 40 44 48 48 48 48 53 54 56 56 60 60 60 60 65 68 68 68 75 81 81 81



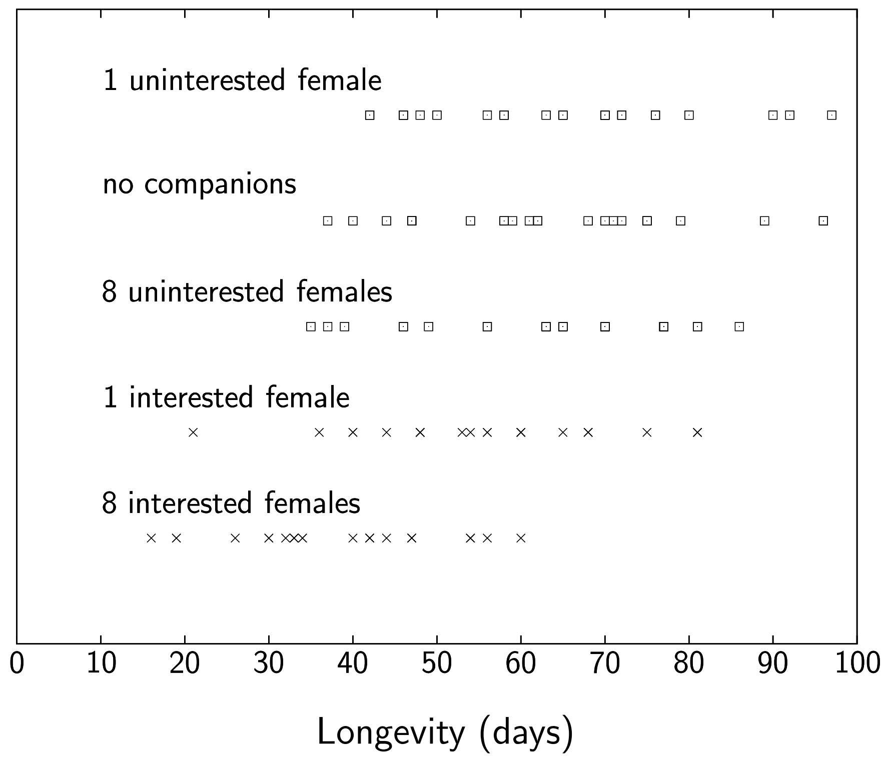
Jogging and pulse rate

Does jogging lead to a reduction in pulse rate? Eight non-jogging volunteers engaged in a one-month jogging programme. Their pulses were taken before and after the programme.

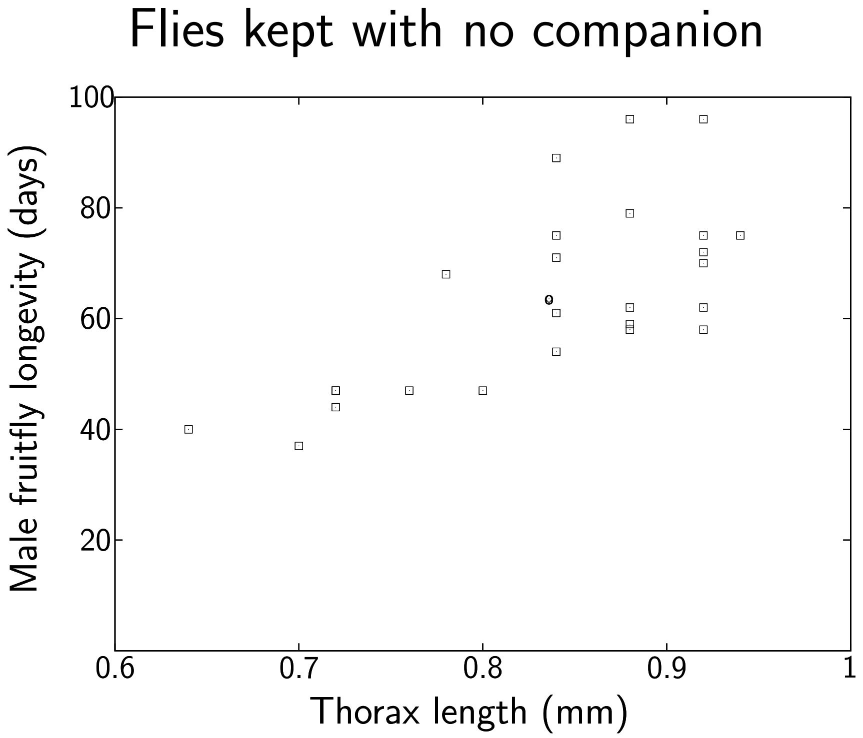
pulse rate before	74	86	98	102	78	84	79	70
pulse rate after	70	85	90	110	71	80	69	74
decrease	4	1	8	-8	7	4	10	-4

Fruitfly data

Groups of 25 males kept with	mean life (days)	s.e.	size (mm)	s.e.	sleep (%/day)	s.e.
no companions	63.56	16.4522	0.8360	0.084261	21.56	12.4569
1 uninterested female	64.80	15.6525	0.8256	0.069886	24.08	16.6881
1 interested female	56.76	14.9284	0.8376	0.070550	25.76	18.4465
8 uninterested females	63.36	14.5398	0.8056	0.081552	25.16	19.8257
8 interested females	38.72	12.1021	0.8000	0.078316	20.76	10.7443

