Intro to Analysis of Multi-variable Data: 12-hour course, 1995

Outline

Session 1

- Y Scales
  Summary Statistics / Parameters

- X & Y Various Configurations
  Measures of relationship Y <--> X

- X_1, X_2 & Y Objectives of M.V.A.

  - Fairer Comparisons
  - Sharper Comparisons
  - "Determinants of ..."
  - Prediction
  - Effect modification
    ("Different Slopes for Different Folks")

- "Non-regression" Methods

  - Risk/Rate Differences/Ratios
  - Differences in Means
  - Slopes

Session 2

- Multiple Regression

- Adjustment / Noise-reduction :- Means

- Adjustment:- Proportions

Session 3

- Several determinants

- Collinearity

- Effect Modification / Interaction

Session 4

- Computing

- Case Studies

Session 1:

Outline

- Y — Scales
  Summary Statistics / Parameters
- X & Y — Various Configurations / Displays
  Measures of relationship Y <--> X

  references: M&M Ch 2
- X_1, X_2 & Y — Roles of X_1 & X_2 :
  - Fairer Comparison of levels of X_1
    BIAS REDUCTION --> X_2 is a "Confounder"
  - Sharper Comparison of levels of X_1
    MORE PRECISION --> X_2 not necessarily confounder but produces considerable addnl. variation in Y
  - Interest in Both X_1, X_2 as determinants of Y
    X_1 and X_2 have same SYMMETRICAL status
  - X_2 "modifies" relationship between X_1 and Y
    DIFFERENT Y<->X_1 relationship
    for DIFFERENT levels (subgroups) of X_2

Examples of (Y, X_1, X_2 ...) Data

- Admissions of Males & Females to Berkeley Graduate Schools
  - overall and faculty by faculty
- Birthweight - Gestational Age ; Gender
- Fatalities & Speed Limit Change - Time
- Low Birthweight - Alcohol ; Smoking ; Social Class
- Intelligence Quotient (IQ) - Mother's Milk; Other Variables
- Stature(height) of Children on Tetracycline -
- Lung Function of Vanadium Factory Workers
  - vs. reference group (matched for smoking and age) that was 3.4 cm different in average height
- Blood Pressure and Altitude - age; height; weight; country
- Weight - Age ; Social Class
- longevity - sexual Activity; Size

X_1, X_2 & Y:

If primary interest is in X_1 contrast, and X_2 is either a "Confounder" or produces considerable additional variation in Y that acts as 'noise'.

Simplest case: X_1 is measured on a 2-point scale (binary) so compare Y in those with X_1 = 0 vs. in those with X_1 = 1;

NON-REGRESSION METHODS

Paired / Less Finely Stratified Observations (X_2 : pair / stratum)

X_2	$X_1 = 0$	$X_1 = 1$	ΔResponse *
1	(ave.) response	(ave.) response	d
2	(ave.) response	(ave.) response	d
..	…	…	…
..	(ave.) response	(ave.) response	d
∑			$\frac{\sum w\cdot d}{\sum w}$

* using d generically to represent any comparison (could be difference, ratio, etc...)

Key: (Weighted) Average of "Within-stratum" or "other-factors-being-equal" comparisons.

Confounding:

Δ of aggregated responses NOT SAME AS aggregate of Δ's

References:

counted and measured Y's:	Smith & Morrow, §14.6
	AAHOVW
	Miettinen §11-16
counted Y's:	Walker §8 & 13
	KKM § 13

Session 2: Multiple Regression: Making Comparisons FAIRER

e.g. BREAST MILK AND SUBSEQUENT INTELLIGENCE QUOTIENT IN CHILDREN BORN PRETERM (Lucas et al Lancet 1992; 339: 261-64)

There is considerable controversy over whether nutrition in early life has a long-term influence on neurodevelopment. We have shown previously that, in preterm infants, mother's choice to breast milk was associated with higher developmental scores at 18 months. We now report data on intelligence quotient (IQ) in the same children seen at 7.5 - 8 years.

