Modelo de Regresión Lineal Múltiple de Porcentaje de Participación de Mujeres mayores de 16 años en Mano de Obra

Reyna Vargas Antonio

11 de junio de 2016

Resumen

El presente documento realiza un modelo de Regresión Lineal Múltiple de datos obtenidos de Gretl, los cuales fueron recolectados por Luis Cruz para conocer la Tasa de participación de Mujeres involucradas en laborar como Mano de Obra, los datos fueron obtenidos en el CENSO de 1990, con el objetivo de conocer que factores influyen para que mujeres de 16 años en adelante opten por el trabajo de Mano de Obra, así mismo se realiza una descripción de los Estadisticos de las variables, de tal forma que sean interpretados correctamente los datos al realizar el mejor Modelo de Regresón Lineal.

Índice

Índice de cuadros

1. Marco teórico

1.1. Del contexto de las variables

A continuación se hace mención de las variables implicadas en el Modelo de Regresión Lineal:

WLFP **(Variable endógena).** Porcentaje de Mujeres mayores de 16 años que trabajan en Mano de Obra.

YF: Mediana de Ingresos de Mujeres mayores de 15 años con Ingresos de 1989.

YM: Mediana de Ingresos de Hombres mayores de 15 años con Ingresos de 1989.

EDUC: Porcentaje de Mujeres mayores de 25 años graduadas de la secundaria.

UE: Porcentaje de desempleo en Mano de Obra

MR: Porcentaje de pobalción de mujeres casadas mayores de 15 años.

DR: Porcentaje de Mujeres divorciadas mayores de 15 años.

URB: Porcentaje de población que vive en zonas urbanas.

WH: Porcentaje de Mujeres blancas mayores de 16 años

Con las variables mencionadas se desea conocer que factores influyen en que la población femenil participe en empleos de Mano de Obra, además de que podemos conocer a que edad empiezan a laborar las mujeres y que es lo que denota a que opten por este tipo de trabajo.

1.2. Marco Teórico del Modelo

1.2.1. Análisis descriptivo de la variable

El objetivo del análisis descriptivo es resumir un conjunto de datos, extrayendo las características e información más relevante para el estudio. Las medidas de posición dan una idea de la situación o centro del conjunto de puntos. La ***media*** es el valor promedio. Si disponemos de N datos de una variable x_1,x_2,..., x_n, la media, o también momento muestral de primer orden, se define como:

$$\bar{x} = \frac{x_1+x_2+...x_N}{N} = \frac{1}{N}\sum x_i$$

La media es un estadístico poco robusto frente a la presencia de valores extremos: las observaciones anómalas van a tener una gran influencia en el valor que tome. En general, interesan estadísticos cuyo valor no varíe mucho ante cambios en los valores de unas pocas observaciones, por muy grandes que sean esas variaciones. La ***mediana***, que es el valor central de la distribución, posee esta propiedad.

Las medidas de posición proporcionan un valor representativo del conjunto de datos que debe complementarse con una medida del error asociado. Para valorar la representatividad de este único valor se utilizan las medidas de dispersión, que informan de si las observaciones están poco concentradas (o muy dispersas) alrededor de su centro. Por ejemplo, la desviación típica, que es la raíz cuadrada positiva de la varianza. La varianza de un conjunto de datos se define como un promedio de los cuadrados de las desviaciones de los datos a la media. Gretl calcula la varianza, S^{*2} o S_x^{*2} , como:

$$S_x^{*2} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...(x_N-\bar{x})^2}{N-1} = \frac{1}{N-1}\sum(x_i-\bar{x})^2$$

Por tanto, la desviaciónn típica, S_x^*se calcula según:

$$S_x^* = \sqrt{\frac{1}{N-1}\sum(x_i-\bar{x})^2}$$

Varianza y desviación típica son medidas de la dispersión de los datos alrededor de la media. Tiene el valor mínimo cero cuando todos los datos de la variable toman el mismo valor. La ventaja de la desviación típica es que tiene las mismas unidades de medida que la variable original. En general, cuanto más próxima a cero esté S_x^*, más concentrados estarán los datos alrededor de la media y ésta será más representativa del conjunto de observaciones. Sin embargo, al dependerS_x^* de las unidades de medida, no es fácil comparar su representatividad en dos conjuntos de datos. Para solucionar este problema se utiliza el coeficiente de variación, C.V.