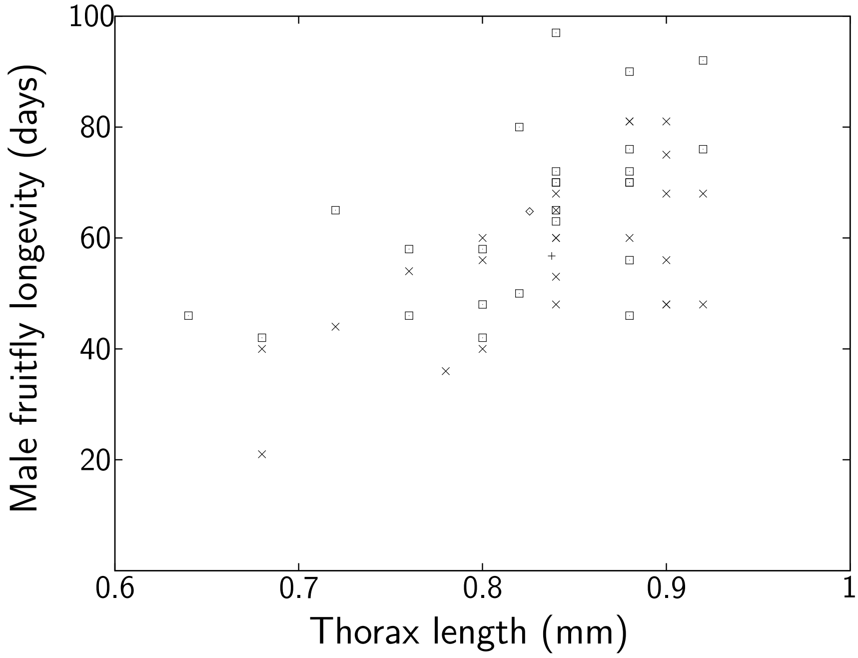


Fruitfly data

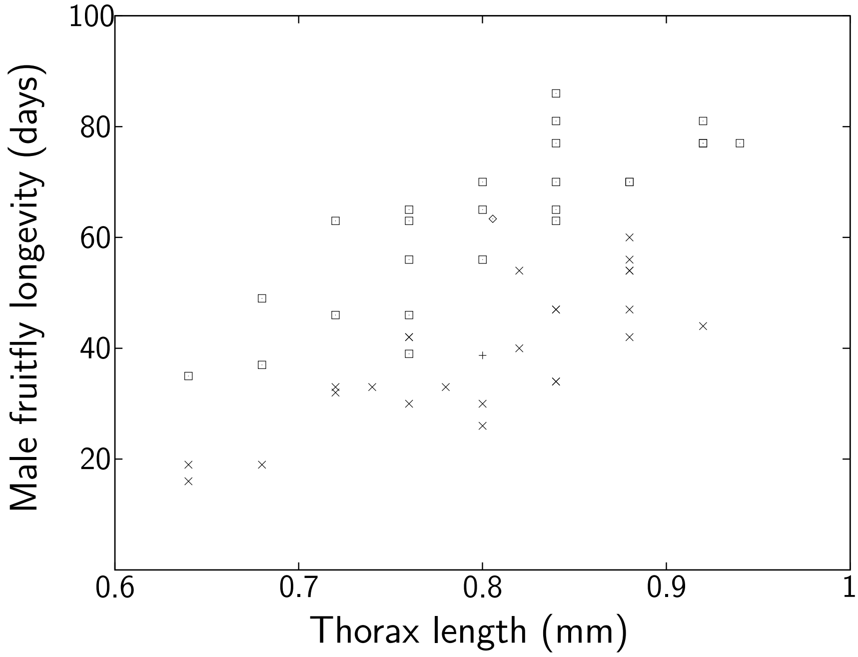


Fruitfly data

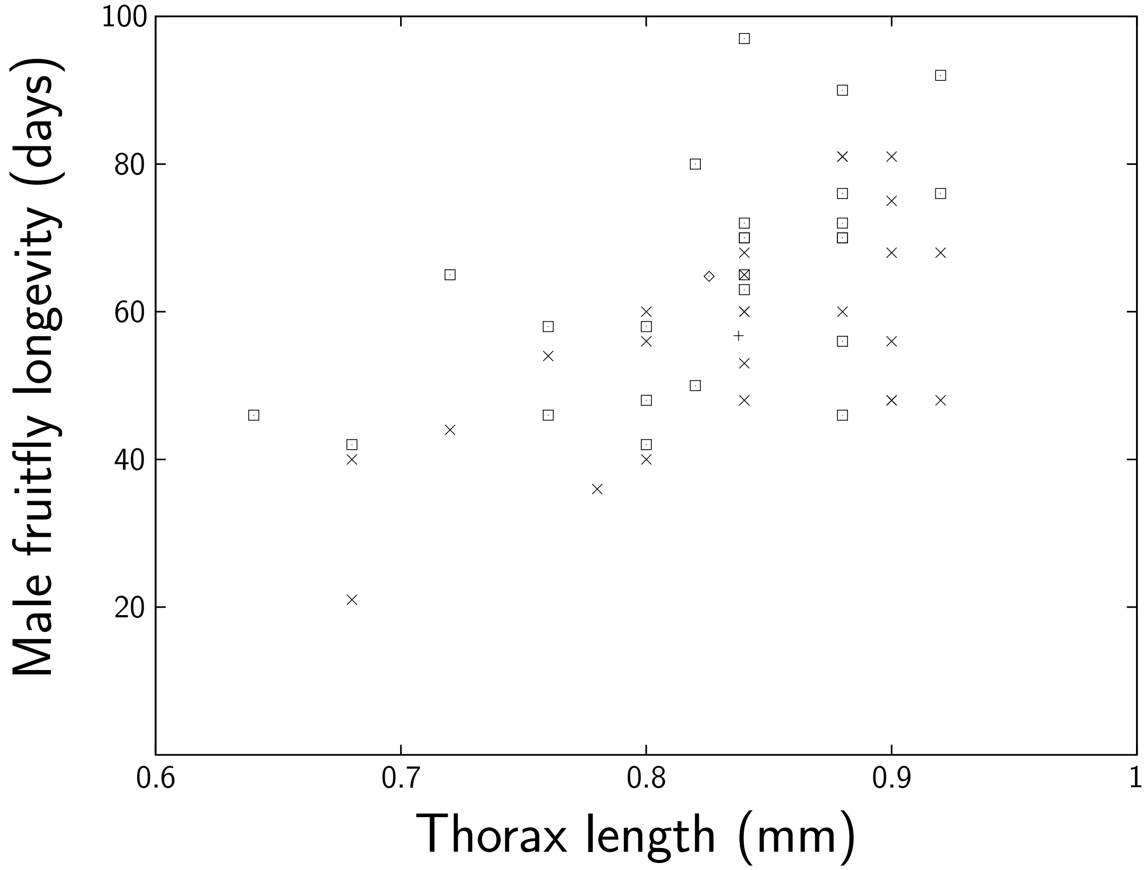
Flies kept with 1 female



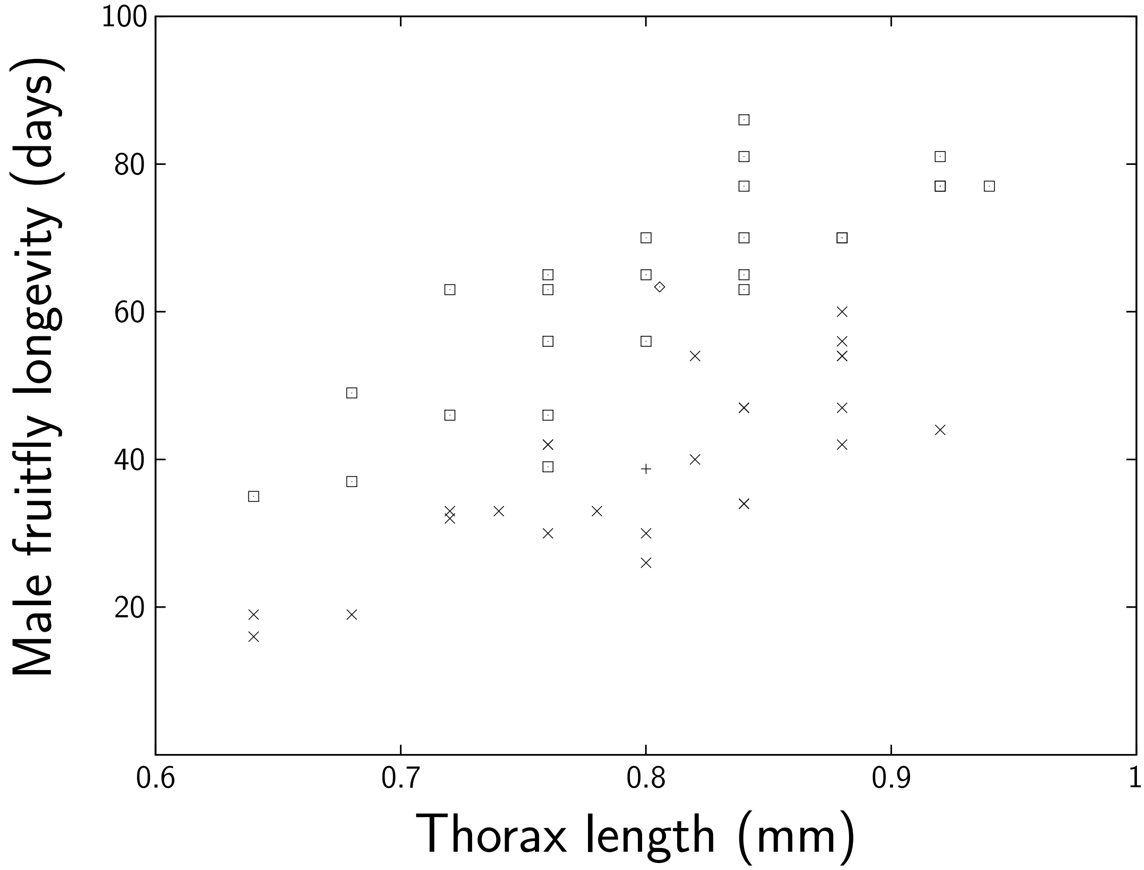
