EXERCISES UD5 - INFERENCE PART 3:

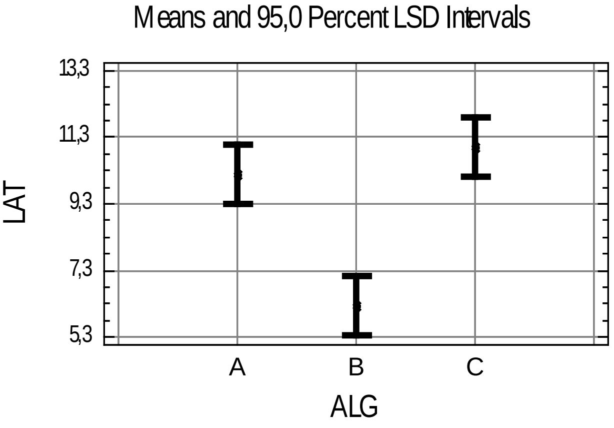
ANALYSIS OF VARIANCE

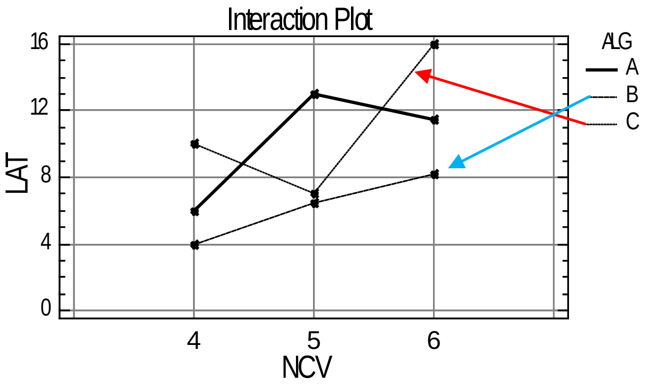
1) One chemical company wants to study the effect of two factors (type of catalyst and concentration of certain additive called NCV) on the final quality of certain polymer used in the manufacturing of electronic devices. For this purpose, an experimental design has been conducted by testing three different catalysts: A, B and C (factor CAT) combined with three concentrations of additive: 4, 5 and 6 mg/l (factor NCV). Each one of the nine treatments was tested twice, and certain quality parameter (variable LAT) was measured in each trial. Analysis of variance was applied to analyze the resulting data. The summary table of ANOVA is the following:

```
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) Complete the summary table of ANOVA and indicate which effects are statistically significant (α=0,05). Justify your answer and all calculations.

b) What information is provided by the following plot on the left? Is that information consistent with the conclusions derived from the ANOVA table? Why?





c) What information is provided by the plot on the right? What would be the interpretation of this plot if the double interaction had not resulted statistically significant?

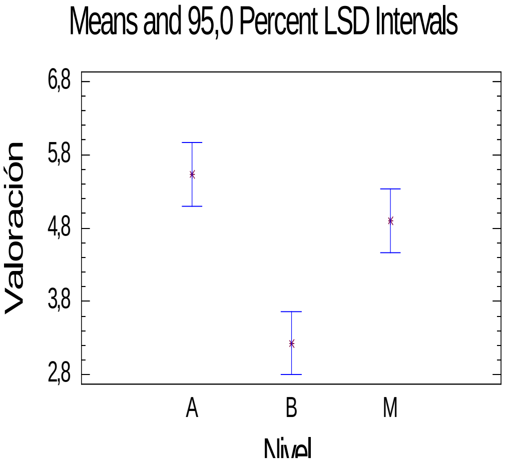
d) What would be the optimum treatment in order to maximize the final quality of the polymer?

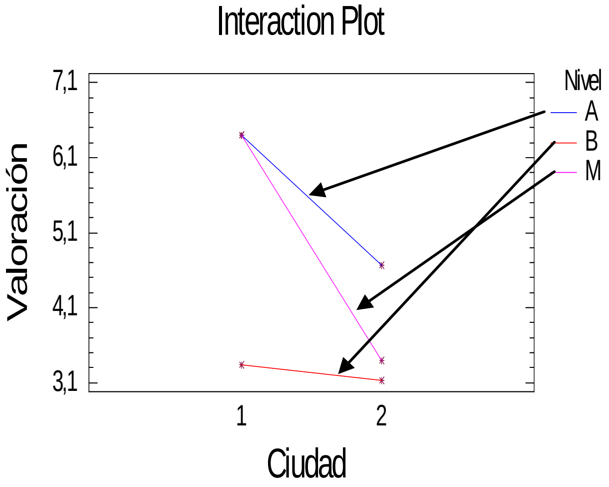
2) One political party has conducted a survey to assess the support for certain leader (measured in a 0-10 scale) in two cities, 1 and 2. Each city is divided in three neighborhoods according to the level of income: high (A), medium (M) and low (B). One computer engineer is requested to conduct the statistical analysis of the resulting data with ANOVA in order to study the effect of both factors on the assessed variable.

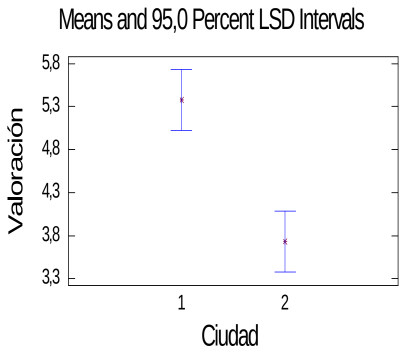
a) Results obtained with Statgraphics are the following. What conclusions can be derived?