1.2.2. Modelo de Regresión Lineal

El modelo simple relaciona dos variables de forma lineal,

$$Y_i = \alpha + \beta X_i + u_i,\ i = 1, ..., N$$

donde:

- Y es la variable a explicar, variable dependiente o endógena, es decir, la variable que estamos interesados en explicar.
- X es la variable explicativa, variable independiente o exógena.
- La ordenada α y la pendiente β del modelo son los coefcientes de la regresión. Si defnimos K como el núumero de coefcientes desconocidos a estimar, en el modelo de regresión simple tenemos $k = 2$ coefcientes a estimar.
- u es el término de error, variable aleatoria o perturbación.
- El subíndice i denota observación, será empleado cuando la muestra contenga datos de sección cruzada y el subíndice t cuando tengamos observaciones correspondientes a series temporales, aunque esto no es de especial relevancia.
- N es el tamaño muestral, número de observaciones disponibles de las variables de estudio (Y, X).

El error u_i se introduce por varias razones, entre las cuales tenemos:

1. Efectos impredecibles, originados por las características de la situación económica o del contexto de análisis, y efectos no cuantificables derivados de las preferencias y los gustos de los individuos o entidades económicas.
2. Errores de medida producidos a la hora de obtener datos sobre las variables de interés.
3. Errores de especificación ocasionados por la omisión de alguna variable explicativa o bien, por las posibles no linealidades en la relación entre X e Y.

En principio no sabemos si las variables en cuestión están relacionadas o no, o si en caso de haber dependencia es significativa o no, una forma de determinarlo y en caso de haberla deducir de qué tipo puede ser, es gráficamente representando los pares de valores observados. A dicho gráfico se le llama nube de puntos o diagrama de dispersión. Ejemplos de casos que podrían darse (Figura 1.1):

1.2.3. Hipótesis del Modelo de Regresión Lineal

1. Linealidad. La ecuación de regresión adopta una forma particular, la variable dependiente es la suma de un conjunto de elementos: el origen de la recta, una combinación lineal de variables independietes o predictoras y los residuos. El incumplimiento del supuesto de linealidad suele denominarse error de especificación. Algunos ejemplos son: omisión de variables independientes importantes, inclusión de variables independientes irrelevantes, no linealidad, parámetros cambiantes, no aditividad, etc. La relación existente entre X e Y es lineal:

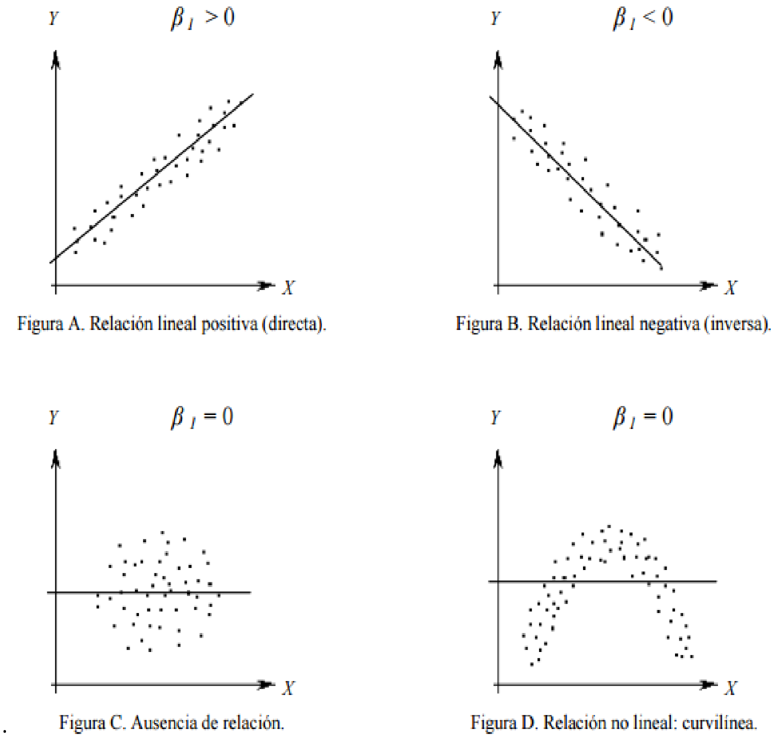


Figura A. Relación lineal positiva (directa).

Figura B. Relación lineal negativa (inversa).

Figura C. Ausencia de relación.

Figura D. Relación no lineal: curvilínea.

Figura 1.1:

$$y = \beta_0 + \beta_1 x$$

2. Independencia. los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria (recordemos que los residuos son las diferencias entre los valores observados y los pronosticados).

$$E[u_i u_j] = 0$$

3. Homocedasticidad. Para cada valor de la variable independiente (o combinacion de valores de las variables independientes), la varianza de los residuos es constante.

$$Var(u_i) = \sigma^2$$

4. Normalidad. Para cada valor de la variable independiente (o combinacion de valores de las variables independientes), los residuos se distribuye normalmente con media cero.

$$u_i \backsim N(0, \sigma)$$

5. No colinealidad. No existe relación lineal exacta entre ninguna de las variables independientes. el incumpliente de este supuesto da origen a colinealidad o multicolinealidad.