Flies kept with 8 females



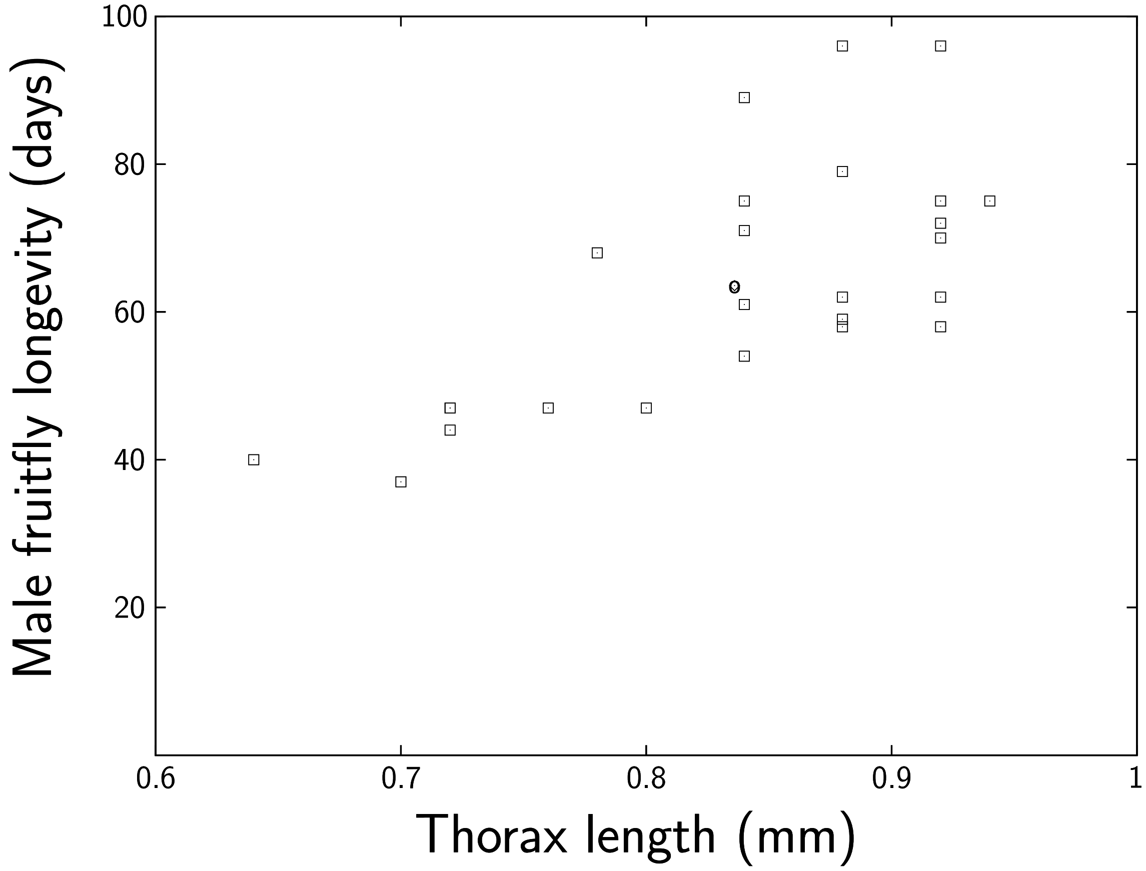
Flies kept with 1 female



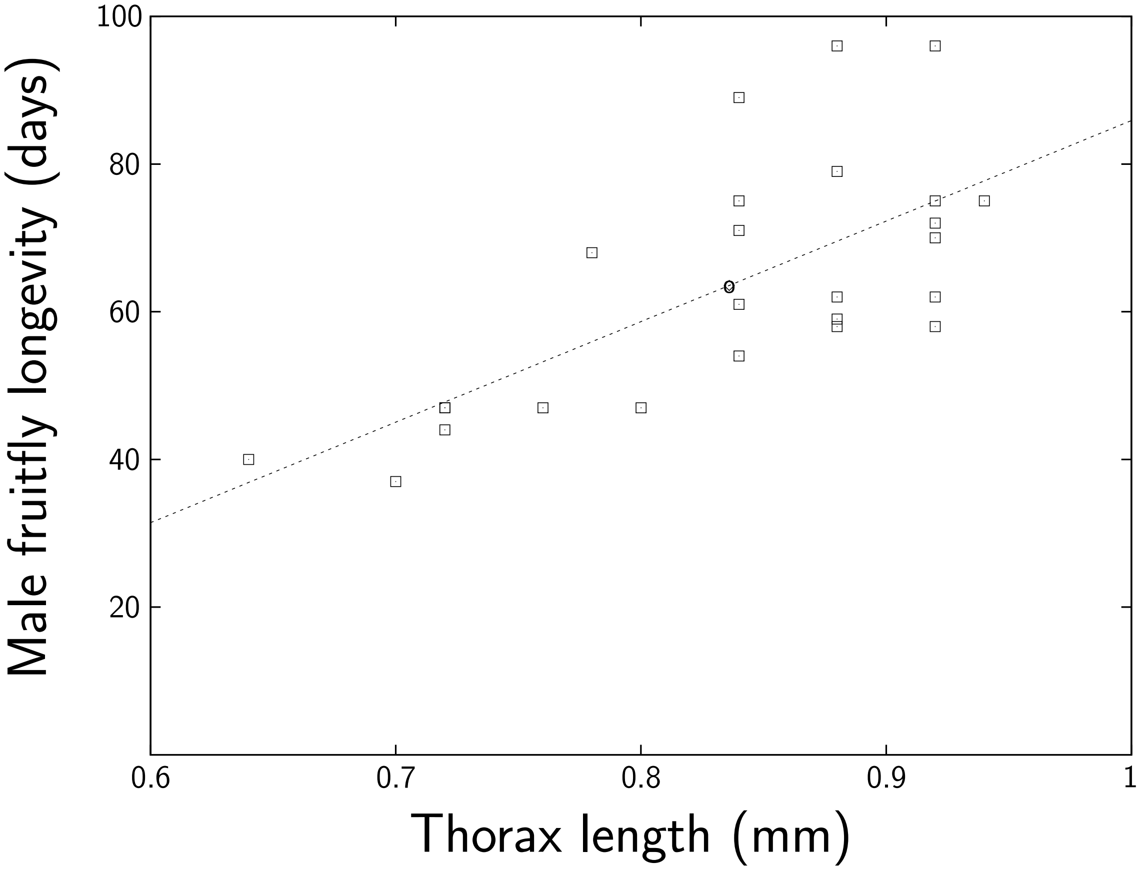
Flies kept with 8 females



Flies kept with no companions



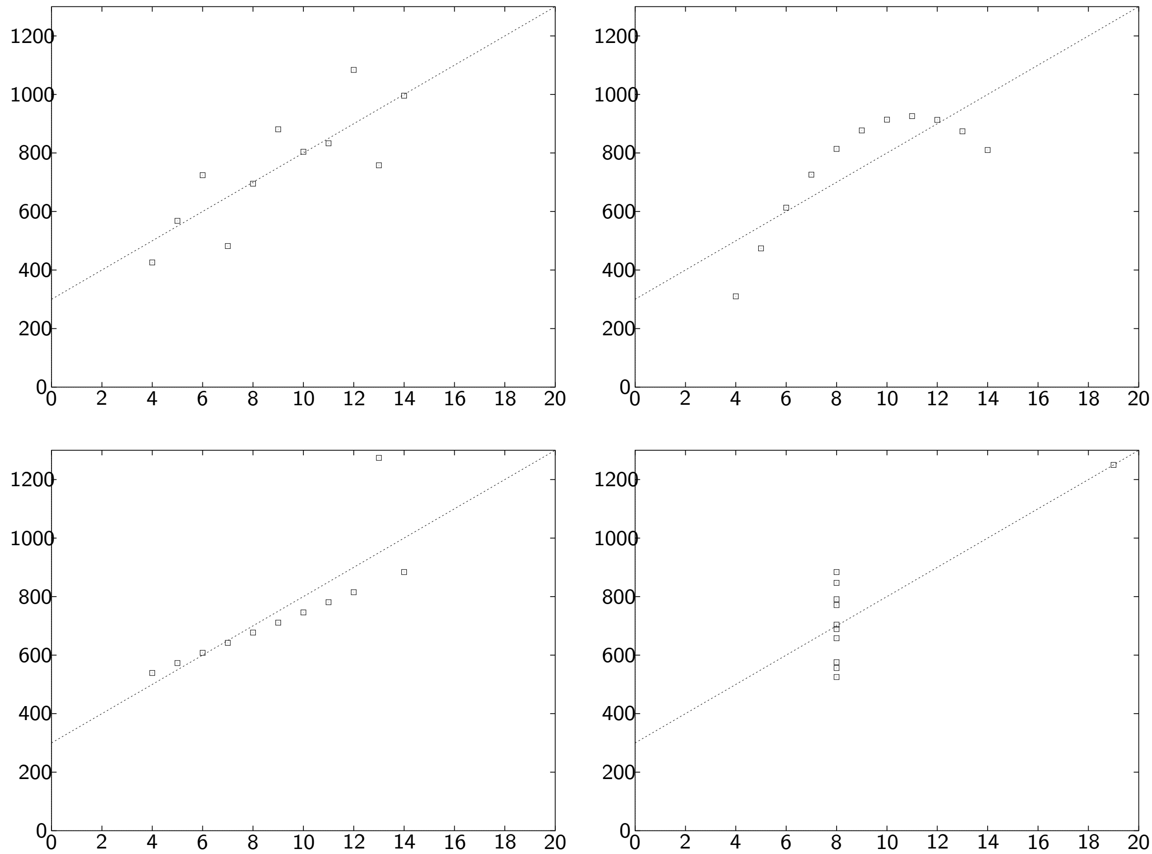
Flies kept with no companions



The regression line of longevity (y) against thorax size (x) is

$$y = -50.242 + 136.1268x\,.$$

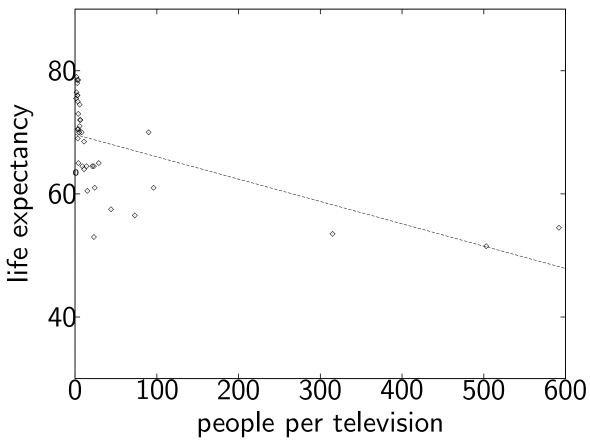
Data sets with the same summary statistics

