EJERCICIOS - UD5: INFERENCIA PARTE 3:

ANÁLISIS DE LA VARIANZA

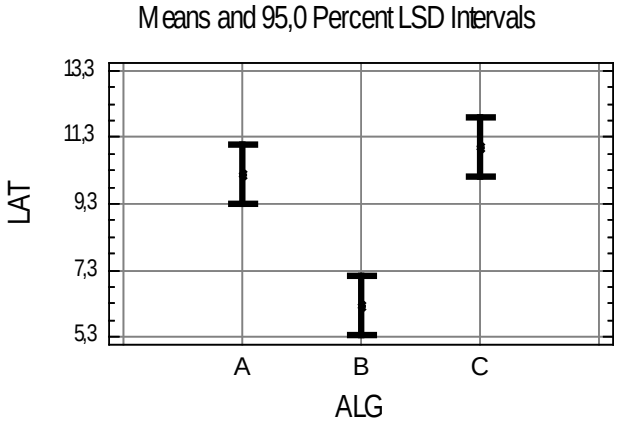
1) Una industria química desea estudiar el efecto del tipo de catalizador y de la concentración de un cierto aditivo denominado NCV en la calidad final de un polímero utilizado en la fabricación de equipos electrónicos. Para ello se ha diseñado un experimento ensayando tres catalizadores diferentes: A, B y C (factor CAT) combinados con tres concentraciones de aditivo: 4, 5 y 6 (factor NCV). Cada uno de los nueve tratamientos se ensayó dos veces, midiéndose en cada prueba un parámetro de calidad final (variable LAT). Tras la realización del experimento y la recogida de datos se llevó a cabo un Análisis de la Varianza cuya tabla resumen se muestra a continuación:

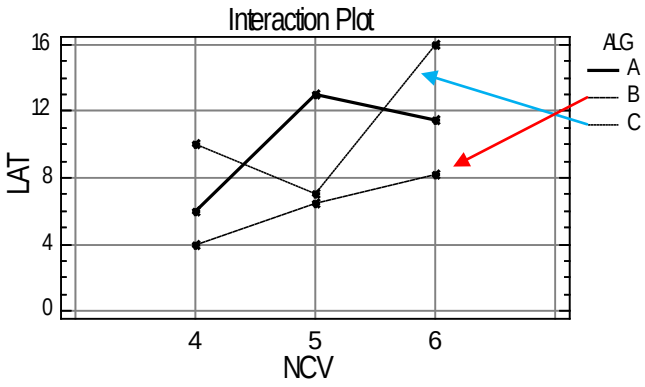
Analysis of Variance for LAT - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:CAT	77,7733	___	_____	_____
B:NCV	_____	___	41,4867	_____
INTERACTIONS				
AB	_____	___	_____	_____
RESIDUAL	16,56	___	_____	
TOTAL (CORRECTED)	250,52	___		

a) Completa la tabla resumen del ANOVA, indicando qué efectos son estadísticamente significativos (α=0,05). Justificar la respuesta, así como los cálculos realizados.

b) ¿Qué información aporta el siguiente gráfico de la izquierda? ¿Dicha información es coherente con las conclusiones del apartado anterior? ¿Por qué?





c) ¿Qué información aporta el gráfico anterior de la derecha? ¿Cuál sería la interpretación del gráfico si la interacción doble no hubiese resultado estadísticamente significativa?

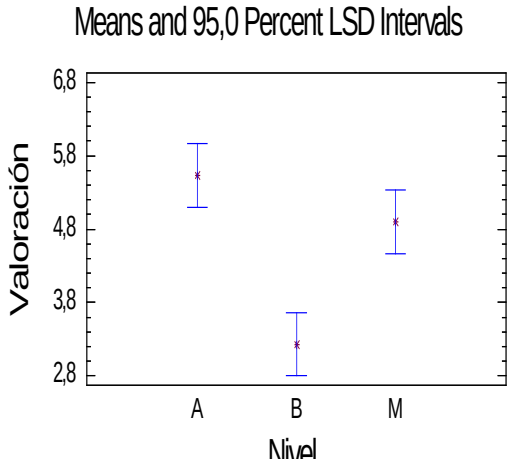
d) ¿Cuál crees que sería el tratamiento óptimo si se desea maximizar la calidad del producto elaborado?

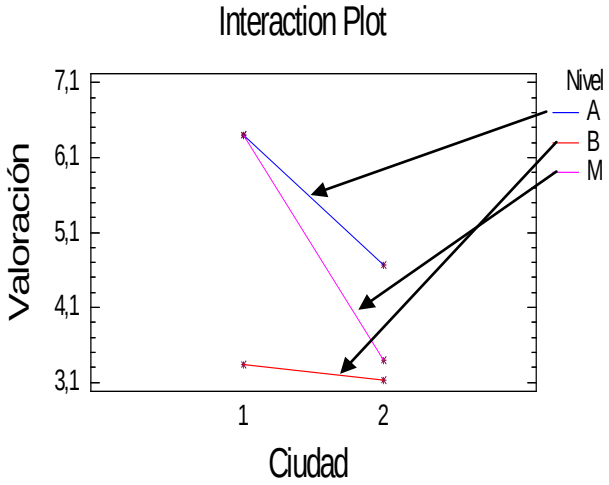
2) Se ha recogido la valoración de un líder político (medida en una escala de 0 a 10) en dos ciudades distintas 1 y 2, dividida cada una en tres barrios según su nivel adquisitivo (Alto, Medio y Bajo). Un ingeniero informático realiza el análisis de los datos resultantes con ANOVA para ver si existe influencia de estos dos factores sobre dicha valoración:

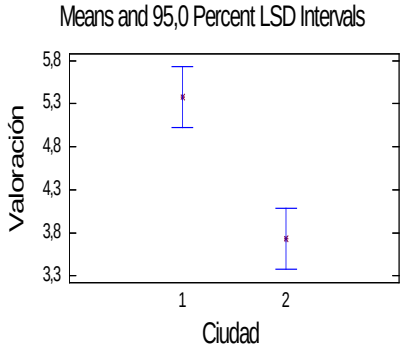
a) Los resultados obtenidos con el Statgraphics han sido los siguientes. ¿Qué conclusiones se deducen?

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Ciudad	60,8444	1	60,8444	21,37	0,0000
B:Nivel	84,6889	2	42,3444	14,87	0,0000
INTERACTIONS					
AB	29,4889	2	14,7444	5,18	0,0076
RESIDUAL	239,2	84	2,84762		
TOTAL (CORRECTED)	414,222	89			

b) A la vista de las gráficas siguientes y, teniendo en cuenta las conclusiones obtenidas en el apartado anterior, determinar cuál es la ciudad en la que se tiene una valoración más alta del líder político y qué nivel adquisitivo tienen, justificando en qué gráficas has obtenido dichas conclusiones.







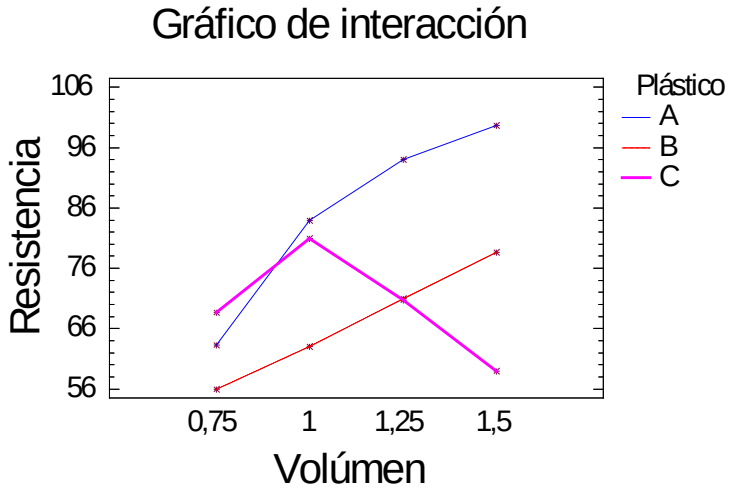
3) En una fábrica de botellas de plástico se quiere decidir qué producto resulta más resistente. Para ello se supone que tanto el tipo de plástico utilizado como materia prima, como el volumen de las botellas, pueden afectar a dicha resistencia. Se estudiaron tres tipos de plástico (*A;B;C*) y 4 volúmenes diferentes (0,75; 1; 1,25 y 1,5), midiendo la resistencia de 3 botellas elegidas al azar para cada posible combinación de tipo de plástico y volumen (se analizaron 36 botellas en total).

a) Completar la tabla del ANOVA de dos factores que resultó de dicho experimento.

