Analisi della varianza del dataset ”ASFALTO”

D’Andrea Gabriele, Iannozzi Vanessa, Naclerio Michele,

Petrone Daniele, Rosi Claudio, Scoglio Melania

Questo dataset contiene dati derivanti da un esperimento a blocchi randomizzati (sono presenti fattori sub-sperimentali che contribuiscono a spiegare la variabilità), completo (almeno un’osservazione, in questo caso una sola, per ciascuna combinazione possibile dei fattori) e bilanciato (lo stesso numero di osservazioni per ciascuna combinazione). Si assume che la variabile risposta Y sia a componenti indipendenti, normali e omoschedastiche. Il dataset contiene 4 variabili esplicative e una variabile risposta per un totale di 48 osservazioni. Le variabili sono le seguenti:

- *Y* : variabile risposta. Indica la variazione dell’elasticità dell’asfalto;
- *A*: tipo di asfalto. L’asfalto può essere di 4 tipi diversi con livelli 1, 2, 3 e 4;
- *C* : presenza o assenza di cemento. Nel primo caso ha valore 1, nel secondo ha valore 2;
- *M* : tipi differenti di mescolatura. Questa variabile si presenta con due classi: 1 e 2;
- *T* : temperatura di lavorazione dell’asfalto. La variabile si presenta con 3 Classi : 1, 2 e 3.

Prima di passare all’analisi della varianza, vengono calcolati alcuni indici di statistica descrittiva:

	Asfalto															
	1				2				3				4			
	Cemento				Cemento				Cemento				Cemento			
	1		2		1		2		1		2		1		2	
	Mescola		Mescola		Mescola		Mescola		Mescola		Mescola		Mescola		Mescola	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.	Var. Elast.
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
Temper.																
1	286.60	287.90	294.10	288.60	235.50	262.80	256.20	252.70	275.00	259.60	273.10	278.70	264.60	257.60	293.80	265.70
2	281.90	282.90	291.70	288.30	262.50	257.60	268.60	260.30	252.10	253.40	281.30	277.90	264.70	263.40	293.70	271.40
3	251.40	237.70	280.50	277.30	238.40	258.80	268.50	277.30	204.90	198.50	225.40	245.90	205.30	219.30	270.10	208.10

The MEANS Procedure

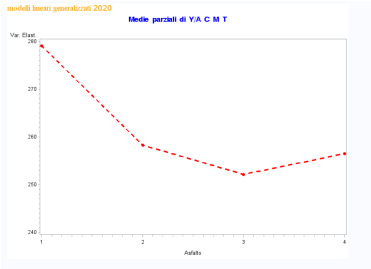
Analysis Variable : Y Var. Elast.			
Temper.	N Obs	Mean	Std Dev
1	16	270.78	16.80
2	16	271.98	13.61
3	16	241.71	28.80

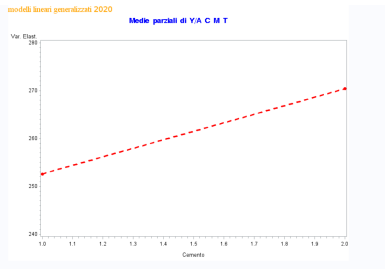
Analysis Variable : Y Var. Elast.			
Mescola	N Obs	Mean	Std Dev
1	24	263.33	25.86
2	24	259.65	24.05

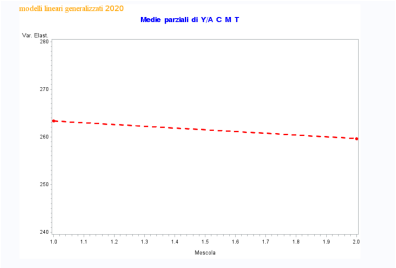
Analysis Variable : Y Var. Elast.			
Cemento	N Obs	Mean	Std Dev
1	24	252.60	25.28
2	24	270.38	21.22

Analysis Variable : Y Var. Elast.			
Asfalto	N Obs	Mean	Std Dev
1	12	279.08	17.06
2	12	258.27	11.93
3	12	252.15	28.76
4	12	256.48	29.84

Ricordando che la media della variabile risposta per tutte le osservazioni è pari a 261.49, possiamo notare da questa tabella che la media dell'elasticità in corrispondenza della terza temperatura (241.714) è molto più bassa della media generale; inoltre in corrispondenza di questa modalità si nota una delle maggiori deviazioni standard.
Per una migliore comprensione sono riportati anche i grafici delle medie parziali:







E' ora possibile analizzare il modello; il primo proposto è il seguente:
$y_{ijkr} = \mu_{ijkr} + \varepsilon_{ijkr}$
$y_{ijkr} \sim N(\mu_{ijkr}, \sigma^2)$
$\varepsilon_{ijkr} \sim N(0, \sigma^2)$ con ε_i e ε_j indipendenti $\forall\ i \neq j$
Si ha quindi che $\mathrm{E}(y_{ijkr}) = \mu_{ijkr}$.

$$\mu_{ijkr} = \mu_{....} + \alpha_i + \beta_j + \gamma_k + \eta_r + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\eta)_{ir} + (\beta\gamma)_{jk} + (\beta\eta)_{jr} + (\gamma\eta)_{kr} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\eta)_{ijr} + (\alpha\gamma\eta)_{ikr} + (\beta\gamma\eta)_{jkr} + (\alpha\beta\gamma\eta)_{ijkr}$$
$$\text{con } i = 1,2,3,4 \quad j = 1,2 \quad k = 1,2 \quad r = 1,2,3.$$

