UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

INFORME

TAREA IV

"REGRESION MULTIPLE"

ECONOMETRIA

PROFESOR: RODRIGO ORTEGA
AYUDANTE: DIEGO BASCUÑAN

INTEGRANTE:

BARBARA LIZAMA

24 DE MAYO DE 2021

PROBLEMA

Se adjunta un set de datos precios de casas y un set de datos de azúcar, para generar diversas pruebas en relación a la población de estos data set.

OBJETIVO

Realizar diversas pruebas de hipótesis con diversos análisis como modelos de regresión múltiple, interpretando sus resultados viendo los valores observados y estimados, para establecer diversas conclusiones en relación a estas hipótesis planteadas.

I. DATA SET PRECIO DE CASAS

1) Cree un modelo con todas las variables en nivel para explicar el precio de una casa en función de sus atributos. Interprete los coeficientes.

Regression Analysis

OVERALL FIT			
Multiple R	0,8197	AIC	1868,6573
R Square	0,6718	AICc	1869,39804
Adjusted R Square	0,6600	SBC	1878,52093
Standard Error	45114,8404		
Observations	87		

ANOVA					Alpha	0,05
	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	3	3,4585E+11	1,1528E+11	56,6401636	5,0047E-20	yes
Residual	83	1,6893E+11	2035348822			
Total	86	5,1478E+11				

	coeff	*std err*	*t stat*	*p-value*	*lower*	*upper*	*vif*
Intercept	-16320,7326	22224,3870	-0,7344	0,4648	-60524,1445	27882,6794	
Sup.sitio (m2)	16,6232	5,2157	3,1871	0,0020	6,2493	26,9971	1,0369
Sup.const (m2)	988,2625	107,7964	9,1679	3,0129E-14	773,8598	1202,6652	1,4253
n°dormitorios	10604,8309	6824,7899	1,5539	0,1240	-2969,4016	24179,0635	1,4052

Ilustración 1 Modelo Regresión (Real Statistics)

Precio de Casa = -16.320,7326 + 16,6232 * X1 + 988,2625 * X2 + 10.604,8309 * X3

$\beta 1$	=	16,6232	Por cada m^2 adicional en la superficie del sitio el precio aumentara en M$ 16,6232
$\beta 2$	=	988,2625	Por cada m^2 adicional en la superficie construida el precio aumentara en M$ 988,2625
$\beta 3$	=	10.604,8309	Por cada dormitorio adicional en la casa el precio aumentara en M$ 10.604,8309

Ilustración 2 Interpretación Coeficientes

2) Establezca las pruebas de hipótesis respectivas y determine la significancia global del modelo y de cada coeficiente. No olvide las unidades de cada variable.

H0: $\beta 1 = 0, \beta 2 = 0, \beta 3 = 0$

Ha: H0 no es verdadera

Ilustración 3 Test de Hipótesis Significancia Global

	df	*SS*	*MS*	*F*	*p-value*	*sig*
Regression	3	3,4585E+11	1,1528E+11	56,6401636	**5,0047E-20**	**yes**
Residual	83	1,6893E+11	2035348822			
Total	86	5,1478E+11				

Ilustración 4 Resultados Real Statistics

Se puede concluir que se rechaza la hipótesis nula, dado que el P-value es menor al alpha, por lo cual si es significativo globalmente el modelo.

H0: $\beta 1 = 0$	H0: $\beta 2 = 0$	H0: $\beta 3 = 0$
Ha: $\beta 1 \neq 0$	Ha: $\beta 2 \neq 0$	Ha: $\beta 3 \neq 0$

Ilustración 5 Test de Hipótesis Significancia Local

	coeff	*std err*	*t stat*	***p-value***	*lower*	*upper*
Intercept	-16.320,7326	22.224,3870	-0,7344	0,4648	-60.524,1445	27.882,6794
Sup.sitio (m2) β1	16,6232	5,2157	3,1871	0,0020	6,2493	26,9971
Sup.const (m2) β2	988,2625	107,7964	9,1679	3,0129E-14	773,8598	1.202,6652
n°dormitorios β3	10.604,8309	6.824,7899	1,5539	0,1240	-2.969,4016	24.179,0635