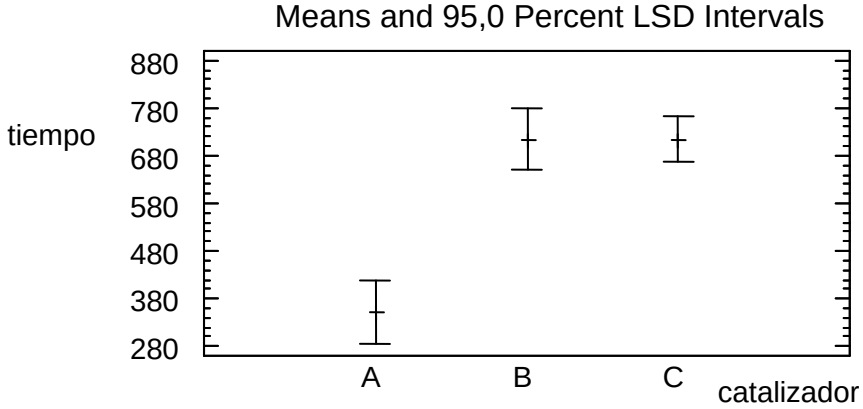
Fuente	SC	Gl	CM	F_c
Plástico				
Volúmen	1613,64			
Plástico x Volúmen	2284,61			
Residual	639,33			
Total	6824,75			

b) En vista de la tabla anterior, ¿qué se puede decir de la significación de los efectos de los factores estudiados? ¿Qué significado tiene en este caso concreto de estudio la interacción? Tomar $\alpha = 0.05$.

c) A partir del gráfico de interacción mostrado a continuación, ¿qué combinación de tipo de plástico y volumen produce una botella más resistente? Si por motivos económicos el único tipo de plástico que se puede utilizar es el *C*, ¿qué volumen de botella produce una mayor resistencia?



4) Una empresa petroquímica puede utilizar tres tipos de catalizadores (A, B y C) en el transcurso de una cierta reacción química en la cual se elabora un polímero usado en la fabricación de CDs. Para determinar cuál de los tres catalizadores es el más efectivo, se ha realizado un experimento consistente en 12 pruebas. En tres de ellas se utiliza el catalizador A, en otras tres reacciones el catalizador B y en otras seis, el C. Los resultados obtenidos, medidos en milisegundos (variable "tiempo") son los indicados en la tabla. Los datos se analizan con ANOVA utilizando Statgraphics, obteniéndose el gráfico que se muestra a continuación.



catalizador A			catalizador B			catalizador C		
373	365	312	739	711	695	615 648	844 809	711 663

Indica cuál de las siguientes afirmaciones es correcta, justificando convenientemente la respuesta.

a) Teniendo en cuenta que $\overline{x}_A = 350$, $\overline{x}_B = \overline{x}_C = 750$, ¿qué catalizador es el menos eficaz?

a.1) El de tipo B, ya que la longitud de su intervalo LSD es mayor que la de C lo cual sugiere que tiene una mayor probabilidad de que se alcancen valores mayores de tiempo.
a.2) El de tipo C, ya que la longitud de su intervalo LSD es menor que el de B lo cual sugiere que su desviación típica es menor.
a.3) El de tipo B o C.
a.4) Cualquiera de los tres, ya que se acepta la hipótesis nula H_0: $m_A = m_B = m_C$.

b) Una de las hipótesis del ANOVA es que la población de datos de la variable tiempo se ajusta a un modelo Normal en cada uno de los tres catalizadores ensayados. ¿Cómo se podría verificar si esta hipótesis es admisible?

b.1) La hipótesis de normalidad es admisible dado que los intervalos LSD son simétricos.
b.2) Habría que calcular los residuos del ANOVA y estudiar si éstos se ajustan bien a un modelo Normal.
b.3) No hay suficientes datos para estudiar si el modelo Normal es admisible.
b.4) No es cierto que el ANOVA asuma una distribución Normal de los datos.

c) ¿Cuál de las siguientes afirmaciones es correcta?

c.1) A la vista de la gráfica se deduce que el p-valor del test del ANOVA es superior a 0,05.
c.2) A la vista de la gráfica se deduce que el p-valor del test del ANOVA es inferior a 0,05.
c.3) A partir de la gráfica no es posible deducir ninguna de las dos respuestas anteriores.
c.4) Depende del nivel de significación del test, el cual no se puede deducir del gráfico.

5) Cierto antibiótico se fabrica por medio de un proceso de fermentación. La temperatura habitual del proceso es 35ºC y el pH es 7, pero los técnicos especulan que posiblemente una temperatura de 30ºC y un pH de 8 podrían incrementar el rendimiento del proceso, lo cual es de gran interés. Para estudiar esta hipótesis, se realiza un diseño de experimentos con dos factores (temperatura y pH) a dos niveles con tres repeticiones. Los resultados obtenidos de rendimiento (medidos en mg/l) son los siguientes:

	pH=7			pH=8			
Temperatura 30ºC	194	186	174	190	189	194	$\overline{x}_{30}$=187,83
Temperatura 35ºC	173	179	166	182	172	177	$\overline{x}_{35}$=174,83
	$\overline{x}_{pH7}$ =178,67			$\overline{x}_{pH8}$ =184			

Analysis of Variance for RENDIMIENTO - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:Temperatura	507,0	1	[illegible]	[illegible]	0,0059
B:pH	[illegible]	1	85,3333	2,17	0,1750
RESIDUAL	354,333	[illegible]	39,3704		
TOTAL (CORRECTED)	946,667	11			

La tabla de resultados del ANOVA es la mostrada anteriormente. En esta tabla, se han ocultado 4 valores. No se incluye la interacción porque no es estadísticamente significativa (p-valor=0.8).

a) Calcular el valor de la F-ratio asociada al factor temperatura.

b) Dado que el p-valor asociado a pH es mayor de 0,05 se puede considerar que el pH no ejerce un efecto estadísticamente significativo en el rendimiento (asumiendo α=0,05). Justificar cómo se puede llegar a la misma conclusión a partir de los datos de la tabla si no se conociese el p-valor.

c) Teniendo en cuenta los resultados del ANOVA y considerando un nivel de significación del 5%, ¿qué temperatura y pH deberían emplearse para maximizar el rendimiento del proceso?