IQ was assessed in 300 children with an abbreviated version of the Weschler Intelligence Scale for Children (revised Anglicised). Children who had consumed mother's milk in early weeks of life had a significantly higher IQ at 7.5 - 8 years than did those who received no maternal milk. An 8.3 point advantage (over half a standard deviation) in IQ remained even after adjustment for differences between groups in mother's education and social class (p < 0.0001). This advantage was associated with being fed mother's milk by tube rather than with the process of breastfeeding. There was a dose- response relation between the proportion of mother's milk in the diet and subsequent IQ. Children whose mothers chose to provide milk but failed to do so had the same IQ as those whose mothers elected not to provide breast milk.

Although these results could be explained by differences between groups in parenting skills or genetic potential (even after adjustment for social and educational factors), our data point to a beneficial effect of human milk on neuro-development.

TABLE I - CHARACTERISTICS OF STUDY POPULATION

	No mother's milk	Mother's milk
	(group I)	(group II)
Characteristics	(n = 90)	(n = 210)
Mean (SEM) birthweight (g)	1420 (30)	1440 (20)
Mean (SEM) gestation (wk)	31.4 (0.3)	31.4 (0.2)
% males (no)	42 (38)	55 (116)*
Days in study: median (quartiles)	30 (22,45)	28 (20,40)
Days to full enteral feeds: "	8 (6,11)	7 (6,9)
% ventilated > 5 days (no)	12 (11)	12 (26)
% in social class I and II (no)	11 (10)	30 (63)+
% mothers higher educ. status (no)@	24 (22)	52 (109)+

*p < 0.05. +p < 0.001 @ GCE O levels or above (see text).

Table II - IQ AT 7.5 - 8 YEARS IN THE TWO GROUPS

	Mean (SEM) scores		Advantage for group II
Abbreviated WISC-R	Group I	Group II	babies (95% CI)
Verbal scale	92.0(2.0)	102.1(1.3)	10.1 (4.7, 15.5)*
Performance scale	93.2(1.7)	103.3(1.2)	10.1 (6.0, 14.2)*
Overall IQ	92.8(1.6)	103.0(1.2)	10.2 (6.3, 14.1)*

*p < 0.001, group 1 vs group II CI = confidence interval

Table III - ADJUSTED ADVANTAGE IN WISC IQ SCORES FOR GROUP II BABIES

	Mean (SEM) scores	Advantage for group II
	Advantage	95% CI
Whole Group*		
Verbal scale	7.7	(3.3, 12.1)
Performance scale	7.9	(3.9, 11.9)
Overall IQ	7.6	(4.0, 11.2)
Successful**		
Verbal scale	7.7	(3.3, 12.1)
Performance scale	7.9	(3.9, 11.9)
Overall IQ	7.6	(4.0, 11.2)

* All 210 babies in Group II (compared with 90 in Group I)
** 193 babies from Group II who received breast milk (compared with infants from Group I plus those from Group II who received no breast milk: n=107)

p < 0.001, group 1 vs group II CI = confidence interval

Table IV- FACTORS RELATING TO IQ AT 7.5–8 YEARS

Factor	Increase in IQ	95% CI	p value
Received mother's milk	8.3	(4.9, 11.7)	<0.0001
Social Class	–3.5/class*	(–1.5,–5.5)	0.0004
Mother's education	2.0/group**	(0.5, 3.5)	0.01
Female sex	4.2	(1.0,7.4)	0.01
Days of ventilation	–2.6/wk	(–3.7,–1.5)	0.02

* Social class recorded as 4 categories: I/II, III non-manual, III manual, IV/V
** Mother's education coded on 5-point scale from 1 (no educational qualifications) to 5 (degree or other professional qualification)

