EVALUACIÓN DE LA ROBUSTEZ DE UN MODELO DE REGRESIÓN MÚLTIPLE PARA PREDECIR LAS VENTAS DIARIAS DE UN HIPERMERCADO EN PEREIRA, RISARALDA.

Evaluation of the Strength of a Multiple Regression Model to Predict the Daily Sales in a Hypermarket in Pereira, Risaralda

.

RESUMEN

En este trabajo se evalúa la robustez de un modelo de regresión lineal múltiple, usado para predecir las ventas diarias en un departamento de un almacén hipermercado en la ciudad de Pereira. Se evalúa el nivel de adecuación de esta técnica para el caso de estudio a partir de la verificación de supuestos, el nivel de explicación del R^2, y validación de la hipótesis: $\beta_{k \neq} 0$.

PALABRAS CLAVES: Regresión Múltiple, Hipermercado, Ventas Diarias

ABSTRACT

This paper assesses the robustness of a multiple linear regression model, used to Predict the daily sales in a department of a supermarket store in the city of Pereira. It assesses the adequacy of this technique to the case study from the verification of assumptions, the level of explanation of R^2, the validation of the hypothesis: $\beta_k \neq 0$

KEYWORDS: Multiple Regression, hypermarket, Daily Sales

JORGE ANDRÉS URRUTIA MOSQUERA
M. Sc.
Profesor Auxiliar
Universidad Tecnológica de Pereira
jurrutia@utp.edu.co

HEVER DARÍO SALAZAR
M. Sc Candidato
Profesor
Universidad Agrícola y rural de Colombia. UNISARC.
hedasa@hotmail.com

EDUARDO ARTURO CRUZ TREJOS
Ingeniero Industrial, M. Sc.
Profesor Asociado
Universidad Tecnológica de Pereira
ecruz@utp.edu.co

Grupo: ADMINISTRACION ECONOMICA Y FINANCIERA

1. INTRODUCCIÓN

Muchos problemas de de investigación y de la industria, requieren la estimación de las relaciones existentes entre el patrón de variabilidad de una variable aleatoria y los valores de una o más variables aleatorias o no de la que puede depender la primera, así como los parámetros que describen dichos comportamientos. La predicción y estudio de variación de las ventas diarias de un almacén de cadena, para la planeación del abastecimiento del mismo, es uno de los problemas que se puede estudiar mediante modelos de regresión u otras medidas de asociación como las correlaciones parciales y matrices de covarianzas. Para nuestro caso estudiaremos las bondades de los modelos de regresión múltiple para la predicción de las ventas diarias en un almacén de cadena en la ciudad de Pereira.

2. DESARROLLO TEÓRICO

2.1 EL MODELO DE REGRESIÓN LINEAL MULTIPLE.

La regresión lineal múltiple, es un método matemático que modeliza la relación entre una variable dependiente o explicada **Y**, y un grupo de variables independientes o regresoras $\mathbf{X}_i$ y un término aleatorio ε, llamado error aleatorio [1], [2].

Los objetivos de los modelos de regresión son:

Determinar si la variable explicada, está correlacionada con las variables regresoras o explicativas.

Predecir el valor de la variable dependiente (Variable explicada), dado unos determinados valores de las variables regresoras.

Fecha de Recepción: 25 de Enero de 2011
Fecha de Aceptación: 28 de Abril de 2011

Valorar el nivel de concordancia entre los valores de las variables regresoras y la variable explicada.

[2] El modelo de regresión lineal múltiple se expresa como:

$$y_{i=}\ \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

Donde

$y_{i=}$ Es la í- ésima observación de la variable aleatoria explicada.

$X_{i1}\ldots X_{ik}$ son las la í- ésima observación de la variable aleatoria regresoras.

$\beta_0, \beta_1, \ldots \beta_k$, son los coeficientes de regresión.

εi es la variable aleatoria que se supone presenta los siguientes supuestos

a. $E(\varepsilon_i) = 0$
b. Los errores tienen varianza igual pero desconocida.
c. Los errores nos son correlacionados.

2.1.1 Supuestos del modelo de regresión:

1. La variable aleatoria ϵ (error) debe ser estadísticamente independiente de los valores de X_i y tener una distribución normal con una media igual a cero (supuesto 1 y 2).

