Práctica Regresión

Curso 2019/20

Regresión y ANOVA

Grado en Ingeniería Informática

Universidad **de** Valladolid

Grupo 10

González Caminero, Juan
Rodríguez Arroyo, Diego
Castro Caballero, Manuel
Sáenz Niño, Héctor
Valdunciel Sánchez, Pablo

DUREZA PLÁSTICA

Ejercicio 1.22

a) Obtén la función de regresión estimada. Representar los valores estimados y los datos. ¿Parece que una función de regresión lineal se ajusta bien?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

Root MSE	3.23403	R-Square	0.9731
Dependent Mean	225.56250	Adj R-Sq	0.9712
Coeff Var	1.43376		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	168.60000	2.65702	63.45	<.0001
tiempo	1	2.03438	0.09039	22.51	<.0001

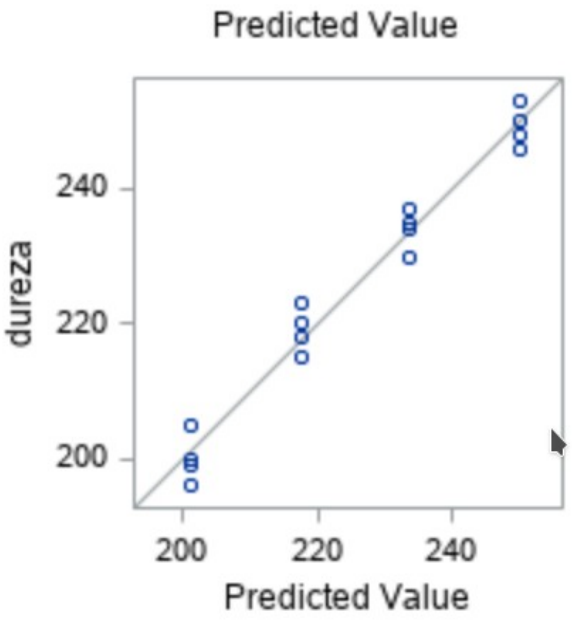
La estimación de los parámetros del modelo es:

- $\beta_0 = 168{,}6$
- $\beta_1 = 2{,}03438$

por lo que la función de regresión estimada será:

$$\mathbf{Y = 168{,}6 + 2{,}03438(x^*)}$$

donde *Y = dureza en unidades Brinell* y *X = tiempo transcurrido en horas*.



Al observar el gráfico de los valores de la variable dependiente Y = dureza frente a los valore predichos vemos como una función de regresión lineal sí parece ajustarse bien al problema. Si tenemos en cuenta el análisis de la varianza, vemos cómo el R-cuadrado es muy elevado, hasta el 97,31% de la variabilidad del modelo está explicada por la variable *X = tiempo transcurrido en horas*. El p-valor para la hipótesis nula H_0: $\beta_1 = 0$ es <0,0001, por lo que se rechaza la hipótesis nula a todos los niveles de confianza habituales y existen evidencias para afirmar que existe una asociación entre el *tiempo transcurrido* y la *dureza*.

b) Obtén la estimación puntual de la dureza media cuando X = 40 horas.

Para obtener la estimación puntual de la dureza cuando X = 40 horas aplicamos el modelo de regresión:

$$Y = 168{,}6 + 2{,}03438*(40) = 168{,}6 + 81{,}3752 = 249{,}9752$$

Cuando el tiempo transcurrido sean 40 horas (X = 40), la dureza media será 249,9752 unidades Brinell.

c) Obtén la estimación puntual de la variación en la dureza media cuando X se incrementa en 1 hora.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	168.60000	2.65702	63.45	<.0001	162.90125	174.29875
tiempo	1	2.03438	0.09039	22.51	<.0001	1.84050	2.22825

Cuando X se incrementa en 1 hora, la dureza media se incrementará en **2,03438** unidades Brinell.

Ejercicio 1.26

a) Obten los residuos e_i. ¿Es su suma igual a 0?

Sí, la suma de los residuos es igual a 0.

