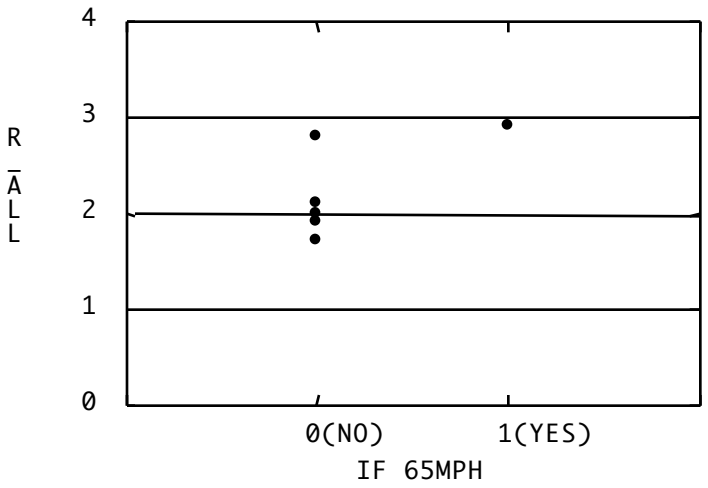
Using Multiple Regression to Make Comparisons FAIRER

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 R^2: 0.630

ADJUSTED MULTIPLE R^2: 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		

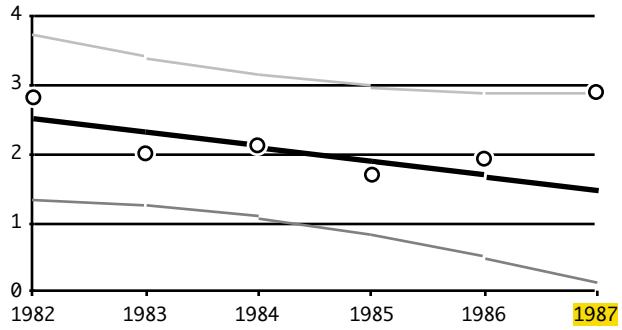
"fitted" rate for 1987 [generically: $\hat{y} = \hat{b}_0 + \hat{b}_1 * x$]
= 418.740 -0.210*1987 = 1.47
(slightly different from authors' because of rounding)

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.



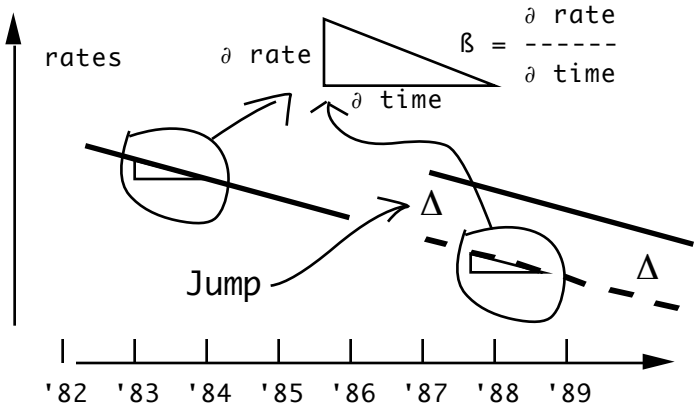
In the diagram, the solid black line is the regression line fitted to the points 1982-1986. The dotted lines represent the 95% limits for individual values [not to be confused by the 95% CI for the regression line (the line of means) itself!].

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

$$t = \frac{2.9 - 1.47}{0.4205} = 3.40$$ 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 R^2 = 0.790

ADJUSTED MULTIPLE R^2 = 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

$$R\_ALL = ß_0 + ß*YEAR + \Delta*IF65MPH$$

'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

$$R\_ALL = ß_0 + ß_1*YEAR + ß_2*YEAR*IF65MPH$$

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.