SPLIT PLOT REGRESSION METHODS

Barry Kurt Moser*

Texas Research and

Development Foundation

Austin, Texas 78746

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #80-16

July 1980

---

* M.S. Applied Statistics, Purdue University (1975)

SPLIT PLOT REGRESSION METHODS

Barry Kurt Moser*
Texas Research and
Development Foundation
Austin, Texas 78746

Two equivalent methods for split plot regression analysis are presented. The first method requires only one regression model and can be easily applied to small data sets. The second method uses one regression model for each plot in the design. The calculations for Method 2 are more straight forward if a computer regression program is available. For large data sets Method 2 will generally use less memory and less computer time than Method 1. Therefore, it is often more efficient to use Method 2.

Key Words: Completely Randomized Design, Split Plot Design, Whole Plot, Split Plot

---

* M.S. Applied Statistics, Purdue University (1975)

1. INTRODUCTION

The regression theory presented in many text books is based on the completely randomized design (CRD) with one random component. A few references are Draper and Smith (1966), Neter and Wasserman (1979) and Chatterjee and Price (1977). Split plot design regressions include more than one random component and therefore the CRD theory is not directly applicable, Daniel and Wood (1971, p. 59) and Anderson and McLean (1974, p. 206). Two split plot regression methods are presented in this paper. The first method is for designs with small samples and few splits. The second method is recommended for designs with large samples or many splits. An actual experimental example will be used to explain the two methods.

2. THE DESIGN

It was of interest to predict the space mean speeds of cars operating on roads of the type that existed in Brazil. It was important that the final equations be easy to apply. Therefore, the independent variables were defined as simple aggregate indices of the road characteristics.

A traffic and fuel algorithm (TAFA) was developed by Zaniewski, Moser and Swait (1979). The algorithm predicted speeds of vehicles on routes of any type and length. TAFA was used to generate a data set of space mean speeds for cars.

A sample of twenty-eight routes was chosen from the population of routes in Brazil. Each route in the sample had a defined geometric profile. The vertical and horizontal geometries of the routes were quantified with an average rise plus fall and average degree of curvature index, respectively.

On each route TAFA was run with two surface type and two roughness values. Car speeds were then generated for each combination within each route.

The randomization scheme for this experiment occurs in two phases. The routes were randomly chosen within the levels of the geometric indices. The factors surface type and roughness were then applied in a random order within each route. The effects of vertical and horizontal geometry from the whole plot portion of the design. The effects of surface type and roughness form the split plot portion. The effect of routes within the levels of the geometric characteristics is the whole plot random component. The residual error is the split plot random component.

The design layout is given in Table 1. There are three levels of average rise plus fall and average degree of curvature, two levels of surface type and roughness. The design is slightly unbalanced since there can be 2, 3, or 4 routes within the levels of the geometric characteristics. The routes are coded from 1 to 28 and the values of the dependent and independent variables are given in Table 2.

3. THE ANOVA MODEL

The ANOVA model for the split plot design example is:

$$Y_{ijk\ell m} = \mu+A_i+B_j+AB_{ij}+C_{(ij)k}+D_\ell+E_{(\ell)m}+AD_{i\ell}+BD_{j\ell}+AE_{i(\ell)m}+BE_{j(\ell)m}+\varepsilon_{ijk\ell m} \quad (1)$$

where: $Y_{ijk\ell m}$ = the space mean speed for cars in the ijkℓm combination

μ = overall mean space mean speed

A_i = fixed effect of the ith level of average rise plus fall; $i = 1,\ldots,I$

B_j = fixed effect of the ith level of average degree of curvature; $j = 1,\ldots,J$

$C_{(ij)k}$ = random effect of the kth route within the ijth combination of AB

D_ℓ = fixed effect of the ℓth level of surface type; $\ell = 1,\ldots,L$

$E_{(\ell)m}$ = fixed effect of the mth level of roughness within the ℓth surface type; $m = 1,\ldots,M$

$AB_{ij}, AD_{i\ell}, \ldots, BE_{j(\ell)m}, \varepsilon_{ijk\ell m}$

= pooled residual of the interactions $CD_{(ij)k\ell}$, $CE_{(ij)k(\ell)m}, ABD_{ij\ell}, ABE_{ij(\ell)m}$ random effect.