Output Statistics

Obs	tiempo	Dependent Variable	Predicted Value	Residual
1	16	199	201.1500	-2.1500
2	16	205	201.1500	3.8500
3	16	196	201.1500	-5.1500
4	16	200	201.1500	-1.1500
5	24	218	217.4250	0.5750
6	24	220	217.4250	2.5750
7	24	215	217.4250	-2.4250
8	24	223	217.4250	5.5750
9	32	237	233.7000	3.3000
10	32	234	233.7000	0.3000
11	32	235	233.7000	1.3000
12	32	230	233.7000	-3.7000
13	40	250	249.9750	0.0250
14	40	248	249.9750	-1.9750
15	40	253	249.9750	3.0250
16	40	246	249.9750	-3.9750

Sum of Residuals	0
Sum of Squared Residuals	146.42500
Predicted Residual SS (PRESS)	194.01315

b) Estima σ^2 y σ. ¿En qué unidades está expresada σ?

La estimación de σ^2 es igual a la Suma Media de la Suma de Cuadrados del Error (MSE), que en este caso es:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

Estimador de σ^2 = MSE = SSE / (n -2) = 146,425 / (16 -2) = **10,45893**

La estimación de σ será igual a la raíz cuadrada de la estimación de σ^2:

Estimador de σ = MSE^(1/2) = **3,2340269**

σ está expresada en **unidades Brinell**.

Ejercicio 2.7

a) Estima la variación de la dureza media cuando el tiempo transcurrido se incrementa en 1 hora. Utiliza un intervalo de confianza al 99%. Interpreta el intervalo estimado.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	168.60000	2.65702	63.45	<.0001	160.69046	176.50954
tiempo	1	2.03438	0.09039	22.51	<.0001	1.76529	2.30346

Podemos afirmar con una confianza del 99% que cuando el tiempo transcurrido se incremente en 1 hora, la dureza se incrementará entre **1,76529** y **2,30346** unidades Brinell.

b) El fabricante de plástico ha afirmado la dureza media debería incrementarse en 2 unidades Brinell por hora. Realiza un test de dos colas para decidir si este estándar se está cumpliendo; utiliza alpha=0.1. Indica las alternativas, la regla de decisión y la conclusión. ¿Cuál es el p-valor del test?

#TODO

Ejercicio 2.16

a) Obtén un I.C. del 98% para la dureza media de objetos moldeados con un tiempo transcurrido de 30 horas. Interpretar el intervalo de confianza.

El I.C. para la dureza media de objetos moldeados con un tiempo transcurido de 30 horas es

[227’4589 , 231’8056]

Este intervalo se puede interpretar de la siguiente forma: podemos afirmar con una confianza del 98% que los objetos moldeados con un tiempo transcurrido de 30 horas tendrán una dureza media comprendida entre 227’4589 y 231’8056 unidades Brinell.

b) Obtén un Intervalo de predicción del 98% para la dureza de un nuevo test de un objeto moldeado con un tiempo transcurrido de 30 horas.

El I.C. para la dureza media de objetos moldeados con un tiempo transcurido de 30 horas es

[220’8695 , 238’3930]

Este intervalo se puede interpretar de la siguiente forma: podemos afirmar con una confianza del 98% que los objetos moldeados con un tiempo transcurrido de 30 horas tendrán una dureza media comprendida entre **220’8695** y **238’3930** unidades Brinell.