Nel modello appena descritto sono inseriti tutti gli effetti interattivi (6 effetti doppi, 4 effetti tripli e un effetto quadruplo) e i 4 effetti singoli; per questo motivo il modello ottenuto è il modello saturo. Tale modello risulta essere utile poiché permette di studiare la quota di devianza spiegata da ogni parametro.
Prima di mostrare l'output di questo primo modello, come per i prossimi, è importante ricordare che il *type I* e il *type III* saranno i medesimi poiché l'esperimento è stato effettuato in una situazione sperimentale, perciò controllata; ciò significa che la matrice dei dati è fatta a blocchi ortogonali.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	47	28848.05667	613.78844	.	.
Error	0	0.00000	.		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
1.000000	.	.	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	.	.
C	1	3794.963333	3794.963333	.	.
A*C	3	265.891667	88.630556	.	.
M	1	162.067500	162.067500	.	.
A*M	3	965.100833	321.700278	.	.
C*M	1	307.040833	307.040833	.	.
A*C*M	3	1185.624167	395.208056	.	.
T	2	9400.690417	4700.345208	.	.
A*T	6	5161.521250	860.253542	.	.
C*T	2	927.187917	463.593958	.	.
A*C*T	6	260.187083	43.364514	.	.
M*T	2	13.703750	6.851875	.	.
A*M*T	6	432.767917	72.127986	.	.
C*M*T	2	9.432917	4.716458	.	.
A*C*M*T	6	777.782083	129.630347	.	.

Figure 1: modello saturo

In Figura 1 si può vedere, come era da aspettarsi, che il modello assume tutti i gradi di libertà; proprio per questo motivo le colonne dei valori F e dei p-value sono vuote. Ciò è dovuto al fatto che l'errore assume 0 gradi di libertà, quindi il valore F, calcolato come rapporto tra il *Mean Square* del modello e il *Mean Square* dell'errore, non è definito. Per questo non è possibile fare affermazioni probabilistiche poiché non è presente la componente accidentale; è però possibile fare affermazioni riguardo all'importanza dei singoli parametri nella scomposizione della devianza totale.
Possiamo infatti notare che gli effetti singoli hanno i *Mean Square* più grandi tra tutti i parametri, fatta eccezione per la Mescola. Riguardo agli effetti interattivi, è possibile notare che l'effetto C*M*T spiega una piccolissima parte di devianza, per questo motivo si potrebbe pensare di escludere questo parametro dal modello. Facendo ciò, però, si incorrerebbe in un modello non gerarchico, perché il *Mean Square* dell'effetto triplo sarebbe assorbito dall'effetto quadruplo. In conclusione possiamo quindi affermare che per una valutazione dei parametri basata su F-value e quindi effettuare un'analisi probabilistica, si rende necessaria l'esclusione dell'effetto interattivo quadruplo dal modello.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	41	28070.27458	684.64084	5.28	0.0219
Error	6	777.78208	129.63035		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.973039	4.354071	11.38553	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	13.33	0.0046
C	1	3794.963333	3794.963333	29.28	0.0016
A*C	3	265.891667	88.630556	0.68	0.5938
M	1	162.067500	162.067500	1.25	0.3063
A*M	3	965.100833	321.700278	2.48	0.1583
C*M	1	307.040833	307.040833	2.37	0.1747
A*C*M	3	1185.624167	395.208056	3.05	0.1138
T	2	9400.690417	4700.345208	36.26	0.0004
A*T	6	5161.521250	860.253542	6.64	0.0183
C*T	2	927.187917	463.593958	3.58	0.0949
A*C*T	6	260.187083	43.364514	0.33	0.8958
M*T	2	13.703750	6.851875	0.05	0.9490
A*M*T	6	432.767917	72.127986	0.56	0.7531
C*M*T	2	9.432917	4.716458	0.04	0.9645

Figure 2: modello senza interazione quadrupla

Il modello presente in Figura 2 assume 41 gradi di libertà su 47 totali, poiché i 6 gradi di libertà che assumeva precedentemente l'effetto quadruplo sono passati all'errore, come anche il suo *Mean Square*. Anche il *Mean Square* del modello è aumentato, seppur di poco, passando da 613.79 a 684.64. Il valore F (che ricordiamo essere una realizzazione della variabile casuale F di Fisher, calcolata come rapporto di due chi-quadro con 41 e 6 gradi di libertà, divisi per i rispettivi gradi di libertà) è pari a 5.28; ciò significa che un grado di libertà del modello spiega 5.28 volte un grado di libertà dell'errore. Possiamo però notare che il modello non è fortemente significativo, poiché il p-value del test F è pari a 0.02; l'ipotesi nulla di questo test è che la somma dei parametri al quadrato è pari a 0.
Anche se non strettamente necessario all'analisi condotta, non può passare inosservato il valore R^2, pari a 0.97.
Analizzando i parametri possiamo notare che gli effetti singoli A, T e C sono gli unici fortemente significativi (ricordiamo che il p-value presente in output corrisponde all'ipotesi nulla che la somma dei parametri relativi alle componenti al quadrato sia pari a zero). Tutti gli altri effetti, sia i doppi sia i tripli, hanno p-value piuttosto alti, in particolare C*M*T e M*T; ciò ci induce a pensare di escluderli entrambi dal modello. Non potendo eliminare due parametri contemporaneamente, è preferibile eliminare il parametro C*M*T che risulta essere quello meno significativo.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	39	28060.84167	719.50876	7.31	0.0030
Error	8	787.21500	98.40187		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.972712	3.793533	9.919772	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	17.56	0.0007
C	1	3794.963333	3794.963333	38.57	0.0003
A*C	3	265.891667	88.630556	0.90	0.4820
M	1	162.067500	162.067500	1.65	0.2353
A*M	3	965.100833	321.700278	3.27	0.0801
C*M	1	307.040833	307.040833	3.12	0.1153
T	2	9400.690417	4700.345208	47.77	<.0001
A*T	6	5161.521250	860.253542	8.74	0.0037
C*T	2	927.187917	463.593958	4.71	0.0445
M*T	2	13.703750	6.851875	0.07	0.9333
A*C*M	3	1185.624167	395.208056	4.02	0.0514
A*M*T	6	432.767917	72.127986	0.73	0.6375
A*C*T	6	260.187083	43.364514	0.44	0.8331

