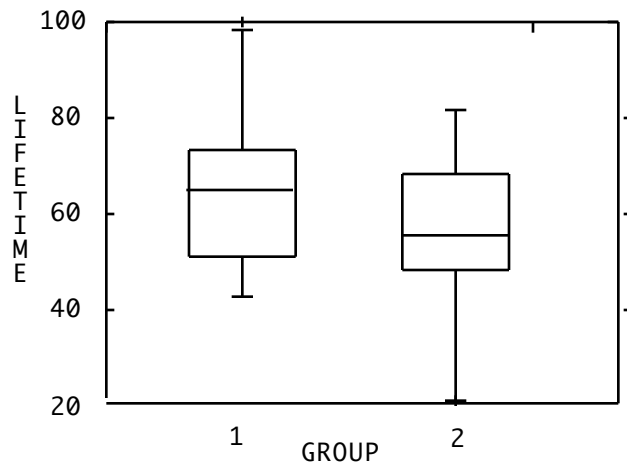


Using Multiple Regression to Make Comparisons SHARPER & FAIRER

Illustration: Effect of sexual activity on male longevity
Longevity (days) of male fruit-flies randomized to live with either uninterested (GROUP 1) or interested females (GROUP 2). Also measured: size of the fruit-fly (thorax, measured in mm) and the percentage of each day he spent sleeping.

GROUP		LIFETIME	THORAX	SLEEP
1	N OF CASES	25	25	25
	Range	42 - 97	0.64 - 0.92	4 - 66
	MEAN	64.8	0.826	24.1
	STANDARD DEV	15.6	0.070	16.7
	STD. ERROR	3.1	0.014	3.3
2	N OF CASES	25	25	25
	Range	21 - 81	0.68 - 0.92	5 - 73
	MEAN	56.8	0.838	25.8
	STANDARD DEV	15.0	0.071	18.4
	STD. ERROR	3.0	0.014	3.7



- t-test comparing GROUPS 1 and 2
(difference in means is 56.76 - 64.8 = -8.04 days)
[Pooled variance is approx 233.92]

> by hand ...

$$t_{48} = \frac{56.760 - 64.8}{\sqrt{233.92 \left[\frac{1}{25} + \frac{1}{25} \right]}} = \frac{-8.04}{4.326} = 1.86$$

> by SYSTAT...

INDEPENDENT SAMPLES T-TEST ON LIFETIME

GROUP	N	MEAN	SD
1	25	64.800	15.652
2	25	56.760	14.928

POOLED VARIANCES T = 1.859
DF = 48
PROB = 0.069

- EQUIVALENTLY:- ANALYSIS OF VARIANCE OF LIFETIME

SOURCE	SS	DF	MS	F-RATIO	P
GROUP	808.02	1	808.020	3.454	0.069
ERROR	11228.56	48	233.928		

N=50 MULTIPLE R = 0.259 MULTIPLE R² = 0.067

- Another way :- CI {difference in mean lifetime}

$$\begin{aligned} CI_{95} &= -8.04 \pm t_{48,95} SE(\text{observed difference}) \\ &= -8.04 \pm t_{48,95} \sqrt{\{SE(64.8)\}^2 + \{SE(56.76)\}^2} \\ &= -8.04 \pm 2.01 (4.326) = -8.04 \pm 8.69 \\ &= -16.74 \text{ to } 0.655, \text{ which just overlaps zero.} \end{aligned}$$

• Yet another way ... Regression analysis

GROUP 1 represented by X = 0 and GROUP 2 by X = 1

Fit: lifetime = CONSTANT + X + random variation

i.e. Mean(lifetime) = $\beta_0 + \beta X$ [β "times" X in "computerese"]

VARIABLE	COEFFICIENT	STD ERROR	T	P(2 TAIL)
CONSTANT	$\hat{\beta}_0 = 64.800$	3.059	0.000	
X	$\hat{\beta} = -8.040$	4.326	1.859	0.069