Sobre el incumpliento del primer supuesto puede obtenerse informacióon a partir de una inspección del diagrama de dispersión; si tenemos intención de utiliza el modelo de regresión lineal, lo razonable es que la relación entre la variable dependiente y las independientes sea dde tipo lineal. Los supuestos de independencia, homocedasticidad y normalidad estan estrechamente asociados al comportamiento de los residuos.

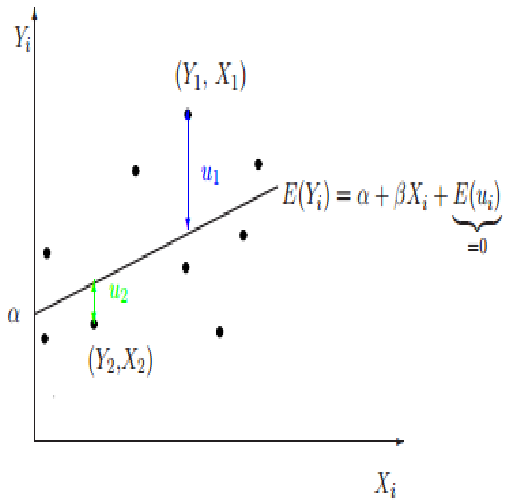


Figura 1.2:

1.2.4. Estimación por Mínimos Cuadrados Ordinarios

Una vez descrito el ámbito en el que nos vamos a mover, vamos a obtener un estimador adecuado de los coeficientes del modelo de regresión simple: el estimador de mínimos cuadrados ordinarios. En primer lugar, obtendremos el estimador y a continuación justificaremos su uso en base a sus propiedades. El modelo simple nos indica que cada observación Y_i es una realización de una variable que tiene dos componentes: uno que depende del valor del regresor X_i, cuyo valor observamos, y un componente residual que no observamos. Esto significa que tenemos N igualdades con una misma estructura:

$$Y_1 = \alpha + \beta X_1 + u_1$$

$$\vdots$$

$$Y_i = \alpha + \beta X_i + u_i$$

$$\vdots$$

$$Y_N = \alpha + \beta X_N + u_N$$

En el gráfico (Figura 1.2) se puede observar que Los puntos (Y_i, X_i) se sitúan o distribuyen alrededor de la recta $\alpha + \beta X_i$. La desviación de cada punto respecto a esta recta central viene dada por el valor que tome el término de error no observable u_i. Así, la recta central será aquella recta que se obtiene cuando el valor de la perturbación es cero. Teniendo en cuenta que suponemos que la perturbación tiene media cero, es decir, que no tiene efectos sistemáticos sobre Y , la recta central recoge el comportamiento medio de la variable de interés. La estimación de un modelo de regresión pretende obtener una aproximación a esta recta central no observable. En términos econométricos, queremos calcular el comportamiento medio de la variable de interés, $\alpha+\beta X_i$, a partir de observaciones provenientes de una muestra $(Y1, X1), (Y2, X2), ..., (YN, XN)$. Gráficamente, la estimación consiste en calcular la pendiente y la ordenada que mejor se ajusta a la nube de puntos.

La *perturbación* del modelo recoge todo aquello que no ha sido explicado por la parte sistemática del modelo y se obtiene como la diferencia entre la variable a explicar y la recta de regresión poblacional:

$$u_i = Y_i - \alpha - \beta X_i.$$

El resultado final obtenido a partir de la información que ofrece una muestra dada se define como la Función de Regresión Muestral (FRM). Se obtiene una vez que los coeficientes de la regresión hayan sido estimados $(\hat{\alpha}, \hat{\beta})$ y también se conoce como modelo estimado:

$$\hat{Y}_i = Ed(Y_i) = \hat{\alpha} + \hat{\beta} X_i$$

El residuo mide el error cometido al estimar la variable endógena y se define como la diferencia entre la variable a explicar y la recta de regresión muestral:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i = \alpha + \beta X_i + u_i - \hat{\alpha} - \hat{\beta} X_i = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) X_i + u_i$$

Este error proviene de dos fuentes: la primera, por el hecho de no poder obtener los valores de la perturbación (u_i) y la segunda se debe a que la estimación de los coeficientes desconocidos (α, β) introduce un error adicional. Es importante, por tanto, diferenciar y no confundir el residuo con la perturbación.