Figure 3: modello senza interazione tripla C*M*T

In Figura 3 si può vedere il nuovo modello ottenuto. La prima cosa che possiamo notare è che i gradi di libertà e il *Mean Square* del parametro eliminato sono "passati" all'errore. Mentre il *Mean Square* dell'errore è diminuito, quello del modello è aumentato; da ciò consegue che il valore F è maggiore di quello ottenuto precedentemente. Anche il p-value del modello è diminuito, quindi il nostro modello può essere accettato.
Il p-value di tutti i parametri è diminuito, anche se alcuni di essi risultano essere ancora fortemente non significativi. Primo tra tutti troviamo l'effetto doppio M*T; tale effetto è però contenuto nell'effetto triplo A*M*T, perciò la sua esclusione genererebbe un modello non gerarchico. Per questo motivo si decide di eliminare l'effetto triplo A*C*T.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	33	27800.65458	842.44408	11.26	<.0001
Error	14	1047.40208	74.81443		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.963692	3.307767	8.649534	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	23.10	<.0001
C	1	3794.963333	3794.963333	50.73	<.0001
A*C	3	265.891667	88.630556	1.18	0.3512
M	1	162.067500	162.067500	2.17	0.1632
A*M	3	965.100833	321.700278	4.30	0.0240
C*M	1	307.040833	307.040833	4.10	0.0623
T	2	9400.690417	4700.345208	62.83	<.0001
A*T	6	5161.521250	860.253542	11.50	0.0001
C*T	2	927.187917	463.593958	6.20	0.0118
M*T	2	13.703750	6.851875	0.09	0.9130
A*C*M	3	1185.624167	395.208056	5.28	0.0120
A*M*T	6	432.767917	72.127986	0.96	0.4833

Figure 4: modello senza interazione tripla A*C*T

Riguardo al modello presente in Figura 4 possiamo fare le stesse osservazioni fatte in precedenza: i gradi di libertà e il *Mean Square* del parametro eliminato sono passati all'errore, il valore F è aumentato e il modello è statisticamente significativo. I p-value dei singoli parametri sono diminuiti ulteriormente, rendendone addirittura alcuni significativi o almeno quasi. Il p-value più alto risulta essere ancora quello in corrispondenza dell'effetto A*M, ma ancora una volta la sua eliminazione creerebbe problemi di gerarchia; per questo motivo il parametro rimosso dal modello è A*M*T.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	27367.88667	1013.62543	13.70	<.0001
Error	20	1480.17000	74.00850		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.948691	3.289902	8.602819	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	23.35	<.0001
C	1	3794.963333	3794.963333	51.28	<.0001
A*C	3	265.891667	88.630556	1.20	0.3361
M	1	162.067500	162.067500	2.19	0.1545
A*M	3	965.100833	321.700278	4.35	0.0164
C*M	1	307.040833	307.040833	4.15	0.0551
T	2	9400.690417	4700.345208	63.51	<.0001
A*T	6	5161.521250	860.253542	11.62	<.0001
C*T	2	927.187917	463.593958	6.26	0.0077
M*T	2	13.703750	6.851875	0.09	0.9120
A*C*M	3	1185.624167	395.208056	5.34	0.0072

Figure 5: modello senza interazione tripla A*M*T

Le medesime conclusioni fatte per i modelli precedenti, possono essere fatte riguardo il modello in Figura 5. Anche riguardo ai parametri possono essere tratte le stesse conclusioni, infatti tutti i p-value sono diminuiti, rimanendo però ancora molto alto quello in corrispondenza di M*T. Avendo già eliminato l'effetto triplo che lo conteneva, è ora possibile eliminare l'effetto doppio.

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	27354.18292	1094.16732	16.11	<.0001
Error	22	1493.87375	67.90335		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.948216	3.151286	8.240349	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	25.45	<.0001
C	1	3794.963333	3794.963333	55.89	<.0001
M	1	162.067500	162.067500	2.39	0.1366
T	2	9400.690417	4700.345208	69.22	<.0001
A*C	3	265.891667	88.630556	1.31	0.2978
A*M	3	965.100833	321.700278	4.74	0.0107
A*T	6	5161.521250	860.253542	12.67	<.0001
C*M	1	307.040833	307.040833	4.52	0.0449
C*T	2	927.187917	463.593958	6.83	0.0049
A*C*M	3	1185.624167	395.208056	5.82	0.0044

Figure 6: modello senza interazione doppia M*T

Il modello presente in Figura 6 è il modello finale ottenuto. In definitiva il modello assume 25 gradi di libertà, mentre l'errore ne assume 22, per un totale di 47. Il *Mean Square* di questo modello è il più alto ottenuto fin'ora, mentre quello per l'errore è il più basso; per questo motivo il valore F anche risulta essere il migliore. E' importante notare come tale valore sia passato da 5.28 (ottenuto nel primo modello) all'attuale valore di 16.11.
Anche se non fondamentale per la nostra analisi, non può essere tralasciato il valore dell'indice R^2, che seppur diminuito (da 0.97 a 0.94), rimane tutt'ora molto alto.
Infine, osservando i p-value dei parametri possiamo notare che tutti risultano essere statisticamente significativi, fatta eccezione per M e A*C. Essi sono pero contenuti nell'effetto triplo A*C*M che non può essere assolutamente eliminato dal modello poiché fortemente significativo.