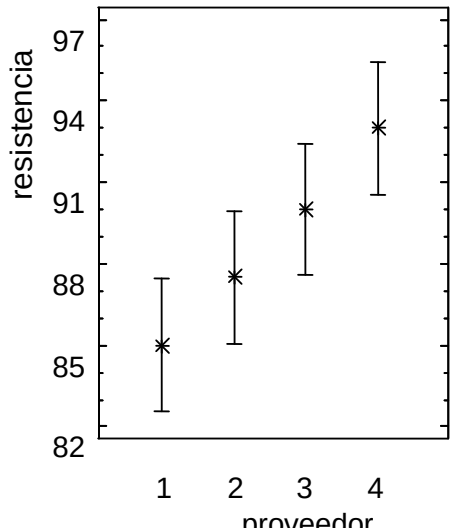
6) Un supermercado decide cambiar las bolsas de plástico que ofrece a sus clientes. Por este motivo, contacta con 4 proveedores distintos (proveedor "1", "2", "3" y "4") que ofrecen bolsas de plástico de características similares. El supermercado pretende contratar el proveedor cuyas bolsas sean las más resistentes. Para ello se seleccionan al azar 5 bolsas de cada proveedor y se realiza un ensayo con cada bolsa para determinar su resistencia. Con los datos obtenidos se realiza un ANOVA cuya tabla de resultados es la siguiente (en la cual se han ocultado 5 valores):

Analysis of Variance for RESISTENCIA - Type III Sums of Squares

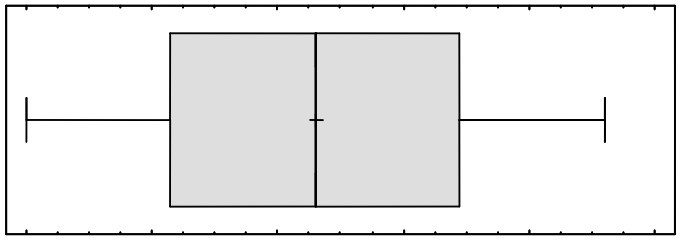
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:proveedor	175,938	[redacted]	58,6458	[redacted]	[redacted]
RESIDUAL	211,6	16	[redacted]		
TOTAL (CORRECTED)	387,537	[redacted]			

a) Teniendo en cuenta que la hipótesis nula es que la media poblacional de la resistencia de las bolsas es la misma para los 4 proveedores, realizar los cálculos necesarios para determinar si la conclusión del ANOVA es aceptar o rechazar la hipótesis nula, considerando un nivel de significación del 5%.

b) El gráfico de medias con intervalos LSD (obtenidos con un nivel de confianza del 95%) se muestra en la figura de la derecha. ¿Qué proveedor debería contratar la empresa teniendo en cuenta que el objetivo es utilizar las bolsas con máxima resistencia?



c) El diagrama Box-Whisker mostrado a continuación se ha obtenido con los residuos del ANOVA. Teniendo en cuenta que los residuos se calculan como diferencia entre cada valor de resistencia y su correspondiente media muestral, indicar si las siguientes afirmaciones son verdaderas o falsas, justificando convenientemente la respuesta:



c.1) La distribución de los residuos es generalmente asimétrica positiva y depende de los grados de libertad del estadístico de contraste, pero en este caso dicha distribución se asemeja razonablemente bien a una Normal.

c.2) En este caso la distribución de los residuos sigue aproximadamente un modelo Normal, lo cual sugiere que los valores de resistencia también siguen una distribución Normal.

c.3) Si el factor proveedor no hubiese resultado estadísticamente significativo, la varianza residual sería similar a la varianza calculada con los 20 valores de resistencia.

7) Una explotación agrícola dispone de 4 tipos de abono (A, B, C y D) que puede aplicar a un mismo cultivo. Para determinar cuál es el abono que más interesa, se escogen al azar 24 parcelas del mismo cultivo (4 bloques de 6 parcelas) y se aplica a cada bloque un tipo de abono distinto a la misma dosis, obteniéndose los rendimientos de cultivo (kg/ha) que se indican en la siguiente tabla. La suma de cuadrados del factor abono vale 3112,5 y el cuadrado medio residual vale 29,4. Realizar los cálculos necesarios para determinar cuál de las siguientes afirmaciones es la única verdadera considerando α=0,05, siendo m_A, m_B, m_C y m_D el rendimiento medio a nivel poblacional obtenido con los abonos de tipo A, B, C y D, respectivamente. Asumir que los datos siguen una distribución Normal, que no hay datos anómalos y que $\sigma^2_A=\sigma^2_B=\sigma^2_C=\sigma^2_D$.

Abono A			Abono B			Abono C			Abono D		
107	101	99	113	121	123	121	120	127	124	126	131
102	93	98	120	113	130	132	125	125	128	130	141
$\bar{x}_A=100$			$\bar{x}_B=120$			$\bar{x}_C=125$			$\bar{x}_D=130$		

a) $m_A < m_B < m_C < m_D$
b) $m_A < m_B < m_D$ y además $m_A < m_C$, pero m_C no difiere significativamente de m_B ni de m_D.
c) $m_A < (m_B = m_C = m_D)$
d) $m_A < (m_C = m_D)$, pero m_B no difiere significativamente de m_A ni de m_C ni de m_D.
e) Se acepta la hipótesis nula H_0: $m_A = m_B = m_C = m_D$.

8) En un proceso de fermentación se elabora un cierto antibiótico. La temperatura de fermentación habitual es de 40ºC y el pH es de 7, pero los técnicos sospechan que posiblemente una temperatura de 35ºC y un pH de 8 podrían aumentar el rendimiento del proceso, lo cual tiene gran interés. Para estudiar esta cuestión, se lleva a cabo un diseño de experimentos con dos factores (temperatura y pH) a dos niveles, con tres repeticiones. Los resultados obtenidos del rendimiento (medido en mg/l) son los indicados en la siguiente tabla. Teniendo en cuenta que la suma de cuadrados de la interacción entre los dos factores vale 176,33, SC_{pH}=85,33 y SC_{total}= 556,67, determinar las condiciones operativas óptimas con las cuales se recomienda fabricar para conseguir en promedio un mayor rendimiento (considerar un nivel de significación del 5%).

	pH=7			pH=8			
Temperatura 35ºC	183	177	174	190	189	194	$\bar{x}_{35}=184,5$
Temperatura 40ºC	173	179	182	172	177	178	$\bar{x}_{40}=176,83$

Indicar cuál de las siguientes respuestas es la única correcta:

a) Se recomienda fabricar con temperatura 35º, da igual emplear pH 7 o pH 8.
b) Se recomienda fabricar con temperatura 35º y pH 8.
c) Se recomienda fabricar con pH 8, da igual emplear temperatura 35º o 40º.
d) Se recomienda fabricar con temperatura 35º y pH 7.
e) Da igual utilizar temperatura 35º o 40º, y también da igual emplear pH 7 o pH 8.

9) Una empresa farmacéutica desea mejorar el rendimiento (Y) de un proceso de fermentación en el cual se elabora un cierto antibiótico. El diseño de experimentos elegido es el indicado a continuación, que permite estudiar 7 factores a dos niveles. Los factores son los siguientes. A: temperatura del circuito de refrigeración; B: temperatura dentro del tanque de fermentación; C: presión en el tanque; D: pH; E: concentración de nitrógeno; F: concentración de calcio; G: tipo de materia prima. Los resultados obtenidos y el diseño utilizado se indican en la siguiente tabla:

A	B	C	D	E	F	G	Y
1	1	1	1	1	1	1	11
1	1	1	2	2	2	2	19
1	2	2	1	1	2	2	23
1	2	2	2	2	1	1	17
2	1	2	1	2	1	2	18
2	1	2	2	1	2	1	18
2	2	1	1	2	2	1	12
2	2	1	2	1	1	2	21

Un ingeniero que no es experto en diseño de experimentos ha estudiado solamente los factores C y D, habiéndose obtenido la siguiente tabla del ANOVA:

Analysis of Variance for Y - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
C	21,125	1	21,125		
D	15,125	1	15,125		
INTERACTIONS					
CD	[illegible]	1	[illegible]	[illegible]	0,0145
RESIDUAL	[illegible]	4	[illegible]		
TOTAL (CORRECTED)	117,875	7			

a) Calcular la suma de cuadrados de la interacción C·D.

b) Indicar cuáles son las tres hipótesis generales cuando se analizan los datos de un diseño de experimentos.

c) En este diseño experimental, la interacción C·D está confundida con el factor G. Teniendo en cuenta este hecho y a la vista de resultados de la tabla del ANOVA, ¿qué conclusión se puede deducir considerando que el objetivo es maximizar el rendimiento?