```
--------------------------------------------------------------------------------
Source                     Sum of Squares     Df     Mean Square     F-Ratio     P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:City                          60,8444       1         60,8444       21,37      0,0000
 B:Level                         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) According to the following plots and taking into account the conclusions obtained in the previous question, what is the city with highest support for the leader and what is the level of income?







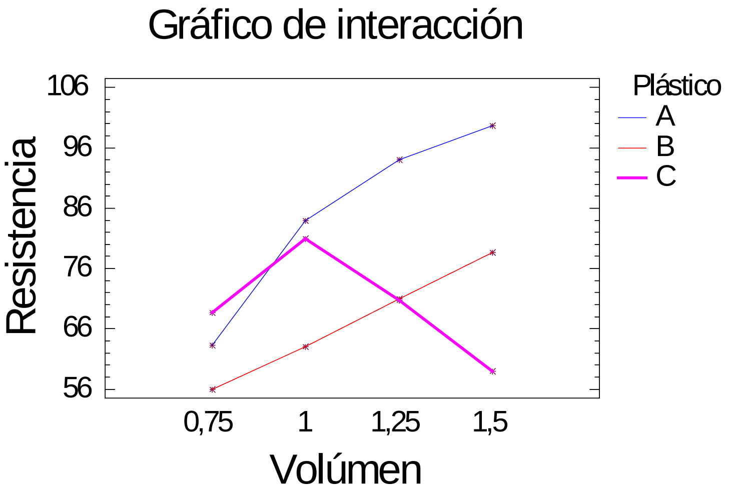
3) One company that manufactures plastic bottles wants to determine what factors lead to the highest resistance. It is assumed that the type of plastic used as raw material and the bottle volume may affect the resistance. Three types of plastic (A, B, C) are studied as well as 4 different volumes (0.75; 1; 1.25 and 1.5 liters). For each combination of plastic and volume, the resistance was measured in 3 bottles randomly chosen (i.e., 36 bottles were analyzed in total).

a) Complete the ANOVA table with the two factors considered in the experiment.

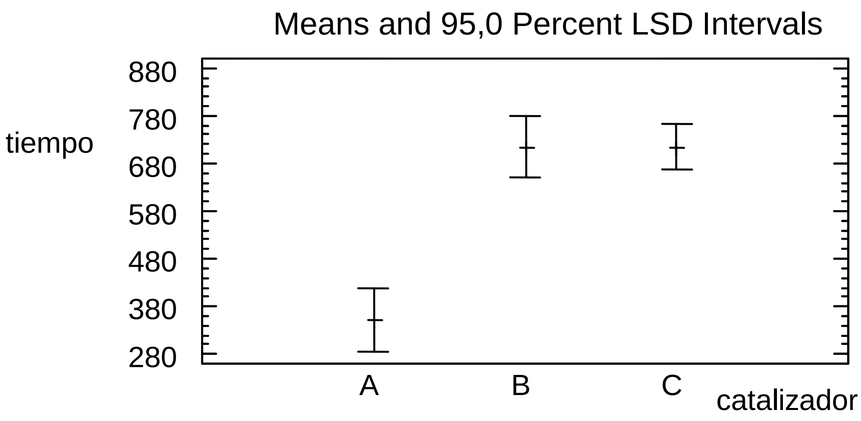
SOURCE	SS	df	MS	F_{ratio}
Plastic				
Volume	1613.64			
Plastic x Volume	2284.61			
Residual	639.33			
Total	6824.75			

b) According to the resulting ANOVA table, what can we say about the statistical significance of the studied effects? What is the meaning in this case of the interaction? Consider α = 0.05.

c) According to the interaction plot shown below, what combination of plastic type and volume leads to bottles with a higher resistance? If the only plastic type that can be used is *C* due to economic reasons, what bottle volume would lead to the highest resistance in this case?



4) One petrochemical company can use three types of catalyst (A, B and C) for certain chemical reaction used in the production of a polymer. In order to determine which catalyst is more efficient, one experiment with 12 trials was conducted. Catalyst A was used in 3 trials, B was used in 3 trials and C was applied in 6 cases. Results obtained, measured in milliseconds (variable "time") are indicated in the table below. Data were analyzed with ANOVA using Statgraphics, and the following plot was obtained:



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

Indicate which of the following statements is correct, justifying conveniently the reply.

a) Taking into account that $\bar{x}_A$ = 350, $\bar{x}_B = \bar{x}_C$ = 750, what catalyst is less efficient?

a.1) Catalyst B, because the LSD interval of B is greater than in the case of C, which suggests that B is more likely of reaching higher values of time.
a.2) Catalyst C, because the LSD interval of C is shorter than in the case of B, which suggests a lower standard deviation.
a.3) Catalyst B or C.
a.4) Any of the 3 catalysts because we have to accept the null hypothesis: $m_A = m_B = m_C$.

b) One hypothesis of ANOVA is that the variable under study follows a Normal distribution in each one of the three tested catalysts. How could we verify if this hypothesis is admissible?

b.1) The hypothesis of normality is admissible because LSD intervals are symmetric.
b.2) We should calculate residuals of ANOVA and study if they follow a Normal distribution.
b.3) There are not enough data to study if the Normal model is admissible.
b.4) It is not true that ANOVA assumes a Normal data distribution.

c) Which of the following statements is the correct one?

c.1) According to the plot it can be deduced that the p-value of the ANOVA test is > 0.05.
c.2) According to the plot it can be deduced that the p-value of the ANOVA test is < 0.05.
c.3) According to the plot it is not possible to deduce any of the previous two statements.
c.4) It depends on the significance level of the test, which cannot be deduced from the plot.

