[CONTRIBUTION FROM THE DEPARTMENT OF CHEMISTRY, UNIVERSITY OF CINCINNATI]

A Combined Analysis of Variance and Regression Treatment in the Evaluation of the Effects of Substituents on Reactivity[1]

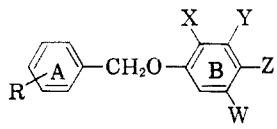
H. H. JAFFÉ

Received December 13, 1957

A technique is described for the statistical analysis of linear regression data. Particular emphasis is placed on tests of the hypotheses that a set of linear regressions have equal slope, or that a *significant* improvement is obtained by inclusion of an additional variable in a set of multiple regressions. As a most sensitive test of fit of data to a regression equation it is proposed to use the significance of the deviations from regression.

In the study of the effect of substituents on the reactivity of chemical compounds, a great deal of use is being made of linear regression,[2] and, to a lesser extent, of multiple regression involving two, and sometimes even more independent variables.[3] Owing to the relatively approximate nature of the correlations often observed and to the relative paucity of data usually available, although a good correlation is usually obvious, it is frequently difficult to evaluate adequately certain possible hypotheses: *e.g.* the hypotheses that various regression coefficients (slopes) differ, or that inclusion of a further independent variable results in a *significant* improvement of the fit. In an attempt to aid in the solution of some such problems we have utilized analysis of variance to provide statistical tests of significance.[4] The methods developed will be illustrated with several sets of data chosen from the literature.

Sets of linear regressions. An old set of data involving the rates of chlorination of a long series of variously substituted benzyl phenyl ethers (I)[5]



I

forms the simplest example. In these data the question arises whether the extent to which substituents in ring A affect the rate of chlorination depends on the substituents already present in ring B. Since the chlorination occurs only on ring B, the reaction behaves as a side chain reaction with respect to ring A, and effects of substituents in this ring can be treated by the use of the Hammett Equation (Equation 1) with good precision. A

$$\log k = Y_{ij} = \sigma_i\sigma_j + Y_{00} \qquad (1)$$

group of eight series of five compounds each was chosen in such a way that the same substituents in ring A recurred in each series, and each series was characterized by a different set of substituents in ring B.[6] The question posed then is equivalent to asking whether differences between the reaction constants (ρ-values) for these eight series are significant. Comparison of any pair of ρ-values and their standard deviations indicates no significant differences, but this comparison is not the most sensitive criterion as long as more than two ρ-values are available. An analysis of the total variance of the entire set of data will be used to obtain further information.

The total variance of each series having 5 degrees of freedom (DF) can be broken up into a contribution from the mean (1 DF) and from deviations from the mean (4 DF). The first of these terms is of little interest, but the latter one can be further divided into 1 DF due to regression and 3 DF due to deviations from regression. There is nothing unusual in this analysis, which is implied in the standard regression analysis, except the expression in the formalism of analysis of variance.

In the present example, we have, however, eight parallel series, and it is possible to add the corresponding terms from the analysis of each series to arrive at a composite analysis. This process is indicated in the first two columns of Fig. 1. The analysis of the over-all variance can be made, however, in terms of over-all terms, and a close correspondence in terms is observed; the third column in the "correlation diagram," Fig. 1, brings out this correlation. The eight individual degrees of freedom for deviations from individual series means split up into a single DF for deviation from the over-all mean, and 7 DF expressing the failure of the individual series means to coincide; in other words, measuring the average effect of substituents in ring B on the chlorination rates (in the customary terminology of analysis of variance, these 7 DF are referred to as "between series"). The eight

(1) Work supported by the Office of Ordnance Research, U. S. Army.

(2) (a) L. P. Hammett, *Physical Organic Chemistry*, McGraw-Hill Book Co., Inc., New York, 1940, Chapter VII. (b) H. H. Jaffé, *Chem. Revs.*, **53**, 191 (1953).