Esto implica que: $\beta 0$ β_1 Son constantes

$$E(\beta_0) = \beta_0 \qquad E(\beta_1) = \beta_1 \qquad (2)$$

Así para determinados valores de X_i se tiene que

$$E(Y) = \beta_0 + \beta_{i1} X_{i1} + \beta_{i2} X_{i2} + \ldots \beta_{ik} X_{ik} \qquad (3)$$

2. Cualquier par de errores, ϵ_i y ϵ_j deben ser estadísticamente independientes entre sí, es decir que su covarianza debe ser igual a 0 (supuesto 3)

3. Las variables aleatorias ϵ_j deben tener una varianza finita σ^2 que es constante para todos los valores de X_j . (Supuesto 4 o de homocedasticidad)

2.1.2 Estimación de los parámetros

La estimación de los parámetros en los modelos de regresión múltiple se realiza, mediante el método de mínimos cuadrados, cuyo propósito es minimizan la suma de los residuos al cuadrado, que se produce al momento de estimar los parámetros [3], [4].

Las ecuaciones de este método son:

2.1.3 Control de Supuestos

[5] Los supuestos a controlar en el modelo son:

Multicolinealidad: a través de matrices de correlación simple entre las variables independientes. Solución: Seleccionar variables independiente con baja correlación entre sí y/o transformar en variables dummy no colineales.

Normalidad De Los Residuos: a través de un gráfico de de distribución de los residuos. Solución: eliminación de datos outliers.

Heteroscedasticidad: a través de gráficos de residuos ϵ para cada valor de $\hat{y}$. Solución: Eliminación de casos outliers, transformación de las variables independientes y/o estandarización de la variable dependiente *Y*.

Auto correlación De Errores: a través de la prueba Durbin-Watson /. Solución: Corrección de observaciones o eliminación de datos.

3. DESARROLLO

Para el caso de estudio se determinó la variable dependiente y aquellas variables explicativas de acuerdo a la filosofía de los métodos de regresión lineal Múltiple, entendiendo en este caso como variable dependiente o variable de respuesta las ventas diarias y como variables explicativas *facturas, Ticket y números de clientes*. Es de anotar que se tomaron como observaciones el total vendido en dinero de los 26 productos que se ofrecen en el departamento seleccionado del Hipermercado, teniendo así las siguientes variables de estudio:

Variable De Respuesta	**Variables Explicativas**
Ventas Diarias(Millones)	*Facturación (Unidades Expedidas)*
	Ticket (Unidades enteras)
	Número de Clientes (Unidades enteras)

Tabla 1. Descripción de variables.

En el estudio se consideraron las ventas concernientes al año 2010, sin embargo para nuestro caso de estudio mostraremos el análisis para las ventas de un mes. Para efectos de estudiar la robustez del modelo de regresión a partir de estas variables de estudio, los datos se analizaron con el soler de Excel 2007 y el software XLSTAT 2010. El nombre del Hipermercado y los datos utilizados para nuestro estudio no

son presentados en este trabajo, por razones de confidencialidad.

3.1 ANÁLISIS DE RESULTADO

Al examinar la tabla 2 de los estadísticos de la regresión, se observa que el valor del coeficiente de correlación múltiple presenta un valor de 0.929, lo que significa que existe una alta asociación positiva entre la variable de respuesta y las variables explicativas. Del mismo modo, el valor del R^2 es de 0.864, lo traduciéndose en que el modelo permite explicar el 86% de la variabilidad de las ventas diarias del departamento seleccionado del hipermercado estudiado a partir de las variables *facturas, Ticket y números de clientes.* El valor del R^2 –ajustado es de 0.845, valor que expresa que hay buen ajuste entre los datos reales y los datos modelados de predicción [6]. El error típico que presenta el modelo es de *2767.32*, valor que se puede considerar pequeño, dadas las unidades del problema, lo que se traduce en poca dispersión de los datos; sin embargo una mejor medida de este aspecto no los dará el gráfico de probabilidad normal y el gráfico de regresión [7][8].

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.929763684
Coeficiente de determinación R^2	0.864460507
R^2 ajustado	0.845977849
Error típico	2767.32537
Observaciones	26

Tabla 2. Estadísticos de la regresión.

El valor crítico de *F* o valor *P* de la tabla 3 del análisis de varianza, muestra un valor de $1.02667E^{-09}$, valor que contrasta el criterio de prueba para el modelo de regresión. Como $p < \alpha$, se valida el modelo, puesto que el estadístico *p* de la prueba f es menor que α, para una confianza del 95%.

ANÁLISIS DE VARIANZA	*Grados de libertad*	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	3	1074539616	358179872.1	46.77143857	1.02667E-09
Residuos	22	168477973.5	7658089.704		
Total	25	1243017590			

Tabla 3. Análisis de varianza para el modelo de regresión.