1.2.5. El Criterio de Estimación de Mínimo-Cuadrático

Dados el modelo y una muestra, debemos decidir cómo obtener la función de regresión muestral, es decir, cómo calcular las estimaciones $\hat{\alpha}$ y $\hat{\beta}$ a partir de los datos. Un método muy utilizado por su sencillez y buenas propiedades es el método de mínimos cuadrados ordinarios. El estimador de Mínimos Cuadrados Ordinarios, o MCO, de los parámetros α y β se obtiene de minimizar la suma de los residuos al cuadrado:

$$\underset{\hat{\alpha},\hat{\beta}}{Mín} \sum_{i=1}^{N} \hat{u}_i^2 = \underset{\hat{\alpha},\hat{\beta}}{Mín} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 = \underset{\hat{\alpha},\hat{\beta}}{Mín} \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

Las expresiones del estimador de α y β se obtienen de las condiciones de primer orden, para lo cual igualamos las primeras derivadas a cero:

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0$$

Las expresiones de los estimadores MCO para los coeficientes poblacionales α y β se obtienen de resolver las ecuaciones para $\hat{\alpha}$ y $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{N} (X_i - \overline{X})^2} = \frac{Sxy}{S_X^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta} \overline{X}$$

1.2.6. Propiedades de los Estimadores MCO

Necesitamos saber cuáles son las propiedades que justifican el uso de los estimadores MCO en el modelo de regresión simple bajo las hipótesis básicas. Los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son lineales en la perturbación, es decir, pueden expresarse como una combinación lineal de las perturbaciones $u_1, ..., u_N$. En segundo lugar, los estimadores MCO son variables aleatorias cuya distribución está centrada alrededor del valor poblacional, esto es

$$E(\hat{\alpha}) = \alpha \ E(\hat{\beta}) = \beta$$

y, por tanto, son estimadores insesgados. Y en cuanto a la precisión, el Teorema de Gauss- Markov prueba que los estimadores MCO tienen mínima varianza dentro del conjunto de los estimadores lineales (en u) e insesgados. Las varianzas y covarianza para los estimadores son las siguientes:

$$Var(\hat{\alpha}) = \sigma^2 \left(\frac{\sum_{i=1}^{N} X_i^2}{N \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \right) = \sigma^2 \left(\frac{1}{N} + \frac{\overline{X}^2}{N S_X^2} \right)$$

$$Var(\hat{\beta}) = \sigma^2 \left(\frac{1}{N \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \right) = \frac{\sigma^2}{N} \frac{1}{S_X^2}$$

$$Cov(\hat{\alpha},\hat{\beta}) = \sigma^2 \left(- \frac{\overline{X}}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \right) = - \frac{\sigma^2}{N} \frac{\overline{X}}{S_X^2}$$

Ambas varianzas dependen de la dispersión de la perturbación $var(u_i) = \sigma^2$, del tamaño muestral y de la dispersión del regresor X. En ambos casos, cuanto mayor sea N o la variabilidad de X, S_x^2, menor es la varianza de los estimadores MCO. En cuanto a la covarianza sería no nula a no ser que la media aritmética de la variable explicativa sea cero.

1.2.7. Precisión de la Estimación y la bondad de Ajuste

Una vez realizada las estimaciones de los coeficientes del modelo, la siguiente etapa del análisis consiste en el análisis y evaluación de los resultados. Por ejemplo,

1. Obtener una medida de la precisión en la estimación de α y β.

2. Evaluar la calidad del ajuste a los datos, es decir, si la función de regresión muestral, $\hat{Y}i = \hat{\alpha} + \hat{\beta} X_i$, resume bien el comportamiento observado de la variable endógena.

3. Evaluar si el modelo propuesto es correcto o si hay algún error en la especificación del modelo, en las hipótesis planteadas.

La Precisión de la estimación La desviación típica de la distribución muestral de los estimadores es un buen indicador de la precisión, sin embargo, habitualmente la desviación típica de los estimadores tiene algún elemento desconocido. Esto sucede en este caso, como puede comprobarse en la expresión de las varianzas, que dependen de la varianza de la perturbación $var(u_i) = \sigma^2$. Podemos obtener una estimación de la desviación típica sustituyendo el parámetro poblacional σ por un estimador insesgado, $\hat{\sigma}$. El resultado se conoce como errores típicos de los coeficientes de la regresión, es decir,

$$Error\,Típico(\hat{\alpha}) = \hat{desv}\,(\hat{\alpha}) = \frac{\hat{\alpha}}{\sqrt{N}} \sqrt{1 + \frac{\bar{X}^2}{N S_X^2}}$$

$$Error\,Típico(\hat{\beta}) = \hat{desv}\left(\hat{\beta}\right) = \frac{\hat{\alpha}}{\sqrt{N}} \frac{1}{S_X}$$

Un estimador insesgado de la varianza σ^2 es:

$$\hat{\sigma^2} = \frac{1}{N-2} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

donde $P_i \hat{u}_i^2$ es la suma de cuadrados residual, (o SCR), y $N-2$ son los grados de libertad que tenemos tras estimar α y β. Su raíz cuadrada $\hat{\sigma}$ se conoce como error típico de los perturbaciones o error típico de la regresión. Por tanto, la precisión de las estimaciones de los coeficientes aumenta con el número de observaciones N y la dispersión del regresor S_X y disminuye cuando crece el error típico $\hat{\sigma}$.

