IDENTIFICACIÓN DE VARIABLES ECONÓMICAS QUE INCIDEN EN EL PRECIO DE LA LECHE EN EL ESTADO DE CHIHUAHUA

POR:

ING. ARACELI ALVARADO HOLGUÍN

Tesina presentada como requisito parcial para obtener el grado de Maestría Profesional en Estadística Aplicada

Universidad Autónoma de Chihuahua

Facultad de Zootecnia y Ecología

Secretaría de Investigación y Posgrado

Chihuahua, Chih., México **Octubre de 2015**

Identificación de variables económicas que inciden en el precio de la leche en el estado de Chihuahua. Tesina presentada por Araceli Alvarado Holguín como requisito parcial para obtener el grado de Maestría Profesional en Estadística Aplicada, ha sido aprobada y aceptada por:

M. A. Luis Raúl Escárcega Preciado
Director de la Facultad de Zootecnia y Ecología

M. C. Antonio Humberto Chávez Silva
Secretario de Investigación y Posgrado

D. Ph. Pablo Fidel Mancillas Flores
Coordinador Académico

Joel

D. Ph. Joel Domínguez Viveros
Presidente

DICIEMBRE 04-2015

Fecha

Comité:

D. Ph. Joel Domínguez Viveros
Dr. Juan Ángel Ortega Gutiérrez
Dr. Nicolás Callejas Juárez

Araceli Alvarado Holguín
PERIFÉRICO FRANCISCO R. ALMADA KM. 1, CHIHUAHUA, CHIH., MÉXICO C.P. 31453

DICIEMBRE 2015

AGRADECIMIENTOS

Por sobre todos agradezco al único y verdadero Dios, al autor y consumador de mi fe, por su amor y fidelidad expresados en mi vida.

Agradezco a mis hijos a quienes he pretendido dejar una herencia espiritual y un ejemplo de vida para no rendirse ante los obstáculos que se presentan, sino luchar hasta alcanzar su propósito y destino.

A mis amigos, gracias por sus palabras de aliento cuando surgieron los deseos de abandonar la tarea y tirar la toalla.

A mis maestros y compañeros que me tuvieron paciencia y colaboraron entusiastamente en ésta labor.

Agradezco a la maestra M. E. María Teresa Pérez Piñón por haberme impulsado para iniciar ésta maestría, al D. Ph. Joel Domínguez Viveros por su paciencia y colaboración en la realización de éste trabajo, a las autoridades de mi alma máter, que me brindaron su apoyo y comprensión para realizar mis estudios: al M. C. Javier Martínez Nevárez y al M. A. Luis Raúl Escárcega Preciado; a todos, muchas gracias por contribuir a mi desarrollo profesional.

DEDICATORIA

Éste trabajo lo dedico a mis padres que me inculcaron el amor al conocimiento, por su entrega y su cariño. A mis hijos Isaí e Israel, que son mi tesoro más preciado.

CURRICULUM VITAE

La autora nació el 26 de enero de 1964 en la ciudad de Parral, Chihuahua, México.

1982-1986	Estudios de Licenciatura en la Facultad de Zootecnia de la Universidad Autónoma de Chihuahua obteniendo el título de Ingeniero Zootecnista en Sistemas de Producción en el año 2002.
2009-2011	Estudiante graduada del Programa de Maestría Profesional en Estadística Aplicada. Facultad de Zootecnia y Ecología de la Universidad Autónoma de Chihuahua. Chihuahua, México.
2005-2015	Docente en la Facultad de Enfermería y Nutriología de la Universidad Autónoma de Chihuahua. Colaborando en la impartición de la asignatura de bioestadística, en programas de Maestría, Licenciatura, Nivelatorio y Postécnico.

RESUMEN

IDENTIFICACIÓN DE VARIABLES ECONÓMICAS QUE INCIDEN EN EL PRECIO DE LA LECHE EN EL ESTADO DE CHIHUAHUA

POR:

ING. ARACELI ALVARADO HOLGUÍN

Maestría Profesional en Estadística Aplicada

Secretaría de Investigación y Posgrado

Facultad de Zootecnia y Ecología

Universidad Autónoma de Chihuahua

Presidente: D. Ph. Joel Domínguez Viveros

En el ámbito agropecuario, los precios de insumos y productos relacionados pueden incidir en el comportamiento a futuro del precio de la leche al productor. El objetivo fue identificar aquellos productos agropecuarios cuyos precios inciden en el comportamiento del precio de la leche de vaca en el estado de Chihuahua. Del Sistema de Información Agropecuaria (SIAP) se conformó una base de datos con el comportamiento histórico (1980 a 2011). Se realizaron tres análisis estadísticos: la estimación de correlaciones de Pearson ($\mathbf{r_e}$) entre los precios de los productos, un análisis de componentes principales para identificar las variables dependientes y el patrón de asociación de estas con el precio de la leche mediante una regresión múltiple. Los análisis se realizaron con el programa SAS. Todas las correlaciones de Pearson (r_e) fueron diferentes de cero ($P < 0.01$) y positivas, con estimaciones de 0.49 a 0.99 y media de 0.85. La ecuación estimada de regresión múltiple quedó conformada por diez productos, de ellos

siete son insumos y tres productos relacionados. En el análisis de componentes principales cuatro eigenvalores explican la variabilidad total.

ABSTRACT

IDENTIFICATION OF ECONOMIC VARIABLES AFFECTING THE PRICE OF MILK IN THE ESTATE OF CHIHUAHUA

BY:

ARACELI ALVARADO HOLGUÍN

In the agricultural field, the prices of different dairy products or inputs can affect the behavior of milk sale prices. The research objectives were to identify agricultural products that affect the behavior of prices. From Sistema de Information Agropecuaria (SIAP) a database of historic behavior (1980-2011) was gathered. Three statistical analyzes were performed; estimation of Pearson correlations ($\mathbf{r_e}$), multiple regression, and an analysis of the principal components to reduce the dimension of the dependent variables structure and to identify a pattern of association. Analyzes were performed with SAS statistical analysis program. All $\mathbf{r_e}$ was different than zero ($P < 0.01$) and positive, with estimates of 0.49 to 0.99 and average of 0.85. The results of multiple regression indicated that the final model was composed of ten products; seven were inputs and three related products. In the principal component analysis four eigenvalues explicate the total variability.