The two random components of model (1) are $C_{(ij)k}$ and $\varepsilon_{ijk\ell m}$. The former is used to test the whole plot factors, the latter tests the split plot factors. The results of the ANOVA were the following:

(1) The mean square of $C_{(ij)k}$ is significantly larger than the mean square of $\varepsilon_{ijk\ell m}$. Therefore the two random components are not pooled.

(2) The whole plot factors A_i and B_j are significant at $\alpha = 0.05$. Tests of the mean speeds per level of A_i and B_i showed that no significant change in speed occurs after a rise plus fall of 27 (m/km) and the effect of average degree of curvature before 125 (degrees/km) is greater than the effect after 125.

(3) The split plot factors D_{ℓ}, $E_{(\ell)m}$, $BD_{j\ell}$ and $AE_{i(\ell)m}$ are significant at $\alpha = 0.05$. Tests of the mean speeds per level of the significant factors were run. The tests indicated that the roughness influence on paved roads was significantly larger than on unpaved roads, the average degree of curvature influence was greater on paved roads than on unpaved roads and the roughness influence increased with average rise plus fall up to a value of 27 (m/km).

Since the means squares of $C_{(ij)k}$ and $\varepsilon_{ijk\ell m}$ are heterogeneous the regression model or models must identify two random components. Two regression methods are presented below. Each approach includes the independent variables that were identified as significant in the ANOVA.

4. METHOD 1 - SMALL SAMPLES

The first method uses one regression model. For the example problem the following model is hypothesized:

$$Y_{ijk\ell m} = a_0+a_1RPF11+a_2ADC11+a_3ADC12+b_1ST+b_2QI1+b_3ST\ QI1 + b_4ADC11\ ST+b_5ADC12\ ST+b_6RPF11\ QI1+\varepsilon'_{ijk\ell m} \qquad (2)$$

where: RPF11 = RPF if RPF < = 27; 27 if RPF > 27

ADC11 = ADC if ADC < = 125; 125 if ADC > 125

ADC12 = 0 if ADC < = 125; (ADC-125) if ADC > 125

ST = -1 if Paved Surface; 1 if Unpaved Surface

QI1 = (QI-65) if Paved Surface; (QI-135) if Unpaved Surface.

Any linear regression computer package can easily be applied to calculate the regression table values and the standard errors of the estimates. These values are presented in Table 3. The equation is:

$$\hat{Y} = 75.6-.223RPF11-.105ADC11-.052ADC12-6.21ST-.076QI1+.023ST\cdot QI1$$
$$+ .0092ADC11\cdot ST+.0196ADC12\cdot ST-.0020RPF11\cdot QI1 \qquad (3)$$

The mean square error of $\varepsilon'_{ijk\ell m}$ is the pooled estimate of the whole plot and split plot residuals. Tests of significance on the whole plot coefficients must be based on the whole plot residual mean square. The split plot residual mean square forms the basis for the tests on the split plot coefficients. The error $\varepsilon'_{ijk\ m}$ can be divided into its two components in the following way:

$$\varepsilon' = \varepsilon_1 + \varepsilon_2$$

$$SS(\varepsilon') = SS(\varepsilon_1) + SS(\varepsilon_2)$$

$$\Sigma\, \varepsilon'^{\,2}_{ijk\ell m} = LM \sum_i^I \sum_j^J \sum_k^K \varepsilon'^{\,2}_{ijk..} + \sum_i^I \sum_j^J \sum_k^K \sum_\ell^L \sum_m^M (\varepsilon'_{ijk\ell m} - \varepsilon'_{ijk..})^2 \qquad (4)$$

ε_1 is the whole plot residual and ε_2 is the split plot residual. The calculations for the example are done using the formula (4) produce the following results:

$$SS(\varepsilon_1) = 1185.1 \qquad (5)$$

$$SS(\varepsilon_2) = 512.4 \qquad (6)$$

There are 28 routes and 4 coefficients estimated in the whole plot. Therefore ε_1 has 28-4 = 24 degrees of freedom. From Table 3, ε' has 102 degrees of freedom. Therefore ε_2 has 102-24 = 78 degrees of freedom. The mean square can be directly calculated.