Fit means for two GROUPS by substituting X values.

gp 1 $\hat{\beta}_0 + \hat{\beta} * X = 64.800 + -8.040 * 0 = 64.80$

gp 2: $\hat{\beta}_0 + \hat{\beta} * X = 64.800 + -8.040 * 1 = 64.80 - 8.040 = 56.76$

i.e. the coefficient $\hat{\beta}$ associated with the "dummy" variable X estimates the difference in the means of the two populations i.e. we can represent the "GROUPS" by variables that take on numerical values, just like any other numerical predictor variable. [This is in contrast to the use of the the variable "GROUP" which simply uses the integers 1 and 2 as labels for groups]. Note that $\hat{\beta}$ divided by its SE of 4.326 gives the identical t-value as in the previous analyses. Also, the MEAN-SQUARE RESIDUAL of 233.928 based on 48 degrees of freedom is the "pooled variance" used in the t-test. The anova table that goes with the regression analysis (below) is identical to the one that goes with the classical anova table used above.. the only differences are the interchangeable uses of the terms RESIDUAL in place of ERROR and REGRESSION (i.e. X) in place of GROUP.

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	808.020	1	808.020	3.454	0.069
RESIDUAL	11228.560	48	233.928		

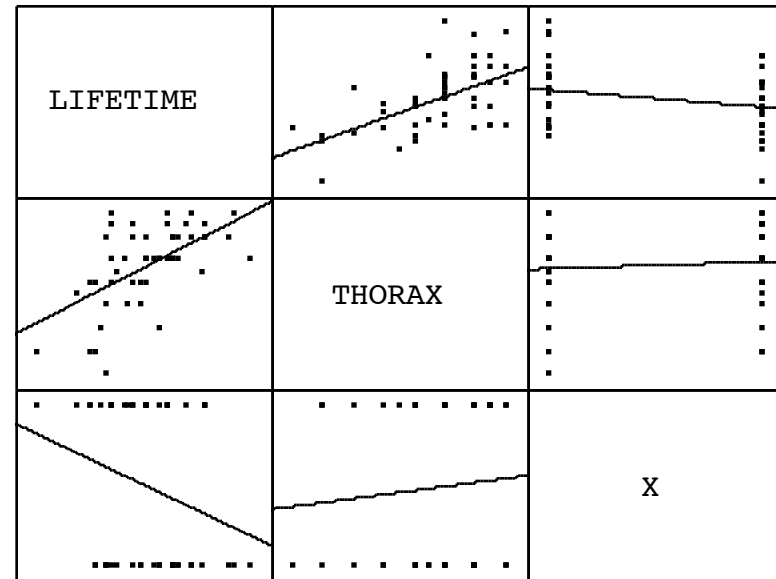
[$\sqrt{233.928} = 15.3 =$ SD of residuals, sometimes called "SE of estimate"]

LIFETIME N=50 MULTIPLE R = 0.259 MULTIPLE R² = 0.067

• Should one worry about the distribution of the other variables thorax size and sleep?

The crude differences in lifetime between the two study groups are shown in the top right panel of the scatter plot matrix [groups are represented by X=0 and X=1].

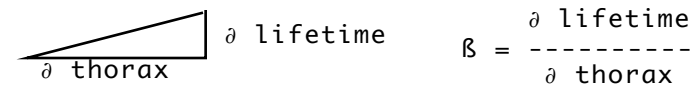
One can see from the middle panel in the top row of the scatter plot matrix that thorax size has an important influence on longevity, with larger flies living considerably longer than smaller ones. [Ideally, one should look at this in each group separately, but the data (not shown) show the same relationship in each one]