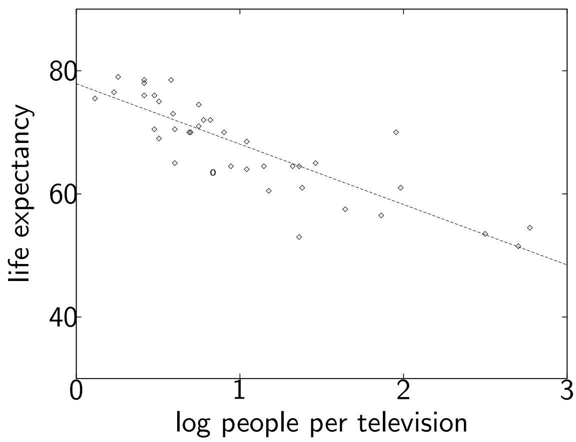


4	426	4	310	4	539	8	525
5	568	5	474	5	573	8	556
6	724	6	613	6	608	8	576
7	482	7	726	7	642	8	658
8	695	8	814	8	677	8	689
9	881	9	877	9	711	8	704
10	804	10	914	10	746	8	771
11	833	11	926	11	781	8	791
12	1084	12	913	12	815	8	847
13	758	13	874	13	1274	8	884
14	996	14	810	14	884	19	1250

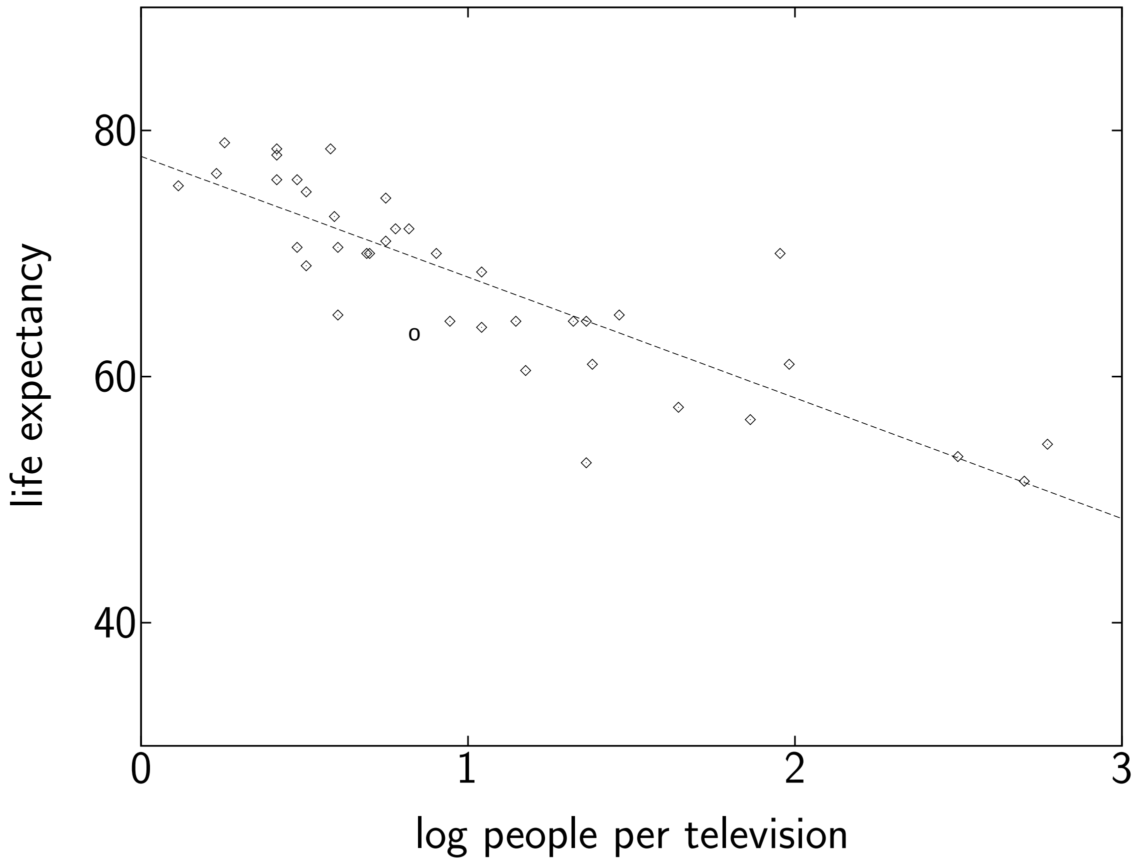
Life expectancy and people per television

country	mean life expectancy, y	people per television, u	people per doctor, v
Argentina	70.5	4.0	370
Bangladesh	53.5	315.0	6166
Brazil	65.0	4.0	684
⋮			⋮
United Kingdom	76.0	3.0	611
United States	75.5	1.3	404
Venezuela	74.5	5.6	576
Vietnam	65.0	29.0	3096
Zaire	54.0	*	23193