Ilustración 6 Resultados Real Statistics

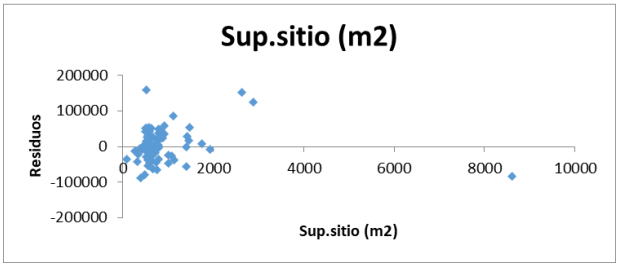
En el caso del **β1** que es la superficie del sitio en m^2, se puede concluir que se rechaza la hipótesis nula, dado que el p value (**0,0020**) es menor al alpha (**0,05**), por lo cual si es significativo localmente.

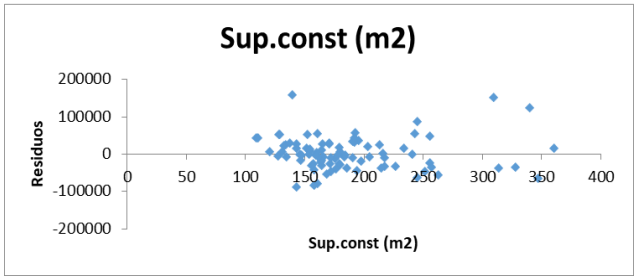
En el **β2** que es la superficie construida en m^2 se puede concluir que se rechaza la hipótesis nula, ya que el p value (**3,0129E-14**) es menor al alpha (**0,05**) por lo cual es significativo localmente.

Y por último el **β3** es el número de dormitorios en la casa, donde no se puede rechazar la hipótesis nula debido a que el p value (**0,1240**) es mayor al alpha (**0,05**), por consiguiente no es significativo localmente.

3) Visualmente determine si el modelo tiene problemas de heterocedasticidad.

A la vista se podría indicar que no hay problemas de heterocedasticidad.





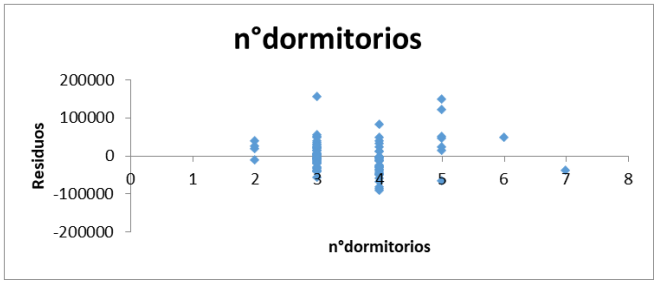


Ilustración 7 Gráficos de los residuales

4) Determine si existen problemas de multicolinealidad. Utilice el estadístico adecuado.

Para ver el tema de multicolinealidad se utilizó el estadístico VIF, con un valor de corte de 10.

	vif	*1-VIF*
Sup.sitio (m2) b1	1,03694902	0,96436757
Sup.const (m2) b2	1,42533436	0,70158977
n°dormitorios b3	1,40515538	0,71166507

Ilustración 8 Calculo estadístico VIF

Al revisar los VIF de cada *B*eta se puede concluir que ninguno de ellos tiene problema de multicolinealidad ya que el estadístico es menor a 10 en todas las variables, debido a que todas tienen un valor ínfimo.

5) Aplique logaritmo a todas las variables, excepto número de dormitorios. Determine la significancia global y de cada coeficiente. Interprete cada coeficiente.

$H0: \beta1 = 0, \beta2 = 0, \beta3 = 0$

Ha: H0 no es verdadera

Ilustración 9 Test de Hipótesis Significancia Global

ANOVA				Alpha	0,05	
	df	*SS*	*MS*	*F*	***p-value***	***sig***
Regression	3	5,134871479	1,711623826	49,64369631	**1,78381E-18**	**yes**
Residual	83	2,861688152	0,034478171			
Total	86	7,996559631				

Ilustración 10 Resultados Real Statistics

Se puede concluir que se rechaza la hipótesis nula, dado que el P-value es menor al alpha, por lo cual si es significativo globalmente el modelo.