5) Certain antibiotic is manufactured in a fermentation process. The fermentation temperature is usually 35ºC and pH is 7, but the technicians speculate that perhaps a temperature of 30ºC and pH=8 might increase the yield of the process, which is of high interest. In order to study this issue, a design of experiments is carried out with two factors (temperature and pH) at two levels with three replicates. Results obtained of the yield (measured in mg/l) are the following:

	pH=7			pH=8			
Temperature 30ºC	194	186	174	190	189	194	$\bar{x}_{30}$=187.83
Temperature 35ºC	173	179	166	182	172	177	$\bar{x}_{35}$=174.83
	$\bar{x}_{pH7}$ =178.67			$\bar{x}_{pH8}$ =184			

The table of results from ANOVA is the following. In this table, 4 values have been hidden. The interaction is not included because it is not statistically significant (p-value=0.8).

```
Analysis of Variance for YIELD - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                    Sum of Squares     Df    Mean Square     F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:Temperature                     507,0      1     [redacted]  [redacted]     0,0059
 B:pH                         [redacted]      1        85,3333        2,17     0,1750

RESIDUAL                         354,333 [redacted]    39,3704
--------------------------------------------------------------------------------
TOTAL (CORRECTED)                946,667     11
--------------------------------------------------------------------------------
```

a) Calculate the F-ratio associated to factor temperature.

b) Since the p-value associated to pH is higher than 0.05 we can consider that pH does not present a statistically significant effect in the yield (assuming α=0.05). Justify how we could reach the same conclusion from the data in the table if the p-value would be unknown.

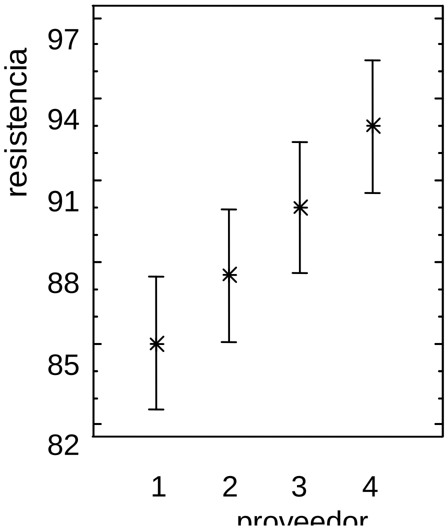
c) Taking into account the ANOVA results and considering a significance level of 5%, what temperature and pH should be used in order to maximize the process yield?

6) One supermarket decides to change the model of plastic bags that offers to its customers. For this reason, it contacts with 4 different suppliers ("1", "2", "3", and "4") which offer plastic bags with similar features. The supermarket wants to determine what plastic bags are the most resistant. For this purpose, 5 bags from each supplier are randomly selected and one test is performed with each bag to determine its resistance. Analysis of variance is applied to study the results. In the following summary table of ANOVA, five values have been hidden.

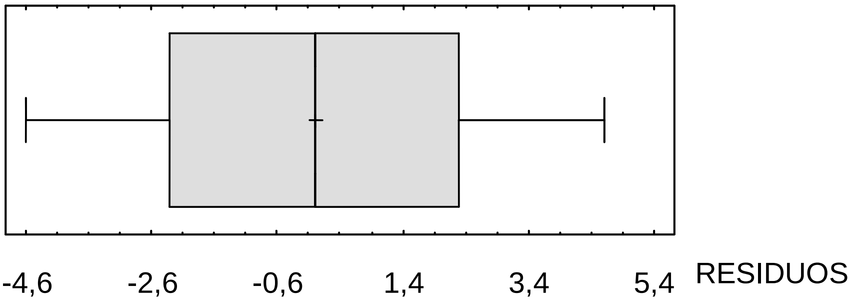
```
Analysis of Variance for RESISTANCE - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:supplier                   175,938     [redacted]      58,6458    [redacted]   [redacted]

RESIDUAL                        211,6     16     [redacted]
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             387,537     [redacted]
--------------------------------------------------------------------------------
```

a) Taking into account the null hypothesis that the population average of the resistance is the same for the four suppliers, conduct the appropriate calculations in order to decide whether the conclusion of the ANOVA test is to accept or reject the null hypothesis, considering a significance level of 5%.

b) The plot on the right shows the averages and LSD intervals (obtained with a confidence level of 95%). Which supplier should the supermarket purchase the plastic bags from, taking into account that the target is to use the bags with highest resistance?



c) The box-whisker diagram shown below has been obtained with the residuals of ANOVA. Taking into account that residuals are calculated as the difference between each observed value of resistance minus the corresponding sample mean, indicate whether the following statements are true or false, justifying conveniently the answer:



c.1) Generally speaking, the distribution of residuals is positively skewed and depends on the degrees of freedom of the statistical parameter to test, but in this case such distribution can be regarded as Normal.

c.2) In this case the distribution of residuals follows approximately a Normal model, which suggests that values of resistance will also follow a Normal distribution.

c.3) If factor supplier had not resulted statistically significant, then the residual variance would have been similar to the variance calculated with the 20 values of resistance.

7) One farm has four types of fertilizer (A, B, C and D) that can be applied to certain crop. In order to determine which fertilizer is the most convenient, 24 fields with the same crop are randomly chosen (4 blocks of 6 fields) and a different type of fertilizer is applied on each block to the same dose, resulting the crop yields (kg/ha) shown in the following table. The sum of squares of factor fertilizer is 3112.5 and the residual mean square is 29.4. Perform the appropriate calculations to determine which of the followings statements is the only one true considering α=0.05, being m_A, m_B, m_C and m_D the average yield at the population level obtained with fertilizers A, B, C and D, respectively. Assume that data follow a Normal distribution, that there are no outliers and that $\sigma^2_A = \sigma^2_B = \sigma^2_C = \sigma^2_D$.