El análisis de varianza de la tabla 4, para los parámetros del modelo, permite verificar que las variables que han sido significativas para el modelo de regresión son: *Fac* y *Clientes.* Dado que su valor $p < \alpha$ con significancia del 5%. De igual forma los intervalos de confianzas, para cada uno de los coeficientes del modelo, muestran los valores mínimos y máximos que pueden tomar cada uno de ellos con el fin de modelar las ventas diarias. De los parámetros del modelo podemos construir la ecuación de regresión es:

*VENTAS=5655,82318774058+7,39855762007813*FAC +319,860594303288*CLIENTES* (4)

De la cual se puede decir que si el número de clientes permanece constaste, el valor de las ventas aumenta o disminuyen en *7,39855762007813*, por cada unidad de factura expedida o no, de igual modo es interpretada la variable cliente. La figura 1, muestra la tendencia del modelo:

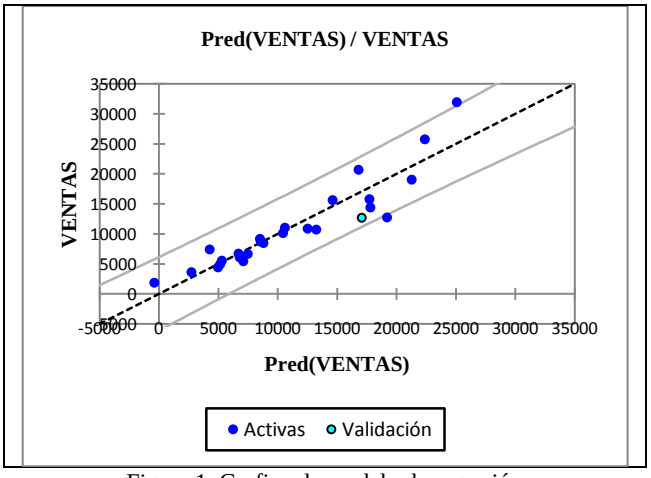

Figura 1. Grafico de modelo de regresión.

Parámetros del modelo:

Fuente	Valor	Desviación típica	t	Pr > |t|	Límite inferior (95%)	Límite superior (95%)
Intersección	-5655.823	1577.833	-3.585	0.002	-8928.049	-2383.597
FAC	7.399	5.869	1.261	0.021	-4.773	19.570
TICKET	0.000	0.000				
CLIENTES	319.861	28.483	11.230	< 0,0001	260.790	378.931

Tabla 4. Análisis de varianza para los parámetros del modelo de regresión.

La tabla 5, de correlaciones entre las variables del modelo, se puede observar que las variables regresoras presentan baja correlación entre ellas mismas, lo que garantiza que se cumple el supuesto de *Multicolinealidad* [5].

Matriz de correlación:

Variables	FAC	TICKET	CLIENTES	VENTAS
FAC	**1.000**	-0.399	0.225	0.304
TICKET	-0.399	**1.000**	0.652	0.640
CLIENTES	0.225	0.652	**1.000**	0.925
VENTAS	0.304	0.640	0.925	**1.000**

Tabla 5. Correlaciones entre las variables del modelo de regresión.

El grafico de residuos estandarizados de la figura 2, y 3, muestran que las observaciones se encuentran a menos de 3 desviaciones estándar, con respecto a la línea media de regresión, significando así que no existen valores atípicos, y se cumple con el supuesto de *Normalidad De Los Residuo y el supuesto de Homocedasticidad*.

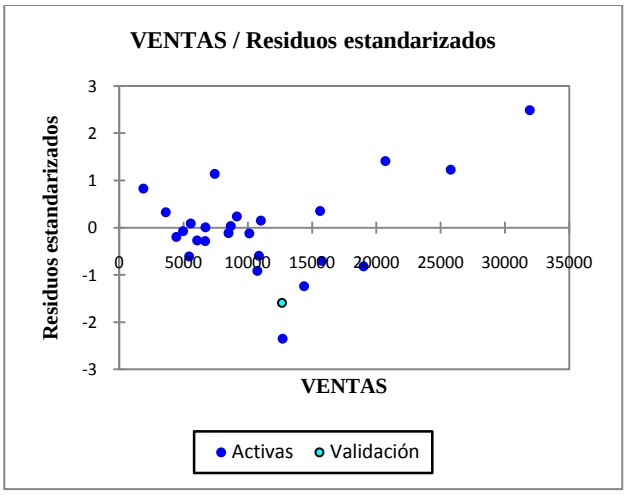

Figura 2. Grafico los residuos estandarizados.