$$\sigma^2(\varepsilon_1) = MS(\varepsilon_1) = 1185.1/24 = 49.3 \qquad (7)$$

$$\sigma^2(\varepsilon_2) = MS(\varepsilon_2) = 512.4/78 = 6.6 \qquad (8)$$

F tests on the independent variables can now be performed using 49.3 for the whole plot tests and 6.6 for the split plot. The correct standard errors for the coefficients can be calculated using the following formula:

$$\text{Whole Plot} \quad \sigma(a) = \sigma^*(a)\cdot\sigma(\varepsilon_1)/\sigma(\varepsilon') \tag{9}$$

$$\text{Split Plot} \quad \sigma(b) = \sigma^*(b)\sigma(\varepsilon_2)/\sigma(\varepsilon') \tag{10}$$

where: $\sigma(a)$ = correct standard error for any whole plot coefficient.

$\sigma^*(a)$ = corresponding standard error for the whole plot coefficient from Table 3

$\sigma(b)$ = correct standard error for any split plot coefficient

$\sigma^*(b)$ = corresponding standard error for the split plot coefficient from Table 3.

T statistics to test the hypothesis that the coefficient equals zero can now be calculated using the correct standard error values.

$$\text{Whole Plot} \quad t_{1,24} = a/\sigma(a) \tag{11}$$

$$\text{Split Plot} \quad t_{1,78} = b/\sigma(b) \tag{12}$$

Table 4 gives the F statistics, the correct standard errors and the t statistics for each coefficient.

These calculations can become tedious as the sample size or number of split plots increases. Thus an equivalent alternative method is presented.

5. METHOD 2 - LARGE SAMPLES

The number of random components determines the number of regression models needed for Method 2. For the example two models are used. The first is designed to predict the effects tested by the random component

ε_1, the second for the effects tested by ε_2. The two models are:

$$Y_{ijk...} = a_0 + a_1 RPF11 + a_2 ADC11 + a_3 ADC12 + \varepsilon_1 \qquad (13)$$

$$Y' = (Y_{ijk\ell m} - Y_{ijk...})$$

$$= b_1 ST + b_2 QI1 + b_3 ST\ QI1 + b_4 ADC11 \cdot ST + b_5 ADC12 \cdot ST + b_6 QI1 \cdot RPF11 + \varepsilon_2 \ . \qquad (14)$$

Model (13) is run as a weighted regression. The weight is the number of observation per mean value of $Y_{ijk...}$, LM. For the example LM=2.2=4. Model (14) is an unweighted regression. It is run without an intercept since the split plot variables ST and QI1 are centered around zero.

The regression table values and standard error estimates of the coefficients for Model (13) are given in Table 5. The equation is:

$$\hat{Y}_{ijk...} = 75.6 - .223\ RPF11 - .105\ ADC11 - .052\ ADC12 \qquad (15)$$

The information reported in Table 5 were produced directly by the computer regression program used for this example. The values do not need adjustment. All estimates and statistics are equal to the equivalent whole plot information given for Method 1. The result of model (14) is:

$$\hat{Y}' = -6.21ST - .076QI1 + .023ST \cdot QI1 + .0092ADC11 \cdot ST + .0196ADC12 \cdot ST - .0020RPF11 \cdot QI1 \qquad (16)$$

A change has to be made to the usual computer regression output for equation (16). The twenty eight mean values $Y_{ijk...}$ were subtracted from the original dependent variable $Y_{ijk\ell m}$ to define Y'. The degrees of freedom of the error in equation (16) must be adjusted by the number of mean values subtracted, in this case twenty eight. The mean square error, the standard error of the estimates, and the statistics

must be recalculated. All of these values are given in Table 6. The formulas for the standard error of the estimate and t statistic calculations are:

$$\sigma(b) = \sigma^{\delta}(b) \cdot \sigma^{\delta}(\varepsilon_2)/\sigma(\varepsilon_2) \qquad (17)$$

$$t_{1,\ 106-28=78} = b/\sigma(b) \qquad (18)$$

where $\sigma^{\delta}(b)$ = std. error of estimate from the output given in Table 6

$\sigma^{\delta}(\varepsilon_2)$ = std. error of the equation form the output before correction by the whole plot degrees of freedom (for the example $\sigma^{\delta}(\varepsilon_2) = [512.4/106]^{\frac{1}{2}} = 2.2$)

6. CONCLUSIONS

The two regression methods presented can be applied to split plot design data with any number of splits. A set of speed data is used to describe the calculation procedures for the two methods. The methods are shown to produce equivalent results with the second method being slightly easier to apply.