De forma similar, se construye el siguiente estimador insesgado de la matriz de las varianzas y la covarianza de los estimadores MCO:

$$\hat{V}\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \widehat{var(\hat{\alpha})} & c\hat{o}v(\hat{\alpha},\hat{\beta}) \\ c\hat{o}v(\hat{\alpha},\hat{\beta}) & var(\hat{\beta}) \end{pmatrix} = \hat{\sigma}^2 \begin{pmatrix} \left(\frac{1}{N} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right) & \frac{-\bar{X}}{\sum(X_i - \bar{X})^2} \\ & \frac{1}{\sum(X_i - \bar{X})^2} \end{pmatrix}$$

Bondad de Ajuste La medida de la bondad del ajuste que vamos a utilizar es el coeficiente de determinación,R^2 o $R-cuadrado$. Este coeficiente, tiene la siguiente expresión en el modelo de regresión lineal simple:

$$R^2 = r_{XY}^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Este coeficiente mide la ganancia obtenida al pasar de un modelo sin variable explicativa X:

$$Y_i = \alpha + u_i$$

a otro en el que se incluye esta variable: $Y_i = \alpha + \beta X_i + u_i$

Por tanto el R-cuadrado mide la proporción de la variabilidad observada de la variable dependiente Y que se ha podido explicar por incluir de forma lineal en el modelo la variable explicativa X. Normalmente se interpreta en porcentajes, por ejemplo, se dice que la regresión explica el $100\times$ R^2 por ciento de la variación observada en Y . Es fácil comprobar que:

- El criterio mínimo-cuadrático equivale a maximizarR^2.
- $R^2 = r_{\hat{Y}Y}^2$, mide la correlación entre el valor observado y el valor predicho o ajustado con la regresión. Como $0 \leq r_{\hat{Y}Y}^2 \leq 1$, si $R^2 = 0$ diremos que el ajuste es pobre y, por el contrario, sería un buen ajuste cuando este estadístico esté próximo a la unidad. Esta propiedad no se cumple en modelos sin término independiente, es decir, $Y_i = \beta X_i + u_i$.

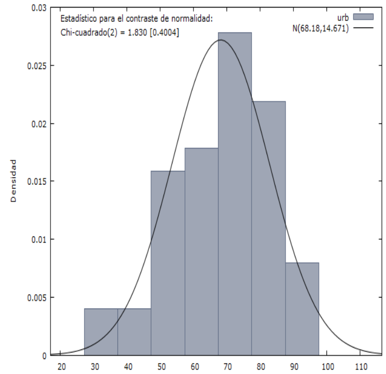
2. Variables

2.1. Características Naturales

Los datos utilizados para la realizacion del Modelo de Regresión Lineal fueron obtenidos de las muestras de Gretl de Ramanathan, los cuáles hablan del Porcentaje de Mujeres mayores de 16 años que laboran en actividades de Mano de Obra por lo cual es necesario conocer la naturaleza de nuestros datos.

1. En primera instancia los datos son de **Corte Transversal**, es decir, se dispone de una observación por individuo y se refieren a un punto determinado en el tiempo. En la mayoría de los estudios, los individuos encuestados son personas, en este caso son estudiantes. Las encuestas son una fuente típica para datos de corte transversal. En muchos estudios econométricos contemporáneos de corte transversal el tamaño muestral es bastante elevado. En los datos de corte transversal, las observaciones deben ser obtenidas mediante un muestreo aleatorio, lo que implica que las observaciones sean independientes entre sí.
2. **Unidades de Observación:** son Mujeres mayores de 16 años que participan en trabajos de Mano de Obra.

Figura 2.1: Distribución de la Variable Urb (Porcentaje de Población que vive en zonas Urbanas



3.

Variable	**Unidades de Medida**
WLFP	Porcentaje de Mujeres
YF	Miles de dólares
YM	Miles de dólares
EDUC	Porcentaje de Mujeres
UE	Porcentaje de desempleados
MR	Porcentaje de Mujeres
DR	Porcentaje de Mujeres
College	Porcentaje de Mujeres

2.2. Análisis Exploratoria

En el cuadro se puede observar el Resumen de Estadística Descriptiva de las variables del Modelo de Regresión Lineal.