(3) H. H. Jaffé, *Science*, **118**, 246 (1953); *J. Am. Chem. Soc.*, **76**, 4261 (1954).

(4) G. W. Snedecor, *Statistical Methods*, Iowa State College Press, Ames, Ia., 4th ed. 1946.

(5) B. Jones, *J. Chem. Soc.*, 2903 (1931); 1835 (1935); 1414 (1938); 267, 358 (1941).

(6) In three series, data for one substituent were missing, and were supplied using the Hammett Equation. Three DF were subtracted from the total number to account for this filling in of missing data.

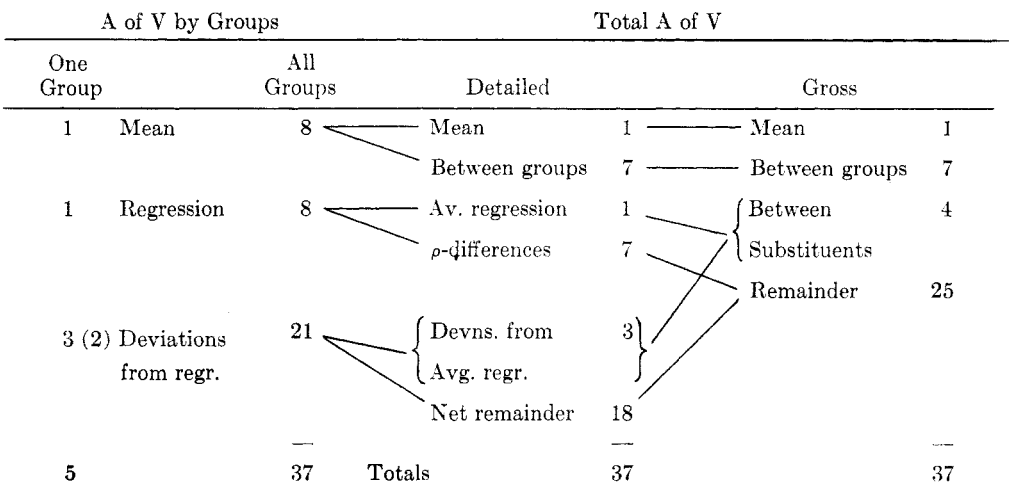


Fig. 1. Correlation diagram of degrees of freedom

DF for individual correlations can similarly be broken up into one DF for average correlation and 7 DF which measure the differences between the eight individual regression coefficients, *i.e.* between the eight ρ-values. The sum of squares for the average regression is most readily obtained by correlating the average of the eight individual rate constants (or rather their logarithms) with the corresponding sigma-values. Finally, the 21 DF for deviation from regressions contain three DF for deviations from the average regression, and the remaining DF are not specified in our model. A direct analysis of variance of the complete set of data (*cf.* right hand column of Fig. 1), divides the total variance into contributions from mean (1 DF), differences between series means (7 DF), between substituent means (4 DF), and a gross remainder (25 DF). The relations between the various terms in the third and fourth columns in Fig. 1 are readily apparent. Thus, the 4 DF for deviations from substituent means (in the fourth column) obviously contain 1 DF for average regression and 3 DF for deviations from it. Similarly, the 25 DF in the gross remainder contain the 7 DF for the differences in ρ-values, and then leave 18 DF of net remainder which are not otherwise identified in the present model. These 18 DF actually represent higher order interaction terms, and could conceivably be used for statistical tests, but, in the absence of duplicate rate data they are a measure of experimental error.