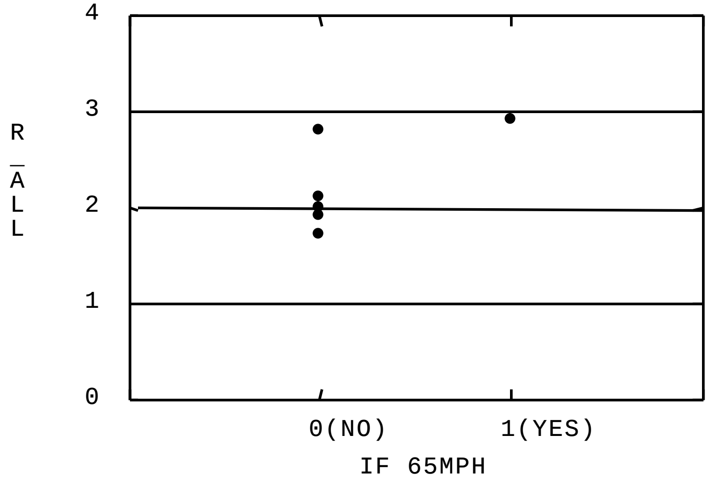
Using Multiple Regression to Make Comparisons FAIRER

Illustration: Analysis of Rates of Fatal Crashes on rural interstate highways in New Mexico in the 5 years 1982-1986 (55 mph limit) and in 1987 (65 mph limit). See Oct. 27 article in JAMA by Gallaher et al. 1989;262:2243-2245.

DATA:	55 mph					65 mph
	1982	1983	1984	1985	1986	1987
Rates per 10^8 v-m*	2.8	2.0	2.1	1.7	1.9	2.9

*vehicle miles; Variable named "R_ALL" below.

SUMMARIES	IF65MPH = 0 (coded "TYPE" = 1)	IF65MPH = 1 (coded "TYPE" = 2)
N OF CASES	5	1
MEAN	2.100	2.900
VARIANCE	0.175	0.000



Two simple (but - at least in this case - cruder, less sensitive and more biased) analyses (2 are equivalent).

(1) t-test The only estimate of the common variance is from the 1st 5 years; in fact, some statistical packages will not compute the t test in this situation.

$$t_4 = \frac{2.9 - 2.1}{\sqrt{s^2 \left[\frac{1}{5} + \frac{1}{1} \right]}} = \frac{0.8}{\sqrt{0.175 \left[0.2 + 1.0 \right]}} = 1.746$$

(2) ANOVA

DEP VAR: R_ALL N: 6 MULTIPLE R: 0.66 MULTIPLE R^2: 0.43

```
SOURCE SUM-OF-SQUARES DF MEAN-SQUARE F-RATIO  P(2-sided)
TYPE*       0.533     1    0.533      3.048    0.156
ERROR**     0.700     4    0.175
------      -----    ---   -----
 Total      1.233     5    0.246
```

* Note: The "BETWEEN TYPES" SS is a weighted sum [weights 5:1 or 1:0.2] of the squared devns. of the mean, for each of the 2 types of years, from the $\bar{\bar{y}}$ of all 6 years

i.e. as $5[\bar{y}_1 - \bar{\bar{y}}]^2 + [\bar{y}_2 - \bar{\bar{y}}]^2 = 0.533$

As such, apart from a divisor, it has the form of a variance. [notice the ratio of 5 :1 or 1/0.2:1/1 i.e. the same ones which appear in the denominator of the t-test]

Compare the 0.533 with the $\frac{[2.9 - 2.1]^2}{[\ 0.2 + 1.0\ ]}$ one would get by squaring the numerator and part of the denominator of the t-test statistic. Squaring the entire t_4 statistic of 1.746 yields the $F_{1,4}$ ratio test statistic of 3.048.

**Note: The "ERROR" is calculated by pooling the variances "within" each of the two types of years. In this e.g. the estimate of error is contributed entirely by the "TYPE" = 1 years . The "mean square error" is the same as the within group variance in the t-test.