$H0: \beta1 = 0$ $H0: \beta2 = 0$ $H0: \beta3 = 0$

$Ha: \beta1 \neq 0$ $Ha: \beta2 \neq 0$ $Ha: \beta3 \neq 0$

Ilustración 11 Test de Hipótesis Significancia Local

	coeff	*std err*	*t stat*	***p-value***	*lower*	*upper*
Intercept	7,39227	0,44441	16,63392	3,69304E-28	6,50836	8,27618
Ln (X1)	0,16753	0,03860	4,33955	**3,99924E-05**	0,09074	0,24431
Ln (X2)	0,69938	0,09356	7,47520	**7,14297E-11**	0,51330	0,88547
n°dormitorios X3	0,03735	0,02780	1,34349	**0,18278**	-0,01795	0,09265

Ilustración 12 Resultados Real Statistics

En el caso del $\beta1$ que es la superficie del sitio en m^2, se puede concluir que se rechaza la hipótesis nula, dado que el p value (**3,999E-05**) es menor al alpha (**0,05**), por lo cual si es significativo localmente.

En el **β2** que es la superficie construida en m^2 se puede concluir que se rechaza la hipótesis nula, ya que el p value (**7,142E-11**) es menor al alpha (**0,05**) por lo cual es significativo localmente.

Y por último el $\beta3$ es el número de dormitorios en la casa, donde no se puede rechazar la hipótesis nula debido a que el p value (0,18278) es mayor al alpha (0,05), por consiguiente no es significativo localmente.

β1 = 0,1675 Por cada aumento de 1% en m^2 en la superficie del sitio el precio aumentara en 0,1675%

β2 = 0,6994 Por cada aumento de 1% en m^2 en la superficie construida el precio aumentara en 0,6994%

β3 = 0,0374 Por cada dormitorio adicional en la casa el precio aumentara en 3,735%

Ilustración 13 Interpretación Coeficientes

6) Pruebe que en la población β2=60*β1. Establezca la prueba de hipótesis correspondiente y explique sus procedimientos.

$$H0 = \beta2 = 60 * \beta1 \quad V/S \quad H0 = \beta2 \neq 60 * \beta1$$

Ilustración 14 Test de Hipótesis

Para poder realizar el cálculo se deberá ocupar un nuevo parámetro que sería θ1, por lo cual se cambia el test de hipótesis.

$$\beta2 = 60 * \beta1 + \theta1$$
$$\theta1 = \beta2 - 60 * \beta1$$

$$H0 = \theta1 = \beta2 - 60 * \beta1 \qquad H0 = \theta1 \neq \beta2 - 60 * \beta1$$

Ilustración 15 Nuevo test de hipótesis con theta

precio casa = β0 + β1* sup. Sitio + β2* sup. const + β3* n° dormitorio
precio casa = β0 + β1* sup. Sitio + (60* $\beta1 + \theta1$)* sup. const + β3* n° dormitorio
precio casa = β0 + β1* sup. Sitio + 60* β1* sup. const + θ1 sup const + β3* n° dormitorio
precio casa = β0 + β1* (sup. Sitio + 60* sup. Const)+ θ1 sup const + β3* n° dormitorio

Ilustración 16 Reemplazo en el modelo original

Resumen

Estadísticas de la regresión	
Coeficiente de correlación	0,8197
Coeficiente de determinaci	0,6718
R^2 ajustado	0,6600
Error típico	45114,8404
Observaciones	87

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	3	3,45847E+11	1,15282E+11	56,64016362	5,00475E-20
Residuos	83	1,68934E+11	2035348822		
Total	86	5,14781E+11			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*
Intercepción	-16320,7326	22224,38703	-0,7344	0,4648	-60524,1445
sup sitio + 60* sup cons	16,6232	5,215728386	3,1871	0,0020	6,2493
Sup.const (m2) X2	**-9,1301**	343,8765894	-0,0266	0,9789	-693,0868
n° dormitorio	10604,8309	6824,789866	1,5539	0,1240	-2969,4016

Ilustración 17 Resultado Real Statistics

Generando la nueva regresión se puede concluir que no se puede rechazar la hipótesis nula debido a que el p value (**0,9789**) es mayor al alpha (**0,05**).