Fertilizer A	Fertilizer B	Fertilizer C	Fertilizer 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
$\overline{x}_A$ =100	$\overline{x}_B$ =120	$\overline{x}_C$ =125	$\overline{x}_D$ =130

a) $m_A < m_B < m_C < m_D$
b) $m_A < m_B < m_D$ and also $m_A < m_C$, but m_C does not differ significantly from m_B nor m_D.
c) $m_A < (m_B = m_C = m_D)$
d) $m_A < (m_C = m_D)$, but m_B does not differ significantly from m_A nor m_C nor m_D.
e) We should accept the null hypothesis H_0: $m_A = m_B = m_C = m_D$.

8) Certain antibiotic is produced by a fermentation process. The usual fermentation temperature is 40ºC and pH is 7, but control engineers suspect that perhaps a temperature of 35ºC and pH=8 might increase the process yield, which is of high interest. In order to study this issue, one experimental design is carried out with two factors (temperature and pH) at two levels, with three replicates. Results obtained of the fermentation yield (measured in mg/l) are indicated below. Taking into account that the sum of squares of the double interaction is 176.33, SS_{pH} = 85.33 and SS_{total} = 556.67, determine which are the recommended optimal operative conditions in order to achieve on average a higher yield (consider a significance level 0.05).

	pH=7			pH=8			
Temperature 35ºC	183	177	174	190	189	194	$\bar{x}_{35}$=184.5
Temperature 40ºC	173	179	182	172	177	178	$\bar{x}_{40}$=176.83

Indicate which one of the following is the true answer:

a) It is recommended to operate with temperature=35ºC, no matter using pH=7 or pH=8.
b) It is recommended to operate with temperature=35ºC and pH=8.
c) It is recommended to operate with pH=8, no matter using temperature=35º or 40ºC.
d) It is recommended to operate with temperature=35º and pH=7.
e) It doesn't matter to use temperature=35º or 40º, and it doesn't matter to use pH=7 or 8.

9) One chemical company wants to improve the final quality parameter (Y) of a fermentation process for the production of antibiotics. The experimental design is shown in the following table, which allows the study of 7 factors at two levels. Factors are the following. A: temperature of the cooling circuit; B: temperature inside the tank; C: pressure in the tank; D: pH; E: concentration of nitrogen; F: concentration of calcium; G: type of raw material. The results obtained and the experimental design are indicated in the table:

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

One engineer who is not an expert in design of experiments has only studied factors C and D and has obtained the following ANOVA table:

```
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                       [redacted]      1     [redacted] [redacted]     0,0145

RESIDUAL                  [redacted]      4     [redacted]
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            117,875      7
--------------------------------------------------------------------------------
```

a) Calculate the sum of squares of the interaction C·D.

b) What are the three general assumptions in the analysis of results obtained in a design of experiments?

c) In this experimental design, interaction C·D is confounded with factor G. Taking into account this fact and the results shown in the ANOVA table, what conclusion can be derived considering that the target is to maximize the quality parameter?

EXERCISES - UD5 INFERENCE PART 4:

REGRESSION

10) In order to study the effect of concentration and type of catalyst in the yield of a chemical reaction, one experiment was performed using three different concentrations: 20 mg/l, 30 mg/l and 40 mg/l. Each concentration was tested four times. The effect of concentration on the reaction yield was analyzed with multiple linear regression, and the following results were obtained:

```
Multiple Regression Analysis
------------------------------------------------------------------------------
Dependent variable: YIELD
------------------------------------------------------------------------------
                                       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
```

Taking into account the results shown above, what catalyst concentration should be used in order to maximize the reaction yield under the conditions of the experiment? Choose the correct answer, considering α=0.05.

a) conc = 20 b) conc = 30 c) conc = 40
d) conc = 20 or conc = 40 e) none of the rest

11) Certain pharmaceutical drug is obtained by fermentation using genetically modified microorganisms. The drug concentration at the end of the fermentation (mg/l) is an index of the process yield. In order to study what factors affect the yield, the company compiles different data about 30 fermentation batches produced in the last month. The following variables are available from each one: mean temperature (variable “temperature” measured in ºC), mean pH (varible “pH”), initial concentration of sugars (variable “sugar”) and initial concentration of proteins (variable “protein”), both measured in gr/liter. Multiple linear regression is applied to analyze the data, and results are shown below. According to the results, answer the following questions:

a) Write the mathematical equation that should be used to predict the yield obtained at the end of the fermentation as a function of those variables that present a statistically significant effect. Justify conveniently which variables present a statistically significant effect, considering a type I risk of 5%.

b) Provide a practical interpretation for the values 156.827 and 2.73502 that appear in the column *Estimate*.

c) Calculate the expected yield when temperature=26, pH=7.6, sugar=23 and protein=7. What is the probability to obtain a yield lower than 90 in such conditions?

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: yield
-----------------------------------------------------------------------------
                                       Standard         T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 156,827        36,5803        4,28722         0,0002
temperature              2,73502       0,709492        3,85489         0,0007
pH                      -27,1323        4,14406       -6,54728         0,0000
sugarr                   1,91988       0,212637        9,02891         0,0000
protein                  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) Certain liquid product is manufactured by a chemical company. Viscosity is the main quality parameter. Process engineers suspect that viscosity may depend on the reaction temperature and the amount of catalyst. In order to explore this hypothesis, one multiple linear regression is performed with the temperature and amount of catalyst corresponding to 50 batches. Results are shown below.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: viscosity
-----------------------------------------------------------------------------
                                       Standard         T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                -24,8334        19,7837       -1,25525         0,2156
temperature              3,32293       0,306009        10,8589         0,0000
catalyst               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) Calculate the coefficient of determination. What is the practical interpretation of this parameter?

b) Obtain the mathematical equation that should be applied to predict the viscosity as a function of those variables exerting a statistically significant effect (consider α=0.05).

c) Provide a practical interpretation for the coefficient associated to temperature.

d) Engineers suspect that temperature may present a quadratic effect. How could we verify this hypothesis? What would be the null hypothesis H_0 and the alternative hypothesis of the test?