Two more complex [but also more sensitive and less biased] analyses. (The two methods are equivalent in the example here)

The aim is to take compare 1987 with the most relevant period; the average of 1982-1986 is probably too high (rates seem to have been falling over that time). Also one should take out the systematic variation in the 5 years that, in the s^2 used in the t-test or 1-way anova, appears as "unexplained noise". In other words, the idea is to make the comparison both FAIRER and SHARPER.

(1) What the authors did... Fit a regression line to the 5 years, estimate the "expected" value for 1987 and the expected range of variation around this fitted mean, and determine where, relative to this predicted range of variation, the observed value in 1987 lies.

```
DEP VAR: R_ALL  N:5  MULTIPLE R:0.794  MULTIPLE R²: 0.630

ADJUSTED MULTIPLE R²: 0.507
STANDARD ERROR OF ESTIMATE:  0.294 (This is a misnomer; It is
really the √  of the average squared residual [0.086]  and
could be called an "average residual")

VARIABLE  COEFF.  STD ERROR     T     P(2 TAIL)

CONSTANT  418.740  184.345    2.272     0.108
YEAR       -0.210    0.093   -2.260     0.109

                  ANALYSIS OF VARIANCE
SOURCE   SUM-OF-SQUARES  DF  MEAN-SQUARE  F-RATIO   P

REGRESSION    0.441      1      0.441     5.108   0.109
RESIDUAL      0.259      3      0.086
```

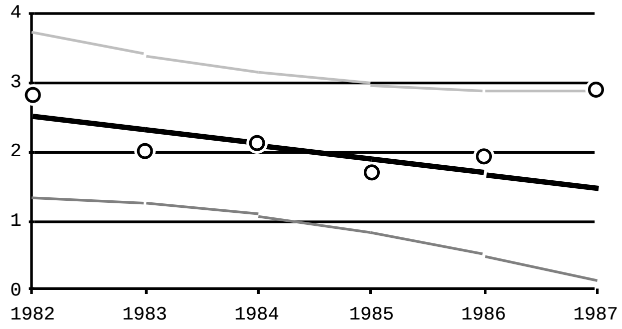
"fitted" rate for 1987 [generically: $\hat{y} = \hat{b}_0 + \hat{b}_1 * x$]
= 418.740 -0.210*1987 = 1.47
(slightly different from authors' because of rounding)

Range of variation of individual point about 1.47 :

$$1.47 \pm t_{3,95} \times 0.294 \times \sqrt{1 + \frac{1}{5} + \frac{[1987 - 1984]^2}{\sum[\text{year} - 1984]^2}}$$

$$1.47 \pm 3.182 \times 0.294 \times \sqrt{1 + \frac{1}{5} + \frac{9}{10}} = 1.47 \pm 1.33$$

0.14 to 2.80.



In the diagram, the solid black line is the regression line fitted to the points 1982-1986. The dotted lines represent the 95% limits for individual values [not to be confused by the 95% CI for the regression line (the line of means) itself!].

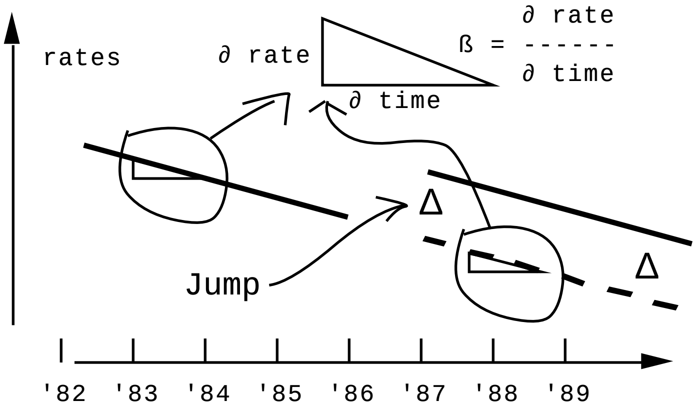
The observed point of 2.9 (not shown) is just outside the 95% range of random variation about the mean predicted for 1987. In fact, using the SD of 1.45 [the 0.4205 obtained by multiplying the 0.294 by the radical, the 2.9 is

$$t = \frac{2.9 - 1.47}{0.4205} = 3.40 \text{ SD's above expected, and since}$$

the estimated SD is based on only 3 df, this deviate is somewhere between the 97.5% and the 99%ile. It is not clear whether the p-value in the article is 1- or 2-sided, or indeed whether the authors calculated it in the same way as here.