Il modello finale è quindi il seguente: $y_{ijkr} = \mu_{ijkr} + \varepsilon_{ijkr}$
$y_{ijkr} \sim N(\mu_{ijkr}, \sigma^2)$
$\varepsilon_{ijkr} \sim N(0, \sigma^2)$ con ε_i e ε_j indipendenti $\forall\ i \neq j$
$E(y_{ijkr}) = \mu_{ijkr}$
$\mu_{ijkr} = \mu_{....} + \alpha_i + \beta_j + \gamma_k + \eta_r + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\eta)_{ir} + (\beta\gamma)_{jk} + (\beta\eta)_{jr} + (\alpha\beta\gamma)_{ijk}$
con $i = 1,2,3,4 \quad j = 1,2 \quad k = 1,2 \quad r = 1,2,3.$

Dopo aver scelto il modello finale, possiamo esaminare le stime dei parametri utilizzando le metodologie del coding effect e del corner point.

CORNER POINT

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	223.6583333	B	6.06473268	36.88	<.0001
A 1	49.7583333	B	8.24034904	6.04	<.0001
A 2	48.2916667	B	8.24034904	5.86	<.0001
A 3	16.4000000	B	8.24034904	1.99	0.0591
A 4	0.0000000	B	.	.	.
C 1	-13.7000000	B	7.52237508	-1.82	0.0822
C 2	0.0000000	B	.	.	.
M 1	37.4666667	B	6.72821682	5.57	<.0001
M 2	0.0000000	B	.	.	.
T 1	34.3812500	B	6.51456792	5.28	<.0001
T 2	39.8437500	B	6.51456792	6.12	<.0001
T 3	0.0000000	B	.	.	.
A*C 1 1	-13.6000000	B	9.51513547	-1.43	0.1670
A*C 1 2	0.0000000	B	.	.	.
A*C 2 1	-2.0666667	B	9.51513547	-0.22	0.8301
A*C 2 2	0.0000000	B	.	.	.
A*C 3 1	-28.7000000	B	9.51513547	-3.02	0.0064
A*C 3 2	0.0000000	B	.	.	.
A*C 4 1	0.0000000	B	.	.	.
A*C 4 2	0.0000000	B	.	.	.
A*M 1 1	-33.4333333	B	9.51513547	-3.51	0.0020
A*M 2 1	-36.4666667	B	9.51513547	-3.83	0.0009
A*M 2 2	0.0000000	B	.	.	.
A*M 3 1	-45.0333333	B	9.51513547	-4.73	0.0001
A*M 3 2	0.0000000	B	.	.	.
A*M 4 1	0.0000000	B	.	.	.
A*M 4 2	0.0000000	B	.	.	.
A*T 1 1	-17.1500000	B	8.24034904	-2.08	0.0493
A*T 1 2	-23.1250000	B	8.24034904	-2.81	0.0103
A*T 1 3	0.0000000	B	.	.	.
A*T 2 1	-53.6750000	B	8.24034904	-6.51	<.0001
A*T 2 2	-46.1000000	B	8.24034904	-5.59	<.0001
A*T 2 3	0.0000000	B	.	.	.
A*T 3 1	8.2000000	B	8.24034904	1.00	0.3305
A*T 3 2	-0.1000000	B	8.24034904	-0.01	0.9904
A*T 3 3	0.0000000	B	.	.	.
A*T 4 1	0.0000000	B	.	.	.
A*T 4 2	0.0000000	B	.	.	.
A*T 4 3	0.0000000	B	.	.	.
C*M 1 1	-39.3666667	B	9.51513547	-4.14	0.0004
C*M 1 2	0.0000000	B	.	.	.
C*M 2 1	0.0000000	B	.	.	.
C*M 2 2	0.0000000	B	.	.	.
C*T 1 1	20.6875000	B	5.82680668	3.55	0.0018
C*T 1 2	15.5125000	B	5.82680668	2.66	0.0142
C*T 1 3	0.0000000	B	.	.	.
C*T 2 1	0.0000000	B	.	.	.
C*T 2 2	0.0000000	B	.	.	.
C*T 2 3	0.0000000	B	.	.	.
A*C*M 1 1 1	39.1333333	B	13.45643363	2.91	0.0082
A*C*M 1 1 2	0.0000000	B	.	.	.
A*C*M 1 2 1	0.0000000	B	.	.	.
A*C*M 1 2 2	0.0000000	B	.	.	.
A*C*M 2 1 1	24.1000000	B	13.45643363	1.79	0.0871
A*C*M 2 1 2	0.0000000	B	.	.	.
A*C*M 2 2 1	0.0000000	B	.	.	.
A*C*M 2 2 2	0.0000000	B	.	.	.
A*C*M 3 1 1	53.7666667	B	13.45643363	4.00	0.0006
A*C*M 3 1 2	0.0000000	B	.	.	.
A*C*M 3 2 1	0.0000000	B	.	.	.
A*C*M 3 2 2	0.0000000	B	.	.	.
A*C*M 4 1 1	0.0000000	B	.	.	.
A*C*M 4 1 2	0.0000000	B	.	.	.
A*C*M 4 2 1	0.0000000	B	.	.	.
A*C*M 4 2 2	0.0000000	B	.	.	.