The computations of all the terms are performed readily, making free use of the principle of additivity of sums of squares, and are listed in Table I. Comparison of the mean square of the net remainder with the corresponding values for the various other types of variations shows that the effects due to substituents in either ring on the rate of chlorination, and the correlation are highly significant. The very high F values (variance ratios) are characteristic of analysis of variance of well known and readily observable effects. No statistical treatment was necessary to recognize these effects and failure to have attained large F values for these terms would have done more to throw doubt on the analysis than on the phenomena. The degrees of freedom for differences between ρ-values, however, turn out not to be significant (at the 95% level) so that it appears that ρ-values are not significantly affected by substituents in ring B. Further, we might have asked whether there exists a significant difference in ρ-values depending on the number and positions of the sites available for chlorination (some of the groups X, Y, Z, W in I are hydrogen atoms). Such a question could be examined by dividing the set of eight series in groups according to the nature and number of available sites, and calculating average regressions for each group. Comparison of the sum of the sums of squares for these separate groups with the overall average regression sum of squares then permits a test of significance of such differences. Such a procedure represents a division of the sum of squares corresponding to the 7 DF for ρ-value differences into portions between and within groups. In the present case the total sum of squares corresponding to these 7 DF is so small that, even if all of it were accounted for in a single degree of freedom, no significance would result. Consequently no such analysis is carried out, and it can be concluded that the number and type of sites available do not affect the ρ-values.

Sets of correlations with two independent variables. Tirouflet and co-workers[7] have recently measured the rates of alkaline hydrolysis of four 5-substituted phthalimides (II), of their conjugate bases, and of their *N*-methyl and *N*-ethyl derivatives. The authors suggest that the data may be correlated with the corresponding σ_p-values (σ-values for the para position), or slightly better with an average between the σ-values for meta and para position. Since the mechanism of the reaction most

(7) R. Dabard and J. Tirouflet, *Bull. soc. chim. France*, 565 (1957); J. Tirouflet, R. Dabard, and E. Laviron, *Bull. soc. chim. France*, 570 (1957).

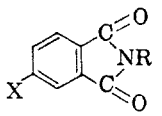
TABLE I

ANALYSIS OF VARIANCE OF THE RATES OF CHLORINATION OF SUBSTITUTED BENZYL PHENYL ETHERS (I)[a]

Source	D.F.	Sum of Squares	Mean Square	F^b
A: Total	37[c]	49.942088	—	
Mean	1	44.685732	—	
Between ring A substituents	4	3.092759	0.773190	6970**
Between ring B substituents	7	2.161072	0.308725	2780**
Gross remainder	25	0.002525	—	
B: Deviations from means for ring B substituents	29	3.095284	—	
Individual regressions	8	3.081962	0.385245	3470**
Deviations from individual regressions	21	0.013322	0.000634	5.71**
C: Deviations from means for ring B substituents	29	3.095284	—	
Average regression	1	3.081434	3.081434	2780**
Differences between ρ's	7	0.000528	0.000075	<1
Deviations from average regression	3	0.013850	0.004465	40**
Net remainder	18	0.001997	0.000111	—

[a] In each subtable the first entry is either contained in one of the preceding subtables, or is the sum of several terms in one of them. The remaining entries in each subtable are the breakdown of the entry in the first line. [b] One asterisk indicates significance at the 95%, a double asterisk, at the 99% level. [c] Corrected for data supplied by use of the Hammett Equation.

likely involves attack by OH^- on *one* of the carbonyl carbon atoms, the more reactive of the car-



II

bonyl groups, *i.e.* the one for which the σ-value is higher (more positive), should be the reactive group.[8] In this case, correlation with a single σ-value should be made with the larger one of the two values ($\sigma_>$), and correlations with two values should make use of the equation:

$$\log k = Y_{ij} = \sigma_> \rho_{1j} + \sigma_< \rho_{2j} + Y_{oj} \qquad (2)$$

When such an analysis is attempted it is again found, as expected, that the correlation, either with one or two independent variables, is excellent. But the critical question is whether a *significant* improvement can be achieved by the use of Equation 2 over the simple linear correlation, and whether there are significant differences between the regression coefficients for the four series of data.