EJERCICIOS - UD5: INFERENCIA PARTE 4:

REGRESIÓN

10) Para estudiar el efecto de la concentración de catalizador en el rendimiento de una reacción química, se ha realizado un experimento ensayando tres tipos de concentración: 20 mg/l, 30 mg/l y 40 mg/l. Cada una de las concentraciones se ha ensayado 4 veces. El efecto de la concentración sobre el rendimiento obtenido se analiza con regresión lineal múltiple, obteniéndose los siguientes resultados:

```
Multiple Regression Analysis
----------------------------------------------------------------------------
Dependent variable: rendimiento
----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
----------------------------------------------------------------------------
CONSTANT                -258,333       125,178       -2,06372         0,0691
conc                       73,0        11,1942        6,52125         0,0001
conc^2                  -1,21667       0,211749        -5,7458         0,0003
----------------------------------------------------------------------------

                          Analysis of Variance
----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
----------------------------------------------------------------------------
Model                    299756,0      2     149878,0      29,40       0,0001
Residual                  45876,0      9      5097,33
----------------------------------------------------------------------------
Total (Corr.)            345632,0     11

R-squared = 86,7269 percent
R-squared (adjusted for d.f.) = 83,7774 percent
Standard Error of Est. = 71,3956
Mean absolute error = 47,5
```

Teniendo en cuenta los resultados del modelo, ¿qué concentración de catalizador debería utilizarse para maximizar el rendimiento de la reacción, bajo las condiciones del experimento? (elegir la respuesta correcta de entre las siguientes). Considerar α=0,05.

a) conc = 20 b) conc = 30 c) conc = 40
d) conc = 20 o bien conc = 40 e) ninguna de las anteriores

11) La materia activa de un determinado medicamento se obtiene por fermentación con microorganismos modificados genéticamente. La concentración de la materia al terminar la fermentación (mg/l) es un índice del rendimiento del proceso. Con el objetivo de determinar qué variables son las que afectan al rendimiento, se recopila información de 30 lotes de fermentación obtenidos en el último mes. De cada uno de ellos se dispone de los siguientes datos: temperatura media (variable "temperatura" medida en ºC), pH medio (variable "pH"), concentración inicial de azúcares (variable "azúcar") y concentración inicial de proteínas (variable "proteína"), ambas medidas en gramos/litro. Con estos datos se realiza un análisis de regresión lineal múltiple, cuyos resultados se muestran a continuación.

A la vista de estos resultados, responder a las siguientes preguntas:

a) Escribir la ecuación del modelo que se debería utilizar para predecir el rendimiento obtenido al finalizar la fermentación en función de las variables que ejercen un efecto estadísticamente significativo. Nota: justificar convenientemente cuáles son las variables con un efecto estadísticamente significativo, considerando un riesgo de primera especie del 5%.

b) Interpretar qué significado práctico tiene el valor 156,827 y 2,73502 que aparecen en la columna *Estimate*.

c) Calcular el rendimiento medio esperado cuando temperatura=26, pH=7.6, azúcar=23 y proteína=7. En dichas condiciones, ¿cuál es la probabilidad de obtener un rendimiento inferior a 90?

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: rendimiento
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 156,827       36,5803        4,28722         0,0002
temperatura              2,73502       0,709492       3,85489         0,0007
pH                      -27,1323        4,14406      -6,54728         0,0000
azúcar                   1,91988       0,212637       9,02891         0,0000
proteína                 3,22501        0,76551        4,2129         0,0003
-----------------------------------------------------------------------------

                         Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    5372,01      4       1343,0      44,64       0,0000
Residual                 752,181     25      30,0873
-----------------------------------------------------------------------------
Total (Corr.)             6124,2     29

R-squared = 87,7179 percent
R-squared (adjusted for d.f.) = 85,7527 percent
Standard Error of Est. = 5,48519
Mean absolute error = 4,13418
Durbin-Watson statistic = 1,68767 (P=0,2086)
Lag 1 residual autocorrelation = 0,140306
```

12) En un determinado proceso químico se elabora un cierto producto líquido. La viscosidad resultante es el principal parámetro de calidad. Los técnicos sospechan que la viscosidad puede depender de la temperatura de reacción y de la cantidad de catalizador. Para estudiar esta hipótesis, se toman los datos de viscosidad, temperatura y cantidad de catalizador correspondientes a 50 lotes del producto y se realiza un análisis de regresión lineal múltiple cuyos resultados se muestran a continuación.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: viscosidad
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                -24,8334       19,7837       -1,25525         0,2156
temperat                 3,32293       0,306009       10,8589         0,0000
cataliz                0,0272425       0,010429       2,61218         0,0120
-----------------------------------------------------------------------------

                         Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    6029,79      2      3014,89      65,15       0,0000
Residual                 2175,11     47       46,279
-----------------------------------------------------------------------------
Total (Corr.)             8204,9     49

R-squared = 73,4901 percent
R-squared (adjusted for d.f.) = 72,362 percent
Standard Error of Est. = 6,80287
Mean absolute error = 5,24693
Durbin-Watson statistic = 1,81172 (P=0,2549)
Lag 1 residual autocorrelation = 0,0658777
```

a) Calcular el valor del coeficiente de determinación. ¿Cómo se interpreta en la práctica este parámetro?

b) Obtener la ecuación matemática que se recomendaría para predecir la viscosidad en función de las variables que ejercen un efecto estadísticamente significativo (considerar α=0,05).

c) Interpretar el significado práctico del coeficiente asociado a la variable temperatura.

d) Se sospecha que pueda existir un efecto cuadrático de la temperatura. ¿Cómo se puede verificar esta hipótesis? ¿Cuál sería H_0 y H_1 del contraste de hipótesis a plantear?

13) Un cierto polímero se elabora en un determinado proceso químico en continuo. Uno de los índices de calidad de dicho polímero es el parámetro K. Los técnicos desconocen cuáles son las variables del proceso cuyo efecto en dicho parámetro K es estadísticamente significativo (considerando α = 0,05). Para averiguarlo, se toman los datos de calidad correspondientes a 30 días de producción en los cuales las condiciones de proceso (temperatura y presión en el interior del reactor) han variado. La temperatura se mide en ºC y la presión en bars. Con estos datos se realiza un análisis de regresión lineal múltiple cuyos resultados se muestran a continuación.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: param_K
-----------------------------------------------------------------------------
                                    Standard          T
Parameter               Estimate        Error      Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 12,8577      3,01256        4,26802         0,0002
Temperat                0,284645    0,0269679         10,555         0,0000
Presion                  2,72455      0,73552        3,70425         0,0010
-----------------------------------------------------------------------------

R-squared = 81,4591 percent
Standard Error of Est. = 0,912361
Mean absolute error = 0,701048
```

El responsable del control de calidad opina que la velocidad de agitación y el pH (variables velocidad y pH) pueden afectar también al parámetro de calidad. Tras incorporar los datos de estas variables en el modelo, se obtienen los siguientes resultados.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: param_K
-----------------------------------------------------------------------------
                                    Standard          T
Parameter               Estimate        Error      Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 12,0343       10,651        1,12988         0,2693
Temperat                0,286278    0,0272758        10,4957         0,0000
pH                     -0,0226495    0,0194075       -1,16705         0,2542
Presion                  2,65545     0,762342        3,48327         0,0018
Velocidad                0,905893      2,21905       0,408234         0,6866
-----------------------------------------------------------------------------

R-squared = 82,6351 percent
Standard Error of Est. = 0,917591
Mean absolute error = 0,637539
```

a) A la vista de los resultados, el responsable del control de calidad opina que el segundo modelo es más adecuado ya que el coeficiente R^2 es mayor. ¿Estás de acuerdo con este razonamiento?

b) Obtener la ecuación matemática que se debería utilizar para predecir el parámetro K.

c) En la ecuación del primer modelo, interpretar el significado práctico del coeficiente asociado a la variable temperatura.

d) Se considera que el producto es de mala calidad si el parámetro K es superior a 43. Calcular la probabilidad de obtener producto de mala calidad si se ha fabricado en las siguientes condiciones: temperat = 72 ºC, velocidad = 150 rpm, presión = 3,2 bar, pH = 8.