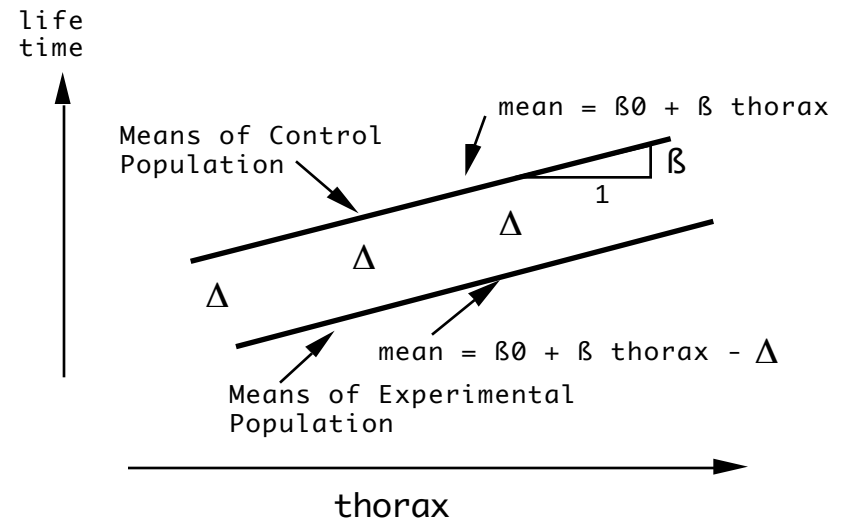
However, this was a randomized study and one can also see from the middle panel in the rightmost column that the flies are quite evenly distributed with respect to size. For what it is worth, the experimental group (x=1) is just slightly larger on average than the control group (x=0) and as such is starting out with a slight survival advantage; thus the final adjusted estimate of "days lost" by the experimental group needs to be enlarged to compensate for the fact that, had it started out without this advantage, it would have lost even more.

"Correcting for" imbalances with respect to size will enhance the observed difference only slightly. As we will see later, the adjustment moves the shortened lifetime from an average of 8.04 to an average of 9.65 days. If we still use the margin of error of ± 8.69 that we calculated at the beginning, this new estimate moves the significance level from $P=0.06$ to $P=0.025$ approx.

However, there is another important reason to take the variation due to size into account {An even better term might be "to take the effect of size out of the account"}. I deliberately use the term "take into account" rather than "correct for" since we often think of the latter only when there is an imbalance. In fact, taking the extraneous factor into account can have an important impact even if the two groups are perfectly balanced by design or by good fortune. The logic is the same as the one that says we should perform a paired t-test, rather than an independent samples test, when measurements come from matched pairs: using only the intra-pair differences removes what could be a large variation [even between members of the same group] due to the extraneous matching factor. We can think of regression [or as it is sometimes called analysis of covariance] as using "synthetic" or "poor-person's" matching in order to remove noise from the comparison of two groups [see article on appropriate uses of multivariate analysis]. This is done by fitting regression lines to the data from each of the two groups and calculating the vertical distances between them. Just as in the example of the effect of liberalizing speed limits, the idea is depicted schematically as follows:



$$\beta = \frac{\Delta \text{ lifetime}}{\Delta \text{ thorax}}$$



- Regression analysis: GROUP 1 coded $X = 0$ and GROUP 2 $X = 1$, and with linear effect of thorax

Fit: lifetime = CONSTANT + X + THORAX + random variation

i.e. Mean(lifetime) = $\beta_0 + \beta_T \cdot \text{THORAX} + \beta_X \cdot X$
 [using β_X for Δ]

VARIABLE	COEFFICIENT	STD ERROR	T	P(2 TAIL)
CONSTANT	$\hat{\beta}_0 = -46.038$	20.799	-2.214	0.032
THORAX	$\hat{\beta}_T = 134.252$	25.019	5.366	0.000
X	$\hat{\beta}_X = -9.651$	3.456	-2.793	0.008

SOURCE	SS	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	5073.677	2	2536.839	17.124	0.000
RESIDUAL	6962.903	47	148.147		

ESTIMATE = $\sqrt{148.147} = 12.2$ [vs 15.295]

$\hat{\beta}_x = -9.651$ is the "adjusted" mean difference between the groups; this is a 20% adjustment on the "crude" estimate of 8.04 days. Furthermore, the uncertainty about the new estimate, as measured by its SE, is 3.456, which is also a 20% reduction from the previous SE of 4.326, which was based on the overall or crude variation within each group. [that the two corrections are both 20% is just a coincidence].