7. ACKNOWLEDGEMENTS

The author wishes to thank Dr. Virgil Anderson for his support and Mr. Leonard Moser for his participation in many discussions.

REFERENCES

ANDERSON, V. L. and McLEAN, R. A. (1974). Design of Experiments, A Realistic Approach. Marcel Dekker, Inc., New York.

CHATTERJEE, S. and PRICE, B. (1977). Regression Analysis by Example. John Wiley & Sons, New York.

DANIEL, C. and WOOD, F. S. (1971). Fitting Equations to Data. Wiley Interscience, New York.

DRAPER, N. R. and SMITH, H. (1966). Applied Regression Analysis. Wiley, New York.

NETER, J. and WASSERMAN, W. (1974). Applied Linear Statistical Models. Richard D. Irwin, Inc., Homewood, Illinois

ZANIEWSKI, J. P., MOSER, B. K., SWAIT, J. D. (1979). Predicting Travel Time and Fuel Consumption for Vehicles on Low-Volume Roads. Low Volume Roads: Second International Conference, Transportation Research Record, 702, p. 335.

TABLE 1 - Design Layout for Car Speed Aggregation Example

RISE PLUS FALL	0-15(mm/km)									15-35									35+									
AV. DEG. CURV.	0-100 deg/km			100-180			180+			0-100		100-180			180+				0-100			100-180			180+			
ROUTE CODE	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
ROUGHNESS																												
SURFACE TYPE																												
PAVED 100																												
PAVED 50																												
UNPAVED 200																												
UNPAVED 70(QI*)																												

TABLE 2 - Data for Car Speed Example

Route	ADC	RPF	ST	QI	Y	ST	QI	Y	ST	QI	Y	ST	QI	Y
1	5.2	11.5	-1	50	81.9	-1	100	75.2	1	70	71.7	1	200	70.3
2	5.3	11.4	-1	50	81.8	-1	100	74.5	1	70	70.3	1	200	65.8
3	0.0	11.2	-1	50	81.4	-1	100	75.9	1	70	74.3	1	200	73.9
4	104.0	9.8	-1	50	70.4	-1	100	63.1	1	70	63.5	1	200	47.9
5	104.9	6.8	-1	50	75.7	-1	100	68.3	1	70	67.1	1	200	55.1
6	100.9	10.0	-1	50	61.9	-1	100	54.5	1	70	58.4	1	200	42.5
7	240.6	13.9	-1	50	58.3	-1	100	51.2	1	70	54.3	1	200	43.9
8	206.2	13.7	-1	50	61.3	-1	100	54.2	1	70	56.9	1	200	45.9
9	240.4	14.4	-1	50	62.3	-1	100	54.9	1	70	56.9	1	200	44.5
10	16.5	27.2	-1	50	80.3	-1	100	73.3	1	70	69.0	1	200	51.6
11	17.8	26.7	-1	50	79.3	-1	100	71.7	1	70	67.9	1	200	47.9
12	153.1	30.4	-1	50	62.0	-1	100	54.7	1	70	57.9	1	200	43.9
13	132.3	25.9	-1	50	60.1	-1	100	53.1	1	70	57.3	1	200	43.6
14	138.6	26.0	-1	50	68.5	-1	100	60.9	1	70	60.8	1	200	44.8
15	191.1	26.4	-1	50	59.8	-1	100	53.1	1	70	56.7	1	200	44.9
16	319.8	32.1	-1	50	49.3	-1	100	41.6	1	70	49.3	1	200	39.0
17	329.3	32.3	-1	50	55.1	-1	100	48.1	1	70	52.9	1	200	42.3
18	198.8	26.7	-1	50	54.2	-1	100	46.9	1	70	52.4	1	200	41.1
19	47.3	40.8	-1	50	76.0	-1	100	69.2	1	70	67.1	1	200	52.7
20	34.2	40.2	-1	50	78.1	-1	100	70.7	1	70	67.7	1	200	51.2
21	30.7	40.9	-1	50	77.8	-1	100	70.7	1	70	67.3	1	200	50.5
22	131.4	40.6	-1	50	59.4	-1	100	52.1	1	70	57.2	1	200	42.8
23	130.8	40.5	-1	50	69.8	-1	100	52.5	1	70	61.6	1	200	45.4
24	112.1	40.9	-1	50	74.4	-1	100	67.1	1	70	64.5	1	200	47.0
25	117.1	40.9	-1	50	70.2	-1	100	63.0	1	70	61.9	1	200	46.1
26	202.0	39.3	-1	50	63.7	-1	100	56.5	1	70	57.6	1	200	43.9
27	412.1	39.4	-1	50	45.0	-1	100	38.2	1	70	46.2	1	200	37.2
28	326.3	38.3	-1	50	46.6	-1	100	39.4	1	70	47.4	1	200	38.3