Le informazioni contenute nelle tabelle presenti nella pagina precedente, sono state originate proprio con questa metodologia. La prima cosa che si può notare è che il valore della stima dell'intercetta, cioè la media generale, è cambiato rispetto al valore precedente, ciò è dovuto al fatto che la stima risente della scelta del parametro preso come corner point. Cambiando il parametro di riferimento, infatti, si potrebbero ottenere risultati anche molto diversi da quelli ottenuti con altri corner point; nel nostro caso però non è stato così, infatti la variazione del parametro di riferimento non ha portato cambiamenti significativi nei risultati ottenuti.
Successivamente si può notare che nella tabella è presente il valore della statistica test t (ricordiamo che al denominatore SAS per default inserisce la migliore stima della varianza a disposizione, che in questo caso è proprio la stima della varianza dell'errore) e p-value del t-test; tale test confronta il valore del parametro con il valore del parametro selezionato come corner point.
Per una interpretazione pratica, possiamo prendere in riferimento il valore della stima del parametro A_1. Tale stima ha valore 49.76; ciò sta a significare che l'utilizzo del primo tipo di asfalto, conduce a un incremento pari a circa 50 unità dell'elasticità dell'asfalto che si sarebbe prodotto se fosse stato utilizzato il quarto tipo di asfalto. E' però importante notare cosa succede quando questa prima tipologia di asfalto viene usata in corrispondenza della prima tipologia di cemento; infatti il risultato che si ottiene è un decremento dell'elasticità media, che ricordiamo essere in questo caso pari a 223.66.
Analoghe considerazioni possono essere fatte per i restati valori presenti nelle tabelle.
Riguardo ai p-value, ricordiamo che se tale valore è compreso tra 0 e α (solitamente posto uguale a 0.05), il test è statisticamente significativo, quindi i dati a nostra disposizione ci inducono a rifiutare l'ipotesi nulla di uguaglianza tra le medie; se invece il valore p non è compreso nell'intervallo, il test è non statisticamente significativo, quindi non è possibile rifiutare l'ipotesi nulla.

CODING EFFECT

Questa metodologia permette di identificare il legame che intercorre tra parametri e media generale; nello specifico permette di leggere i parametri come scostamenti della media generale della variabile risposta. Di seguito vengono riporti i vincoli d'identificabilità dei parametri:

$\sum_{i=1}^{4} \alpha_i = 0$

$\sum_{j=1}^{2} \beta_j = 0$

$\sum_{k=1}^{2} \gamma_k = 0$

$\sum_{r=1}^{3} \eta_r = 0$

$\sum_{i=1}^{4} (\alpha\beta)_{ij} = 0 \ \forall \ j$

$\sum_{j=1}^{2} (\alpha\beta)_{ij} = 0 \ \forall \ i$

$\sum_{i=1}^{4} (\alpha\gamma)_{ik} = 0 \ \forall \ k$

$\sum_{k=1}^{2} (\alpha\gamma)_{ik} = 0 \ \forall \ i$

$\sum_{i=1}^{4} (\alpha\eta)_{ir} = 0 \ \forall \ r$

$\sum_{r=1}^{3} (\alpha\eta)_{ir} = 0 \ \forall \ i$

$\sum_{j=1}^{2} (\beta\gamma)_{jk} = 0 \ \forall \ k$

$\sum_{k=1}^{2} (\beta\gamma)_{jk} = 0 \ \forall \ j$

$\sum_{j=1}^{2} (\beta\eta)_{jr} = 0 \ \forall \ r$

$\sum_{r=1}^{3} (\beta\eta)_{jr} = 0 \ \forall \ j$

$\sum_{i=1}^{4} (\alpha\beta\gamma)_{ijk} = 0 \ \forall \ j, \ \forall \ k$

$\sum_{j=1}^{2} (\alpha\beta\gamma)_{ijk} = 0 \ \forall \ i, \ \forall \ k$

$\sum_{k=1}^{2} (\alpha\beta\gamma)_{ijk} = 0 \ \forall \ j, \ \forall \ i$

Stima degli effetti della risposta **A*C**

A\C	1	2		A_effect
1	1.22	-1.22		17.58
2	3.23	-3.23		-3.23
3	-2.68	2.68		-9.34
4	-1.77	1.77		-5.02
C_effect	-8.98	8.98		261.49

Da questa matrice, ricordando che la variabile C assume valore 1 in presenza di cemento e 2 in assenza, si può osservare che l'utilizzo di cemento rende più rigido l'asfalto poiché la media dell'elasticità diminuisce di quasi 9 unità. La tipologia di asfalto più elastica risulta essere la prima, mentre quella più rigida è la terza.
Combinate insieme quest'ultima tipologia di asfalto, con l'assenza di cemento, è possibile così aumentare la media generale della variabile risposta di 2.32($=\hat{A}_3 + \hat{C}_2 + A_3\hat{C}_2$ = - 9.34 + 8.98 + 2.68) unità, mentre se fosse presente anche il cemento, la media si abbasserebbe di ben 21 unità.
Il miglior risultato, cioè la combinazione che comporta l'incremento maggiore dell'elasticità, risulta essere quando si utilizza l'asfalto di primo tipo, ma non il cemento. In questo caso infatti l'elasticità aumenta di ben 25.34 unità.