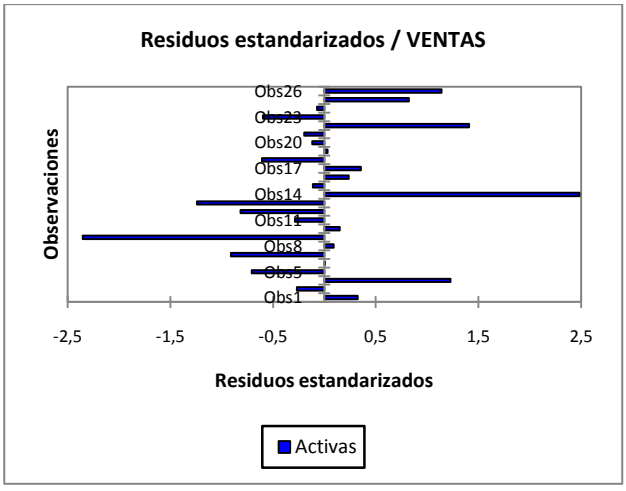

Figura 3. Grafico los residuos estandarizados con las observaciones.

Una vez validados los supuestos y constatado la hipótesis del modelo de regresión lineal múltiple, procederemos al objetivo de este trabajo predecir las ventas diarias. Las tablas 6, 7, 8 y . Muestran el mejor modelo de regresión, las predicciones de las ventas, para las medias y los intervalos de confianzas de las predicciones.

El mejor modelo para el criterio de selección seleccionado se exhibe en azul

o. de variables	Variables	MEC	R^2	R^2 ajustado	Cp de Mallows
2	FAC / CLIENTES	**7602389.084**	0.865	0.847	8.623

Tabla 6.parámetros del mejor modelo de regresión.

Observación	Peso	VENTAS	Pred(VENTAS)	Residuo	Residuo estd.
Obs1	1	3616.000	2718.018	897.982	0.326
Obs2	1	6064.000	6802.220	-738.220	-0.268
Obs4	1	25766.530	22375.763	3390.767	1.230
Obs5	1	15762.330	17711.028	-1948.698	-0.707
Obs6	1	6695.920	6678.675	17.245	0.006
Obs7	1	10728.860	13242.609	-2513.749	-0.912
Obs8	1	5547.450	5297.376	250.074	0.091
Obs9	1	12716.710	19205.029	-6488.319	-2.353
Obs10	1	11009.600	10598.895	410.705	0.149
Obs11	1	6681.800	7472.044	-790.244	-0.287
Obs12	1	19011.830	21266.995	-2255.165	-0.818
Obs13	1	14374.900	17794.135	-3419.235	-1.240
Obs14	1	31928.577	25075.220	6853.357	2.486
Obs15	1	8491.200	8797.600	-306.400	-0.111
Obs16	1	9147.200	8497.704	649.496	0.236
Obs17	1	15609.800	14626.846	982.954	0.356
Obs18	1	5429.400	7103.839	-1674.439	-0.607
Obs19	1	8678.800	8588.718	90.082	0.033
Obs20	1	10116.413	10441.294	-324.881	-0.118
Obs21	1	4426.100	4968.394	-542.294	-0.197
Obs22	1	20689.500	16805.975	3883.525	1.408
Obs23	1	10851.450	12498.800	-1647.350	-0.597
Obs24	1	4966.700	5162.479	-195.779	-0.071
Obs25	1	1874.180	-399.204	2273.384	0.825
Obs26	1	7416.410	4271.207	3145.203	1.141
Obs3	1	12663.195	17059.955	-4396.760	*-1.595*

Tabla 7.Predicciones de las ventas diarias y residuos estandarizados.

Des. estd sobre la pred. (Media)	Límite inferior 95% (Media)	Límite superior 95% (Media)
972.119	701.967	4734.069
863.602	5011.220	8593.220
1235.976	19812.506	24939.020
847.580	15953.256	19468.800
678.016	5272.556	8084.795
833.513	11514.008	14971.210
757.237	3726.962	6867.790
897.686	17343.342	21066.715
1257.514	7990.971	13206.819
801.111	5810.642	9133.446
1207.175	18763.467	23770.524
944.353	15835.667	19752.602
1403.622	22164.287	27986.153
628.867	7493.409	10101.791
794.380	6850.261	10145.148
912.704	12734.014	16519.677
648.668	5758.583	8449.094
675.925	7186.935	9990.501
596.400	9204.436	11678.152
760.691	3390.817	6545.971
1739.089	13199.325	20412.624
761.179	10920.211	14077.389
744.647	3618.176	6706.782
1135.761	-2754.628	1956.220
821.257	2568.023	5974.391
1121.271	14734.581	19385.329

Tabla 8.Intervalos de confianzas para el promedio de las predicciones de las ventas diarias.