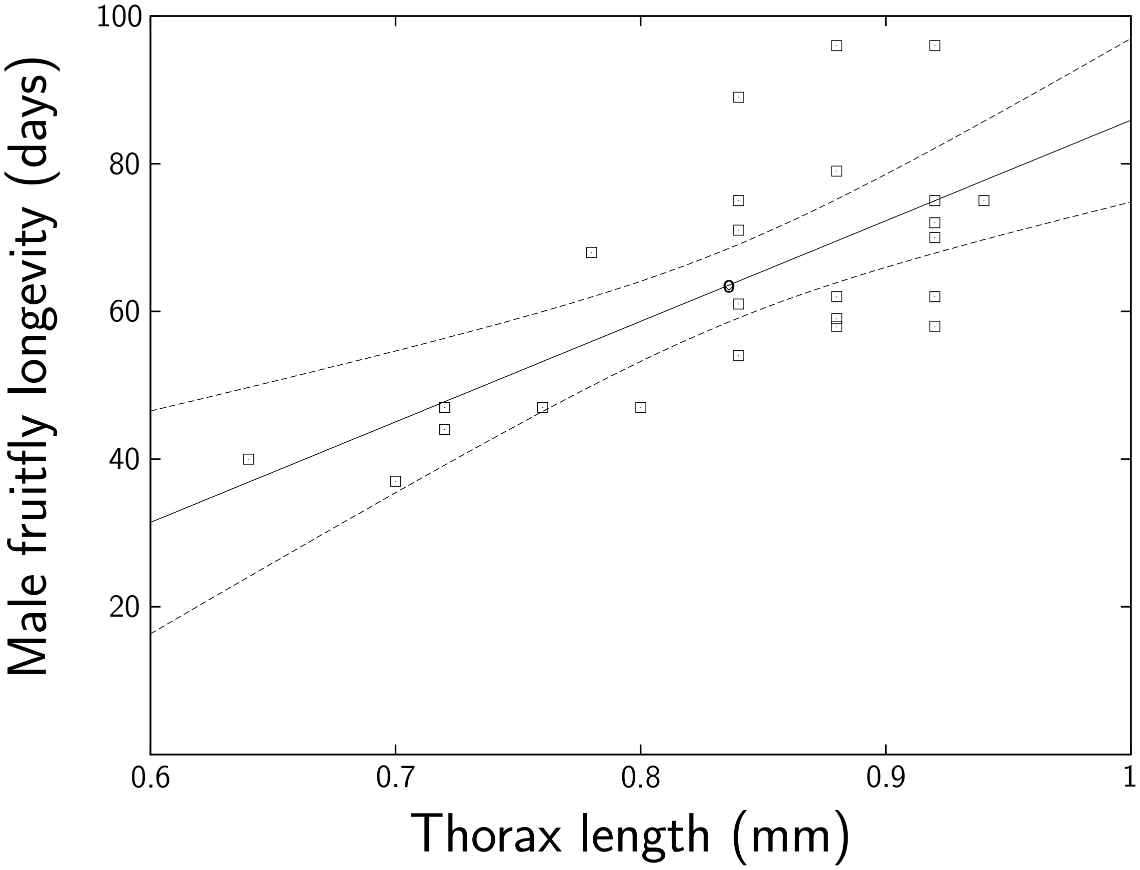


Life expectancy against log people per television



Flies kept with no companions

95% confidence bands for $\boldsymbol{a + \beta x}$

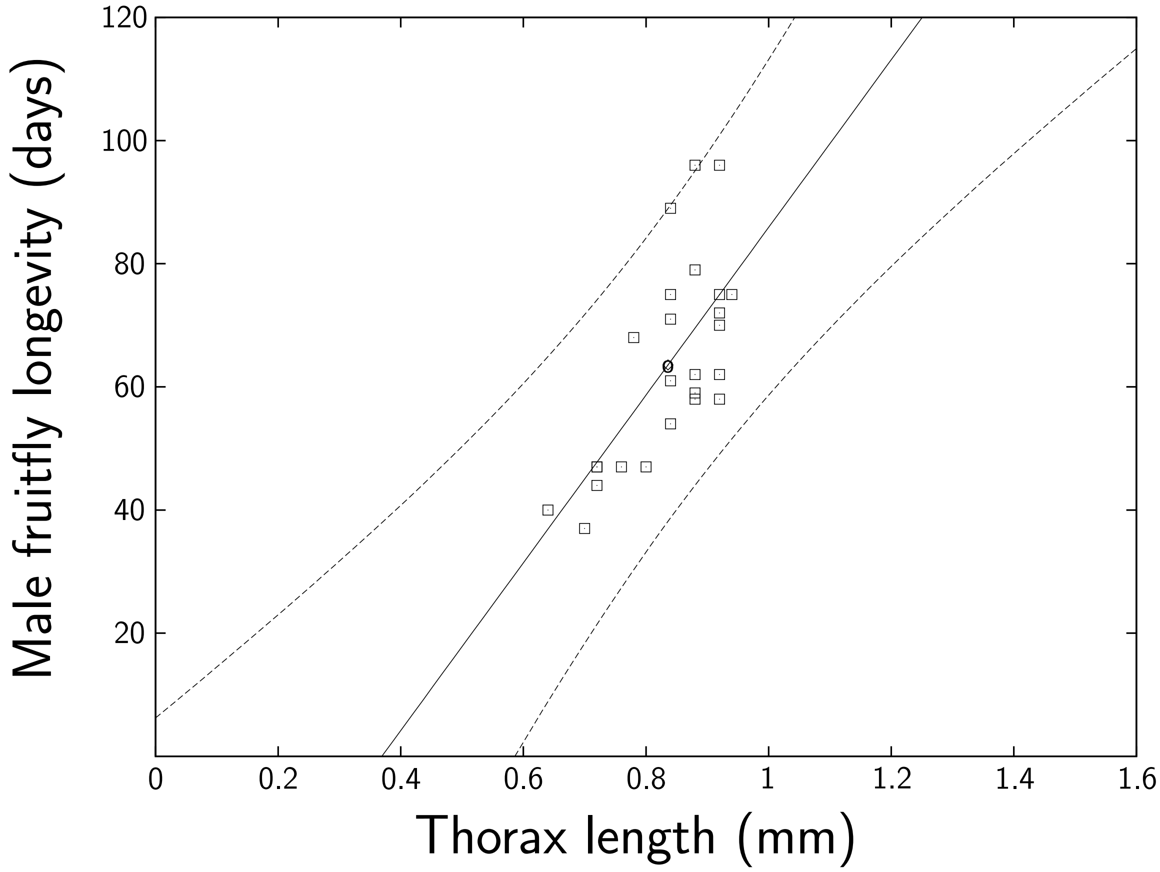


$$\hat{a} + \hat{\beta} x \pm t_{0.025}^{(n-2)} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Flies kept with no companions

95% predictive confidence bands for

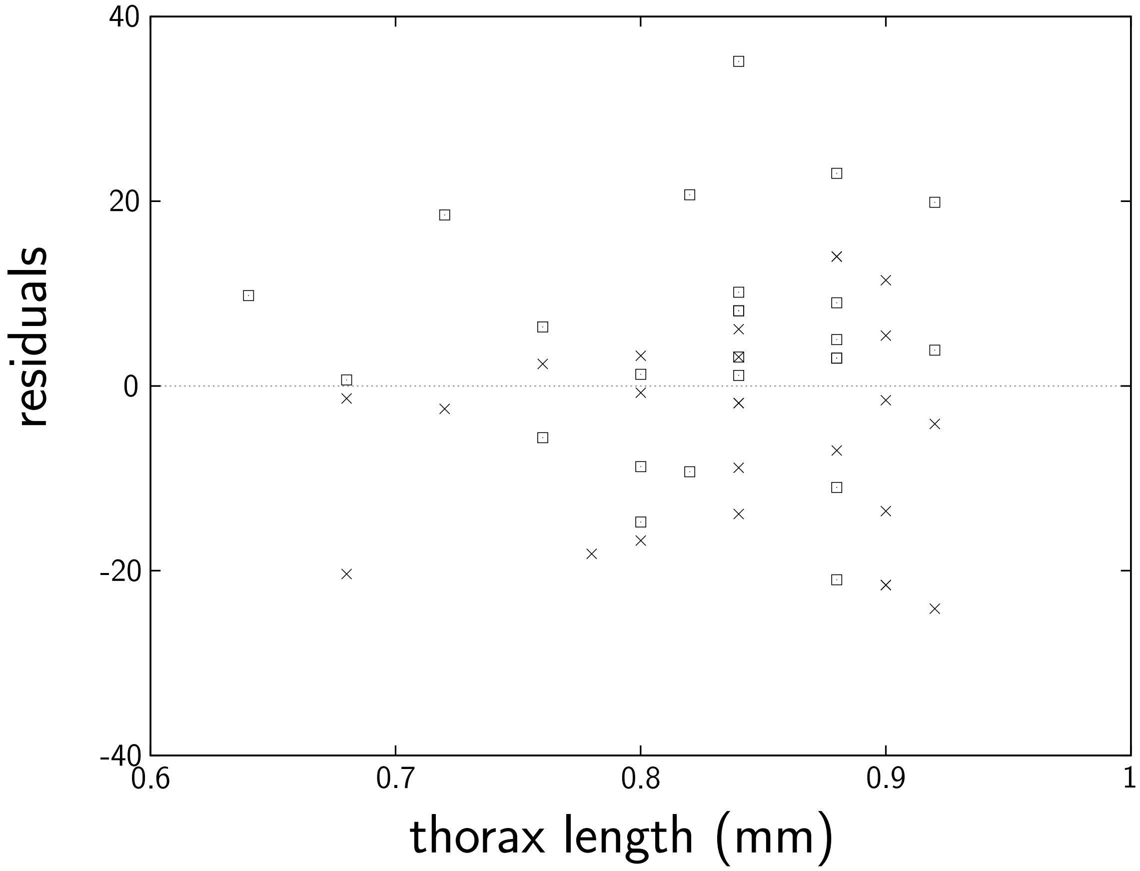
$$\boldsymbol{Y = a + \beta x_0 + \epsilon_0}$$