(2) Another equivalent multivariate method..both this and the author's methods are multivariate -- in the sense that they deal with 3 (i.e. > 2) variables (the rates and the two "explanatory" variables of year and the status of the law).

The idea is to estimate simultaneously both the trend over years and the apparent "effect" (in terms of a jump in the fatal crash rates) that the relaxing of the law had. The data points could be thought of as two series with the same trends but with the second series, starting in 1987, have a higher level. e.g.



One could represent these two lines by two equations:

- expected rate = $ß_0$ + ß*year   ('82-'86: 55 mph)
- expected rate = $ß_0$ + ß*year + D ('87:   65 mph)

If we want to be compact about it, and define an "indicator variable" which takes on the value 0 if the limit is 55 mph and 1 if 65 mph, we can write the two equations in one as:

- expected rate = $ß_0$ + ß*year + D*indicator_variable

In the computer run below, because of limitations on the number of letters in the name, the indicator variable has been called IF65MPH.

By fitting the multiple regression equation:

R_ALL = CONSTANT + YEAR + IF65MPH ,

we obtain the estimates $\hat{ß}_0$, $\hat{ß}$ and $\hat{D}$ as the coefficients accompanying the variables named CONSTANT, YEAR and IF65MPH.

```
DEP VAR = R_ALL N=6  MULTIPLE R=0.889 MULTIPLE R2 = 0.790

ADJUSTED MULTIPLE R2 = 0.650
STANDARD ERROR OF ESTIMATE = 0.294 (see comment above)

VARIABLE  COEFFICIENT  STD ERROR   T      P(2 TAIL)

CONSTANT    418.740     184.345   2.272    0.108
YEAR         -0.210       0.093  -2.260    0.109
IF65MPH       1.430       0.426   3.358    0.044
```

i.e. the estimates are

$\hat{ß}_0$ = 418.74 ; $\hat{ß}$ = -0.210 and $\hat{D}$ = 1.430, with SE's

184.345;   0.093 and   0.426 respectively.

The one of direct interest is $\hat{D}$ = 1.430, which is

$$t_3 = \frac{1.430 - 0}{0.426} = 3.358 \text{ SE's greater than } 0$$

[which, apart from the rounding errors, is just like it was in the previous analysis].

What we did do to get the same answer? We introduced one more observation directly into the analysis, but it went entirely to estimating D; the residual variation is still based on the variance of the 5 first years from their trend (the estimated trend also remains the same). Year is a covariate here.

Usually, analyses of covariance involve covariates which overlap within the two or more groups of direct interest and one has some chance to test whether it is reasonable to assume common slopes for the lines. Also, one is usually more interested in estimating the D within the middle of the range of the covariate, not at its extreme, as was the case here. For completeness, the partition of the overall 5 df variation of s^2 = 1.233 in the 6 datapoints is given below.

Note that the MULTIPLE R^2 = 0.790 comes from dividing the portion "explained by a jump from a linear trend by the total variation of 1.233 is .7899, or 0.790 when rounded.