14) Circulando por autopista, la velocidad X (km/h) y el consumo de gasoil Y (litros/100 km) de un cierto modelo de vehículos están correlacionados según el siguiente modelo de distribución normal bivariante, del cual se indica el vector de medias y la matriz de varianzas-covarianzas. El coeficiente de correlación es 0,9. ¿A qué velocidad debería circular el vehículo para que el consumo de gasoil fuera inferior a 7 litros/100 km en el 60% de los casos?

$$(X,\ Y) \approx N\left(m = \begin{Bmatrix} 110 \\ 6 \end{Bmatrix};\ V = \begin{bmatrix} 49 & \text{cov}_{xy} \\ \text{cov}_{xy} & 4 \end{bmatrix} \right)$$

15) En una biblioteca se ha estudiado la relación entre el número de usuarios que utiliza un cierto sistema informático y el tiempo de respuesta (en milisegundos), habiéndose obtenido los siguientes resultados:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: T_RESPUESTA
Independent variable: N_USUARIOS
------------------------------------------------------------------------------
                            Standard          T
Parameter       Estimate      Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        12,7442      4,74559       2,68548         0,0151
Slope           0,345851     0,234634         1,474         0,1578
------------------------------------------------------------------------------

                         Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                    37,6192      1      37,6192       2,17        0,1578
Residual                 311,663     18      17,3146
------------------------------------------------------------------------------
Total (Corr.)            349,282     19

Correlation Coefficient = 0,328184
Standard Error of Est. = 4,16108
Mean absolute error = 3,00625
Durbin-Watson statistic = 1,7033 (P=0,1714)
Lag 1 residual autocorrelation = 0,0734933
```

La matriz de varianzas-covarianzas es la siguiente: $\begin{pmatrix} s_{xx}^2 & s_{xy}^2 \\ s_{yx}^2 & s_{yy}^2 \end{pmatrix} = \begin{pmatrix} 16{,}55 & 5{,}72 \\ 5{,}72 & 18{,}38 \end{pmatrix}$

a) ¿Cuánto vale el coeficiente de determinación? ¿Cuál es la interpretación práctica de dicho parámetro?

b) ¿Qué se entiende por "residuo" en un análisis de regresión? Describe un procedimiento eficiente para estudiar la existencia de residuos anómalos.

c) ¿La correlación entre las dos variables es estadísticamente significativa? (α=0,05)

d) Considerando α=0.05 y teniendo en cuenta que $\bar{x} = 19{,}83$ y que $\bar{y} = 19{,}6$, ¿cuál es la probabilidad de obtener un tiempo de respuesta mayor que 25 milisegundos cuando el número de usuarios en el sistema es de 20?

16) Un estudio ha determinado que la relación entre el número de usuarios y el tiempo de respuesta (en milisegundos) de un sistema informático es la indicada por el siguiente modelo de regresión:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: T_RESPUESTA
Independent variable: N_USUARIOS
------------------------------------------------------------------------------
                            Standard          T
Parameter       Estimate      Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept         13,841      4,32372       3,20117         0,0047
Slope           0,315361     0,204745       1,54027         0,1400
------------------------------------------------------------------------------

                          Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                     56,199      1       56,199       2,37       0,1400
Residual                 450,082     19      23,6885
------------------------------------------------------------------------------
Total (Corr.)            506,281     20
```

La matriz de varianzas-covarianzas es la siguiente: $\begin{pmatrix} s_{xx}^2 & s_{xy}^2 \\ s_{yx}^2 & s_{yy}^2 \end{pmatrix} = \begin{pmatrix} 28.25 & 8.91 \\ 8.91 & 25.31 \end{pmatrix}$

Considerando α=0.05 y teniendo en cuenta que $\bar{x} = 20.47$ y $\bar{y} = 20.3$, ¿cuál es la probabilidad de obtener un tiempo de respuesta superior a 25 ms cuando hay 20 usuarios en el sistema?

17) Cierto antibiótico se fabrica por medio de un proceso de fermentación. La temperatura del proceso habitualmente es de 35ºC y el pH es 7, pero los técnicos opinan que posiblemente una temperatura de 30ºC y un pH de 8 podría incrementar el rendimiento del proceso. Para estudiar dicha cuestión, se lleva a cabo un diseño de experimentos con dos factores (temperatura y pH) a dos niveles con tres repeticiones. Los resultados obtenidos del rendimiento (medidos en mg/l) se indican en la tabla. Se ha ajustado un modelo de regresión lineal múltiple con estos 12 valores, habiéndose obtenido los resultados indicados a continuación:

	pH=7			pH=8		
Temperatura 30ºC	194	186	174	190	189	194
Temperatura 35ºC	173	179	166	182	172	177

```
Multiple Regression Analysis
------------------------------------------------------------------------------
Dependent variable: RENDIMIENTO
------------------------------------------------------------------------------
                                    Standard          T
Parameter               Estimate      Error       Statistic        P-Value
------------------------------------------------------------------------------
CONSTANT                225,833      35,9992        6,27329         0,0001
Temperatura                -2,6     0,724526       -3,58855         0,0059
pH                      5,33333      3,62263        1,47223         0,1750
------------------------------------------------------------------------------

R-squared = 62,5704 percent
R-squared (adjusted for d.f.) = 54,2527 percent
Standard Error of Est. = 6,27458
Mean absolute error = 4,22222
```

Indicar si las siguientes afirmaciones son ciertas o faltas, justificando convenientemente la respuesta:

- a) En este caso es preferible usar ANOVA en lugar de regresión porque ANOVA permite estudiar la interacción entre los dos factores, lo cual no puede estudiarse con regresión.

- b) La tabla de resultados de regresión sugiere que si el pH se incrementa en una unidad, el rendimiento se incrementará en promedio en 5.333 unidades si la temperatura se mantiene constante.

18) La relación que existe entre la potencia de un coche (horsepower) y su consumo (mpg) y su país de procedencia se ha estudiado a través de un modelo de regresión cuyos resultados obtenidos con Statgraphics se muestran a continuación:

```
-----------------------------------------------------------------------------
Dependent variable: horsepower
-----------------------------------------------------------------------------
                                     Standard          T
Parameter               Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  161,85       6,75849        23,9477         0,0000
mpg                      -2,56657      0,196964       -13,0306         0,0000
Pais                      1,43398         2,909       0,492947         0,6228
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   55446,7       2      27723,3     121,33       0,0000
Residual                33589,4     147      228,499
-----------------------------------------------------------------------------
Total (Corr.)           89036,1     149

R-squared = 62,2744 percent
R-squared (adjusted for d.f.) = 61,7611 percent
Standard Error of Est. = 15,1162
Mean absolute error = 11,615
Durbin-Watson statistic = 1,46866 (P=0,0005)
```

a) ¿De qué variables depende de forma significativa para un valor de α=0,05 la potencia? Justifica la respuesta.

b) ¿Cuánto vale el coeficiente de determinación?

19) Para estudiar la relación entre el tamaño de una matriz de datos (variable X) y el tiempo (en ms) que tarda cierto algoritmo en procesar dicha matriz (variable Y), se han obtenido experimentalmente los siguientes valores: (X=2; Y=5.6); (X=3; Y=6.3); (X=4; Y=7.9); (X=5; Y=8.6); (X=6; Y=10.3); (X=7; Y=11.1). Si se ajusta un modelo de regresión lineal simple con estos datos, calcular:

a) Ecuación matemática del modelo de regresión.
b) Coeficiente de correlación.
c) Valor del residuo correspondiente a X=4.

20) Se han medido dos variables, X: tamaño de un fichero de datos (en MB) e Y: tiempo necesario en llevar a cabo cierta operación con el fichero (en ms). De la población de posibles observaciones de X e Y, se ha tomado la siguiente muestra: (X=3; Y=5.7); (X=4; Y=6.4); (X=5; Y=7.8); (X=6; Y=8.9); (X=7; Y=10.1); (X=8; Y=11.4).

a) Obtener la ecuación del modelo de regresión lineal simple que relaciona Y en función de X.
b) Si un fichero tiene un tamaño de 5 MB, ¿cuál es el valor esperado de Y?

c) Si un fichero tiene un tamaño de 5 MB, ¿cuál es la probabilidad de tardar más de 6,5 ms en llevar a cabo la operación?
d) ¿Cómo se podría estudiar si un modelo cuadrático es más apropiado en este caso para predecir la variable respuesta Y?