The analysis of variance (*cf.* Table II) and the correlation diagram are analogous to the previous case. The only new feature involves the splitting of the 2 DF for multiple regression (either individual or average) into a DF for linear regression and a DF for improvement due to the inclusion of the second independent variable. Similarly, the differences between ρ-values are separated into differences in the linear regression ρ's, and additional differences due to the second set of ρ's. It should be noted, however, that the second set of differences as evaluated in Table II do not represent the differences in the ρ_2's in Equation 2. It is also possible to separate the variance due to multiple regression (either individual or average) into two parts, for ρ_1 and ρ_2 separately, and then the differences between corresponding values for individual and average regressions give the variance due to differences in ρ_1's and ρ_2's separately. This analysis is not carried out in the present case, since there seems no reason to expect such differences in one set of ρ's, but not in the other.

Table II shows that the differences between ρ-values for the various series are not significant but that Equation 2 produces a very appreciable improvement over the simple correlation. As in the previous case, one might have anticipated especially large differences between ρ-values for a particular series (the phthalimide conjugate bases, *e.g.*, might have been expected, due to the presence of the lone pair of electrons, to have a ρ-value different from the other 3 reaction series). Such a hypothesis can be tested most sensitively by examining correlations for average of groups believed to be similar. Again, if all the differences between slopes were concentrated in two degrees of freedom, the extreme situation possible, this term would not be significant. Hence the hypothesis is rejected with no need for computation of the division of the 6DF for slope differences. The uncertainties in the ρ_1- and ρ_2- values, owing to the small number of data available, are too large to make the actual values obtained of much interest. The average values, obtained in the present treatment, however, are better estimates than could have been obtained otherwise.

Sets with two-way correlations. Whenever reaction rates in a series of reactions satisfying the Hammett equation are determined at various temperatures, ρ-values can be calculated for each temperature. Hammett[2a] suggested long ago that ρ-values should be proportional to $1/T$, and much evidence has ac-

(8) Actually, probably some reaction must occur at both carbonyl groups, but the analysis of this situation is complicated since it involves additivity in k, not log k, and will not be attempted, particularly as long as data are not available for much longer series of substituents.

TABLE II

ANALYSIS OF VARIANCE OF RATES OF HYDROLYSIS OF SUBSTITUTED PHTHALIDES (II)[a]

Source	D.F.	Sum of Squares	Mean Squares	F^b
A: Total adjusted for mean	15	9.554063	—	
Between N-substituents	3	2.249285	0.749762	165**
Between ring A substituents	3	7.236849	2.412283	530**
Gross remainder	9	0.067929	—	
B: Deviations from means for N-substituents	12	7.304778	—	
Individual linear regressions	4	6.992836	1.748209	384**
Improvement by individual multiple regressions	4	0.298079	0.074520	16.4**
Deviations from individual multiple regressions	4	0.013863	0.003466	<1
C: Deviations from means for N-substituents	12	7.304778	—	
Average linear regression	1	6.961088	6.961088	1530**
Improvement by average multiple regression	1	0.275551	0.275551	60.5**
Differences between slopes	6	0.054276	0.009046	1.99
Deviations from average multiple regression	1	0.000210	0.000210	<1
Net remainder	3	0.013653	0.004551	—

[a] and [b] *cf.* Table I.

cummulated to indicate that this prediction was correct.[2b] However, owing to the large uncertainties usually accompanying ρ-values (with a median standard error of about ± 10%), this conclusion is based on the existence of rough trends, and significant differences are rarely demonstrated between any two ρ-values at two temperatures.[2b]

The analysis of variance technique developed permits ready demonstration of the significance of such differences. A series of reactions at four temperatures was chosen at random from our files and subjected to a combined analysis of variance and regression. The mathematical model implied by the correlation of rates on σ-values and on $1/T$ is given in Equation 3, where Y_{00} is an intercept in a

$$\log k = Y_{ij} = (\sigma - \bar{\sigma})\rho + (1/T - \overline{1/T})K + (\sigma - \bar{\sigma})(1/T - \overline{1/T})\alpha + Y_{00} \quad (3)$$