The purposes, then, of the analysis of covariance (using thorax size as a covariate) are two-fold: (1) to correct, by an arithmetic adjustment, for any imbalances in important variable(s) between the groups being compared and (2) to sharpen the contrasts by removing noise due to these same variables and thereby reducing the SE of the estimated between-group differences in the response variable. Note that (2) will occur even if (1) is unnecessary from a "bias-correction" point of view.

• A simpler example of this might be the following data which one could imagine resulting from a comparison of the fuel consumption of two makes of automobile. Shown are the measurements in 4 runs of 1000, 2000, 3000, and 4000Km for each make. Consumption measured as litres/run (B) or litres/100Km, shortened to 1/100Km (C).

Make 1			Make 2		
(A)	(B)	(C)	(A)	(B)	(C)
Km	litres	1/100Km	Km	litres	1/100Km
1000	81	8.1	2000	182	9.1
3000	231	7.7	1000	89	8.9
4000	328	8.2	3000	270	9.0
2000	160	8.0	4000	360	9.0

xbar 200<----->225.3
s 105 t = 0.322 116.3

xbar 8<----->9
s 0.22 t = 8.66 0.08

In theory, one might want to weight each of the 4 observations according to its precision [e.g. fuel might not be measured equally precisely; even if fuel can be measured precisely, a longer trip might be less likely to be influenced by various short-term fluctuations] However, the main point is that, even though the 2 sets of observations are "balanced" with respect to distance, the variations in the "litres per run" index are very large and due more to variations in distance than to variations in fuel consumption between makes. Such noise makes it difficult to

see that make 2 is a bigger consumer of fuel -- something that can be seen clearly (and if need be backed up with a statistical test, which quantifies the limits of random variation) if one uses the less noisy index of 1/100Km. In this particular example, the runs could be matched and one could use a paired analysis to bring out the signal. But what if the 8 runs were all of different distances? A regression approach would handle this. In this e.g. below, distances are in units of 100Km, and remain balanced.

Fuel	DIST1	DIST2 (unit = 100Km)
81	10	0
231	30	0
328	40	0
160	20	0
89	0	10
270	0	30
360	0	40
182	0	20

FIT: average(FUEL) = β_1 *DIST1+ β_2 *DIST2 (NO CONSTANT)

VARIABLE	COEFFICIENT	STD ERROR	T	P(2 TAIL)
DIST1	$\hat{\beta}_1 = 8.02$	0.091	88.00*	0.000
DIST2	$\hat{\beta}_2 = 9.01$	0.091	98.86*	0.000

(* proof that it takes a non-zero amount of gasoline to drive 100KM !)

SE(9.01- 8.02) = $\sqrt{0.091^2 + 0.091^2} = 0.13$ (approx), so difference of 0.99 1/100Km is 7.5 SE's beyond zero (t=8.66 above arrived at by slightly different method).

ANALYSIS OF VARIANCE					
SOURCE	SS	DF	MS	F-RATIO	P
REGRESSION	436501.5	2	218250.75	8759.227	0.000
RESIDUAL	149.5	6	24.92		

Collinearity

Example of the issue: Suppose that in a study of workers aged 45-65 to quantify the degree to which hearing loss was affected by their exposure to the noise from heavy machinery, the number of years of exposure to this noise and the extent of hearing loss were determined for each person. A multiple regression is planned to assess the effect and to take the person's age into account (hearing loss generally becomes worse with age, even if there is no unusual occupational exposure).

What is the correlation between age and cumulated exposure likely to be?

If it is very high, what will it do to the estimate of the regression slope of loss on exposure?.

If it is low, what will it do? If you think it will do very little, would you bother to include age in the regression? [This question has to do with reduction of noise and making comparisons sharper].