Ejercicio 2.26

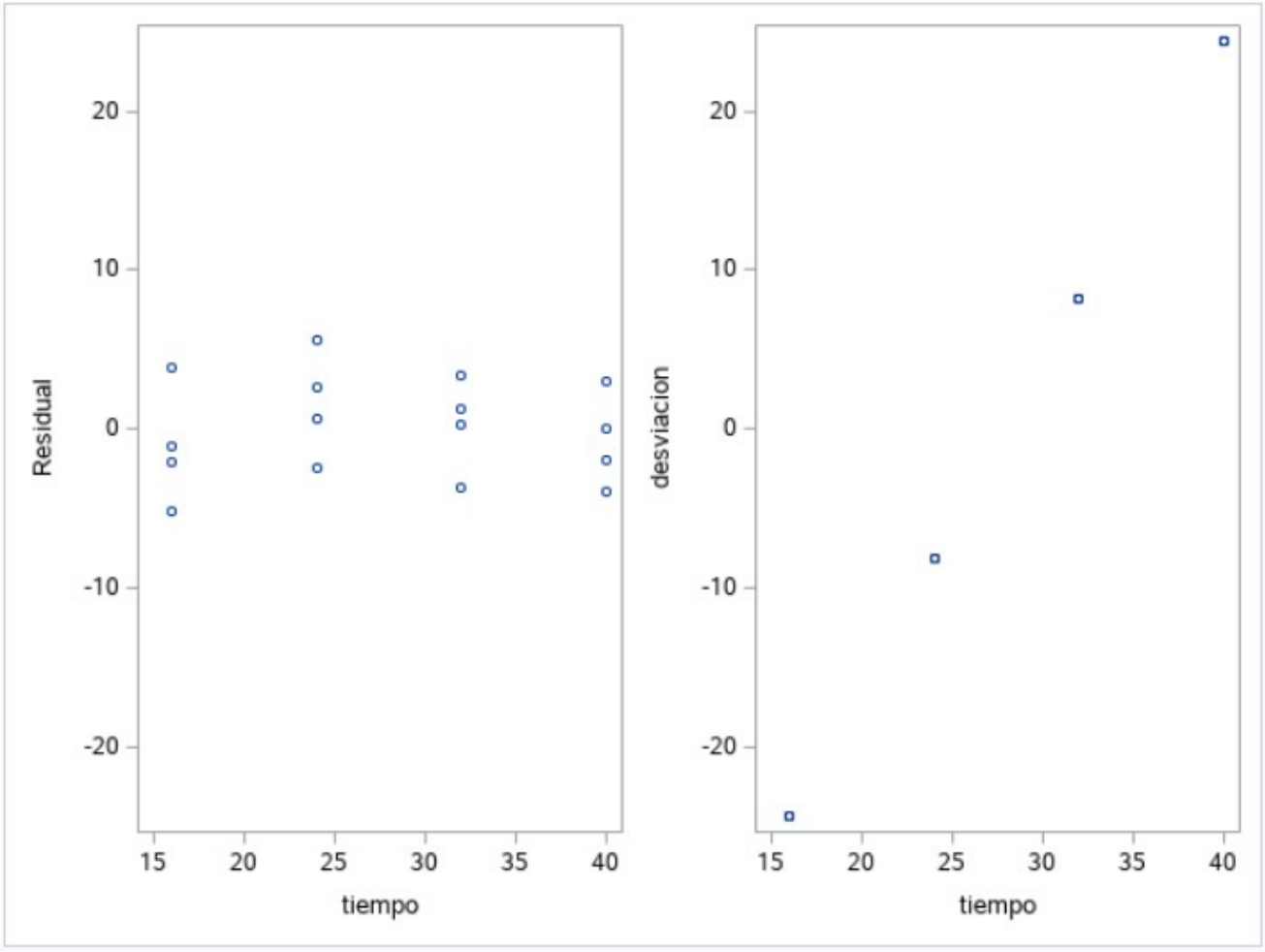
b) Determina mediante el test F si existen asociación lineal entre la dureza del plástico y el tiempo transcurrido. Utiliza alpha=0.01. Indica las alternativas, la regla de decisión y la conclusión.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

El valor del estadístico F es **506,51** y el p-valor es **2,158718E-12**, por lo que rechazamos la hipótesis nula a los niveles de confianza habituales. Existen por lo tanto evidencias para afirmar que hay una asociación entre la variable independiente *tiempo transcurrido* y la variable dependiente *dureza.*

c) Representa los residuos frente a los valores de X. Representa los ($\text{predicho}_i - \overline{Y}$) frente a los valores de X en otro gráfico utilizando la misma escala. A partir de los dos gráficos, ¿cuál parece el componente más grande de SST? ¿SSE o SSR? ¿Qué implica esto sobre la magnitud de R-cuadrado?