21) Se han medido dos variables, X: tamaño de un mensaje de texto, e Y: tiempo que tarda el mensaje en llegar a su destino a través de una red informática (en ms). De la población de todos los posibles valores de X e Y, se ha tomado la siguiente muestra: (X=2; Y=5.7); (X=3; Y=6.4); (X=4; Y=7.8); (X=5; Y=8.9); (X=6; Y=10.2); (X=7; Y=11.6).

a) Obtener la ecuación del modelo de regresión lineal que relaciona Y en función de X.
b) Calcular el coeficiente de determinación.
c) Si X=3, calcular el intervalo que contendrá el valor de Y en el 95% de casos.

22) En cierta red informática, se han medido dos variables, X: tamaño de un mensaje de texto, e Y: tiempo que tarda el mensaje en llegar a su destino a través de una red informática (en ms). Se han obtenido los siguientes pares de valores: (X=2; Y=1.4); (X=3; Y=5.3); (X=4; Y=3.7); (X=5; Y=4.6); (X=6; Y=8.4); (X=7; Y=8). La siguiente tabla muestra los resultados de un modelo de regresión lineal que se ha ajustado con los datos utilizando Statgraphics:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Y
Independent variable: X
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept      -0,320952       1,67454      -0,191666         0,8573
Slope            1,23429      0,347907        3,54774         0,0238
------------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                     26,6606      1      26,6606      12,59       0,0238
Residual                  8,47276      4      2,11819
------------------------------------------------------------------------------
Total (Corr.)             35,1333      5