WLFP (Porcentaje de Mujeres que trabajan en Mano de Obra mayores de 16 años) presentan una media y mediana similar, es decir, ambas reafirman que el 57 % de la mujeres trabajan en Mano de Obra, cuyos valores oscilan entre el 42.6 % y el 66.4 %, con una desviación del 4.2 %, pero presentan una variabilidad del 7.39 %, lo cual nos indica que nuestros datos se encuentran concentrados alredor de la media, lo cual nos reafirma la curtosis que es mayor que 0 por lo tanto presentan mayor concentración, pero por la simetría de los datos los datos se reunen más hacia la derecha de media.

URB (Porcentaje de población que vive en zonas urbanas) nos indica que el 68 % de la poblacion vive en zonas urbanas pero debido a que presenta una variabilidad del 21.52 % de los datos se presenta una mayor desviación por lo tanto el rango es más grande, indicandonos que desde el 32.2 % de la población hasta el 92.6 % vive en zonas urbanas por lo tanto no es tan confiable nuestra media, con lo cual se podría presentar 2 comportamientos de con respecto al porcentaje de población que vive en zonas urbanas de los cuales influyen una variable para que se pueda dividir la información de acuerdo a otro factor.

YM (Media de Ingresos de Hombres mayores de 15 años) nos indica que los hombres aportan una media de ingresos de 27.70 de miles de dólares al año, que puede ser desde un aporte de 21.42 de miles de dólares hasta 35.62 de miles de dólares, ademas de que los datos recolectados tienen una variabilidad del 12.44 % por lo cual se considera que es buen valor nuestra media y que nuestro datos se centran a la izquierda de nuestra media, es decir que la mayoria son menores al valor de la media pero tienden a ser cercanos a 27.70 de miles de dólares que aportan los hombres.

Cuadro 1: Resumen de Estadística Descriptiva de las variables del Porcentaje de Mujeres mayores de 16 años en Mano de Obra

Estadísticos principales, usando las observaciones 1 - 50

Variable	Media	Mediana	Mínimo	Máximo
wlfp	57.4740	57.7500	42.6000	66.4000
yf	18.4160	18.0790	14.2710	25.6200
ym	27.7899	27.2195	21.4250	35.6220
educ	76.1080	77.1000	64.5000	86.1000
ue	6.16000	6.15000	3.50000	9.60000
mr	54.2332	54.2000	46.8800	60.9200
dr	9.50940	9.35000	6.42000	15.0600
urb	68.1800	68.8000	32.2000	92.6000
wh	65.9050	69.1300	24.6900	77.7300

Variable	Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
wlfp	4.24878	0.0739253	−0.695482	1.58764
yf	2.70318	0.146784	0.653440	−0.185426
ym	3.45939	0.124484	0.444564	−0.228542
educ	5.73595	0.0753660	−0.435516	−0.725421
ue	1.36382	0.221399	0.448943	0.316542
mr	3.13442	0.0577953	−0.0509672	−0.231799
dr	1.57863	0.166007	0.830801	1.48274
urb	14.6714	0.215187	−0.334807	−0.529312
wh	9.37887	0.142309	−1.83361	5.56323

Variable	porc. 5 %	porc. 95 %	Rango IQ	Observaciones ausentes
wlfp	50.7500	63.8500	4.97500	0
yf	14.6145	23.6093	3.88875	0
ym	22.1021	35.5462	4.48225	0
educ	65.5500	84.2050	7.97500	0
ue	3.86500	9.16000	1.60000	0
mr	48.5920	59.8520	4.49250	0
dr	7.28050	11.9460	2.10000	0
urb	40.7750	89.1800	24.8750	0
wh	50.8280	77.1530	11.5800	0

Cuadro 2: Distribución de la Variable WLFP

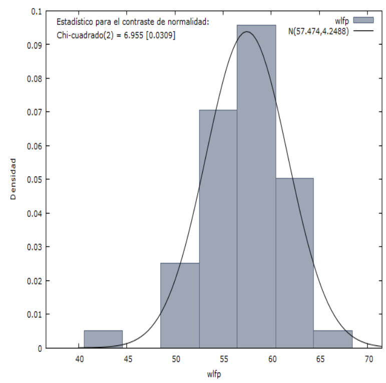


Figura 2.2: Distribución de YM (Media de Ingresos de Hombres mayores de 15 años)