CONTENIDO

	Página
RESUMEN…………………………………………………………………	x
ABSTRACT………………………………………………………………	xii
LISTA DE CUADROS…………………………………………………..	xiii
LISTA DE GRÁFICAS…………………………………………………..	xiv
INTRODUCCIÓN………………………………………………………..	1
REVISIÓN DE LITERATURA…………………………………………..	9
Regresión Lineal…………………………………………………	9
Análisis de la Varianza (ANOVA)……………………………….	11
Coeficiente de Determinación……………………………………..	15
Regresión Lineal Múltiple………………………………………..	16
Cuadro del Análisis de la Varianza……………………………..	18
Análisis de Componentes Principales………………………….	20
MATERIALES Y MÉTODOS…………………………………………...	24
Descripción de Estudio…………………………………………..	24
Datos Experimentales…………………………………………....	24
Análisis Estadístico………………………………………………	24
RESULTADOS Y DISCUSIÓN………………………………………...	26
CONCLUSIONES Y RECOMENDACIONES……..……………………..	24
LITERATURA CITADA……………………………………………….	25

LISTA DE CUADROS

Cuadro		Página
1	Insumos utilizados en la producción estabulada de leche en el estado de Chihuahua, 2012 (cifras dadas en porcentaje)……………………………………………………….	3
2	Componentes del análisis de varianza de la regresión lineal simple………………………………………………………………..	14
3	Componentes del análisis de varianza de la regresión lineal múltiple……………………………..	19
4	Componentes del análisis de varianza de la regresión lineal múltiple considerando la suma de cuadrados reducidos a través de los coeficientes de regresión……………………….	21
5	Estructura de los eigenvalores y eigenvectores producto del análisis de los componentes principales……………………….	28

LISTA DE GRÁFICAS

Gráfica		Página
1	Comportamiento a través del tiempo de las variables generadas con los eigenvectores uno (EG1) y dos (EG2) producto del análisis de componentes principales………..	29

INTRODUCCIÓN

La producción de leche en México se desarrolla en condiciones diversas desde el punto de vista tecnológico, agroecológico y socioeconómico; dado la versatilidad de condiciones climatológicas, se definen características regionales matizadas por la tradición y costumbres de la población. Para el desarrollo de la producción lechera en México, se requiere mejorar el manejo alimenticio y reproductivo, utilizar razas o recursos genéticos que se ajusten a las condiciones agroecológicas, de mercado y de disponibilidad de recursos forrajeros; así como, el cumplimiento de programas sanitarios para el control de enfermedades; se debe impulsar la adopción de nuevas técnicas y sistemas de manejo con el fin de mejorar su eficiencia y productividad (Escalante y Catalán, 2008; CONARGEN, 2010).

El registro ordenado de los eventos que ocurren en la unidad de producción, particularmente de las características de interés económico, es fundamental para que el ganadero diagnostique su situación actual (y a futuro) en cuanto a producción, inversiones y rentabilidad. Los registros son útiles para analizar los resultados técnico-económicos y proporcionan un medio de control para mejorar la eficiencia administrativa de las empresas agropecuarias. La situación económica y rentabilidad de las unidades de producción de bovinos para leche esta defina por los ingresos, a través del comportamiento de los precios de venta del producto; y los egresos, en función del costo de los insumos (Villegas, 1995; Vargas-Leitón *et al.,* 2012).

La toma de decisiones para la administración se fundamenta en la calidad y cantidad de información que se genera en los registros del control de producción; así como en los registros históricos de los precios de venta y costos de los insumos. Se debe de conformar una base de datos que describa los niveles y potenciales productivos; además, de la evolución del mercado y los proveedores de insumos. Como complemento en el contexto agropecuario, se debe de considerar el comportamiento (en precios e insumos) de otros productos de origen animal o vegetal que afectan indirectamente la rentabilidad de la producción de leche. Bases de datos con información completa y de calidad permiten realizar estudios desde el punto de vista retrospectivo y prospectivo.

En México, los insumos agrícolas más importantes utilizados en la producción de leche son el maíz grano, sorgo, trigo, semilla de algodón, soya, heno de alfalfa y ensilaje de maíz (Ochoa *et. al.,* 1998). La producción agropecuaria de Chihuahua compite por el uso de recursos y el precio al productor está definido por el precio de los insumos y productos (sustitutos y complementarios) relacionados, mismos que se cultivan o producen regionalmente; los principales insumos agrícolas de uso en Chihuahua han sido: alfalfa seca, concentrado, grano rolado y semilla de algodón (Cuadro 1).

La rentabilidad de los sistemas de producción es definida por el costo de producción, a través del comportamiento de los precios de venta de la leche y de los insumos. Sin embargo, los precios de productos relacionados pueden incidir en la variación de los precios de venta de la leche; la toma de decisiones

Cuadro 1. Insumos utilizados en la producción estabulada de leche en el estado de Chihuahua, 2012 (cifras dadas en porcentaje)

	Alta	Media	Baja	Secas	Alta	Secas
Insumo	800 vacas				50 vacas	
Silo de maíz	3.2	32.7	21.4	7.2	45.3	82.9
Triticale verde	3.5			4.8		
Alfalfa seca	11.3				23.8	
Alfalfa verde	2.9	32.7	12.8	4.3		
Alfalfa de primera	9	10.2	19.8	21.3		
Alfalfa de tercera						16.6
Concentrado	28.9	14.3	26.5	20.4	26.4	0.5
Grano rolado	21.4	4.1	7.6	20.4		
Semilla algodón	19.9	6.1	12	21.6		

Fuente: Caracterización Socioeconómica de Unidades Representativas de Producción Agropecuarias y Forestales (URPAF) del estado de Chihuahua para la administración de la producción se fundamenta en los registros del control de producción, en los registros históricos de los precios de venta y costos de los insumos, así como de otros productos de origen animal o vegetal que afectan indirectamente la rentabilidad de la producción de leche (CONARGEN, 2010).