Correlation Coefficient = 0,871114
R-squared = 75,884 percent
R-squared (adjusted for d.f.) = 69,855 percent
Standard Error of Est. = 1,4554
Mean absolute error = 1,07778
Durbin-Watson statistic = 2,89218 (P=0,0161)
```

a) ¿Cuál es la ecuación de regresión que se debería utilizar para predecir Y?
b) ¿Hay suficiente evidencia para afirmar que la correlación entre X e Y es estadísticamente significativa, considerando un nivel de significación de 0,05?
c) Obtener el residuo correspondiente a la observación X=5.
d) ¿Cuál es la interpretación práctica del valor del coeficiente *R-squared*?

23) Dada la siguiente matriz, ¿podría tratarse de una matriz de varianzas-covarianzas, o hay alguna característica que permita concluir lo contrario?

$$\begin{pmatrix} 9 & 15 \\ 15 & 16 \end{pmatrix}$$

24) Dada una variable aleatoria bidimensional (X; Y), donde X e Y están expresadas en cm, se sabe que la covarianza entre ambas variables es 5. Calcular la covarianza si X e Y se expresan en metros.

25) En una variable aleatoria bidimensional (X; Y) el coeficiente de correlación es: r (X; Y) = 0.9. Calcular:

a) r (-X; Y)
b) r (-X; X)
c) r (3·X; Y)

26) Un modelo de regresión lineal simple (Y = a + b·X) relaciona dos dimensiones (diámetro y longitud) de ciertas piezas metálicas fabricadas por una empresa. Ambas variables X e Y están medidas en mm. El coeficiente de determinación es 0,95.

a) ¿Cuál es la interpretación práctica de los coeficientes 'a' y 'b' ?
b) ¿Cuál es la interpretación práctica del coeficiente de determinación?
c) Si X e Y se expresan en cm, calcular la ecuación resultante del modelo de regresión.

SOLUCIONES - PROBLEMAS DE ANOVA

1a) Tabla resumen del ANOVA:

Analysis of Variance for LAT - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio
MAIN EFFECTS				
A:CAT	77,7733	2	38,8866	21,13
B:NCV	82,9734	2	41,4867	22,55
INTERACTIONS				
AB	73,2133	4	18,303	9,95
RESIDUAL	16,56	9	1,84	
TOTAL (CORRECTED)	250,52	17		

Dado que se han realizado 18 pruebas experimentales, el número de grados de libertad totales será 18 – 1 = 17. Como en los dos factores hay tres niveles, los grados de libertad de cada factor serán 3 – 1 = 2. La interacción doble tendrá 2 · 2 = 4 grados de libertad, y los residuales se obtienen por diferencia: $Df_{res} = 17 – 2 – 2 – 4 = 9$

Cuadrado medio$_{NCV}$ = suma de cuadrados / gr. Lib

$41,4867 = SC / 2 \rightarrow SC = 82,9734$

$SC_{AB} = SC_{total} – SC_{CAT} – SC_{NCV} = 73,2133$

Dividiendo las sumas de cuadrados por los grados de libertad se obtiene el cuadrado medio. Dividiendo el cuadrado medio de un factor entre el cuadrado medio residual se obtiene la F-ratio.

El F-ratio de CAT (21,13) y el F-ratio de NCV (22,5) superan el valor crítico (α=0,05) de una $F_{2;9}$ que vale 4,26. El F-ratio de la interacción (9,95) supera el valor crítico (α=0,05) de una $F_{4;9}$ que vale 3,63. Por tanto, el efecto simple de los dos factores y de la interacción son estadísticamente significativos.

1b) Este gráfico muestra los intervalos LSD (Least Significant Differences) para el factor CAT, obtenidos con un nivel de confianza del 95%. A la vista del gráfico se deduce que el valor medio de la variable LAT es significativamente distinto entre los catalizadores A y el B, así como también entre B y C ya que sus intervalos LSD no se solapan. Sin embargo, no hay diferencias significativas entre A y C porque sus respectivos intervalos se solapan. Por tanto, se deduce que: $m_B < (m_A = m_C)$.

La información deducida del gráfico es coherente con el hecho de que el factor CAT resulta estadísticamente significativo, pues ello indica que al menos uno de los catalizadores tendrá un valor medio significativamente distinto de los demás.

1c) Teniendo en cuenta que la interacción doble es estadísticamente significativa, según se deduce del apartado a), el gráfico de la interacción muestra que el efecto de NCV en la variable LAT depende del tipo de catalizador. Así pues, el efecto es lineal con el catalizador B, y se observa un efecto cuadrático en los otros dos casos. Con el catalizador A el valor máximo obtenido corresponde a NCV=5, mientras que con el catalizador C, se obtiene el mínimo con NCV=5, lo que indica que el efecto cuadrático es distinto en A y en C. Si la interacción doble no hubiese resultado estadísticamente significativa no se podría concluir que el efecto de NCV sobre LAT depende del tipo de catalizador.

1d) Dado que la interacción doble es estadísticamente significativa, el tratamiento óptimo será utilizar catalizador C con NCV=6, ya que el valor medio obtenido en estas condiciones (LAT=16) es el mayor de los 9 tratamientos ensayados.

2.a) Los dos factores y la interacción son significativas (p-valor<0.05), lo que implica que las ciudades y los distintos niveles adquisitivos valoran al político de diferente manera. La interacción significaría que los distintos niveles no valoran igual en las dos ciudades.

2.b) Dado que la interacción es significativa no podemos mirar en los intervalos LSD ya que nos dan los valores del factor sin tener en cuenta la relación con la otra variable. Si miramos en la gráfica "Interaction Plot" vemos que la valoración más alta es en la ciudad 1 con nivel adquisitivo Alto y Medio, que tienen idéntica valoración.

3.a)

Fuente	SC	Gl	CM	F_c
Plástico	2287,17	2	1143,585	42,929
Volúmen	1613,64	3	537,88	20,191
Plástico x Volúmen	2284,61	6	380,7683	14,294
Residual	639,33	24	26´639	
Total	6824,75	35		

3.b)

Para plástico: F_c=42´929 > F tablas $F_{2,24}^{0,05} = 3'40 \Rightarrow \quad P-valor < 0,05 \Rightarrow$ *Efecto significativo*

Para volumen: F_c=20´191 > F tablas $F_{3,24}^{0,05} = 3'01 \Rightarrow \quad P-valor < 0,05 \Rightarrow$ *Efecto significativo*

Para la interac.: F_c=14´294 > F tablas $F_{6,24}^{0,01} = 2,51 \Rightarrow \quad P-valor < 0,05 \Rightarrow$ *Efecto significativo*

El efecto significativo de la interacción puede interpretarse como que el efecto del volumen no es el mismo en los tres tipos de plastic

3.c) Mejor combinación (mayor resistencia): Plástico A y Volumen 1,5 litros. Para el plástico C el volumen que ofrece mayor resistencia es de 1 litro.

4.a) Solución: a.3), ya que al solaparse los intervalos LSD entre los catalizadores B y C, no hay diferencias significativas entre ellos, mientras que el A tiene una media significativamente inferior y por tanto es más eficaz, pues reduce el tiempo de la reacción.

4.b) Los intervalos LSD son siempre simétricos y por tanto no informan de la normalidad de los datos (por tanto, b.1 es falsa). La respuesta correcta es b.2, pues podemos utilizar técnicas para estudiar si los residuos son Normales aunque sólo haya 12 datos.

4.c) En la figura se indica que los intervalos LSD se han construido con un nivel de confianza (1-α) del 95% ("95,0 Percent LSD Interval"), de modo que α=0,05. Dado que no todos los intervalos se solapan, se rechaza la hipótesis nula de igualdad de medias para un nivel de significación de 0,05, de modo que la respuesta correcta es la c.2).

5.a) $$\text{F-ratio} = \frac{CM_{temp}}{CM_{residual}} = \frac{SC_{temp} / gr.lib._{temp}}{CM_{resid}} = \frac{507/1}{39,37} = 12,88$$

5.b) Grados de libertad residuales = $gr.lib._{totales} - gr.lib._{temp} - gr.lib._{pH} = 11 - 1 - 1 = 9$

Si H_0 es cierta, F-ratio asociada a pH sigue una distribución $F_{1;9}$ (un grado de libertad en el numerador y 9 en el denominador que son los grados de libertad residuales). Según tablas, el valor crítico $F_{1;9}^{0,05} = 5{,}12$. Como F-ratio=2,17 es inferior al valor crítico, se acepta H_0.

5.c) Como el factor pH no es estadísticamente significativo (p-valor>0,05), da lo mismo utilizar cualquiera de los dos valores de pH. El efecto de la temperatura resulta significativo (p-valor<0,05), por lo que la media poblacional del rendimiento obtenido a 30ºC será distinto que a 35ºC. Interesará utilizar 30ºC ya que, como se deduce de la tabla, a esta temperatura el rendimiento obtenido es mayor.

6.a) $\text{F-ratio} = \dfrac{CM_{factor}}{CM_{residual}} = \dfrac{CM_{factor}}{SC_{resid} / gr.lib._{resid}} = \dfrac{58{,}6458}{211{,}6/16} = 4{,}43$

Grados de libertad del factor = nº variantes – 1 = 4 – 1 = 3

Si H_0 es cierta, F-ratio sigue una distribución $F_{3;16}$ (3 grados de libertad en el numerador y 16 en el denominador). Según tablas, el valor crítico $F_{3;16}^{0,05} = 3{,}24$. Como F-ratio es superior al valor crítico (4,43 > 3,24) se **rechaza** H_0.

6.b) Debería contratar el proveedor 4 o bien el proveedor 3, pues dado que sus intervalos LSD se solapan, no hay evidencia suficiente para afirmar que la media poblacional del proveedor 4 sea superior a la del 3.

6.c1) Falso, ya que la distribución de los residuos depende de la distribución de la variable de partida, que frecuentemente se ajusta a un modelo Normal.

6.c2) Verdadero, ya que los residuos se calculan como diferencia entre cada valor y su media muestral, por lo que si su distribución es normal, implica que la variable de partida también lo es.

6.c3) Verdadero, ya que si el factor no es significativo, las medias muestrales de cada proveedor serán similares a la media de los 20 datos, de modo que la varianza resultará similar.

7) Se calculan los intervalos LSD con la siguiente ecuación, siendo J=6 (ya que las medias de cada tratamiento se han obtenido como promedio de 6 valores) y los grados de libertad residuales son 20 (23 que son los totales menos 3 que son los del factor).