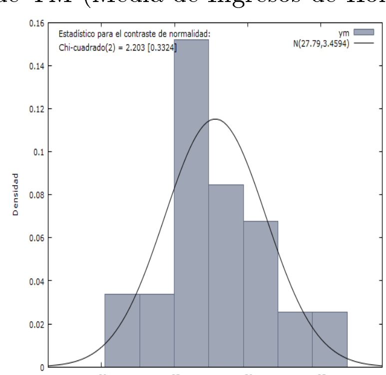


Figura 2.3: Distribución de la Variable YF(Media de Ingresos de Mujeres mayores de 15 años)

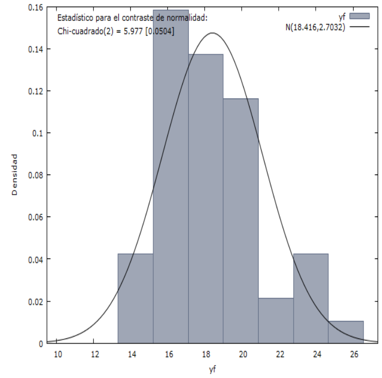
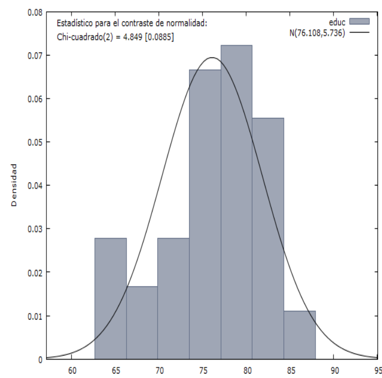


Figura 2.4: Distribución de la Variable EDU (Porcentaje de Mujeres mayores de 15 años que terminaron la secundaria)



YF (Media de Ingresos de Mujeres mayores a 15 años) nos indica que las mujeres solo presentan una aportación con media de 18.41 de miles de dólares, pero presentan una variabilidad de 14.67 % con respecto a la aportación de la media de las mujeres, que su ingreso medio puede ser desde 14.27 de miles de dólares hasta 25.62 de miles de dólares, pero con respecto a los hombre las mujeres tienen menor ingreso medio, pero presentan mayor concentración los datos en el rango, es decir, que existe mayor probabilidad de conocer el ingreso medio de las mujeres.

EDU (Porcentaje de Mujeres mayores de 15 años que terminaron la secundaria), solo el 76.10 % de la mujeres han terminado la secundaria, esto quiere decir que nuestros datos solo presentan una variabilidad del 7.5 %, lo cual nos indica que nuestros datos son muy parecidos, entonces se puede suponer que las mujeres estan interesadas en su mayoría en terminar la secundaria pero que por algunos factores tienen que dejar de estudiar o combinar el trabajo con su educación.

WH (Porcentaje de mujeres de piel blanca), el 65.9 % de las mujeres son de piel blanca pero los datos presentan una variabilidad del 14.23 %, es decir que existe cierto grupo de mujeres en cierta población que representan un menor porcentaje de mujeres con piel blanca en el Estado, por lo tanto puede haber desde un 24.69 % de mujeres de piel blanca hasta el 77.73 %, ademas de que los datos se encuentran más dispersos con respecto a la media.

MR (Porcentaje de mujeres mayores de 15 años que estan casadas), el porcentaje promedio de mujeres casadas es de 54.23 %, es decir que la mitad de mujeres que trabajan estan casadas, lo cual es reafirmado ya que se presenta una variabilidad del 5.7 % de los datos, pero de acuerdo a la curtosis presentan uns baja concentración con respecto a la media, por lo contrario la variable **DR** nos indica

Figura 2.5: Distribución de la Variable WH (Porcentaje de Mujeres de piel blanca)

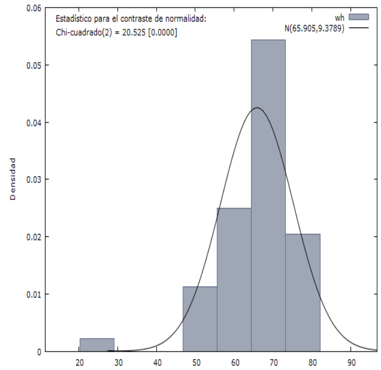
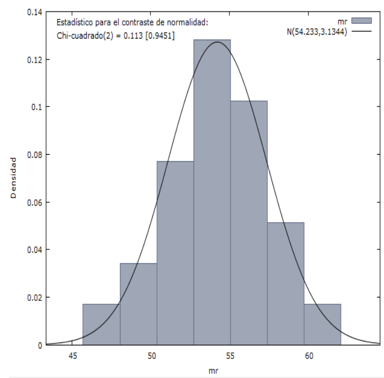


Figura 2.6: Distribución de la Variable MR (Porcentaje de Mujeres mayores de 15 años que estan casadas)



que solo el 9.5 % de la población de las mujeres estan divorciadas, pero no nuestro datos presetan una variabilidad del 16.61 %, pero la curtosis nos indica que los datos cuentan con una mayor concentración con respecto a la media.

