

Cuarta práctica de REGRESIÓN.

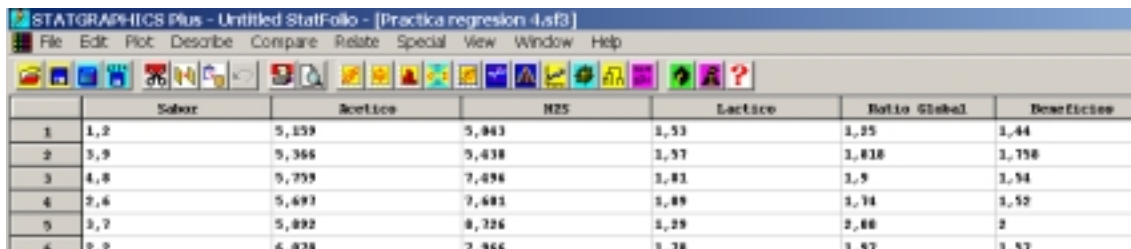
DATOS: fichero “practica regresión 4.sf3”

1. Objetivo:

El objetivo de esta práctica es detectar el problema de Multicolinealidad y ajustar modelos de regresión cuando los datos son colineales.

2. Regresión con datos colineales.

Vamos a utilizar el fichero “practica regresión 4.sf3”. El fichero contiene 8 variables correspondientes a dos ejercicios. El aspecto del fichero es:



	Sabor	Acetico	H2S	Lactico	Ratio Global	Beneficios
1	1,2	5,159	5,843	1,53	1,25	1,44
2	3,9	5,366	5,438	1,57	1,818	1,758
3	4,8	5,759	7,496	1,83	1,9	1,94
4	3,4	5,693	7,683	1,89	1,74	1,52
5	1,7	5,892	8,726	1,29	2,68	2
6	2,2	6,978	7,966	1,78	1,97	1,57

Los dos problemas corresponden a:

Primer ejercicio. Variables SABOR (Y), ACETICO, H2S y LÁCTICO.

Los datos son el resultado de un experimento de cata. Se han probado 24 muestras de quesos y los catadores han valorado el sabor en una escala de 0 a 6. Las variables explicativas son concentraciones de ácido Acético, Láctico y H2S. El objetivo es estudiar de qué depende el sabor para mejorar los quesos.

Segundo ejercicio. Variables RATIO GLOBAL (Y), BENEFICIOS, CLIENTES Y TASA DE INVERSIÓN. Estos datos se utilizarán también en la 5ª Práctica.

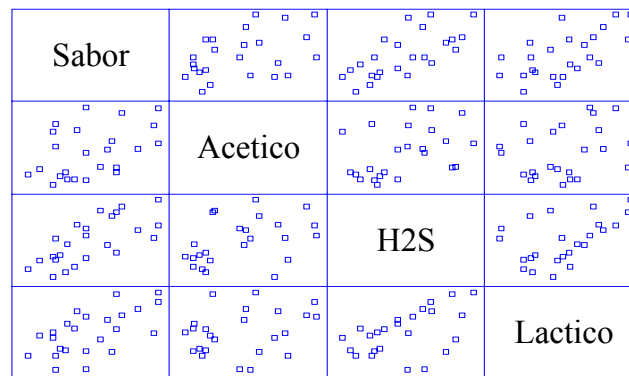
Los datos son los resultados de sucursales de una empresa tras un proceso de fusión. Los nuevos gestores han elaborado unos ratios que miden el desempeño global de cada sucursal (RATIO GLOBAL) en función de un ratio de BENEFICIOS, TASA DE INVERSIÓN, cartera de CLIENTES. El objetivo es estudiar de qué depende básicamente el ratio de desempeño.

Vamos a estudiar el primer problema. El esquema de trabajo será:

- Gráfico de los datos para comprobar las hipótesis
- Ajuste de las regresiones simples para ver si las variables son significativas.
- Ajuste de las regresiones dobles para ver si hay variables colineales.
- Ajuste de la regresión triple.
- Elección del modelo adecuado.