II. DATA SET AZUCAR

1) Estime un modelo para explicar el rendimiento de azúcar_2009 (TM_2009), en función del rendimiento de azúcar_2008 (TM_2008) y el resto de las variables del set de datos, incluidas longitud (X) y latitud (Y). Determine la significancia de cada variable y ordene los coeficientes de mayor a menor (en valor absoluto).

$$\hat{y} = -658{,}1770 + 0{,}0002 * X1 + 0{,}0003 * X2 + 0{,}8019 * X3 + 1{,}7217 * X4 + 2{,}6143 * X5 - 78{,}4421 * X6 + 0{,}3393 * X7 + 0{,}0676 * X8 + 0{,}6856 * X9 - 0{,}6111 * X10 + 1{,}6931 * X11 + 5{,}8543 * X12 - 0{,}2636 * X13 + 4{,}0829$$

Ilustración 18 Modelo Estimado

	Coeficientes
Intercepción (β0)	658,1770
Ntotal (β6)	78,4421
NA_DISP (β12)	5,8543
MO (β5)	2,6143
PH (β4)	1,7217
K_DISP (β11)	1,6931
TM_2008 (β3)	0,8019
CA__DISP (β9)	0,6856
MG_DISP (β10)	0,6111
C_N (β7)	0,3393
CIC (β13)	0,2636
P__DISP (β8)	0,0676
y (β2)	0,0003
x (β1)	0,0002

Ilustración 19 Coeficientes ordenados en valor absoluto

Las variables significativas localmente son Y (0,0461) e TM_2008 (9,496E-18) debido a que su p value es menor a alpha.

Las variables no significativas localmente son x (0,5375), PH (0,1241), MO (0,3345), Ntotal (0,1380), C_N (0,4278), P_DISP (0,0854), CA_DISP (0,0731), MG_DISP (0,5782), K_DISP (0,5319), NA_DISP (0,5121) y CIC (0,2388) ya que su p value es mayor a alpha

2) Transforme todas las variables a Z y corra nuevamente el modelo. Determine la significancia de cada variable y ordene los coeficientes de mayor a menor.

	coeff
β3 Z_TM_2008	0,7056
β6 Z_Ntotal	0,3212
β5 Z_MO	0,2508
β9 Z_CA__DISP	0,1984
β2 Z_y	0,1512
β8 Z_P__DISP	0,1426
β4 Z_PH	0,1088
β13 Z_CIC	0,1041
β7 Z_C_N	0,0925
β10 Z_MG_DISP	0,0652
β1 Z_x	0,0485
β12 Z_NA_DISP	0,0476
β11 Z_K_DISP	0,0435
β0 Intercept	0,0000

Ilustración 20 Coeficiente ordenados en valor absoluto

Las variables significativas localmente son Y (0,0461) e TM_2008 (9,496E-18) debido a que su p value es menor a alpha.

Las variables no significativas localmente son x (0,5375), PH (0,1241), MO (0,3345), Ntotal

(0,1380), C_N (0,4278), P_DISP (0,0854), CA_DISP (0,0731), MG_DISP (0,5782), K_DISP (0,5319), NA_DISP (0,5121) y CIC (0,2388) ya que su p value es mayor a alpha.

3) Aplique logaritmo natural a todas las variables y corra nuevamente el modelo. Determine la significancia de cada variable y ordene los coeficientes de mayor a menor.