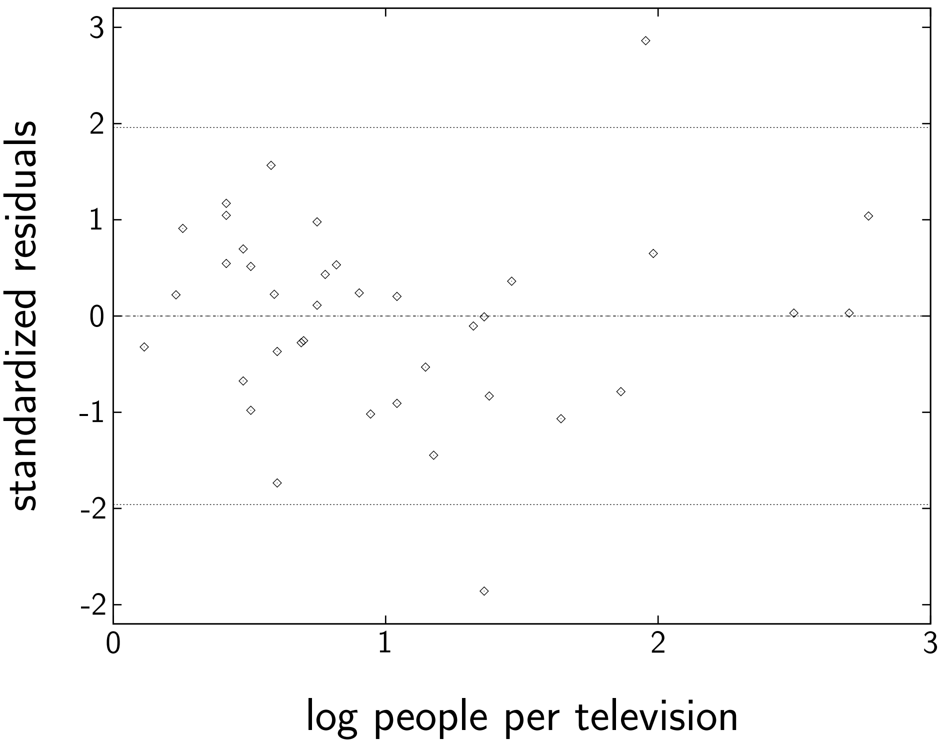
$$\boldsymbol{\hat{a} + \hat{\beta} x_0 \pm t_{0.025}^{(n-2)} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}}$$

Residuals under $\mathbf{H_0} : \boldsymbol{a_2 = a_3}$

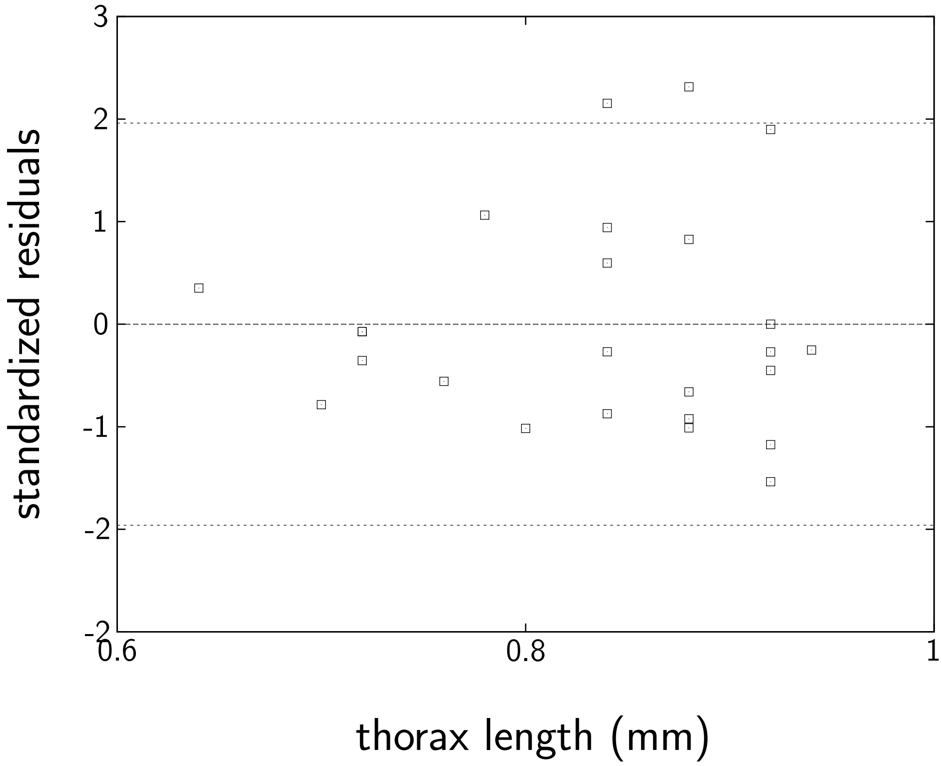
for males kept with 1 female



Residuals plot for regression of life expectancy against log people per television