Note also that neither the 1 df test of a non-zero trend nor the "overall F ratio" for testing whether "two variables are better than none" is statistically significant. However, the inclusion of YEAR in the equation, and therefore the subtraction of the variance explainable by it, is important in letting the signal (estimated at 1.43) shine through the remaining -- now not so large -- unexplained "noise", which we estimate at $s^2_{residual}$ = 0.086. Contrast this with the s^2 = 0.175 in the t-test and anova described at the very beginning.

```
             ANALYSIS OF VARIANCE
SOURCE   SUM-OF-SQUARES  DF  MEAN-SQUARE  F-RATIO    P

REGRESSION     0.974      2    0.487       5.643    0.096
RESIDUAL       0.259      3    0.086
-----------    -----     ---   -----
Total          1.233      5    0.246
```

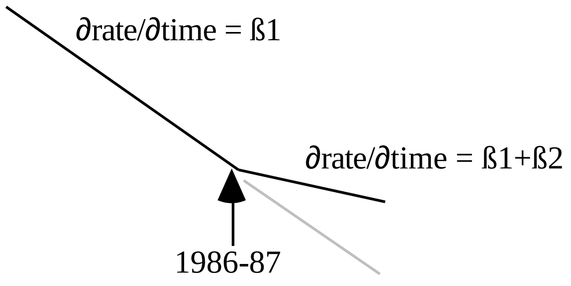
Note: Most would consider the equation

$$R\_ALL = \text{ß}_0 + \text{ß}*YEAR + \Delta*IF65MPH$$

'unnatural' in that it implies a shift to a parallel trend. A more narural one would be a shift to a different slope. This could be represented by an equation of the form

$$R\_ALL = \text{ß}_0 + \text{ß}_1*YEAR + \text{ß}_2*YEAR*IF65MPH$$

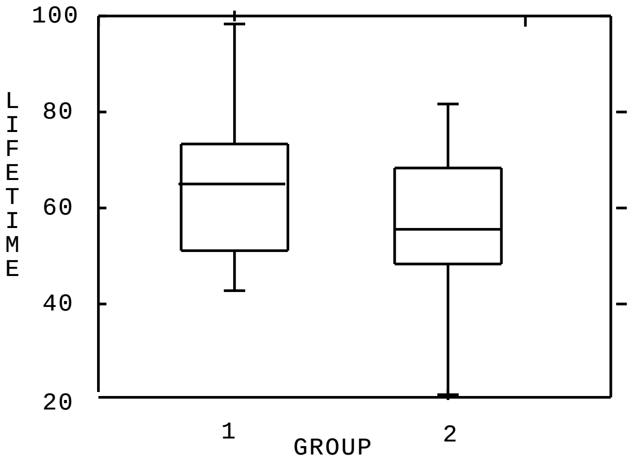
where ß_2 represents the change to the slope with 65MPH (negative ß_2 means a shallower, positive ß_2 a sharper trend. With only 1 datapoint for 65MPH, we cannot judge from the data alone which model fits better.



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}

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

- 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) = $ß_0$ + ß*X [ß "times" X in "computerese"]

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

Fit means for two GROUPS by substituting X values.

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

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

i.e. the coefficient $\hat{ß}$ 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 $\hat{ß}$ 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		

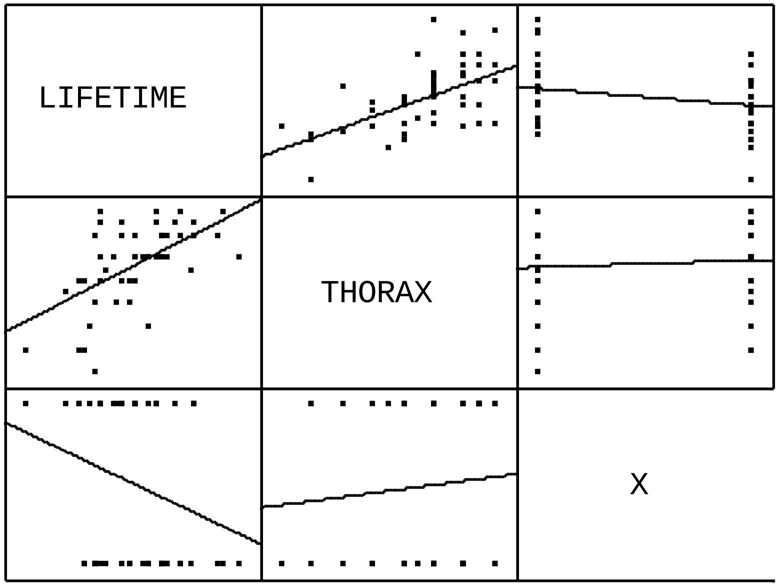
[√233.928 = 15.3 = SD of resuiduals, sometimes called "SE of estimate"]

LIFETIME N=50 MULTIPLE R = 0.259 MULTIPLE R^2 = 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].