	Coeficientes
Intercepción	1897,4411
ln y (β2)	120,3118
Ln x (β1)	12,5123
ln PH (β4)	2,3629
ln C_N (β7)	0,9680
ln Ntotal (β6)	0,7456
ln MG_DISP (β10)	0,6196
ln CA__DISP (β9)	0,5583
ln MO (β5)	0,4590
ln TM_2008 (β3)	0,4010
ln CIC (β13)	0,2388
ln NA_DISP (β12)	0,1610
ln K_DISP (β11)	0,0716
ln P__DISP (β8)	0,0231

Ilustración 21 Coeficientes ordenados en valor absoluto

Las variables significativas localmente son Y (0,0020), TM_2008 (0,0002), PH (0,0044), CA_DISP (0,0067), MG_DISP (0,0062) debido a que su p value es menor a alpha.

Las variables no significativas localmente son x (0,4258), MO (0,5711), Ntotal (0,3355), C_N (0,1264), P_DISP (0,7481), K_DISP (0,5500), NA_DISP (0,4490) y CIC (0,4050) ya que su p value es mayor a alpha

4) Compare los resultados en los tres casos. Cuales es la variable más importante que determina el rendimiento de azúcar del año 2009.

	coeficiente	*p value local*	*varianza*	*R^2*	*p value global*
Ntotal (β6)	78,4421	0,1380	4,082904663	62,437%	2,E-16
NA_DISP (β12)	5,8543	0,5121			
MO (β5)	2,6143	0,3345			
PH (β4)	1,7217	0,1241			
K_DISP (β11)	1,6931	0,5319			
TM_2008 (β3)	0,8019	0,0000			
CA__DISP (β9)	0,6856	0,0731			
MG_DISP (β10)	0,6111	0,5782			
C_N (β7)	0,3393	0,4278			
CIC (β13)	0,2636	0,2388			
P__DISP (β8)	0,0676	0,0854			
y (β2)	0,0003	0,0461			
x (β1)	0,0002	0,5375			

Ilustración 22 Modelo Original

	coeficiente	*p value local*	*varianza*	*R^2*	*p value global*
Z_TM_2008	0,7056	0,0000	0,6507739	62,437%	2,E-16
Z_Ntotal	0,3212	0,1380			
Z_MO	0,2508	0,3345			
Z_CA__DISP	0,1984	0,0731			
Z_y	0,1512	0,0461			
Z_P__DISP	0,1426	0,0854			
Z_PH	0,1088	0,1241			
Z_CIC	0,1041	0,2388			
Z_C_N	0,0925	0,4278			
Z_MG_DISP	0,0652	0,5782			
Z_x	0,0485	0,5375			
Z_NA_DISP	0,0476	0,5121			
Z_K_DISP	0,0435	0,5319			

Ilustración 23 Modelo Normalizado

	coeficiente	*p value local*	*varianza*	*R^2*	*p value global*
ln y (β2)	120,3118	0,0020	0,51398345	40,362%	3,E-07
Ln x (β1)	12,5123	0,4258			
ln PH (β4)	2,3629	0,0044			
ln C_N (β7)	0,9680	0,1264			
ln Ntotal (β6)	0,7456	0,3355			
ln MG_DISP (β10)	0,6196	0,0062			
ln CA__DISP (β9)	0,5583	0,0067			
ln MO (β5)	0,4590	0,5711			
ln TM_2008 (β3)	0,4010	0,0002			
ln CIC (β13)	0,2388	0,4050			
ln NA_DISP (β12)	0,1610	0,4490			
ln K_DISP (β11)	0,0716	0,5500			
ln P__DISP (β8)	0,0231	0,7481			

Ilustración 24 Modelo Variables con Logaritmo

Todos los modelos tienen una significancia global, en el caso de la significancia local se puede decir que existen dos variables, la Y e TM_2008 donde el p value es menor al alpha.

Para el R^2 se puede deducir que el modelo con logaritmo tiene un bajo porcentaje de explicación respecto a los otros dos modelos.

En el caso de la varianza se debería quedar con el modelo normalizado ya que es el que tiene una menor variabilidad.

Para concluir el mejor modelo para explicar el rendimiento de la azúcar del año 2009 es el modelo normalizado ya que su significancia global y su r^2 son mayores a los otros modelos. Y la variable más importante es el TM_2008 (β3), ya que su coeficiente es muy superior a la del resto de las variables del modelo.