Stima degli effetti della risposta **A*M**

A\M	1	2		A_effect
1	0.12	-0.12		17.58
2	-5.15	5.15		-3.23
3	-2.02	2.02		-9.34
4	7.05	-7.05		-5.02
M_effect	1.84	-1.84		261.49

Avendo la variabile M solo due modalità, appare ovvio che gli effetti dovuti al primo o al secondo tipo di mescolatura sono uguali in modulo ma con segno opposto. Per questo motivo con il primo tipo di mescolatura l'elasticità dell'asfalto aumenta di 1.84 unità, mentre con la seconda mescolatura la media generale diminuisce di 1.84 unità. Quando viene considerato l'effetto interattivo doppio tra questa variabile e il fattore A, ci possono essere delle sostanziali variazioni della media generale della variabile risposta. Una grande variazione si può trovare mescolando il quarto tipo di asfalto. Infatti se venisse mescolato nel primo modo, ci sarebbe un aumento in media di 3.87 unità, mentre se si usasse il secondo tipo di mescolatura, la media generale diminuirebbe di 13.91 unità.
Il più grande aumento della media si ha in corrispondenza della mescolatura di primo tipo, con un impasto contenente la prima tipologia di asfalto. Infatti in questa situazione, la media dell'elasticità aumenterebbe di ben 19.54 unità.

Stima degli effetti della risposta **A*T**

A\T	1	2	3		A_effect
1	0.94	-3.36	2.43		17.58
2	-15.76	-6.51	22.26		-3.23
3	10.16	3.54	-13.7		-9.34
4	4.66	6.34	-11		-5.02
T_effect	9.29	10.49	-19.78		261.49

La variabile T apporta grandi variazioni alla media generale. Infatti con i primi due livelli, l'asfalto diviene più elastico, mentre diviene molto più rigido quando viene utilizzato il terzo livello della temperatura.
L'utilizzo del secondo tipo di asfalto, anche se considerato singolarmente fa diminuire la media di 3.23 unità, riesce quasi ad annullare del tutto l'effetto negativo dell'ultimo livello di temperatura disponibile. Infatti in corrispondenza di questa combinazione, si ha un abbassamento della media generale pari solo a 0.75 unità. Se invece a parità di temperatura, venisse utilizzato il terzo tipo di asfalto, la media generale diminuirebbe di circa 43 unità. La combinazione che permette il massimo incremento prevede l'utilizzo del primo tipo di asfalto e l'uso del primo livello della temperatura. In corrispondenza di queste modalità infatti la media generale aumenta di 27.81 unità.

Stima degli effetti della risposta **C*M**

C\M	1	2		C_effect
1	-2.53	2.53		-8.98
2	2.53	-2.53		8.98
M_effect	1.84	-1.84		261.49

Gli effetti interattivi tra le variabili C e M risultano in modulo tutti uguali poiché la matrice degli effetti è quadrata, avendo entrambe le variabili solo due modalità.
L'utilizzo del cemento e del primo tipo di mescolatura, porta a una diminuzione del valore medio dell'elasticità pari a 9,67 unità.
Se l'impasto non contenente il cemento, fosse mescolato nel primo modo, si avrebbe un buon aumento, pari a 13.35 della media generale.

Stima degli effetti della risposta **C*T**

C\T	1	2	3		C_effect
1	4.31	1.72	-6.03		-8.98
2	-4.31	-1.72	6.03		8.98
T_effect	9.29	10.49	-19.78		261.49

In questo caso, in alcun modo è possibile annullare la variazione proveniente dall'utilizzo della temperatura più alta disponibile. In corrispondenza di questa temperatura, infatti, il minimo decremento che si può avere è di 4.77 unità, ottenuto quando nell'impasto non è presente il cemento.

Combinando le due variabili però si possono avere variazioni anche più forti. Mentre l'assenza del cemento tende ad aumentare la media generale, la sua presenza invece tende ad abbassarla. Per questo motivo, in corrispondenza della terza modalità della temperatura, e della presenza di cemento, si ha un decremento pari a 34.79 unità.
Anche in questo caso, come per l'effetto A*T, il massimo incremento è minore, in modulo, del massimo decremento a causa del fortissimo effetto del terzo livello della variabile T. In questo caso infatti si ha che il massimo incremento si può trovare in corrispondenza della temperatura intermedia e dell'assenza di cemento. In queste condizioni la media generale aumenta di 17.75 unità.

Stima degli effetti della risposta **A*C*M**

M = 1				M = 2		
A\C	1	2		A\C	1	2
1	2.47	-2.47		1	-2.47	2.47
2	-1.29	1.29		2	1.29	-1.29
3	6.13	-6.13		3	-6.13	6.13
4	-7.31	7.31		4	7.31	-7.31

Per una migliore rappresentazione, si è scelto di creare una matrice del parametro A*C condizionatamente alla k-esima modalità della variabile M, con k=1,2.
Per una migliore lettura delle tabelle, vengono ripetute le matrici degli effetti doppi A*C, A*M, C*M.