3-dimensional plot, and ρ, K and α are adjustable parameters. The last term represents the variation in individual ρ-values (for different temperatures) and in individual K-values (for different substituents). ρ and K in Equation 3 are the average values for the set of temperatures and substituents,

TABLE III

ANALYSIS OF VARIANCE OF THE RATES OF DISSOCIATION OF *t*-BUTYL PERBENZOATES AT VARIOUS TEMPERATURES[a,b]

Source	D.F.	Sum of Squares	Mean Squares	F^c
A: Total	19[d]	35.128856	—	
Mean	1	25.791747	—	
Between temperatures	3	8.027031	2.675677	2090**
Between substituents	4	1.242866	0.310717	243**
Gross remainder	11	0.067212	—	
B: Total adjusted for mean	18	9.337109	—	
Average regression on σ	1	1.240856	1.240856	969**
Average regression on 1/T	1	8.022765	8.022765	6260**
Average regression on σ/T	1	0.054355	0.051355	42**
Remainder	15	0.019133	0.001276	—
C: Deviations from means for substituents	15	1.310078	—	
Individual regressions on 1/T	4	1.293145	0.323286	173**
Deviations	11	0.014933	0.001357	<1
D: Deviations from means for substituents	15	1.310078	—	
Average regression on 1/T	1	1.240856	1.240856	665**
Deviations from average regression	3	0.002010	0.000670	<1
Variation in slopes	3	0.052289	0.017430	9.35**
Net remainder	8	0.014923	0.001865	—
E: Deviations from means for temperatures	14	8.094243	—	
Individual regressions on σ	5	8.082105	1.616421	1440**
Deviations	9	0.012138	0.001348	1.20
F: Deviations from means for temperatures	14	8.094243	—	
Average deviation on σ	1	8.022765	8.022765	7131**
Deviation from average regression	2	0.004266	0.002133	1.90
Variation in ρ's	4	0.059340	0.014835	13.2**
Net remainder	7	0.007872	0.001125	—

[a] *cf.* Table I. [b] A. T. Blomquist and I. A. Bernstein, *J. Am. Chem. Soc.*, **73**, 5546 (1951). [c] *cf.* Table I. [d] One value was extrapolated, using the Arrhenius equation, and consequently the total number of D.F.'s reduced by one.

TABLE IV

ANALYSIS OF VARIANCE OF THE RATES OF ALKALINE HYDROLYSIS OF SUBSTITUTED PHTHALIDES[a]

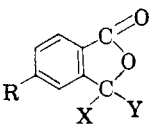
Source	D.F.	Sum of Squares	Mean Squares	F^b
A: Total	16	34.751698	—	
Mean	1	28.156289	—	
Between aromatic ring substituents	3	5.689432	1.896477	379*
Between methylene substituents	3	0.860946	0.286982	57*
Gross remainder	9	0.045031	0.005003	
B: Deviations from over-all means	15	6.595409	—	
Average regression on σ_i	1	5.668356	5.668356	1383*
Average regression on σ^*_j	1	0.834770	0.834770	183*
Regression on $\sigma_i\sigma^*_j$	1	0.022538	0.022538	3.88
Deviations	12	0.069565	0.00580	
C: Deviations from means for single methylene substituents	12	5.734463	—	
Individual regressions on σ_i	4	5.691818	1.422954	183*
Improvement of regression by inclusion of σ_i'	4	0.011261	0.002815	<1
Deviations from multiple regression	4	0.031084	0.007771	1.89
D: Deviations from means for single methylene substituents	12	5.734463	—	
Average regression on σ_i	1	5.668536	5.668536	1383**
Differences between ρ's	3	0.023282	0.007761	1.89
Improvement of average regression by inclusion of σ'_i	1	0.001837	0.001837	<1
Additional differences between and ρ's due to inclusion of σ'_i	3	0.009424	0.003141	<1
Deviations from average multiple regression	1	0.019059	0.019059	4.64
Deviations from regression on σ'_i	2	0.020896	0.010448	2.54
Net remainder	3	0.012325	0.004108	—
E: Deviations from aromatic ring substituents means	12	0.905977	—	
Individual regressions on σ_j*	4	0.853579	0.213395	32**
Deviations from individual regressions on σ_j*	8	0.053398	0.006675	1.47
F: Deviations from means for aromatic ring sub-substituents	12	0.905977	—	
Average regression on σ_j*	1	0.834770	0.834770	183**
Differences between ρ*	3	0.018809	0.006270	1.38
Deviations from average regression on σ_j*	2	0.026176	0.013088	2.87
Net remainder	6	0.026222	0.004555	