El componente más grande del SST es SSR, lo que implica que la mayor parte de al variabilidad de la variable Y está explicada por el modelo de regresión lineal y sólo una pequeña parte de la variabilidad está explicada por el error o las variables no controladas. Esto dará lugar un R^2 elevado, cercano a 1,0.

d) Calcular R-cuadrado y r.

- R^2 = SSR / SST = 0,9730977453
- r = 0,9864571885

MONTGOMERY 7- 8

Ejercicio 7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	148.31296	148.31296	11.47	0.0033
Error	18	232.83436	12.93524		
Corrected Total	19	381.14732			

Root MSE	3.59656	R-Square	0.3891
Dependent Mean	91.81800	Adj R-Sq	0.3552
Coeff Var	3.91705		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	77.86328	4.19889	18.54	<.0001
hydro	hydro	1	11.80103	3.48512	3.39	0.0033

a) Ajustar un modelo de regresión lineal simple a los datos.

Siendo las variables *Y = pureza del oxígeno (%)* y *X = hidrocarburos en el condensador principal (%)*, el modelo de regresión simple ajustado es:

$$\mathbf{Y = 77{,}86328 + 11{,}80103 * X}$$

b) Probar la hipótesis H_0: $\beta_1 = 0$

El p-valor obtenido para la hipótesis H_0: $\beta_1 = 0$ es **0,0033,** por lo que se rechaza la hipótesis H_0: $\beta_1 = 0$ a todos los niveles de confianza habituales. Existen evidencias para afirmar que hay una asociación entre la pureza del oxígeno y el porcentaje de hidrocarburos en el condensador principal de la unidad de procesamiento.

c) Calcular R^2.

- R^2 = SSR / SST = 148,31296 / 381,14732 = **0,3891224002**

Este valor tan bajo del R^2 indica que el modelo de regresión lineal no se ajusta bien al problema. Tan sólo un 38,91% de la variabilidad de la variable Y está explicada por modelo.

d) Determinar un intervalo de confianza del 95% para la pendiente.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	77.86328	4.19889	18.54	<.0001	69.04175	86.68482
hydro	hydro	1	11.80103	3.48512	3.39	0.0033	4.47907	19.12299

La pendiente de la recta se corresponde con la estimación del parámetro β_1, por lo que un I.C. al 95% para la pendiente de la recta es:

[4’47907 , 19’12299]

Esperamos con una confianza del 95% que un aumento del 1% en el porcentaje de hidrocarburos produzca un aumento de entre el 4’47% y el 19’12% en la pureza del oxígeno. Podemos ver que este intervalo es muy amplio, probablemente debido a que, como ya hemos visto antes, el modelo no explica muy bien la variabilidad de la variable respuesta.

e) Determinar un intervalo de confianza del 95% para la pureza media, cuando el porcentaje de hidrocarburos es 1.00.

Añadimos una observación con un 1 en *hydro* y nada en *purity* para hacer la predicción. El I.C. para la pureza media cuando el porcentaje de hidrocarburos es 1.00 es

[87’5102 , 91’8185]

Podemos afirmar con una confianza del 95% que la pureza de oxígeno media cuando el porcentaje de hidrocarburos es 1 estará comprendida entre **87’5102** y **91’8185,**

Ejercicio 8

a) ¿Cuál es la correlación entre la pureza del oxígeno y el porcentaje de hidrocarburos?

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	purity	hydro
purity purity	1.00000	0.62380 0.0033
hydro hydro	0.62380 0.0033	1.00000

La correlación entre las variables *Y = pureza del oxígeno (%)* y *X = hidrocarburos en el condensador principal (%)* es **0,6238.**

b) Probar la hipótesis que rho = 0.

Pearson Correlation Statistics (Fisher's z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits	p Value for H0:Rho=0
purity	hydro	20	0.62380	0.73120	0.01642	0.61367	0.234947 0.830623	0.0026

El p-valor obtenido es **0,0026**, por lo que rechazamos la hipótesis nula H_0: rho = 0 a todos los niveles de confianza habituales. Existen evidencias suficientes para afirmar que existe una correlación lineal entre las variables *Y = pureza del oxígeno (%)* y *X = hidrocarburos en el condensador principal (%).*

c) Establecer un intervalo de confianza del 95% para rho.