Con esta información se pueden realizar estudios retrospectivos y prospectivos. Estudios retrospectivos permiten conocer la evolución de los sistemas de producción e identificar aquellos factores que influyen de manera positiva y/o negativa. Por otro lado, análisis prospectivos permiten predecir con cierto grado de certeza los factores que influyen en los sistemas de producción y/o realizar proyecciones a futuro de la viabilidad y/o rentabilidad económica (Vargas-Leitón y Cuevas-Abrego, 2009). Con base en lo anterior, el objetivo fue identificar y analizar los precios de los productos agropecuarios (relacionados o no con la producción de leche) que pueden incidir en la proyección de los precios de la leche. Los resultados del presente estudio serán de utilidad para realizar estudios prospectivos a través de series de tiempo.

REVISIÓN DE LITERATURA

Regresión Lineal

El objetivo de la regresión lineal es describir la relación entre dos variables, X e Y; donde X se denomina independiente, e Y se determina como dependiente y tiene asociado un error de medición. Cada observación (Y_i) se puede expresar en términos de la función: $Y_i = \beta_0 + \beta_{Y*X} X_i + \varepsilon_i$; donde β_0 representa el intercepto, el punto donde se intercepta la línea recta con el eje de las Y's, esto es, el valor (la mayoría de las veces hipotético) de Y cuando X vale 0; β_{Y*X}, es la pendiente de la línea de regresión, indica cuanto cambia el valor de Y por cada unidad de cambio en X. Se le conoce como el coeficiente de regresión de Y en X, ó simplemente como el coeficiente de regresión. ε_i es un término de error aleatorio con media cero, independiente de X_i ó la covarianza (ε_i, X_i) = 0, y la varianza de (ε_i) = σ^2 (la misma para cualquier ε_i). Dado que β_0 y β_{Y*X} son parámetros, la media de Y_i cuando se conoce X_i, es $\beta_0 + \beta_{Y*X} X_i$; esto es, la media de Y para cierto valor de X cae sobre la línea de regresión. La varianza de Y_i es igual a la varianza de ε_i, la desviación aleatoria de la línea de regresión ó media (Montgomery *et al.,* 2006; Acuña, 2011).Al modelar Y_i, la suposición es que sigue una distribución normal.

En lo práctico, el problema consiste básicamente en tomar una muestra o un conjunto de datos y "ajustar" una línea recta a los mismos. Para hacer esto, se requiere ciertos criterios que señalen qué línea se ajusta a los datos mejor que cualquier otra línea; es necesario desarrollar un método que cumpla con dichos criterios. Uno de los criterios que frecuentemente se considera es que las estimaciones de β_0 y β_{Y*X} sean insesgadas; esto es, al estimar el intercepto y la pendiente, en promedio las estimaciones deben de ser igual a los parámetros. Un segundo criterio que se considera es el de varianza mínima; entre todos los estimadores insesgados, se desea encontrar aquellos con la varianza más pequeña (Tusell, 2008). El procedimiento de mínimos cuadrados proporciona estimadores insesgados de varianza mínima para el intercepto y la pendiente en el análisis de regresión, por lo que es el método que se desarrolla. El método de cuadrados mínimos selecciona las estimaciones de β_0 y β_{Y*X} de tal manera que la suma de las distancias verticales, al cuadrado, de las observaciones originales a la línea de regresión sea mínima (Tusell, 2008; Acuña, 2011).

Los estimadores de los cuadrados mínimos se van a representar cómo b_0 y b_{Y*X}, ó $\hat{\beta}_0$ y $\hat{\beta}_{y*x}$; de esta manera, la ecuación de predicción ó línea de regresión está dada por $\hat{Y}_i = b_0 + b_{Y*X} X_i$ que predice o estima a $\beta_0 + \beta_{Y*X}X_i$. Por lo que $Y_i - \hat{Y}_i = e_i$ estima el término del error, ε_i. El método de los cuadrados mínimos escoge a b_0 y b_{Y*X} de tal manera que $\Sigma(Y_i - b_0 - b_{Y*X}X_i)^2 = \Sigma e_i^2$ sea mínima. Los estimadores de los cuadrados mínimos se encuentran matemáticamente derivando Σe_i^2 con respecto a b_0, y luego con respecto a b_{Y*X} e igualando las dos ecuaciones resultantes a cero. Esto produce lo que se conoce como las “ecuaciones normales”: a) $\Sigma Y_i = nb_0 + b_{Y*X}\Sigma X_i$; y, b) $\Sigma X_iY_i = b_0\Sigma X_i + b_{Y*X}\Sigma X_i^2$. Resolviendo para los estimadores de los parámetros (Montgomery y Runger, 2002; Montgomery *et al.,* 2006):

$$b_{Y*X} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_{Y*X}\bar{X}$$

Algunas de las propiedades de las ecuaciones normales son: a) $\Sigma e_i = 0$; b) $\Sigma X_i e_i = 0$; c) $\Sigma \hat{Y}_i e_i = 0$; d) Σe_i^2 es mínima; e) $\Sigma Y_i = \Sigma \hat{Y}_i$; y, f) la línea de regresión pasa por el punto ($\bar{X}, \bar{Y}$). Una estimación insesgada de σ^2 es Σe_i^2/ (n-2). Donde, n-2 son los grados de libertad asociados con la estimación de la varianza, ya que se pierde un grado de libertad al estimar β_0 y otro al estimar β_{Y*X}.

Análisis de la Varianza (ANOVA)

Con el fin de determinar la relación que existe entre dos variables (X e Y), lo que se hace con el análisis de regresión es representar dicha relación en base a un modelo.Para el caso de la regresión lineal, el modelo usado es:

$Y_i = \beta_0 + \beta_{Y*X}X_i + \varepsilon_i$ (i= 1, 2, . . . , n = Número de observaciones); donde, $E(Y_i) = \beta_0 + \beta_{Y*X}X_i$; $E(\varepsilon_i) = 0$; $Var(Y_i) = Var(\varepsilon_i) = \sigma^2$; $\varepsilon_i NID(0, \sigma^2)$; además de la ecuación se debe definir el primer (la esperanza o media) y segundo (la varianza) momento de la distribución de las variables del modelo; no se conocen los parámetros del modelo, sino que se estiman mediante el conjunto de datos que se analizan; con base en las estimaciones se hacen inferencias acerca de los parámetros, entre las cuales se tienen las pruebas de hipótesis y los intervalos de confianza (Harrell, 2002; Montgomery y Runger, 2002).