TABLE 3 - Regression Table and Statistics for Model (3)

Source	d.f.	Sum of Squares
RPF11	1	1244.4
ADC11	1	5555.1
ADC12	1	1398.1
ST	1	2137.2
QI1	1	2860.9
ST·QI1	1	65.5
ADC11·ST	1	169.6
ADC12·ST	1	199.9
RPF11·QI1	1	57.6
ε'	102	1697.5
TOTAL	111	15385.8

Parameter	Std. Error of Estimate = σ*
INTERCEPT	1.35
RPF11	0.054
ADC11	0.0091
ADC12	0.0057
ST	0.801
QI1	0.026
ST·QI1	0.012
ADC11·ST	0.0091
ADC12·ST	0.0057
RPF11·QI1	0.0011

*The standard errors come directly from the computer output and must be modified by a multiple.

TABLE 4 - Correct t and F Statistics and Standard Errors of the Estimates for Model (3)

Parameter	F	t	Std. Error of the Est.
INTERCEPT		32.5	2.32
RPF11	25.2	- 2.4	0.093
ADC11	112.7	- 6.7	0.016
ADC12	28.4	- 5.3	0.010
ST	323.8	-12.3	0.504
QI1	433.5	- 4.8	0.016
ST•QI1	9.9	3.1	0.0074
ADC11•ST	25.7	1.6	0.0057
ADC12•ST	30.3	5.5	0.00356
RPF11•QI1	8.7	- 3.0	0.00067

TABLE 5 - Regression Table and Statistics for Model (15)

Source	df	Sum of Squares	Mean Square
RPF11	1	1244.4	
ADC11	1	5555.1	
ADC12	1	1398.1	
ε_1	25	1185.1	49.4
TOTAL		9382.6	

Parameter	Std. Error of Estimate	t value
INTERCEPT	2.32	32.5
RPF11	0.093	- 2.4
ADC11	0.016	- 6.7
ADC12	0.010	- 5.3

TABLE 6 - Regression Table and Corrected Statistics for Model (16)

Source	df	Sum of Squares	Mean Square
ST	1	2137.2	
QI1	1	2860.9	
ST·QI1	1	65.5	
ST·ADC11	1	169.6	
ST·ADC12	1	199.9	
QI·RPF11	1	57.6	
ε_2	106-28=78	512.4	6.6
TOTAL	84	6003.1	

Parameter	t from Output	Correct t	$\sigma^{\delta}(b)$ = Std. Error from Output	$\sigma(b)$ = Correct Std. Error
ST	-14.4	-12.3	0.43	0.50
QI1	- 5.4	- 4.8	0.014	0.016
ST·QI1	3.7	3.1	0.0063	0.0074
ST·ADC11	1.9	1.6	0.0049	0.0057
ST·ADC12	6.4	5.5	0.0030	0.0035
QI·RPF11	- 3.5	- 3.0	0.00058	0.00067