El gráfico de los datos es:

(En la 3ª Práctica se explicó cómo se hace)



Como puede observarse los datos cumplen las hipótesis. Vamos a realizar las tres regresiones simples:

Acético:

Multiple Regression Analysis					

Dependent variable: Sabor					

Parameter	Estimate	Standard Error	T Statistic	P-Value	

CONSTANT	-6,83135	4,27102	-1,59947	0,1240	
Acetico	1,68421	0,746254	2,25689	0,0343	

Analysis of Variance					

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value

Model	11,3351	1	11,3351	5,09	0,0343
Residual	48,9583	22	2,22538		

Total (Corr.)	60,2933	23			

R-squared = 18,7999 percent					
R-squared (adjusted for d.f.) = 15,1089 percent					
Standard Error of Est. = 1,49177					
Mean absolute error = 1,21318					
Durbin-Watson statistic = 1,14762					

Láctico:

Multiple Regression Analysis					

Dependent variable: Sabor					

Parameter	Estimate	Standard Error	T Statistic	P-Value	

CONSTANT	-2,98466	1,60231	-1,86272	0,0759	
Lactico	3,76685	1,03182	3,6507	0,0014	

Analysis of Variance					

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value

Model	22,7461	1	22,7461	13,33	0,0014
Residual	37,5472	22	1,70669		

Total (Corr.)	60,2933	23			
R-squared = 37,7258 percent					
R-squared (adjusted for d.f.) = 34,8951 percent					
Standard Error of Est. = 1,3064					
Mean absolute error = 1,08274					
Durbin-Watson statistic = 1,28444					

H2S:

Multiple Regression Analysis					
Dependent variable: Sabor					
Parameter	Estimate	Standard Error	T Statistic	P-Value	
CONSTANT	-0,867054	0,826938	-1,04851	0,3058	
H2S	0,563027	0,122013	4,61447	0,0001	
Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	29,6546	1	29,6546	21,29	0,0001
Residual	30,6387	22	1,39267		
Total (Corr.)	60,2933	23			
R-squared = 49,1839 percent					
R-squared (adjusted for d.f.) = 46,874 percent					
Standard Error of Est. = 1,18011					
Mean absolute error = 0,968839					
Durbin-Watson statistic = 1,32547					

Se observa que las tres regresiones simples son significativas. Vamos a hacer las regresiones con dos variables.

Acético y H2S:

Multiple Regression Analysis				
Dependent variable: Sabor				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-4,3881	3,43092	-1,27898	0,2149
Acetico	0,682432	0,645467	1,05727	0,3024
H2S	0,505225	0,133405	3,78715	0,0011

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	31,203	2	15,6015	11,26	0,0005
Residual	29,0903	21	1,38525		
Total (Corr.)	60,2933	23			

R-squared = 51,7521 percent
R-squared (adjusted for d.f.) = 47,157 percent
Standard Error of Est. = 1,17697
Mean absolute error = 0,952095
Durbin-Watson statistic = 1,24456

Acético ha dejado de ser significativa. Por tanto es colineal con H2S. Obsérvese que el valor de R-squared adjusted es parecido al de H2S. Esto ocurre porque Acético no aporta prácticamente explicación adicional sobre H2S.

Láctico y Acético:

Multiple Regression Analysis				
Dependent variable: Sabor				
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-8,60132	3,62988	-2,36959	0,0275
Acetico	1,11327	0,651816	1,70795	0,1024
Lactico	3,28445	1,02912	3,19152	0,0044

Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	27,3257	2	13,6628	8,70	0,0018
Residual	32,9677	21	1,56989		
Total (Corr.)	60,2933	23			

R-squared = 45,3212 percent
R-squared (adjusted for d.f.) = 40,1137 percent
Standard Error of Est. = 1,25295
Mean absolute error = 0,987229
Durbin-Watson statistic = 1,18031

De nuevo Acético ha dejado de ser significativa. Por tanto es colineal con Láctico. Obsérvese que el valor de R-squared adjusted es prácticamente igual al de Láctico. Esto ocurre porque Acético no aporta ninguna explicación adicional sobre Láctico.