El I.C. para rho al 95% de confianza es

[0’234947 , 0’830623].

CÓDIGO SAS

```
* a2- DUREZA PLASTICA: 1.22, 1.26, 2.7 (a, b), 2.16 (a, b), 2.26 (b,c,d);


** -- Importación de los datos ---;

data durezaPlastica;
infile "/folders/myfolders/Datos/Datos Kutner/Datos_Kutner/Chapter  1 Data Sets/CH01PR22.txt";
input dureza tiempo;
run;


** -----------  1.22 -----------------**;

** a);

proc reg data=durezaPlastica;
model dureza = tiempo;
output out=residuos r=res p=pred student=rst;
run;


proc sgscatter data=residuos;
plot dureza*pred;
run;



**b);
proc reg data=durezaPlastica;
model dureza = tiempo / P CLB;
ID tiempo;
run;


** ------------- 1.26 ---------------**;

**a);

proc reg data=durezaPlastica;
model dureza = tiempo / R;
ID tiempo;
output out=residuos r=res p=pred student=rst;
run;

**b);
```

*** Estimador(sigma²) = MSE = 10.45893
*** Estimador(sigma) = MSE^(0.5) = 3.2340269

```
** ------------- 2.7 ----------------**;
**a);
proc reg data=durezaPlastica;
model dureza = tiempo /CLB alpha=0.01;
run;

**b);

* VERIFICAR QUE ES ASÏ;
data aux;
pvalor=2*(1-probt(9.184,14));
run;


** ------------- 2.16 --------------**;
data durezaPlasticaExtra;
input dureza tiempo;
cards;
"" 30
;
run;

proc append base=durezaPlastica data=durezaPlasticaExtra;
run;

proc print; run;

**a); **b);
proc reg data=durezaPlastica;
model dureza=tiempo/ CLM CLI alpha=0.02;
id tiempo;
run;

* TODO;


** ------------ 2.26 --------------**;

**b); **d);
proc reg data=durezaPlastica;
model dureza = tiempo/alpha=0.01;
output out=residuos2 r=res p=pred student=rst;
run;
```

```
data aux2;
pvalorf=1-probf(506.51,1,14);
run;

**c);

** -- Caldular media de Y --;
proc means data=durezaPlastica;
var dureza;
run;

** -- Crear variable (predicho_i - media_Y) -- ;

data aux3;
set residuos2;
desviacion = pred - 225.5625000;
run;
proc print; run;

** -- Rerpesentar los dos gráficos en la misma escala --;
proc sgscatter data=aux3;
plot (res desviacion)*tiempo / uniscale=all;
run;



** MONTOGOMERY CAPÍTULO 2 - EJERCICIOS 7 Y 8;

** -- Importación de los datos --;


FILENAME REFFILE '/folders/myfolders/Datos/Datos Montgomery Ch 02/DATOS Montgomery ch.
2-20190926/data-prob-2-7.XLS';

PROC IMPORT DATAFILE=REFFILE
        OUT=purezaOxigeno replace
        DBMS=XLS;

        GETNAMES=YES;
RUN;

proc print data=purezaOxigeno;
run;

** ---- Ejercicio 7 ---- ;

** a) b) c);
proc reg data=purezaOxigeno;
```

```
model purity=hydro;
run;


** d);
proc reg data=purezaOxigeno;
model purity=hydro / CLB alpha=0.05;
run;

** e);
data purezaOxigenoExtra;
input purity hydro;
cards;
"" 1
;
run;

proc append base=purezaOxigeno data=purezaOxigenoExtra;
run;
proc print; run;

proc reg data=purezaOxigeno;
model purity=hydro / CLM alpha=0.05;
run;


** ---- Ejercicio 8 --- ;

**a);
proc corr data=purezaOxigeno;
var purity hydro;
run;

**b) c);

proc corr fisher data=purezaOxigeno;
var purity hydro;
run ;
```