13) Certain polymer is elaborated in a continuous chemical process. One quality index of this polymer is parameter K. Engineers are not sure about which are the key process variables that affect the quality parameter. In order to study this issue, the quality parameter is measured during 30 days of production. In this period, the temperature and pressure inside the tank have varied. Temperature is measured in ºC and pressure in bars. One multiple linear regression was performed using these data, and results are shown below. Consider $\alpha = 0.05$.

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

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

The engineer responsible for the quality control considers that the steering speed and pH (variables speed and pH) might also affect the quality parameter. After incorporating both variables in the regression model, the following results were obtained:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: param_K
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 12,0343         10,651        1,12988         0,2693
Temperature             0,286278      0,0272758        10,4957         0,0000
pH                    -0,0226495      0,0194075       -1,16705         0,2542
Pressure                 2,65545       0,762342        3,48327         0,0018
Speed                   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) According to the results, the process engineer concludes that the second model is more convenient than the first one because coefficient R^2 is higher. Do you agree with this reasoning?

b) Obtain the mathematical equation that should be used to predict parameter K.

c) In the first predictive equation, provide a practical interpretation for the coefficient associated to temperature.

d) The product is considered as inappropriate if parameter K is higher than 43. Calculate the probability to obtain such inappropriate polymer if it was obtained under the following conditions: temperature = 72 ºC, speed = 150 rpm, pressure = 3.2 bar, pH = 8.

14) Circulating by highway, the speed X (km/h) and the fuel consumption Y (litres/100km) of certain model of vehicle can be assumed to follow a bivariate normal distribution with a correlation coefficient ρ=0.9. At what speed should circulate the vehicle in order to achieve a fuel consumption lower than 7 litres/100km in 60% of the cases? The matrix of variance-covariances and the vector of means is the following:

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

15) Certain library has studied the relationship between the number of users that use certain computer service and the response time (in milliseconds). The following results were obtained.

```
Regression Analysis - Linear model: Y = a + b*X
----------------------------------------------------------------------------
Dependent variable: T_RESPONSE
Independent variable: N_USERS
----------------------------------------------------------------------------
                               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
```

The matrix of variances-covariances is the following: $\begin{pmatrix} s_{xx}^2 & s_{xy}^2 \\ s_{yx}^2 & s_{yy}^2 \end{pmatrix} = \begin{pmatrix} 16{,}55 & 5{,}72 \\ 5{,}72 & 18{,}38 \end{pmatrix}$

a) Calculate the coefficient of determination. What is the practical interpretation of this parameter?

b) Define the concept of “residual” in regression analysis. Describe an effective procedure to study the presence of abnormal residuals.

c) Is the correlation between both variables statistically significant? (α=0.05).

d) Considering α=0.05 and taking into account that $\bar{x} = 19{,}83$ and $\bar{y} = 19{,}6$ what is the probability to obtain a response time higher than 25 ms if the number of users in the system is 20?

16) One study has established that the relationship between the number of users and the response time (in milliseconds) of a computer system is given by the results shown below. The matrix of variances-covariances is the following: $\begin{pmatrix} \mathrm{s}_{\mathrm{xx}}^2 & \mathrm{s}_{\mathrm{xy}}^2 \\ \mathrm{s}_{\mathrm{yx}}^2 & \mathrm{s}_{\mathrm{yy}}^2 \end{pmatrix} = \begin{pmatrix} 28.25 & 8.91 \\ 8.91 & 25.31 \end{pmatrix}$

Considering α=0.05 and taking into account that $\bar{\mathrm{x}} = 20.47$ and $\bar{\mathrm{y}} = 20.3$, what is the probability to obtain a response time higher than 25 ms when the number of users in the system is 20 ?

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: T_RESPONSE
Independent variable: N_USERS
-----------------------------------------------------------------------------
                               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
```

17) Certain antibiotic is manufactured in a fermentation process. The fermentation temperature is commonly of 35ºC and the pH is 7, but the technicians speculate that perhaps a temperature of 30ºC and a pH of 8 might increase the yield of the process, which is of high interest. In order to study this issue, a design of experiment is carried out with two factors (temperature and pH) at two levels with three replicates. The results obtained of the yield (measured in mg/l) are indicated in the table. A multiple linear regression has been conducted with the 12 values of yield, according to the variables temperature and yield, obtaining the following results.

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

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: YIELD
-----------------------------------------------------------------------------
                                      Standard          T
Parameter                Estimate         Error       Statistic         P-Value
-----------------------------------------------------------------------------
CONSTANT                  225,833        35,9992        6,27329          0,0001
Temperature                  -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
```

Indicate whether the following statements are true or false, justifying conveniently the response:

- a) In this case it is preferable to use ANOVA instead of regression analysis because ANOVA allows studying the interaction between the two factors, which cannot be studied with regression.

- b) The table of results shown above suggests that if pH increases one unit, the yield will increase in average in 5.333 units if the temperature is kept constant.