Las pruebas de hipótesis se realizan con base en pruebas de F; a partir del ANOVA, permite identificar las diferentes fuentes de variación para la variable de interés; permite identificar los factores que están afectando la variable en estudio. En el análisis de regresión, un objetivo es predecir valores de la variable dependiente (Y) basándose en valores de la variable independiente (X) que se conocen; si no conocen los valores de la variable X, la mejor predicción para la variable Y sería su media ($\bar{Y}$); asimismo, si no existiera ninguna relación entre X e Y, entonces $\beta_{Y*X} = 0$ y la media de Y sería la misma para cualquier valor de X, en este caso β_0. Sin embargo, cuando $\beta_{Y*X} \neq 0$, la media de Y es diferente para cada valor de X, a partir del planteamiento $\mu_{Y|Xi} = \beta_0 + \beta_{Y*X}X_i$. En este caso, la predicción que se obtiene con la regresión $\hat{Y}_i = b_0 + b_{Y*X} X_i$, es mejor que la predicción con $\bar{Y}$, la media general sin considerar X. Por lo tanto, el objetivo principal es saber si realmente la regresión de Y en X, ayuda a predecir mejor el valor de Y, en comparación a utilizar únicamente $\bar{Y}$ como predictor; en términos de hipótesis el planteamiento es: H_o: $\beta_{Y*X} = 0$ vs H_a: $\beta_{Y*X} \neq 0$.

La base del ANOVA es dividir la suma de cuadrados total en sus diferentes componentes, de acuerdo a los factores considerados en el modelo. Esto es, $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$; donde, desviación total = desviación debido al error + desviación debido a la regresión. En términos de suma de cuadrados, la suma de cuadrados total, o suma de cuadrados alrededor de la media, está dada por la suma de cuadrados de la regresión y la suma de cuadrados del error o suma de cuadrados alrededor de la regresión; esto es, $\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$. La demostración se obtiene como sigue:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2$$

$$= \Sigma[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2$$

$$= \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2 + 2\Sigma(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

De esta manera, lo que se debe demostrar es que $2\Sigma(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 0$; a partir de $\hat{Y}_i = b_0 + b_{Y*X} X_i$, y sustituyendo $b_0 = \bar{Y} - b_{Y*X} \bar{X}$,

$$\hat{Y}_i = \bar{Y} - b_{Y*X}\bar{X} + b_{Y*X}X_i = \bar{Y} + b_{Y*X}\left(X_i - \bar{X}\right).$$

De esta manera, $\hat{Y}_i - \bar{Y} = \bar{Y} + b_{Y*X}\left(X_i - \bar{X}\right) - \bar{Y} = b_{Y*X}\left(X_i - \bar{X}\right)$.

Así mismo, $Y_i - \hat{Y}_i = Y_i - \bar{Y} - b_{Y*X}\left(X_i - \bar{X}\right)$.

Por lo tanto, $\Sigma(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \Sigma\left[b_{Y*X}\left(X_i - \bar{X}\right)\right]\left[\left(Y_i - \bar{Y}\right) - b_{Y*X}\left(X_i - \bar{X}\right)\right]$

$= b_{Y*X}\left[\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - b_{Y*X}\Sigma\left(X_i - \bar{X}\right)^2\right]$ y puesto que

$$b_{Y*X} = \Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \Big/ \Sigma\left(X_i - \bar{X}\right)^2$$

$$\begin{aligned}
&= b_{Y*X}\left[\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - \left\{\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \Big/ \Sigma\left(X_i - \bar{X}\right)^2\right\}\Sigma\left(X_i - \bar{X}\right)^2\right] \\
&= b_{Y*X}\left[\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - \Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right] = b_{Y*X}(0) \\
&= 0
\end{aligned}$$

En el Cuadro 2 se presentan los componentes del ANOVA. Con base en los componentes del ANOVA la regla de decisión en la prueba de hipótesis si F calculada > F tablas se rechaza Ho; o, si F calculada < F tablas se acepta Ho.

Coeficiente de Determinación

Para investigar si la ecuación de regresión realmente ayuda en la predicción de Y, no únicamente se realiza probando si β_{Y*X} es diferente o igual a

Cuadro 2. Componentes del análisis de varianza de la regresión lineal simple

F.V.	g. l	S.C.	C.M.	Fcal.	F tab.
Regresión	1	$SCR = \Sigma\left(Y_i - \hat{Y}_i\right)^2$	CMR=SCR/1	CMR/CME	F_1
Error	n-2	$SCE = \Sigma\left(Y_i - \hat{Y}_i\right)^2$	CME = SCE/n-2		
Total	n-1	$SCT = \Sigma\left(Y_i - \underline{Y}\right)^2$			

F.V. = fuente de variación; g.l = grados de libertad; S.C. = suma de cuadrados; C.M. = cuadrados medios; Fcal = valor de F calculado; Ftab = valor de F tabular, considerando los grados de libertad del numerador, los grados de libertad del denominador, así como un valor de probabilidad (α); SCR = suma de cuadrados de la regresión; SCE = suma de cuadrados del error; SCT = suma de cuadrados total; CMR = cuadrado medio de la regresión; CME = cuadrado medio del error.

cero, mediante el ANOVA; la prueba de F señala si la relación entre X e Y es estadísticamente importante o no, pero no dice en qué proporción. Al analizar un conjunto de datos, tan grande que al disminuir el error estándar de estimación de β_{Y*X}, la prueba de F va a resultar significativa, aun cuando desde el punto de vista práctico esa relación no es importante. Un estadístico de ayuda al respecto es el coeficiente de determinación, el cual se estima como:

$$R^2 = \frac{SCR}{SCT} = \frac{\Sigma\left(\hat{Y}_i - \bar{Y}\right)^2}{\Sigma\left(Y_i - \bar{Y}\right)^2}$$

Dónde: SCR = suma de cuadrados del residual, y SCT = suma de cuadrados total. El coeficiente de determinación (R^2) indica que proporción de la variación total en la variable Y está explicada por la regresión; puede tomar valores entre 0 y 1, con 1 indicando que toda la variación en Y está explicada por la regresión. Otra forma de calcular R^2 en el caso de la regresión lineal es elevando al cuadrado el coeficiente de correlación (R). Cuando el interés radica en obtener la ecuación de regresión para utilizarla en predicciones futuras de Y a partir de valores conocidos de X, R^2 es uno de los aspectos más importantes a considerar, pues para que la ecuación sea de utilidad esta deberá de explicar una buena proporción de la variación en Y. Por otra parte, cuando únicamente interesa saber qué factores son los que influyen sobre la variable de interés, la prueba de F del análisis de varianza es la principal pieza de información a considerar.