[a] and [b] *cf.* Footnotes to Table I.

respectively. A significant contribution to the variance from α indicates significant differences between the individual ρ's and K's. But since the correlations are only approximate, it is still of interest to test deviations from average ρ- and K-values separately (Table III), and this procedure permits an answer to the question whether differences from the average regression and from the individual regressions, are significant.

In the present case, the last term in Equation 3 is highly significant, providing the demonstration that ρ-values change significantly with temperature. Although the mean square values for differences between ρ- and K-values, and for error, evaluated in two different ways do not agree accurately, the discrepancies are so small that they can be ascribed to accidental fluctuation.

Simultaneous multiple and linear regression. The rates of saponification of 2- and 5-substituted phthalides (III) reported by Tasman[9] provide the most complicated set of data we have treated by the present technique, although further extensions



III

should not produce any new difficulties. The effect of substituents in the aromatic ring in III should be expressable by a two parameter Hammett equation,[3] and the effect of X and Y might be expressed by the Taft equation.[10] Again the differences between ρ-values is of interest; *i.e.*, do 2-substituents affect the ρ-values? The mathematical model is rather complicated:

$$\log k = Y_{ij} = (\sigma_{pi} - \bar{\sigma}_{pi})\rho_1 + (\sigma_{mi} - \bar{\sigma}_{mi})\rho_2 + (\sigma^*_j - \bar{\sigma}_j{}^*)\rho^* + (\sigma_{pi} - \bar{\sigma}_{pi})(\sigma^*_j - \bar{\sigma}^*_j)\alpha_1 + (\sigma_{mi} - \bar{\sigma}_{mi})(\sigma^*_j - \sigma^*_j)\alpha_2 + Y_{oo} \quad (4)$$

The analysis, however, produces no new complications. When it appeared that the ρ_2 term produced no significant improvement, the α_2 term in

(9) A. Tasman, *Rec. trav. chim.*, **46**, 653 (1927).

(10) R. W. Taft, Jr., in M. Newman, *Steric Effects in Organic Chemistry*. John Wiley and Sons, Inc., New York, 1956, Chapter 13, Section VI.

Equation 4 was also ignored. Again, the analysis is carried out in detail in order to segregate deviations from the various linear and multiple regressions from the remainder terms (Table IV). In this case it appears that, the linear regressions are again highly significant, the model

$$\log k = Y_{ij} = (\sigma_{pi} - \bar{\sigma}_{pi})\rho_i + (\sigma^*_j - \sigma^*_j)^*\rho + Y_{oo} \quad (5)$$

is as satisfactory as the much more complicated model of Equation 4.

Discussion. Maybe the most striking features of Tables I–IV are the tremendous values of the variance ratios (F) for the differences in group means, and the manner in which the vast majority of the corresponding variances are accounted for by the simple linear regressions. As noted above, this fact is not surprising since the relations involved are well established and have long been apparent in extensive sets of data without recourse to more than the most rudimentary statistics. It is not a purpose of the present statistical technique to demonstrate their significance; rather, if they turned out not highly significant, the technique would be open to question.