A\C	1	2		A_effect
1	1.22	-1.22		17.58
2	3.23	-3.23		-3.23
3	-2.68	2.68		-9.34
4	-1.77	1.77		-5.02
C_effect	-8.98	8.98		261.49

A\M	1	2		A_effect
1	0.12	-0.12		17.58
2	-5.15	5.15		-3.23
3	-2.02	2.02		-9.34
4	7.05	-7.05		-5.02
M_effect	1.84	-1.84		261.49

C\M	1	2		C_effect
1	-2.53	2.53		-8.98
2	2.53	-2.53		8.98
M_effect	1.84	-1.84		261.49

Ricordando i vincoli presenti all'inizio di questa analisi, appare ovvio che le matrici ottenute in corrispondenza di M = 1 e M = 2 sono uguali ma con segno opposto.
La variazione più grande della media generale, in modulo, avviene quando è utilizzato il quarto tipo di asfalto. In questo caso infatti l'elasticità aumenta di circa 7 unità se vengono utilizzati contemporaneamente il primo metodo di miscelatura e se è assente il cemento oppure se viene utilizzata la seconda metodologia di miscelatura ma il cemento è presente nell'impasto. La più piccola variazione dell'elasticità, in valore assoluto, si ha invece quando l'asfalto utilizzato è il secondo.

Il più grande incremento della media generale della variabile risposta si ottiene quando viene utilizzato il primo tipo di asfalto, quando l'impasto è lavorato con la seconda tipologia di miscelatura ed è presente il cemento, e quando la temperatura di lavorazione è la prima:

$$\Delta\mu = \hat{A}_1 + \hat{C}_1 + \hat{M}_2 + \hat{T}_1 + \hat{A_1T_1} + \hat{C_1T_1} + \hat{C_1M_2} + \hat{A_1M_2} + \hat{A_1C_1} + \hat{A_1C_1M_2} =$$
$$= 17.58 - 8.98 - 1.84 + 9.29 + 0.94 + 4.31 + 2.53 - 0.12 + 1.22 + 7.31 = 32.24$$

Per terminare l'analisi, inseriamo anche il modello nel caso in cui le variabili esplicative non siano qualitative ma quantitative. In questo caso, per selezionare il miglior modello, è stata utilizzata la procedura *stepwise.*

A questo scopo è stato creato un dataset contenente variabili rappresentanti le interazioni risultate significative nel modello selezionato precedentemente. Il primo passo consiste in un'analisi della correlazione tra le variabili; ci aspettiamo che queste risultino correlate tra di loro.

Pearson Correlation Coefficients, N = 48

	Y	A	C	M	T	AC	AM	AT	CM	CT	ACM
Y Var. Elast.	1.00000	-0.33710	0.36270	-0.07495	-0.48408	-0.03423	-0.33257	-0.61174	0.17437	-0.09789	-0.11162
A Asfalto	-0.33710	1.00000	0.00000	0.00000	0.00000	0.77460	0.77460	0.70711	0.00000	0.00000	0.64450
C Cemento	0.36270	0.00000	1.00000	0.00000	0.00000	0.57735	0.00000	0.00000	0.68825	0.61237	0.48038
M Mescola	-0.07495	0.00000	0.00000	1.00000	0.00000	0.00000	0.57735	0.00000	0.68825	0.00000	0.48038
T Temper.	-0.48408	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.64550	0.00000	0.75000	0.00000
AC	-0.03423	0.77460	0.57735	0.00000	0.00000	1.00000	0.60000	0.54772	0.39736	0.35355	0.83205
AM	-0.33257	0.77460	0.00000	0.57735	0.00000	0.60000	1.00000	0.54772	0.39736	0.00000	0.83205
AT	-0.61174	0.70711	0.00000	0.00000	0.64550	0.54772	0.54772	1.00000	0.00000	0.48412	0.45573
CM	0.17437	0.00000	0.68825	0.68825	0.00000	0.39736	0.39736	0.00000	1.00000	0.42146	0.69798
CT	-0.09789	0.00000	0.61237	0.00000	0.75000	0.35355	0.00000	0.48412	0.42146	1.00000	0.29417
ACM	-0.11162	0.64450	0.48038	0.48038	0.00000	0.83205	0.83205	0.45573	0.69798	0.29417	1.00000

I valori riportati nella matrice in Figura? Confermano quanto ipotizzato precedentemente. La correlazione esistente tra alcune delle variabili del nuovo dataset è dovuta all'inserimento di variabili derivanti dall'interazione tra le componenti semplici. Il primo modello con variabili quantitative che analizziamo considera l'ultimo dataset creato che a sua volta è basato sul miglior modello prodotto considerando le variabili come qualitative.

I gradi di libertà del modello, come ci si aspettava, diminuiscono e diventano 13, quindi i gradi di libertà dell'errore di conseguenza aumentano e arrivano a quota 34. La diminuzione dei gradi di libertà del modello produce una quota di devianza spiegata da ognuno di essi più alta. Infine, l'F-value risulta minore rispetto al caso qualitativo, ma comunque significativo dato l'alto numero di gradi di libertà al denominatore (34). Passando all'R^2, come ci si aspettava, risulta minore rispetto al caso qualitativo in quanto le variabili entrano in modo più rigido nel modello. Possiamo notare che in questo caso Type I e Type III non sono più uguali, ciò è dovuto al fatto che la matrice delle variabili esplicative non è più a blocchi ortogonali. Infatti, sono state aggiunte le variabili rappresentanti le interazioni, ciò ha alterato la struttura del dataset che, essendo il risultato di una procedura sperimentale, produceva per costruzione variabili incorrelate. Type I rappresenta una decomposizione della devianza e associa ad ogni variabile la quota di devianza spiegata da essa quando entra nel modello, tenendo conto delle variabili che sono entrate precedentemente. Quindi, l'ordine in cui le variabili vengono inserite nel modello è importante ai fini della quota di devianza che spetta ad ognuna di esse. Al contrario, Type III non è una decomposizione della devianza, indica per ogni variabile la quota di devianza che essa