Regresión Lineal Múltiple

La diferencia de la regresión lineal múltiple en comparación con la regresión lineal simple se debe a que la regresión lineal múltiple tiene más de una variable independiente en el modelo, y por lo tanto, más de un coeficiente de regresión. Por ejemplo, con dos variables independientes el modelo quedaría como sigue: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$; donde: $i = 1, 2, \ldots, n$; y_i es el valor de la variable dependiente para la i-ésima observación; x_{1i} y x_{2i} son los valores para la i-ésima observación de las variables independientes x_1 y x_2, respectivamente; β_0 es el intercepto; β_1 es el coeficiente de regresión de la variable respuesta en la variable x_1 para un valor fijo de la variable x_2; β_2 es el coeficiente de regresión de la variable respuesta en la variable x_2 para un valor fijo de la variable x_1; e_i es el error aleatorio para la i-ésima observación. Las suposiciones en el modelo de regresión lineal múltiple son las mismas que para la regresión lineal simple. Esto es, dado que se conoce x_{1i} y x_{2i}, $E(y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$; $E(e_i) = 0$; $Var(y_i) = Var(e_i) = \sigma^2$; $e_i \sim NID(0, \sigma^2)$. Para analizar el modelo de regresión lineal múltiple el planteamiento en forma matricial es: $y = X\beta + e$; $E(y) = X\beta$; $Var(y) = Var(e) = I\sigma^2$; $e \sim MVN(0, I\sigma^2)$; donde:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_{1n} & x_{2n} \end{bmatrix} \quad \text{y} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Lo anterior se puede generalizar para cualquier número de variables independientes; al tener m variables independientes,

$$X = \begin{bmatrix} \mathbf{1} & \mathbf{x_{11}} & . & . & . & \mathbf{x_{m1}} \\ \mathbf{1} & \mathbf{x_{12}} & . & . & . & \mathbf{x_{m2}} \\ \mathbf{1} & \mathbf{x_{13}} & . & . & . & \mathbf{x_{m3}} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \mathbf{1} & \mathbf{x_{1n}} & . & . & . & \mathbf{x_{mn}} \end{bmatrix} \quad \text{y} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_m \end{bmatrix}$$

Una vez que se han definido las matrices y los vectores correspondientes, los cálculos para estimar los parámetros son: $X'X\hat{\beta} = X'y$ y $\hat{\beta} = (X'X)^{-1}X'y$ desarrollando las ecuaciones de cuadrados mínimos, éstas toman la forma:

$$\begin{bmatrix} n & \Sigma x_{1i} & \Sigma x_{2i} & . & . & . & \Sigma x_{mi} \\ \Sigma x_{1i} & \Sigma x_{1i}^2 & \Sigma x_{1i}x_{2i} & . & . & . & \Sigma x_{1i}x_{mi} \\ \Sigma x_{2i} & \Sigma x_{1i}x_{2i} & \Sigma x_{2i}^2 & . & . & . & \Sigma x_{2i}x_{mi} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \Sigma x_{mi} & \Sigma x_{1i}x_{mi} & \Sigma x_{2i}x_{mi} & . & . & . & \Sigma x_{mi}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ . \\ \hat{\beta}_m \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{1i}y_i \\ \Sigma x_{2i}y_i \\ . \\ . \\ . \\ \Sigma x_{mi}y_i \end{bmatrix}$$

Cuadro del Análisis de la Varianza

Los cálculos para el cuadro del análisis de la varianza son similares que para la regresión lineal simple (Cuadro 3). Esto quiere decir que en la prueba de hipótesis se está probando la significancia de más de un parámetro al mismo tiempo. Las hipótesis ahora son: Ho: $\beta_1 = \beta_2 = . . . = \beta_m = 0$; y, Ha: Al menos un coeficiente de regresión, ajustado por el resto de las variables independientes, es diferente de cero.

Una prueba de F significativa señala que al menos un coeficiente de

Cuadro 3. Componentes del análisis de varianza de la regresión lineal múltiple

F. V.	g. l	S. C.	C. M.	Fcal	F tab
Regresión	p - 1	β'x'y - n $\bar{y}^2$	SCR/p – 1	CMR/CME	$F^{p-1}_{n-p,\alpha}$
Error	n - p	y'y - β'X'y	SCE/n – p		
Total	n - 1	y'y - n $\bar{y}^2$			

F.V. = fuente de variación; g.l = grados de libertad; S.C. = suma de cuadrados; C.M. = cuadrados medios; Fcal = valor de F calculado; Ftab = valor de F tabular, considerando los grados de libertad del numerador, los grados de libertad del denominador, así como un valor de probabilidad (α); SCR = suma de cuadrados de la regresión; SCE = suma de cuadrados del error; SCT = suma de cuadrados total; CMR = cuadrado medio de la regresión; CME = cuadrado medio del error; p =número de parámetros en el modelo.

regresión es diferente de cero, pero no dice cual ni cuántos de ellos; se requiere desglosar los p - 1 grados de libertad de la regresión en grados de libertad individuales; esto es, probar las hipótesis: Ho: $\beta_1 = 0$, Ho: $\beta_2 = 0$,. . ., Ho: $\beta_m = 0$, desglosando la suma de cuadrados de la regresión en lo que se conoce como las sumas de cuadrados reducidas (SCR). Las sumas de cuadrados reducidas se obtienen por diferencia al restar las sumas de cuadrados de la regresión, o las sumas de cuadrados del error, para un modelo y un sub modelo de él. Por ejemplo, si se requiere saber cuánto contribuyó la variable x_2 a la suma de cuadrados de la regresión, además de lo que ya se había logrado con x_1, a la suma de cuadrados de la regresión, incluyendo las dos variables independientes, se le resta la suma de cuadrados de la regresión incluyendo únicamente x_1 en el modelo (SAS, 2001).