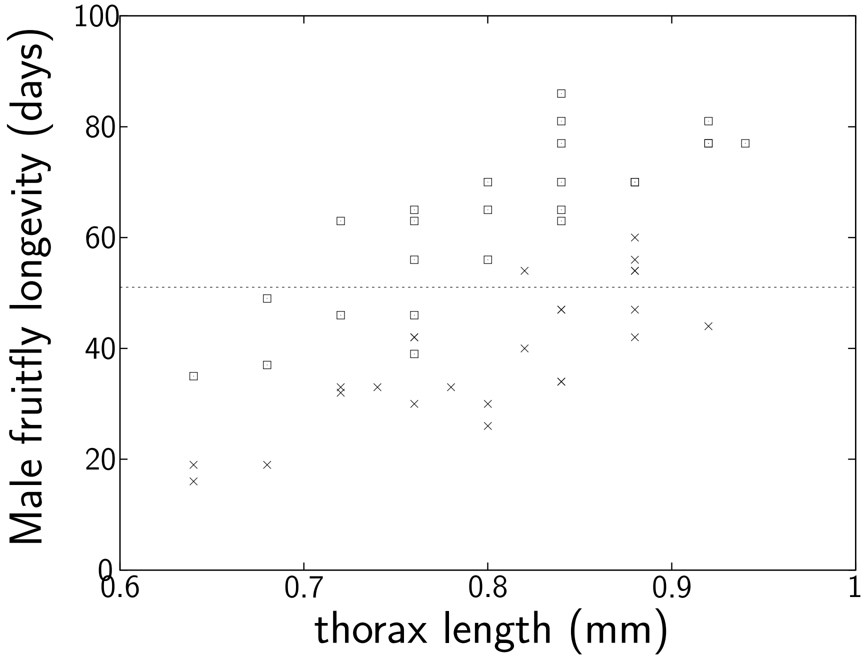


Residuals plot for regression of longevity of male fruitflies kept with no companions against thorax length

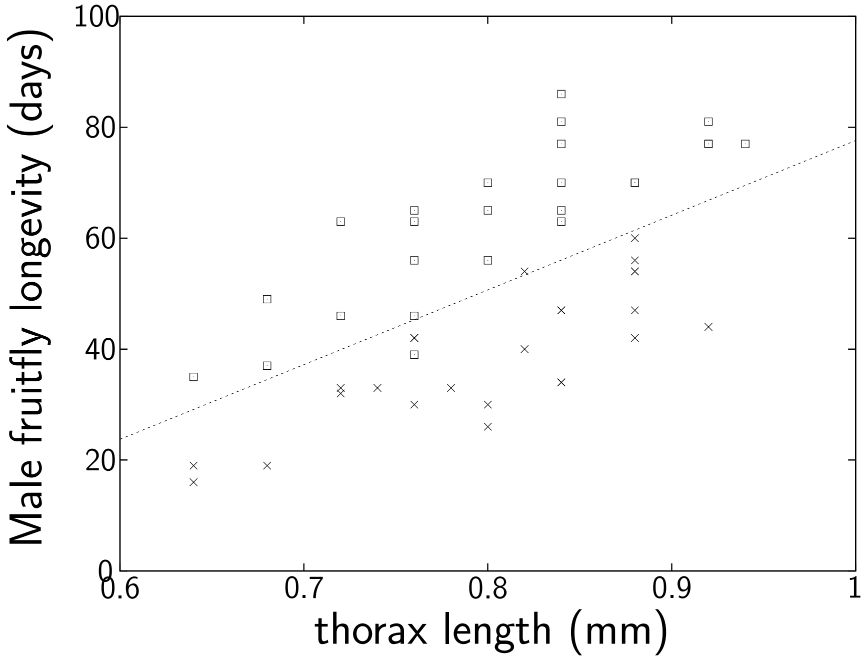


Discriminant analysis between two groups of 25 male flies kept with 8 females

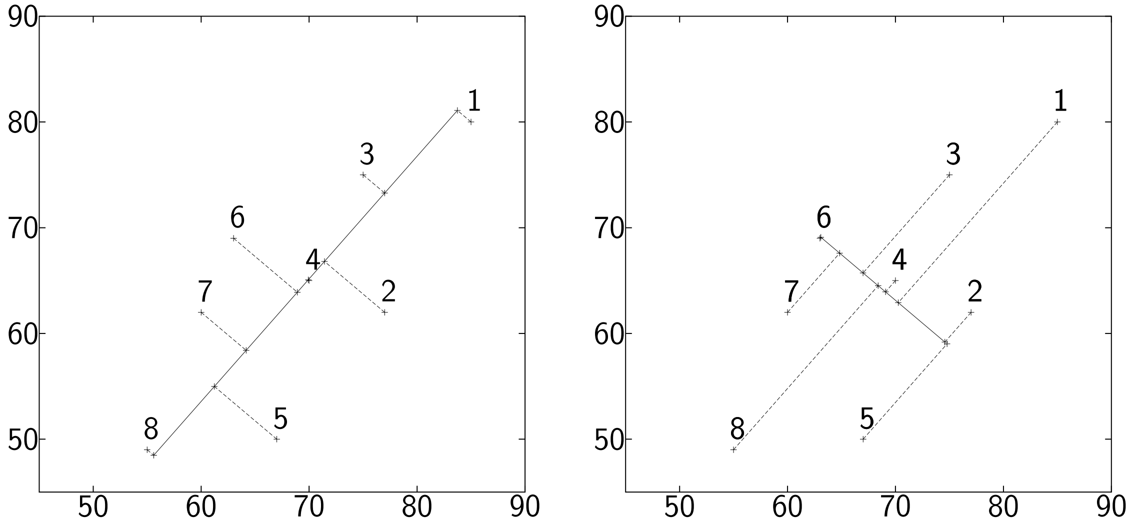
Discriminant based on longevity only:



Discriminant based on longevity and thorax length:



Factor scores



$$\text{IQ factor} = \mathbf{.653}(\text{math score}) + \mathbf{.757}(\text{verbal score})$$
$$\text{mathmo factor} = \mathbf{.757}(\text{math score}) - \mathbf{.653}(\text{verbal score})$$

$$\text{math score} = \mathbf{.653}(\text{IQ factor}) + \mathbf{.757}(\text{mathmo factor})$$
$$\text{verbal score} = \mathbf{.757}(\text{IQ factor}) - \mathbf{.653}(\text{mathmo factor})$$

student	math score	verbal score	IQ factor	mathmo factor
1	85	80	116.1	12.1
2	77	62	97.2	17.8
3	75	75	105.8	7.8
4	70	65	94.9	10.5
5	67	50	81.6	18.1
6	63	69	93.4	2.6
7	60	62	86.1	4.9
8	55	49	73.0	9.6

Histogram of 240 bootstrap samples of $\hat{\theta}$

Output from Excel spreadsheet

to be pasted here.

Example 16.1

In *Nature* (29 August, 1996, p. 766) Matthews gives the following table for various outcomes of Meteorological Office forecasts and weather over 1000 1-hour walks in London.

	Rain	No rain	Sum
Forecast of rain	66	156	222
Forecast of no rain	14	764	778
Sum	80	920	1000

Should one pay any attention to weather forecasts when deciding whether or not to carry an umbrella?

We might present the loss function as

	W^c	W
U^c	L_{00}	L_{01}
U	L_{10}	L_{11}

Here

$W =$ 'it turns out to be wet' and

$U =$ 'we carried an umbrella'.

E.g. $L_{00} = 0$, $L_{10} = 1$, $L_{11} = 2$, $L_{01} = 4$.