18) The relationship between the horsepower of certain car model, petrol consumption (mpg) and its country of origin has been studied by means of a regression model. Results obtained with Statgraphics are shown below.

```
-----------------------------------------------------------------------------
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
Country                   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) What variables produce a statistically significant effect (α=0.05) on horsepower? Justify your anwer.

b) Calculate the coefficient of determination.

19) Two variables are measured, X: size of a data matrix, and Y: time required for processing the matrix by certain algorithm (in ms). If a regression line is fitted with the following pairs of X and Y values: (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), calculate:

a) Equation of the regression model
b) Correlation coefficient
c) Residual corresponding to X=4

20) Two variables are measured, X: size of the data file (in MB), and Y: time required to conduct certain operation with that file (in ms). In the population of all possible X and Y values, the following sample has been obtained: (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) Obtain the equation of the linear regression model that relates X and Y.
b) If a file has a size of 5 MB, what is the expected value of Y?
c) If a file has a size of 5 MB, what is the probability to take more than 6.5 ms to conduct the operation?
d) How would you check if a quadratic model is more suitable in this case to predict Y?

21) Two variables are measured, X: size of a text message, and Y: time required by the message to arrive at destination through a network (in ms). In the population of all possible X and Y values, the following sample has been obtained: (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) Obtain the equation of the linear regression model that relates X and Y.
b) Calculate the coefficient of determination.
c) If X=3, calculate the interval that will contain the value of Y in 95% of the cases.

22) In certain computer network, two variables are measured, X: size of a text message, and Y: time required by the message to arrive at destination through the network (in ms). The following pairs of values were obtained: (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). A regression line is fitted with Statgraphics, resulting the following table:

```
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) What is the regression equation that should be used to predict Y?
b) Is there enough evidence to affirm that the correlation between X and Y is statistically significant, considering a significance level of 0.05?
c) Obtain the residual corresponding to the observation X=5.
d) Provide a practical interpretation of the R-squared value.

23) Given the following covariance matrix, is there anything wrong with it? $\begin{pmatrix} 9 & 15 \\ 15 & 16 \end{pmatrix}$

24) In a two-dimensional random variable (X; Y), being X and Y expressed in cm, the covariance between X and Y is 5. Calculate the covariance if X and Y are expressed in m.

25) In a two-dimensional random variable (X; Y), the correlation coefficient is: r (X; Y) = 0.9. Calculate:
a) r (-X; Y)
b) r (-X; X)
c) r (3·X; Y)

26) A simple linear regression model (Y = a+b·X) relates two dimensions (diameter and length) of certain pieces manufactured in a factory. Both X and Y are measured in mm. The coefficient of determination is 0.95.
a) What is the practical interpretation of coefficients ‘a’ and ‘b’ ?
b) What is the practical interpretation of the coefficient of determination?
c) If X and Y are expressed in cm, calculate the new regression equation.