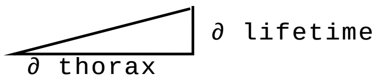
One can see from the middle panel in the top row of the scatter plot matrix that thorax size has an important influence on longevity, with larger flies living considerably longer than smaller ones. [Ideally, one should look at this in each group separately, but the data (not shown) show the same relationship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



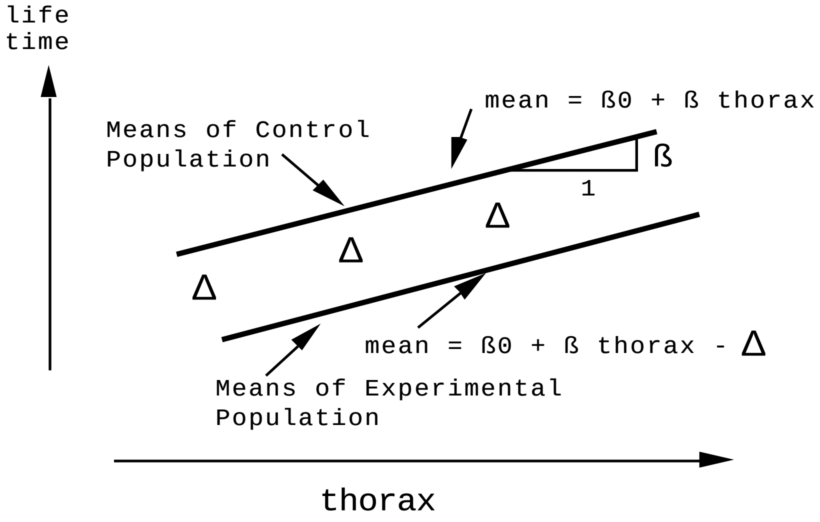
However, this was a randomized study and one can also see from the middle panel in the rightmost column that the flies are quite evenly distributed with respect to size. For what it is worth, the experimental group (x=1) is just slightly larger on average than the control group (x=0) and as such is starting 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:



$$\text{ß} = \frac{\partial \text{ lifetime}}{\partial \text{ thorax}}$$



- 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) = ß_0 + ß_T*THORAX + ß_X*X
[using ß_X for Δ]

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

STANDARD ERROR OF 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 l/100Km (C).

Make 1			Make 2		
(A)	(B)	(C)	(A)	(B)	(C)
Km	litres	l/100Km	Km	litres	l/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 l/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 woul 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 l/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		

Multiple Regression for Proportions

Examples:

Lowbirtweight in relation to alcohol consumption during pregnancy

Asthma trends in Israel

Psychological Stress, Smoking, Alcohol Consumption and susceptibility to the common cold.

Non-regression approaches as before. They depend on the comparative parameter to be used: Δ of proportions; ratio of proportions; ratio of odds.

Likewise, regression approaches depend on the form of comparative parameter.

Suppose we denote the parameters as π's. The general approach is to use a regression model

$$g(\pi) = \sum \beta \bullet x$$

If g() is the "Identity" 'link' ie $g(\theta) = \theta$, we can estimate Δ's of proportions (risk differences).

If g() is the "Log" 'link' ie $g(\theta) = \ln(\theta)$, we can model ratios of proportions ('risk ratios' or 'relative risks').

If g() is the "Logit" 'link' ie $g(\theta) = \ln(\frac{\theta}{1-\theta})$, we can model ratios of odds ('odds ratios' or 'relative odds').