If you had a choice of which workers to select from a larger available group, would you choose on a purely random basis, or on some other basis? Why?

See some examples on next page. The panel on the extreme left shows the distribution of age and exposure (both in years), with a fairly strong positive correlation. An example of a 'stratified sample' is given next to it (upper panel). Here the selection is constrained to obtain persons equally from all 4 quadrants. This makes it easier to separate the effect of age from the effect of exposure. An example of an 'unstratified sample' is given in the lower panel. Here the selection is simply a 'miniature' of the parent distribution and so there will be greater difficulty in separating the effect of age from the effect of exposure.

Suppose that in fact the mean hearing loss for persons of a certain age and exposure is as follows:

$$\text{mean} = 0.3 \cdot (\text{age} - 25) + 0.4 \cdot \text{exposure}$$

and that the inter-individual variation around this mean is Gaussian with a SD of 3. In technical language, we say that $\beta[\text{exposure}] = 0.4$ and that $\beta[\text{age}] = 0.3$, and that the SD of the 'residuals' is 3.0.

On the right hand side of the following page the effects of the collinearity on our estimates of the two β 's are displayed in list and graphic mode for 10 unconstrained and 10 constrained (stratified) random samples. The message from these is that the estimates of the β associated with exposure are more variable (and so less dependable) when the samples have collinearity. (the same is true for the estimates of the β for age). In the extreme, if the collinearity between age and exposure were close to a correlation of 1, the estimates of the β for exposure could oscillate even more, and could go from being quite negative to quite positive. The only thing that would remain reasonably stable is the sum of the estimate of β for exposure and of the β for age (i.e. the sum of the two estimates would be close to $0.4 + 0.3 = 0.7$, but an equation with the estimate of $\beta[\text{exposure}] = -1.2$ and $\beta[\text{age}] = +1.9$ {or for that matter $\beta[\text{exposure}] = +2.3$ and $\beta[\text{age}] = -1.6$ } would do an equally good job of predicting the responses (all the individuals would be spread out along the diagonal in the age vs. exposure diagram). You can see some of this compensatory behaviour of the two estimates in the plot in the panel on the right (estimates from "unstratified" samples), where there is a strong negative correlation between the two estimates.

Effect Modification

In the previous example, if females, because of their longer hair or greater tendency to wear ear-protectors, or because of some biological factor that might make them less susceptible to noise-induced hearing loss, were analyzed separately from males, how would the regression coefficients for hearing loss on years of exposure compare in the two sexes?

Effect Modification = "Different Slopes for Different Folks"

Can we combine the separate equations for males and females into one?

A similar example of combining two equations into one: How to estimate ideal body weight (based on findings of a Harvard study)

For Women: 100 pounds for a height of 5 feet, with five additional pounds for each added inch of height

For Men: 110 pounds for a height of 5 feet, and six additional pounds for every added inch of height

Since 5 feet = 60 inches, and letting H = height in inches - 60, the equations become:

Women: weight = $100 + 5 \cdot H$
Men: weight = $110 + 6 \cdot H$

If denote Women by a variable $G(\text{ender})=0$ and Men by $G=1$, we can combine the 2 equations

$$\text{weight} = 100 + 10 \cdot G + 5 \cdot H + 1 \cdot G \cdot H$$

Terminology: Note that the use of the product $G \cdot H$ as an additional variable in the regression equation is called an 'interaction' term. If the coefficient associated with this variable were 0, we would have 'no statistical interaction' (i.e. we would have the 'same slope for different folks').

Thus the ideas of 'effect modification' and 'statistical interaction' are really the same: epidemiologists tend to use the former and statisticians the latter.

The trouble with the word interaction is that it refers to a purely numerical trick to write the equations for 2 or more non-parallel lines in a single compact equation. Unfortunately, users of the equations sometimes try to give the word a biological meaning. But by suitable transformations, one can sometimes transform non-parallel curves into parallel lines and vice versa, so any 'interaction' term has to be viewed in the context of the scale used.