$$\bar{x}_i \pm \frac{\sqrt{2}}{2} t_{gl.resid}^{\alpha/2} \sqrt{\frac{CM_{res}}{J}} \rightarrow \bar{x}_i \pm \frac{\sqrt{2}}{2} 2{,}086 \cdot \sqrt{\frac{29{,}4}{6}} \rightarrow \bar{x}_i \pm 3{,}265$$

Abono A: 100±3,265 → [96,73 ; 103,26] Abono B: 120±3,265 → [116,73 ; 123,26]
Abono C: 125±3,265 → [121,73 ; 128,26] Abono D: 130±3,265 → [126,73 ; 133,26]

Los intervalos LSD de A y D no se solapan, lo que implica que se rechaza la hipótesis nula, por lo que e) no es correcta. De hecho, teniendo en cuenta que SC_{factor}=3112,5 resulta:

F-ratio= $\dfrac{SC_{factor} / g.l_{factor}}{CM_{res}} = \dfrac{3112{,}5/3}{29{,}4} = 35{,}29$ >> ($F_{3;20}^{0,05}$ =3,1) por lo que se rechaza H_0.

Los intervalos B y D no se solapan, por lo que c) es falsa. Los intervalos A y B no se solapan, por lo que d) es falsa. Los intervalos B y C se solapan, por lo que a) es falsa. Por tanto, la respuesta correcta es la **b)**.

8) Grados de libertad de la interacción = $gr.lib_{pH}$ x $gr.lib_{temp}$ = 1·1 = 1

La F-ratio de la interacción es estadísticamente significativa, ya que:

$$F_{ratio} = \frac{SC_{interac} / g.lib_{interac}}{SC_{res} / g.lib_{res}} = \frac{176{,}33/1}{(556{,}67 - 176{,}33 - 85{,}33 - 176{,}33)/(11-1-1-1)} = 11{,}9 > (F_{1;8}^{0{,}05} = 5{,}32)$$

Por esa razón, la condición operativa óptima será pH 8 y temperatura 35 (respuesta **b**) ya que en estas condiciones se obtendrá un rendimiento medio estimado de 191 que es el mayor de los 4 tratamientos.

9.a)

		$Y_{observado}$		$Y_{predicho}$	$(Y_{pred} - Y_{media})^2$	Y_{media}=17.375
C1	D1	11	12	11.5	$2 \cdot (11.5 - 17.375)^2$	
C1	D2	19	21	20	$2 \cdot (20.0 - 17.375)^2$	
C2	D1	23	18	20.5	$2 \cdot (20.5 - 17.375)^2$	Suma = 102.375
C2	D2	17	18	17.5	$2 \cdot (17.5 - 17.375)^2$	

$SC_{C\cdot D} = 102.375 - SC_C - SC_D = 102.375 - 21.125 - 15.125 =$ **66.125**

9.b) Las hipótesis generales que se asumen cuando se analizan los datos de un diseño de experimentos son:
- Hipótesis de normalidad: la variable respuesta sigue una distribución Normal en todos los tratamientos.
- Hipótesis de homocedasticidad: las poblaciones correspondientes a todos los tratamientos ensayados tienen la misma varianza.
- Hipótesis de independencia: las observaciones de cada tratamiento corresponden a individuos extraídos aleatoriamente de la población considerada.

9.c) En este diseño experimental, la interacción C·D está confundida con el factor G. Por tanto, el hecho que C·D sea estadísticamente significativa debe interpretarse como un efecto relevante del factor G. En este caso será más conveniente usar la materia prima de tipo 2 (es decir, el factor G a nivel 2) porque su media resulta superior:

$\overline{Y}_{G1} = (11+17+18+12)/4 = 14.5$; $\overline{Y}_{G2} = (19+23+18+21)/4 = 20.25$

SOLUCIONES - PROBLEMAS DE REGRESIÓN

10) Las dos variables del modelo son estadísticamente significativas ya que su p-valor es menor a 0,05. La ecuación resultante será: Rendimiento = $-258{,}333 + 73\ conc - 1{,}217\ conc^2$
Para determinar el máximo relativo de esta ecuación, hay que derivar e igualar a cero:
$d(rendim)/dc = 73 - 2 \cdot 1{,}217\ conc = 0 \rightarrow conc = 73/(2\cdot 1{,}217) = 30$ g/l. → Respuesta correcta: **b)**

11.a) Las cuatro variables del modelo tienen un efecto estadísticamente significativo, ya que su p-valor es muy inferior a 0.05 (riesgo de primera especie). Por tanto, el modelo de predicción deberá utilizar la información de las cuatro variables. A partir de los valores estimados de los coeficientes que aparecen en la tabla, el modelo será:
Rendimiento = 156,83 + 2,735 · temperatura – 27,13 · pH + 1,92 · azúcar + 3,22 · proteína

11.b) El valor 156,83 es la constante del modelo y se interpreta como el valor medio del rendimiento que cabe esperar si el valor de las cuatro variables del modelo fuese nulo. El valor 2,73502 es el coeficiente asociado a la variable temperatura y se interpreta como el incremento medio de rendimiento que cabe esperar si la temperatura media durante la fermentación se aumentase en 1º C y el resto de variables permanecieran constantes.

11.c) En dichas condiciones, el rendimiento esperado sería:
Rendimiento = 156,83 + 2,735 · 26 – 27,13 · 7,6 + 1,92 · 23 + 3,22 · 7 = 88,45
Desviación típica residual = estándar error of Est. = 5,48
Por tanto, en dichas condiciones, Rendimiento ~ N (88,45 , 5,48)
P(Rend<90) = P[N(88,45 , 5,48)<90] = P[N(0,1) < (90-88,45)/5,48] = P[N(0,1)<0,28] = **0,61**

12.a) Coeficiente de determinación = R-squared = 73,49%. Este parámetro indica que el modelo explica el 73,49% de la varianza de la viscosidad.

12.b) Las variables que ejercen un efecto estadísticamente significativo son temperatura y catalizador, dado que su correspondiente p-valor es menor que 0,05. Por tanto, ambas variables deberán estar en el modelo. La constante no es estadísticamente significativa y podría eliminarse, con lo cual habría que volver a ajustar el modelo para estimar los coeficientes de las variables. Pero dado que no se dispone de esta información, conviene mantener la constante en el modelo, de modo que la ecuación sería: Viscosidad = -24,83 + 3,323·Temperat + 0,02724·cataliz

12.c) Dicho coeficiente vale 3,32. Interpretación: si la temperatura aumenta en un grado centígrado, la viscosidad aumentará en promedio en 3,32 unidades.

12.d) Habría que introducir en el modelo el término cuadrático: temperatura^2. Es decir, ajustar el modelo: viscosidad = a + b·Temp + c·Temp^2 + d·cataliz
El contraste de hipótesis a plantear será: H_0: c=0 H_1: c≠0
Si el p-valor asociado a este contraste es menor que α se rechazará H_0, concluyéndose que el efecto cuadrático es estadísticamente significativo.

13.a) No, ya que el p-valor de las variables velocidad y pH es mayor a 0,05 y por tanto no ejercen un efecto estadísticamente significativo en el parámetro K. Por esta razón, ambas variables deben eliminarse del modelo a pesar de que provocan un ligero aumento del valor de R^2. Debería utilizarse el primer modelo.

13.b) La ecuación sería la del primer modelo: Param_K = 12,86 + 0,285·Temp + 2,72·Presion

13.c) Dicho coeficiente vale 0,285. Este valor se interpreta como el incremento medio del parámetro K que cabe esperar por cada aumento de 1ºC de la temperatura en el reactor si la presión se mantiene constante.

13.d) $\text{media}_{\text{Param\_K}}$ = 12,86 + 0,285·Temp + 2,72·Presion = 12,86+0,285·72+2,72·3,2 = 42,08
desv típica = standard error of est. = 0,912
P[N(42,08; 0,912) > 43] = P[N(0;1) > (43–42,08)/0,912] = P[N(0;1) > 1,01]=**0,156**

14) Llamando v a la velocidad que se pide, resultará:

$$Var(Y/X=v)=\sigma_Y^2\cdot(1-\rho^2)=4\cdot(1-0{,}9^2)=0{,}76$$

$$P\left[(Y/X=v)<7\right]=0{,}6\ ;\ P\left[N(m_v;\sqrt{0{,}76})<7\right]=0{,}6\ ;\ P\left[N(0;1)<(7-m_v)/\sqrt{0{,}76}\right]=0{,}6$$

Buscando en la tabla de la Normal resulta: $(7-m_v)/\sqrt{0{,}76}=0{,}255 \rightarrow m_v=6{,}78$

$$m_v=E(Y/X=v)=\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(v-\mu_X)\ ;\quad 6{,}78=6+0{,}9\frac{\sqrt{4}}{\sqrt{49}}(v-110)\quad \rightarrow v=\mathbf{113{,}03}\ \text{km/h}$$

15.a) El coeficiente de determinación es el cuadrado del coeficiente de correlación, de modo que: $R^2 = r^2 = 0{,}328^2 = 0{,}108 = 10{,}8\%$

15.b) Residuo es la diferencia entre el valor observado de la variable respuesta y el valor predicho por el modelo de regresión. Para estudiar la existencia de residuos anómalos, un procedimiento eficiente consiste en calcular todos los residuos, representarlos sobre un papel probabilístico Normal y ver si los residuos extremos se separan de la recta, lo que indicaría que se apartan de la normalidad. También podría utilizarse un diagrama Box-Whisker.

15.c) No, ya que el p-valor asociado a la pendiente (0,1578) es mayor que el nivel de significación α.

15.d) Dado que la correlación entre las dos variables no es significativa, la distribución condicional de Y cuando X=20 será la distribución marginal de Y: P[y>25 / x=20] = P[y>25] →

$$P(y>25)=P\left[N(19{,}6;\sqrt{18{,}38})>25\right]=P\left[N(0;1)>\frac{25-19{,}6}{\sqrt{18{,}38}}\right]=P\left[N(0;1)>1{,}26\right]=0{,}104$$

16) La correlación entre ambas variables no es estadísticamente significativa porque el p-valor asociado a la pendiente = 0.14 > 0.05. Por tanto, la distribución condicional de Y cuando X=20 será la distribución marginal de Y: P[y>25 / x=20] = P[y>25] →

$$P(y>25)=P\left[N(20.3,\sqrt{25.31})>25\right]=P\left[N(0;1)>\frac{25-20.3}{\sqrt{25.31}}\right]=P\left[N(0;1)>0.934\right]=0.175$$

17.a) Falso, ya que con regresión también es posible estudiar el efecto de la interacción. En este caso las dos técnicas son equivalentes al tratarse de dos factores cuantitativos a dos niveles, y por esa razón se obtendría el mismo p-valor con ambas tanto para temperatura como para pH.

17.b) Falso, pues el factor pH no es significativo, lo cual implica que si el pH aumenta una unidad, el rendimiento no aumentará en promedio.

18.a) Depende del consumo (mpg), ya que su p-valor es menor que 0,05.

18.b) El coeficiente de determinación, o coeficiente R^2 (*R-squared*) vale 62,27% (se indica en la tabla). Se obtiene también a partir de la tabla "analysis of variance " como el cociente de suma de cuadrados residual / suma de cuadrados total.

19.a) Y = 3.13143 + 1.14857·X ; **b)** 0.9937 ; **c)** 0.1743

20.a) Y = 1.9876 + 1.1629·X ; **b)** 7.8019 ; **c)** 0.957

21.a) Y = 3.0333 + 1.2·X ; **b)** 0.9932 ; **c)** [7.041; 6.225]

22.a) Y = -0.3209 + 1.2343·X ; **b)** Sí ; **c)** -1.2505

23) No puede ser matriz de covarianzas: $\text{cov}(x, y) \leq 12$

24) 0.0005

25.a) -0.9 ; **b)** -1; **c)** 0.9

26.c) Y= (a/10) +b·X