Esto es, SCR $(\beta_2|\beta_o, \beta_1)$ = SCR$(\beta_1, \beta_2|\beta_o)$ – SCR$(\beta_1|\beta_o)$; en otras palabras: suma de cuadrados para β_2 ajustado por β_0 y β_1 es igual a la suma de cuadrados para β_1 y β_2 ajustado por β_0 menos la suma de cuadrados para β_1 ajustado para β_0. Aquí, SCR$(\beta_1, \beta_2|\beta_o)$ es la suma de cuadrados de la regresión incluyendo en el modelo tanto a x_1 como a x_2, mientras que SCR$(\beta_1|\beta_o)$ es la suma de cuadrados de la regresión cuando únicamente se incluyó en el modelo a x_1. Asimismo, la suma de cuadrados reducida para β_1 se obtiene como: SCR $(\beta_1|\beta_o, \beta_2)$ = SCR$(\beta_1, \beta_2|\beta_o)$ – SCR$(\beta_2|\beta_o)$; de esta manera, el cuadro del análisis de la varianza queda como se describe en el Cuadro 4 (SAS, 2001).

Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica estadística

Cuadro 4. Componentes del análisis de varianza de la regresión lineal múltiple considerando la suma de cuadrados reducida a través de los coeficientes de regresión

F. V.	g.l	S. C.	Ho:
Regresión	2	$SCR(\beta_1, \beta_2 \vert \beta_o)$	$\beta_1 = \beta_1 = 0$
$SCR(\beta_1 \vert \beta_o, \beta_2)$	1	$SCR(\beta_1, \beta_2 \vert \beta_o)$ - $SCR(\beta_2 \vert \beta_o)$	$\beta_1 = 0$
$SCR(\beta_2 \vert \beta_o, \beta_1)$	1	$SCR(\beta_1, \beta_2 \vert \beta_o)$ - $SCR(\beta_1 \vert \beta_o)$	$\beta_2 = 0$
Error	n - 3	SCT - $SCR(\beta_1, \beta_2 \vert \beta_o)$	---
Total	n - 1	SCT	---

F.V. = fuente de variación; g.l = grados de libertad; S.C = suma de cuadrados; S.CR. = suma de cuadrados de la regresión; SCT = suma de cuadrados total; β_o = es el intercepto; β_1 = es el coeficiente de regresión para x_1;β_2 = es el coeficiente de regresión para x_2; Ho = es la hipótesis a probar de síntesis de la información, o reducción de la dimensión (número de variables). Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Los nuevos componentes principales o factores serán una combinación lineal de las variables originales, y además serán independientes entre sí. Un aspecto clave en ACP es la interpretación de los factores, ya que ésta no viene dada a priori, sino que será deducida tras observar la relación de los factores con las variables iníciales (Peña, 2002).

Fases de un análisis de componentes principales: a) análisis de la matriz de correlaciones; un análisis de componentes principales tiene sentido si existen altas correlaciones entre las variables, ya que esto es indicativo de que existe información redundante, y por tanto, pocos factores explicarán gran parte de la variabilidad total; b) selección de los factores; la elección de los factores se realiza de tal forma que el primero tome la mayor proporción posible de la variabilidad original; el segundo factor debe tomar la máxima variabilidad posible no acumulada por el primero, y así sucesivamente; del total de factores se elegirán aquéllos que acumulen el porcentaje de variabilidad que se considere suficiente, a éstos se les denominará componentes principales; c) análisis de la matriz factorial; una vez seleccionados los componentes principales, se representan en forma de matriz; cada elemento de ésta representa los coeficientes factoriales de las variables (las correlaciones entre las variables y los componentes principales); la matriz tendrá tantas columnas como componentes principales y tantas filas como variables.

Un problema central en el análisis de datos multivariantes es la reducción de la dimensionalidad: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. El ACP tiene este objetivo: dada n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un número menor de variables construidas como combinaciones lineales de las originales. Por ejemplo, con variables con alta dependencia es frecuente que un pequeño número de nuevas variables (menos del 20 % de las originales) expliquen la mayor parte (más del 80 %) de la variabilidad original. Su utilidad es doble: 1) permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p-dimensional; en este sentido componentes principales es el primer paso para identificar posibles variables ”latentes” o no observadas, que están generando la variabilidad de los datos; y, 2) permite transformar las variables originales, en general correlacionadas, en nuevas variables no correlacionadas, facilitando la interpretación de los datos.

MATERIALES Y MÉTODOS

Descripción de Estudio

La investigación forma parte del proyecto de investigación "Caracterización Socioeconómica de Unidades Representativas de Producción Agropecuarias y Forestales (URPAF) del estado de Chihuahua", con el objetivo de generar los pronósticos de precios de los productos relacionados con el precio de la leche producida en el estado de Chihuahua.

El estudio se desarrolló de acuerdo a las condiciones económicas y de productos agropecuarios del estado de Chihuahua. La información se obtuvo del Servicio de Investigación Agroalimentaria y Pesquera (SIAP, 2012), desarrollado por la Secretaria de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación.

Datos Experimentales

Se conformó una base de datos con el comportamiento histórico (de 1980 a 2011) del precio de la leche y de 23 productos agropecuarios: maíz de grano, maíz forrajero, alfalfa verde, algodón hueso, pastos, avena, sorgo, trigo forrajero, triticale, cebada grano, sorgo grano, avena grano, trigo, cebada, centeno forrajero en verde, sorgo escobero, soya, frijol, chile verde, manzana, nuez, bovinos en pie y bovinos en canal. El 65 % de estos correspondieron a insumos utilizados en la producción de leche y 35 % a productos que compiten por el uso de los recursos (sustitutos).