SOLUTIONS - PROBLEMS OF ANOVA

1a) Summary table of 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
-------------------------------------------------------------------
```

Given that 18 experimental trials were carried out, the total degrees of freedom are: 18 – 1 = 17. As both factors have three levels, degrees of freedom for each factor are 3 – 1 = 2. The double interaction will have 2 · 2 = 4 degrees of freedom, and the residual ones are obtained by difference: $Df_{res} = 17 – 2 – 2 – 4 = 9$. Mean $square_{NCV}$ = sum of squares / deg. fr.

$41.4867 = SS / 2 \rightarrow SS_{NCV} = 82.9734$

$SS_{AB} = SS_{total} – SS_{CAT} – SS_{NCV} = 73.2133$

The mean square is obtained dividing the sum of squares by the degrees of freedom. The F-ratio is obtained dividing the mean square of one factor by the residual mean square. The F-ratio of CAT (21.13) and the F-ratio of NCV (22.5) are higher than the critical value (α=0.05) of $F_{2;9}$ which is 4.26. The F-ration of the interaction (9.95) is higher than the critical value (α=0.05) of $F_{4;9}$ which is 3.63. Therefore, the simple effect of both factors and the interaction are statistically significant.

1b) This plot shows the LSD (Least Significant Differences) intervals for factor CAT, obtained with a confidence level of 95%. The plot shows that the average value of LAT is significantly different between catalysts A and B, as well as between B and C because their LSD intervals are not overlapped. However, the differences between A and C are not statistically significant because their respective intervals are overlapped. Thus, it can be concluded that: $m_B < (m_A = m_C)$. The information deduced from this plot is consistent with the fact that factor CAT is statistically significant, which implies that at least one catalyst will present a mean value significantly different from the rest.

1c) Taking into account that the double interaction is statistically significant, as deduced from question a), the interaction plot shows that the effect of NCV on the variable LAT depends on the type of catalyst. Thus, catalyst B presents a linear effect, but a quadratic effect is observed for the other two ones. In the case of catalyst A, the maximum value corresponds to NCV=5, while in the case of C, the minimum value was achieved with NCV=5, which indicates that the quadratic effect is different in A and C. If the double interaction had not resulted statistically significant, we could not conclude that the effect of NCV on LAT depends on the type of catalyst.

1d) Given that the double interaction is statistically significant, the optimum treatment will correspond to catalyst C with NCV=6, because the mean value obtained in these conditions (LAT=16) is the highest one among the 9 tested treatments.

2.a) Both factors and the interaction are statistically significant (p-value<0.05), which implies that the cities and the level of income affect the support for the leader in a different manner. The interaction implies that the different income levels affected the measurements differently in each city.

2.b) As the interaction is significant, we cannot focus only on the LSD intervals because they give us average values of one factor without taking into account the other factor. The interaction plot indicates that the highest ratings were obtained in city 1 with the high and average income level.

3.a)

Source	SS	d.f.	MS	F_c
Plastic	2287.17	2	1143.58	42.929
Volume	1613.64	3	537.88	20.191
Plastic x Volume	2284.61	6	380.77	14.294
Residual	639.33	24	26.64	
Total	6824.75	35		

3.b)

For plastic: F_c=42.929 > ($F_{2;24}^{0.05}$ = 3.4) => the effect is statistically significant
For volume: F_c=20.191 > ($F_{3;24}^{0.05}$ = 3.01) => the effect is statistically significant
For the interaction: F_c=14.294 > ($F_{6;24}^{0.05}$ = 2.51) => the effect is statistically significant

The interaction is statistically significant, which implies that the effect of volume is not the same in the three types of plastic.

3.c) The best combination that produces the highest resistance is: plastic A and volume 1.5 liters. For plastic C, the highest resistance is achieved with a volume of 1 liter.

4.a) Solution: a.3). As LSD intervals of catalysts B and C are overlapped, the difference between them is not statistically significant. The average of A is significantly lower and consequently it is more effective because it reduces the reaction time.

4.b) LSD intervals are always symmetrical and therefore they do not inform about the data normality (answer b.1 is false). The correct answer is b.2 because we can use techniques to study if residuals are Normally distributed although we only have 12 values in this case.

4.c) The plot indicates that LSD intervals were obtained with a confidence level (1-α) of 95% (“95,0 Percent LSD Interval”), which implies α=0.05. As not all intervals are overlapped, we reject the null hypothesis that all means are equal considering a significance level of 0.05. Thus, the correct answer is 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) Residual degrees of freedom = d.f.$_{total}$ – d.f.$_{temp}$ – d.f.$_{pH}$ = 11 – 1 – 1 = 9
If H_0 is true, F-ratio associated to pH follows a distribution $F_{1;9}$ (one degree of freedom in the numerator and 9 in the denominator which are the residual ones). According to the tables, the critical value $F_{1;9}^{0,05} = 5{,}12$. As F-ratio=2.17 is lower than the critical value, the null hypothesis is accepted.

5.c) As factor pH is not statistically significant (p-value>0.05), any value of pH can be used. The effect of temperature is statistically significant (p-value<0.05), which implies that the population mean of the yield obtained at 30ºC will be different to that at 35ºC. It will be of interest to operate at 30ºC because the yield obtained at this temperature is higher according to the table.

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

Degrees of freedom of the factor = number of variants – 1 = 4 – 1 = 3
If H_0 is true, F-ratio will follow a distribution $F_{3;16}$ (3 degrees of freedom in the numerator and 16 in the denominator). According to the F table, the critical value $F_{3;16}^{0,05} = 3{,}24$. As the F-ratio is higher than the critical value (4.43 > 3.24) then the null hypothesis is **rejected**.

6.b) The company should choose supplier 4 or supplier 3 because their LSD intervals are overlapped, which implies that there is not enough evidence to affirm that the population mean of supplier 4 is higher than supplier 3.

6.c1) False, because the distribution of residuals depends on the distribution of the original variable, which frequently follows a Normal model.

6.c2) True, because residuals are calculated as difference between each value and the sample mean. Consequently, if the distribution is Normal, it implies that the original variable will also follow a Normal distribution.

6.c3) True, because if the effect of supplier is not significant, sample means of each supplier will be similar to the mean of all 20 values, and consequently their variance will be similar.

7) LSD intervals are calculated with the following equation, being J=6 (because the means of each treatment are obtained as average of 6 values) and the residual degrees of freedom are 20 (i.e., 23 which are the total d.f. minus 3 corresponding to the factor).

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

Fertilizer A: 100±3.265 → [96.73 ; 103.26] Fertilizer B: 120±3.265 → [116.73 ; 123.26]
Fertilizer C: 125±3.265 → [121.73 ; 128.26] Fertilizer D: 130±3.265 → [126.73 ; 133.26]
LSD intervals of A and D are not overlapped, which implies that the null hypothesis should be rejected and consequently e) is false. Actually, taking into account that SS_{factor}=3112.5 it results:

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) and consequently H_0 is rejected.

LSD intervals of B and D are not overlapped, which means that c) is false. Intervals of A and B are not overlapped, which also implies that d) is false. Intervals of B and C are overlapped, and hence a) is false. Consequently, the true answer is **b)**.