Des. estd sobre la pred. (Observación)	Límite inferior 95% (Observación)	Límite superior 95% (Observación)	Pred. ajustada
2923.594	-3345.146	8781.182	2590.549
2889.325	810.128	12794.313	6882.518
3021.593	16109.362	28642.164	21523.077
2884.576	11728.783	23693.273	17914.388
2839.383	790.156	12567.195	6677.566
2880.475	7268.870	19216.347	13495.432
2859.335	-632.522	11227.274	5276.975
2899.695	13191.430	25218.627	19974.323
3030.467	4314.091	16883.699	10491.029
2871.266	1517.403	13426.685	7544.905
3009.927	15024.789	27509.202	21801.792
2914.480	11749.874	23838.395	18248.533
3093.953	18658.755	31491.685	22677.919
2828.049	2932.585	14662.616	8814.414
2869.395	2546.943	14448.466	8438.913
2904.379	8603.532	20650.159	14505.885
2832.518	1229.555	12978.122	7201.944
2838.884	2701.232	14476.204	8582.958
2821.007	4590.883	16291.705	10457.241
2860.252	-963.405	10900.193	5013.071
3259.880	10045.397	23566.553	14240.323
2860.382	6566.731	18430.868	12634.705
2856.027	-760.557	11085.516	5177.882
2982.003	-6583.499	5785.092	-863.772
2876.952	-1695.226	10237.640	3965.008
2976.514	10887.042	23232.868	17059.955

Tabla 9. Intervalos de confianzas para la predicción de los datos modelados y ajustados.

Las tablas anteriores suministran información acerca de los parámetros que se deben tener en cuenta en el momento de tomar decisiones, basados en los pronósticos hechos a partir del análisis de regresión lineal múltiple.

.

4. CONCLUSIONES Y RECOMENDACIONES

- El análisis de resultados obtenido a partir del software XLSTAT 2010 y el solver de Excel 2007, Permitió evaluar la robustez de un modelo de regresión múltiple, usado para predecir de las ventas diarias de un departamento de un hipermercado de la ciudad de Pereira. El modelo presentó buen ajuste, explicando el 86% de la variabilidad de las ventas en función de las variables regresoras.

- En la evaluación del modelo, se pudieron validar los supuestos del modelo, lo que corrobora que si es adecuado predecir las ventas diarias a partir de un modelo de regresión lineal múltiple.

- El análisis de los resultados mostró que las variables que resultan ser significativas para predecir las ventas son facturas y Clientes, descartando la variable Ticket.

- La variable que mayor correlación presenta con las venta es la variable cliente con un porcentaje del 92%. Lo que podría considerarse solo hacer uso de esta variable en un modelo de regresión simple y evaluar que tan robusto resulta este modelo para predecir las ventas diarias.

- Los resultados de este análisis que comprobaron que el modelo de regresión lineal múltiple si resulta ser robusto para predecir las ventas diarias, podría compararse con un modelo de series de tiempo con el fin de comprobar que modelo resulta ser más adecuado para este propósito.

5. BIBLIOGRAFÍA

[1], Walpole, Myers. 2007 Probabilidad y Estadística para Ingeniería y Ciencia

[2], Estadística Para Administración Y Economía Anderson 2009 ISBN: 9687529415

[3], Mendenhall. Probabilidad y estadística para ingenieros. Capítulos 11, 12, 13 y 14

[4], Walpole, Myers Myers ye, Probabilidad y estadística para ingeniería y ciencia, 2007, p. 389-402

[5] curso Métodos De Regresión, ofrecido por el Departamento de Estadística de la Universidad Nacional. Disponible en http://www.virtual.unal.edu.co/cursos/ciencias/2007315/index.html

[6], N. Nagelkerke, "una nota sobre una definición general del coeficiente de determinación," Biometrika, vol. 78, no. 3, pp. 691-692, 1991.

[7], Draper, N.R. y Smith, H. (1998). *Análisis aplicado de la regresión*. Wiley-Interscience. ISBN 0-471-17082-8

[8], Pronósticos en los negocios. John E Hanke Arthur G Reitsch