Análisis Estadístico

Se realizaron con el programa de análisis estadísticos SAS (SAS, 2001) tres análisis estadísticos: 1) se estimó el coeficiente de correlación de Pearson

(**r_e**) entre el precio de la leche y los precios de los 23 productos agropecuarios evaluados; 2) se realizó un análisis de regresión múltiple, planteando el precio de la leche como variable dependiente en función de los 23 productos agropecuarios como variables independientes para conformar un modelo final con aquellos productos agropecuarios con valores de probabilidad mayor o igual a 0.15, y que en su conjunto integraran un coeficiente de determinación (R^2) mayor o igual a 0.95; y, 3) con el modelo final producto de la regresión múltiple, se realizó un análisis de componentes principales para reducir la dimensión de la estructura de variables dependientes e identificar un patrón de asociación y/o clasificación.

RESULTADOS Y DISCUSIÓN

Todas las correlaciones fueron estadísticamente diferentes de cero (P<0.01) y positivas, con estimaciones de 0.49 a 0.99 y un promedio de 0.85. En general, el incremento del precio de la leche está asociado de manera positiva con el incremento de los precios de los 23 productos agropecuarios evaluados.

Los resultados de la regresión múltiple indicaron que el modelo final quedó conformado por diez productos: maíz grano (mg), cebada grano (cg), trigo grano (tg), sorgo grano (sg), avena forrajera (af), pastos (ps), triticale (tt), chile verde (cv), manzana (mz) y bovinos en canal (bc). El 70 % correspondieron a insumos utilizados en la producción de leche y 30 % a productos que compiten por el uso de los recursos disponibles en la entidad. El intercepto y los coeficientes de regresión para cada uno de los componentes del modelo final fueron:

$$P\hat{l} = 42.250 - 0.2176mg - 0.626af + 0.228ps - 0.702tt + 0.782cg + 0.460tg$$
$$- 0.395sg + 0.077cv + 0.056mz + 0.102bc$$

Con base en la estructura del modelo se puede observar que cuatro insumos (mg, af, tt y sg) muestran una relación inversa (P<0.15) en la determinación del precio de la leche, lo cual se puede atribuir a que son productos agrícolas utilizados como insumos en la producción de la misma. Dos insumos muestran una relación directa (ps y cg), atribuido a que son consumidos en una menor proporción de la dietas de las vacas lecheras.

Por otro lado, los productos ps, cg, tg, cv, mz y bc presentaron tendencias positivas (P<0.15), lo cual se puede atribuir a que son productos agropecuarios componentes del consumo básico y que su incremento repercute de manera positiva.

Del análisis de componentes principales, cuatro eigenvalores explicaron 97.2 % de la variabilidad total (Cuadro 5 y Gráfica 1). Con los eigenvectores y el histórico de precios de los diez productos agropecuarios se definieron cuatro variables. En la Gráfica 1 se presenta el comportamiento de la variables generadas con los eigenvectores uno y dos (EG1 y EG2) a través del tiempo. El EG1 representa el aumento de precios de todos los productos agropecuarios a través del tiempo, lo cual coincide con las correlaciones (positivas y de mediana a alta magnitud) que se obtienen en el presente estudio.

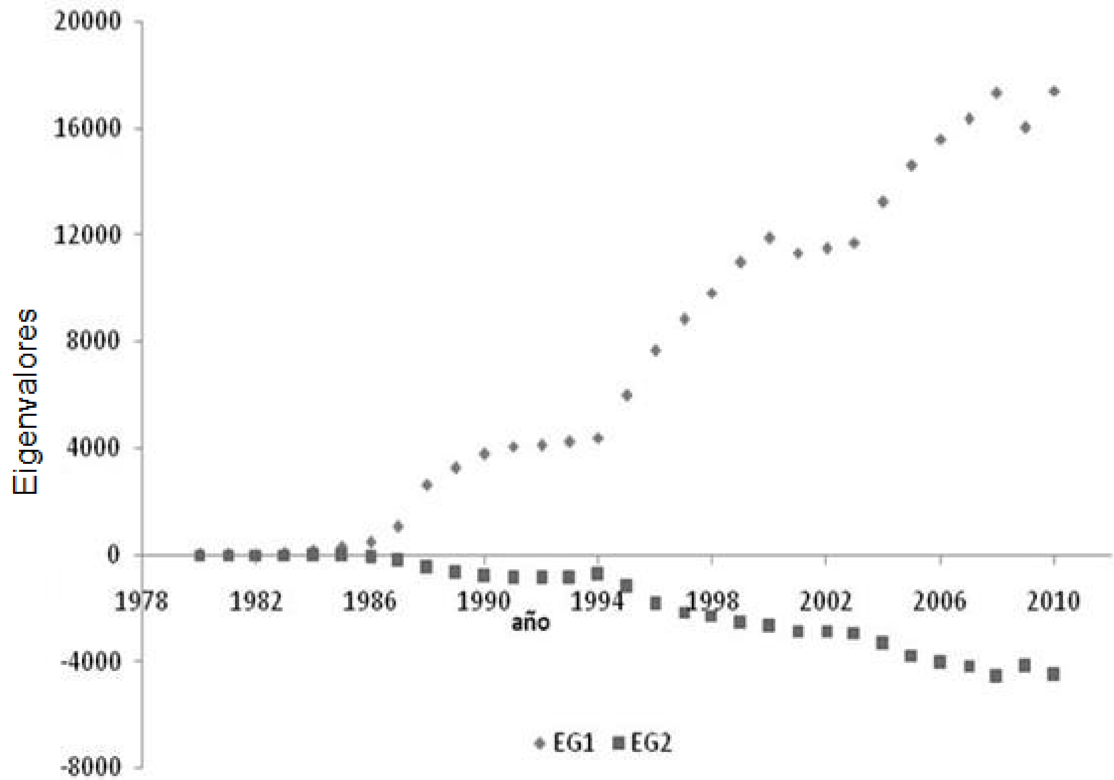
La estructura del eigen vector dos (Cuadro 5) y lo observado en el EG2 de la Gráfica 1 coincide con la estructura del modelo definido con los resultados del análisis de regresión múltiple, donde la avena (+0.956 EG2) muestra la mayor magnitud de los coeficientes de ponderación y con signo positivo, en sentido contrario a todos los demás; que presentaron coeficientes negativos ó positivos de baja magnitud, resaltando su importancia como insumo en los sistemas de producción del Estado.