8) Degrees of freedom of the interaction = $d.f._{pH} \cdot d.f._{temp} = 1 \cdot 1 = 1$
The F-ratio corresponding to the interaction is statistically significant because:

$$F_{ratio} = \frac{SC_{int\,erac} / g.lib_{int\,erac}}{SC_{res} / g.lib_{res}} = \frac{176{,}33/1}{(556{,}67 - 176{,}33 - 85{,}33 - 176{,}33)/(11-1-1-1)} = 11{,}9 > (F_{1;8}^{0,05} = 5.32)$$

Consequently, the optimum operative conditions will be pH= 8 and temperature= 35 (answer **b**). These conditions lead to an estimated average yield of 191 which is the highest one among the 4 tested treatments.

9.a)

		$Y_{observed}$		$Y_{predicted}$	$(Y_{pred} - Y_{mean})^2$	Y_{mean}=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	$2 \cdot (20.5 - 17.375)^2$	Sum = 102.375
C2	D2	17	18	17.5	$2 \cdot (17.5 - 17.375)^2$	

$SS_{C\cdot D} = 102.375 - SS_C - SS_D = 102.375 - 21.125 - 15.125 = \mathbf{66.125}$

9.b) The general hypotheses assumed when data from a design of experiments are analyzed using ANOVA are the following:
- Hypothesis of normality: the response variable follows a Normal distribution in all treatments.
- Hypothesis of homoscedasticity: the populations corresponding to all tested treatments present the same variance.
- Hypothesis of independence: observations from each treatment correspond to individuals randomly extracted from a population.

9.c) In this experimental design, the interaction C·D is confounded with factor G. Thus, the fact that C·D is statistically significant should be interpreted as a relevant effect of factor G. In this case, it would be more convenient to use the raw material of type 2 (i.e., factor G at level 2) because it yields a higher average value:

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

SOLUTIONS - PROBLEMS OF REGRESSION

10) Both variables in the model are statistically significant because their p-value is less than 0.05. The resulting equation will be: Yield = –258.333 + 73 conc –1.217 $conc^2$

To determine the relative maximum of this equation, we should derivate and make it equal to zero:

d(yield)/dc = 73 –2 · 1.217 conc = 0 → conc = 73/(2·1.217) = 30 g/l → Correct answer: **b)**

11.a) The four variables in the model present a statistically significant effect because their p-value is much lower than 0.05 (type I risk). Thus, the predictive model should use the information from the 4 variables. According to the estimated values of the coefficients shown in the table, the model will be:

Yield = 156.83 + 2.735 · temperature – 27.13 · pH + 1.92 · sugar + 3.22 · protein

11.b) The value 156.83 is the constant of the model and it is interpreted as the mean value of the expected yield when all four variables in the model are null. The value 2.73502 is the regression coefficient associated to temperature and it is interpreted as the average increase of yield expected if the mean temperature during the fermentation increases 1ºC and all other variables remain constant.

11.c) In such conditions, the expected yield will be:

Yield = 156.83 + 2.735 · 26 – 27.13 · 7.6 + 1.92 · 23 + 3.22 · 7 = 88.45

Residual standard deviation = standard error of Est. = 5.48

Thus, in such conditions: Yield ~ N (88.45 , 5.48)

P(yield<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) Coefficient of determination = R-squared = 73.49%. This parameter indicates that the model explains 73.49% of the variance of viscosity.

12.b) Temperature and catalyst present a statistically significant effect on the response variable because their p-value is less than 0.05. Thus, both variables should be included in the model. The constant is not statistically significant and it could be eliminated, but in that case a new model should be fitted in order to estimate again the regression coefficients. But given that this information is not available, it is more convenient to leave the constant in the model. Consequently, the predictive equation would be: Viscosity = -24.83 + 3.323·Temperat + 0.02724·catalyst

12.c) The coefficient is 3.32. Interpretation: if temperature increases one degree Celsius, the viscosity will increase in 3.32 units in average if the amount of catalyst remains constant.

12.d) We should introduce in the model the quadratic term (temperature^2) and therefore the following model should be fitted: viscosity = a + b·Temp + c·Temp^2 + d·catalyst

The hypothesis test would be: H_0: c=0 H_1: c≠0

If the p-value associated to this test is lower than α, then we should reject the null hypothesis and conclude that the quadratic effect is statistically significant.

13.a) No, because the p-value of variables *speed* and *pH* are higher than 0.05 and consequently they don't exert a statistically significant effect on parameter K. Thus, both variables should be eliminated from the model despite of producing a slight increase of R^2. The first model should be used.

13.b) The equation to use is the one corresponding to the first model:

Param_K = 12.86 + 0.285·Temp + 2.72·Pressure

13.c) The coefficient is 0.285. This value is interpreted as the average increase of parameter K expected when temperature increases in 1ºC inside the tank and the pressure remains constant.

13.d) mean$_{\text{Param\_K}}$ = 12.86 + 0.285·Temp + 2.72·Pressure = 12.86+0.285·72+2.72·3.2 = 42.08

Standard deviation = 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) If the unknown value of the speed is called *v*:

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

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

According to the Normal table: $(7-m_v)\big/\sqrt{0{,}76}=0{,}255$ → $m_v=6{,}78$

$$m_v=E(Y/X=v)=\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(v-\mu_X)\ ;\quad 6{,}78=6+0{,}9\frac{\sqrt{4}}{\sqrt{49}}(v-110)$$ → *v*=**113.03** km/h

15.a) The coefficient of determination is the squared value of the correlation coefficient. Thus, R^2 = r^2 = 0.328^2 = 0.108 = 10.8%

15.b) Residual is the difference between the observed value and the value predicted by the regression model. In order to study the existence of outliers, one effective procedure is to calculate all residuals and to plot them on a Normal Probability Plot. If the extreme values on this plot are clearly separated from the straight line, it would indicate that they are outliers. A box-whisker plot could also be used though it is not so informative.

15.c) No, because the p-value associated to the slope (0.1578) is higher than the significance level α.

15.d) As the correlation between both variables is not significant, the distribution of Y when X=20 corresponds to the marginal distribution of 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) The correlation between both variables is not statistically significant because the p-value associated to the slope = 0.14 > 0.05. Thus, the conditional distribution of Y when X=20 will be the marginal distribution of Y: P[y>25 / x=20] = P[y>25] →

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

17.a) False. Using regression it is also possible to study the effect of the interaction. In this case both techniques are equivalent because both factors present two levels, and consequently the same p-value will be obtained with ANOVA and regression for factors temperature and pH.

17.b) False, because factor pH is not significant which implies that if pH increases one unit, the yield will not increase in average.

18.a) It only depends on the petrol consumption (mpg), because its p-value is less than 0.05.

18.b) The coefficient of determination, also called coefficient R-squared is 62.27% (it is indicated in the table). It is obtained from the table "analysis of variance" as the ratio of residual sum of squares divided by the total sum of squares.

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) It cannot be a covariance matrix: $\mathrm{cov}(x, y) \le 12$

24) 0.0005

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

26.c) Y= (a/10) +b·X