The GLM Procedure

Dependent Variable: Y Var. Elast.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	21753.69679	1673.36129	8.02	<.0001
Error	34	7094.35988	208.65764		
Corrected Total	47	28848.05667			

R-Square	Coeff Var	Root MSE	Y Mean
0.754078	5.524072	14.44499	261.4917

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	3	5184.095000	1728.031667	8.28	0.0003
C	1	3794.963333	3794.963333	18.19	0.0002
M	1	162.067500	162.067500	0.78	0.3843
T	2	9400.690417	4700.345208	22.53	<.0001
AC	1	132.313500	132.313500	0.63	0.4314
AM	1	343.682667	343.682667	1.65	0.2080
AT	1	1284.255562	1284.255562	6.15	0.0182
CM	1	307.040833	307.040833	1.47	0.2335
CT	1	855.945312	855.945312	4.10	0.0507
ACM	1	288.642667	288.642667	1.38	0.2477

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	3	2002.645834	667.548611	3.20	0.0355
C	1	143.097503	143.097503	0.69	0.4134
M	1	17.797556	17.797556	0.09	0.7720
T	2	3179.021125	1589.510563	7.62	0.0018
AC	1	390.265350	390.265350	1.87	0.1804
AM	1	105.169067	105.169067	0.50	0.4826
AT	1	1284.255562	1284.255562	6.15	0.0182
CM	1	69.816806	69.816806	0.33	0.5668
CT	1	855.945312	855.945312	4.10	0.0507
ACM	1	288.642667	288.642667	1.38	0.2477

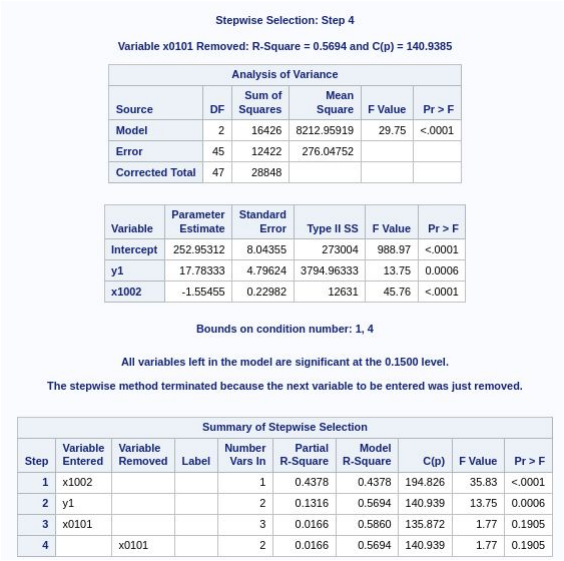
spiegherebbe se fosse inserita per ultima nel modello. In conclusione, nonostante queste differenze, la distribuzione di Type I e Type III non cambia, resta una Fisher con $DF_{Variabile}$, DF_{Errore}. Infine notiamo che alle variabili rappresentanti le interazioni è associato un solo grado di libertà. Come ultima analisi, restando sempre nell'ambito dei modelli con variabili intese come quantitative, utilizziamo una procedura di selezione automatica delle variabili, la procedura stepwise. Questa procedura è stata scelta perché ad ogni step rivaluta tutte le variabili inserite alla luce della significatività che esse assumono nel contesto generale del modello preso in esame in quel particolare step. La strategia scelta consiste nel partire dal miglior modello ottenuto nell'analisi precedente, ossia considerando le variabili come qualitative.

La procedura suggerisce di sfruttare il seguente modello

$$E(y) = \mu + \alpha * cemento + \beta * asfalto * temperatura^2$$

L'F-value risulta elevato e fortemente significativo anche grazie all'elevato numero di gradi di libertà associati all'errore.

Per completezza, inseriamo nel modello quantitativo le variabili ottenute da alcune interazioni che erano state eliminate, perché poco significative, nell'analisi precedente che considerava le variabili come qualitative, allo scopo di individuare eventuali differenze nel modello selezionato dalla procedura stepwise. Si è scelto di inserire le variabili associate all'interazione doppia M*T, ma il modello



Stepwise Selection: Step 4

Variable x0101 Removed: R-Square = 0.5694 and C(p) = 140.9385

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16426	8212.95919	29.75	<.0001
Error	45	12422	276.04752		
Corrected Total	47	28848			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	252.95312	8.04355	273004	988.97	<.0001
y1	17.78333	4.79624	3794.96333	13.75	0.0006
x1002	-1.55455	0.22982	12631	45.76	<.0001

Bounds on condition number: 1, 4

All variables left in the model are significant at the 0.1500 level.

The stepwise method terminated because the next variable to be entered was just removed.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x1002			1	0.4378	0.4378	194.826	35.83	<.0001
2	y1			2	0.1316	0.5694	140.939	13.75	0.0006
3	x0101			3	0.0166	0.5860	135.872	1.77	0.1905
4		x0101		2	0.0166	0.5694	140.939	1.77	0.1905

prodotto risulta essere identico a quello sopracitato suggerendo che la procedura stepwise non riconosca come significative tali variabili.