Sin embargo la avena no aparece como un insumo prioritario en algunos de los sistemas de producción de leche de Chihuahua (Cuadro 5), dado que su uso puede estar relacionado con un ciclo climatológico de sequía de largo plazo (>10 años). Nótese que la avena tiene el coeficiente con menor magnitud en el EG1 (+0.21) aportando una indicación del porqué de su contribución al precio de la leche en el modelo de regresión multivariada resultó negativa con un valor de (-0.626). Por otro lado la alfalfa seca, verde y de primera es usada por los productores en establecimientos estabulados de 800 animales en los diferentes estados fisiológicos del hato, mientras que los establecimientos de menor tamaño

Cuadro 5. Estructura de los eigenvectores y eigenvalores producto del análisis de componentes principales

Producto	EG1§	EG2§	EG3§	EG4§
Maíz	0.331025	-0.08815	-0.33214	0.183164
Sorgo	0.333101	-0.15008	-0.25265	0.170413
Triticale	0.317036	0.061504	0.279470	-0.723000
Pastos	0.302544	0.057677	0.742212	0.481142
Manzana	0.326915	-0.07587	-0.05707	0.305927
Chile	0.332264	-0.07693	0.162833	-0.170530
Avena	0.207874	0.956237	-0.17691	0.049626
Cebada	0.327417	-0.08672	-0.110000	-0.22184
Carne de bovino	0.334253	-0.09147	0.089886	-0.08459
Trigo	0.328648	-0.14369	-0.340030	0.041052
Eigenvalor¶	0.857	0.924	0.953	0.972

§ = estructura de los eigenvectores uno a cuatro; ¶ = eigenvalores acumulados.



Gráfica 1. Comportamiento a través del tiempo de las variables generadas con los eigenvectores uno (EG1) y dos (EG2) producto del análisis de componentes principales.

(50 vacas) solo utilizan la alfalfa seca y exclusivamente para las vacas en condiciones de producción alta (Cuadro 5).

En cuanto a los precios de otros productos no lácteos que se puedan relacionar con el precio de la leche; productos como la manzana y el chile presentan coeficientes negativos en el EG2, en las zonas del estado donde cada uno se produce, compiten fuertemente por el recurso agua con los principales insumos de los sistemas de producción lecheros. Sin embargo, en el caso de la manzana, se están llevando a cabo investigaciones para mejorar los subproductos del proceso de extracción de jugos y así poder integrarlos a las dietas de los bovinos reduciendo los costos de alimentación del ganado en esa zona (Rodríguez-Muela *et al*., 2010).

CONCLUSIONES Y RECOMENDACIONES

En el estudio se identificaron 10 productos clave para ser utilizados en las proyecciones prospectivas de los precios de leche producida en Chihuahua. Con lo que se logró presentar herramientas que pueden ser usadas por otros investigadores y productores para realizar planeaciones y proyectos focalizados que permitan optimizar las decisiones y los resultados del Sistema Producto a futuro.

Se recomienda que en lo sucesivo se continúe este tipo de análisis con bases de datos actualizadas por periodos anuales y desagregar el estudio para cada una de las principales cuencas lecheras en Chihuahua, dado que por lo amplio del territorio presentan condiciones particulares que determinan los insumos clave y los productos no lácteos que compiten por los recursos naturales disponibles en cada zona. Además se puede ampliar a otros Sistemas Producto del estado de Chihuahua.

LITERATURA CITADA

Acuña, F. E. 2011. Análisis de regresión. Departamento de matemáticas. Universidad de Puerto Rico. U.S.A.

CONARGEN. 2010. Guía técnica de programas de control de producción y mejoramiento genético en bovinos lecheros. Consejo Nacional de los Recursos Genéticos Pecuarios.

Escalante, S. R. I. y H. Catalán. 2008. Situación actual del sector agropecuario en México: perspectivas y retos. Secretaria de Agricultura, Ganadería, Pesca, Desarrollo Rural y Alimentación.

Harrell, F. E. 2002. Biostatistical modeling. University of Virginia. Virginia, U.S.A.

Montgomery, D. C. y G. C. Runger.2002. Applied Statistics and Probability for Engineers. John Wiley & Sons, Inc. U.S.A.

Montgomery, D. C., E. A. Peck y G. G. Vining, 2006. Introducción al análisis de regresión lineal. 3ra ed. Editorial Continental. México.

Ochoa R. F., D. P. Anderson, J. O. Outlaw, J. W. Richardso, R. D. Knutson, R. B. Schwartz y J. W. Miller. 1998. Granjas lecheras representativas en México, panorama económico. AFPC Working Paper 98-10. Agricultural and Food Policy Center.

Peña, D. 2002. Análisis de datos multivariantes. Editorial McGraw Hill. México.

Rodríguez-Muela C., A. Becerra-Bernal, H. E. Rodríguez-Ramírez, D. Díaz-Plascencia, C. Hernández-Gómez, F. Gutiérrez-Piña, M. A. Gallegos-Acevedo, S. Romero-Villalobos y F. Lucero-Acosta. 2010. Valor nutricional de la manzarina, obtenida de subproductos de manzana para la alimentación animal. Tecnociencia Chihuahua. Vol IV, No. 3 Sep-Dic 2010.

SAS. 2001. SAS User´s Guide: Statistics (version 9.0) Cary NC, USA: SAS Inst. Inc.

SIAP. 2012. Servicio de Información Agroalimentaria y Pesquera. Fecha de consulta: junio de 2012 Disponible en: http://www.siap.gob.mx

Tusell, F. 2008. Análisis de regresión. Introducción teórica y práctica basada en Regresión. Editorial de la Universidad del País Vasco. España.

Vargas-Leitón, B. y M. Cuevas-Abrego. 2009. Modelo estocástico para estimación de valores económicos de rasgos productivos y funcionales en bovinos lecheros. Agrociencia 43:881-893.

Vargas-Leitón, B., Y. Marín-Marín y J. J. Romero-Zúñiga. 2012. Comparación bioeconómica de grupos raciales Holstein, Jersey y Holstein x Jersey en Costa Rica. Agronomía Mesoamericana 23:329-342.

Villegas, J. 1995. Cómo calcular el costo total de un litro de leche. Revista Frontera Agrícola 3:89-93.