Although more subtle, more important is the fact that in many cases deviations from regression (both individual, and when appropriate, average) are *not* significant. The search for a good criterion of "fit" to an empirical relation has gone on for a long time. It would appear that the fact that deviations from regression are not significant was the best possible criterion. Unfortunately, this criterion is rarely applicable to published data, since, in the analysis of a single regression, no DF remain to make such a test, unless data for duplicate determinations are available to provide an independent estimate of error. In the types of analysis presented in the present paper, also, it has been necessary to use higher order interactions as estimates of error; although these interactions probably provide a reliable estimate, an independent estimate of error would be preferable, since it would also permit the testing of these interactions. Unfortunately, it appears to have become customary *not* to publish duplicate values, except under special circumstances, or in a few instances to illustrate the type of reproducibility obtaine.

Finally, the method developed provides the most sensitive possible test of the significance of differences between regression coefficients (slopes) and of improvements due to inclusion of additional independent variables. The total amount of labor involved in the types of analyses outlined is negligible compared with the work involved in the accummulation of the experimental data. The most complicated of the analyses reported here can be completed with the use of a desk calculator in 1–2 hr.

CINCINNATI 21, OHIO

[CONTRIBUTION FROM THE DEPARTMENT OF CHEMISTRY, MASSACHUSETTS INSTITUTE OF TECHNOLOGY]

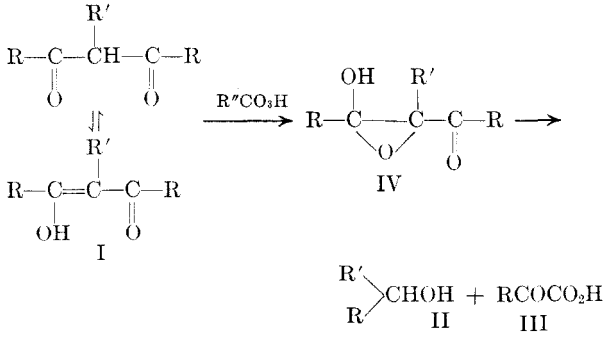
Reaction of β-Diketones with Peracids

HERBERT O. HOUSE AND WALTER F. GANNON[1]

Received December 16, 1957

The reactions of 4-methyl-3,5-heptanedione and 3-benzyl-2,4-pentanedione with monoperphthalic acid have been studied. The heptanedione derivative yielded 4-hydroxy-4-methyl-3,5-heptanedione which underwent thermal isomerization to form the propionic acid ester of 2-hydroxy-3-pentanone. 3-Benzyl-2,4-pentanedione underwent a similar series of transformations.

In 1936 Boeseken and Jacobs reported[2] a study of the reaction of peracetic acid with a series of β-diketones and β-keto esters. The reaction, which occurred only with enolizable β-dicarbonyl compounds I, was said to yield an alcohol II and an α-keto acid III as shown in the accompanying equation when one equivalent of the peracid was employed. With an excess of the peracid a mixture of acids was obtained. In certain cases isolation of the supposed intermediates IV was also reported. Subsequently, a portion of the work claimed to yield the intermediates IV(R′ = H) was repeated by Karrer and co-workers, perbenzoic acid being used as the peracid to facilitate isolation of products.[3–5]



(1) Rohm & Haas Research Assistant, 1957.

(2) J. Boeseken and J. Jacobs, *Rec. trav. chim.*, **55**, 804 (1936).

(3) P. Karrer, J. Kebrle, and R. M. Thakkar, *Helv. Chim. Acta*, **33**, 1711 (1950).

(4) P. Karrer, J. Kebrle, and U. Albers-Schonberg, *Helv. Chim. Acta*, **34**, 1014 (1951).

(5) P. Karrer, U. Albers-Schonberg, and J. Kebrle, *Helv. Chim. Acta*, **35**, 1498 (1952).