If g() is the "Probit" 'link' ie $g(\theta)$ = the Gaussian Z deviate corresponding to a proportion θ, then we can estimate shifts in LD50 (or LDxx) based on a Gaussian distribution of tolerance. The probit curve is a close relative of the logit curve; both can be used for modelling S-shaped 'dose-response' relationships.

References:

Armitage and Berry (3rd ED) §12.8
Healy: GLIM: An Introduction
Hosmer and Lemeshow: Logistic Regression
Kleinbaum: Logistic Regression
Miettinen §18.3
Checkoway H : Res. Methods for Occ. Epi. §8
Kahn HA & Sempos: Sta. Meth. in Epi. : §6
Selvin S: Sta. Anal. of Epi. Data: §7 and 8
Breslow N and Day N: Volume I (Case Control studies)
Schlesselman J: Case Control Studies
Bland and Keirse (2 expository articles: logistic Regression)
AAHOVW

Session 3:
Multiple regression as sequence of simple regressions

E.g.: Regression of Weight(lb) on age(yrs) and Height(in) in 11-16 year olds

3 SIMPLE REGRESSIONS

```
(1)  WEIGHT =  -105.38 + 3.36 * HEIGHT + RESWT

(2)  AGE    = -0.79 + 0.23 * HEIGHT + RESAGE , so that
(2') RESAGE = AGE - {  -0.79 + 0.23 * HEIGHT }

(3)  RESWT  = -0.023 + 2.82 * RESAGE + RESIDUAL
                                    (variance 187.02)
```

Substitute (2') into (3) to get

```
(4)  RESWT  = -0.023 + 2.82 * {AGE - {-0.79 + 0.23 * HEIGHT}}
```

and then (4) into (1) to get ...

```
(5) WEIGHT =

   -105.38  + 3.36 * HEIGHT +
     -0.023 + 2.82 * {AGE - {-0.79 + 0.23 * HEIGHT}}
   + RESIDUAL (variance 187.02)

 =  -105.378  +                   3.36 * HEIGHT + 2.822 * AGE
      -0.023  +          -2.822 * 0.23 * HEIGHT
    + 2.822*{-{-0.789}}
    + RESIDUAL (variance 187.02)

 =  -103.174  +                  2.725 * HEIGHT + 2.822 * AGE
    + RESIDUAL (variance 187.02)
```

This is ***equivalent*** (ignoring rounding errors from not using enough decimal places) to ***performing a multiple linear regression****:*

```
Y=WEIGHT; N=233; MULT. R=0.70; SQ. MULTI. R= 0.49
ADJUSTED SQ. MULT R=0.49; STANDARD ERROR OF ESTIMATE=13.705

VAR.      COEFF. STD ERROR STD COEF    T    P(2 TAIL)

CONST.  -103.150   14.199   0.000    -7.264  0.00000
HEIGHT     2.723    0.294   0.553     9.242  0.00000
AGE        2.822    0.807   0.209     3.498  0.00056

                       ANALYSIS OF VARIANCE
SOURCE     SUM-OF-SQUARES  DF  MEAN-SQ.  F-RATIO   P
REGRESSION  42186.19420    2  21093.09  112.29578 0.000
RESIDUAL    43202.08906   230    187.83517
```

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:

$$\text{mean} = 0.3\bullet(\text{age}–25) + 0.4\bullet\text{exposure}$$

and that the inter-individual variation around this mean is Gaussian with a SD of 3. In technical language, we say that ß[exposure] = 0.4 and that ß[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 ß'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 ß associated with exposure are more variable (and so less dependable) when the samples have collinearity. (the same is true for the estimates of the ß for age).
In the extreme, if the collinearity between age and exposure were close to a correlation of 1, the estimates of the ß 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 ß for exposure and of the ß 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 ß[exposure] = –1.2 and ß[age] = +1.9 {or for that matter ß[exposure] = +2.3 and ß[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.

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 susceptiple 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•H
Men:	weight = 110 + 6•H

If denote Women by a variable G(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•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.