Láctico y H2S

Multiple Regression Analysis					

Dependent variable: Sabor					

Parameter	Estimate	Standard Error	T Statistic	P-Value	

CONSTANT	-3,3004	1,33742	-2,46774	0,0223	
Lactico	2,18239	0,985687	2,21408	0,0380	
H2S	0,422914	0,129032	3,27758	0,0036	

Analysis of Variance					

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value

Model	35,4531	2	17,7266	14,99	0,0001
Residual	24,8402	21	1,18287		

Total (Corr.)	60,2933	23			

R-squared = 58,8011 percent					
R-squared (adjusted for d.f.) = 54,8774 percent					
Standard Error of Est. = 1,0876					
Mean absolute error = 0,831875					

Láctico y H2S son significativas. Comparten información, ya que su t es mucho menor que en las regresiones simples y además R^2 es claramente menor que la suma de las R^2 de las regresiones simples de Láctico y H2S.

Acético, Láctico y H2S

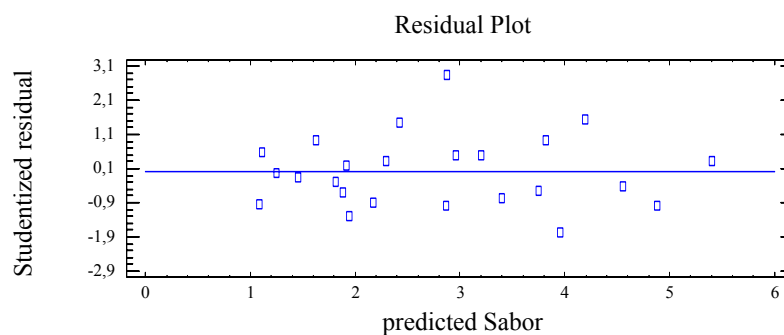
Multiple Regression Analysis					
Dependent variable: Sabor					
Parameter	Estimate	Standard Error	T Statistic	P-Value	
CONSTANT	-6,12043	3,28323	-1,86415	0,0770	
Lactico	2,09606	0,992634	2,11162	0,0475	
H2S	0,380581	0,136983	2,77831	0,0116	
Acetico	0,565218	0,600658	0,940997	0,3579	
Analysis of Variance					
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	36,5063	3	12,1688	10,23	0,0003
Residual	23,787	20	1,18935		
Total (Corr.)	60,2933	23			
R-squared = 60,5478 percent					
R-squared (adjusted for d.f.) = 54,63 percent					
Standard Error of Est. = 1,09057					
Mean absolute error = 0,841667					
Durbin-Watson statistic = 1,32466					

Como era de esperar Acético sigue siendo colineal con Láctico y H2S. El valor de R-squared ajustado es prácticamente el mismo del modelo que contiene únicamente Láctico y H2S.

Primera parte de la práctica: Rellenar la tabla adjunta y determinar qué modelo sería el más adecuado..

Modelo	Acético	Láctico	H2S	R ² R ² Ajustado
1 Coef t	1.78 (2.26)			18.8% 15%
2 Coef t		3.77 (3.65)		37.7% 34.9%
3 Coef t			0.56 (4.61)	49.2% 46.9%
4 Coef t	1.11 (1.71)	3.28 (3.19)		45.3% 40.1%
5 Coef t	0.68 (1.06)		0.51 (3.79)	51.8% 47.2%
6 Coef t		2.18 (2.21)	0.42 (3.27)	58.8% 54.9%
7 Coef t	0.57 (0.94)	2.1 (2.11)	0.38 (2.77)	60.6% 54.6%

El modelo adecuado es aquel que tiene las variables significativas, la diagnosis correcta y el máximo R². Eligiendo el modelo 6, vamos a realizar la diagnosis:
(En la práctica 3 se explica cómo realizar la diagnosis)



El gráfico no muestra estructura por lo que el modelo es adecuado.

3. Segunda parte de la práctica:

Estudiar el efecto sobre el ratio global de los Beneficios, Inversiones y Clientes para las sucursales de la empresa.

Elegir el mejor modelo y realizar la diagnosis.

Modelo	Beneficios	Clientes	Tasa de inversión	R2
1 Coef t				
2 Coef t				
3 Coef t				
4 Coef t				
5 Coef t				
6 Coef t				
7 Coef t				