UE (Porcentaje de desempleados en mano Obra), solo el 6.16 % de la población se encuentra desempleado, es decir que existe oportunidad de trabajo en este campo, aun cuando presentan mayor dispersión nuestros datos con relación a la media pero mayoria de ellos se concentran alrededor de la media, pero con un rango demasiado amplio, es decir que puede haber desde un 3.5 % de personas desempleadas hasta un 9.6 %.

3. Modelo a Estimar

Para el desarrolo del Modelo se desea introducir la regresión, además del término constante, más de una variable explicativa por lo que pasamos del llamado modelo de regresión lineal simple al modelo de regresión lineal múltiple. El modelo de regresión lineal general (MRLG), co k variables explicativas:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i \; i = 1, 2, ..., N$$

se puede escribir de Forma Matricial:

$$\underset{(Nx1)}{Y} = \underset{(NxK)}{X} \underset{(Kx1)}{\beta} + \underset{(Nx1)}{u}$$

Figura 2.7: Distribución de la Variable DR(Porcentaje de mujeres mayores de 15 años divorciadas)

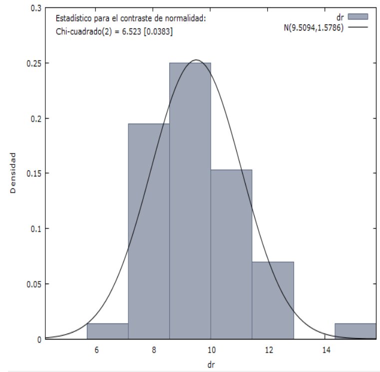
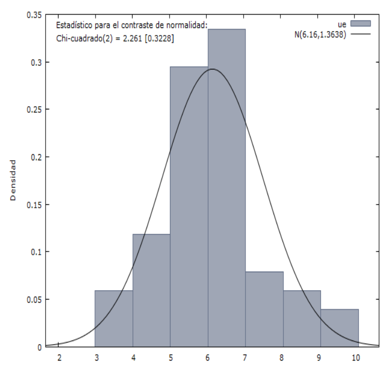


Figura 2.8: Distribución de la Variable UE (Porcentaje de desempleo en Mano de Obra)



Por el momento, seguimos suponiendo las mismas hipótesis básicas sobre el término de perturbación y sobre las variables explicativas o regresoras, a saber:

1. $E(u_i) = 0$, $E\left(u_i^2\right) = \sigma^2$, $E(u_i u_j) = 0$
2. La perturbación sigue una distribución normal
3. Las variables X_2a $X_K$no son estocásticas, son fijas. Esto quiere decir que e muestras repetidas de N observaciones de $Y_i, X_{2i}, ..., X_{Ki,}$las variables $X_{2i}, ...X_{Ki}$, $i = 1, ..., N$ tomaría siempre los mismos valores, implicando que los regresores y el término de perturbación estan incorrelacionados.
4. Los regresores son linealmente independientes, esto quiere decir que el rango de la matriz de los regresores X es K tal que no tiene columnas repetidas ni unas son combinaciones lineales de otras.

Realizando el análisis del Modelo por medio de Mínimos Cuadrados Ordinarios podemos obtener la siguiente ecuación :

$$
\begin{aligned}
\widehat{\text{wlfp}} = \underset{(8.9750)}{44.5096} + \underset{(0.40758)}{0.987983}\,\text{yf} - \underset{(0.30621)}{0.174345}\,\text{ym} + \underset{(0.093165)}{0.285129}\,\text{educ} \\
- \underset{(0.31362)}{1.61058}\,\text{ue} - \underset{(0.17314)}{0.0782145}\,\text{mr} + \underset{(0.25834)}{0.437371}\,\text{dr} - \underset{(0.033335)}{0.0926339}\,\text{urb} \\
- \underset{(0.039845)}{0.0874916}\,\text{wh} \\
T = 50 \quad \bar{R}^2 = 0.7379 \quad F(8, 41) = 18.246 \quad \hat{\sigma} = 2.1751 \\
\text{(Desviaciones típicas entre paréntesis)}
\end{aligned}
$$

En el MCO se puede observar los coeficientes estimados, los cuales nos indican el cambio que presentará WLFP ante un cambio unitario de cada una de las variables explicativasn (EDUC, UE y URB) con valor de significancia del 99 % que presentan con respecto al modelo en conjunto, las variables ym y mr no presentan un valor P demasiado grande por lo que se concluye que las variables no influyen en el cambio de nuestra variable WLFP, lo cual es cierto ya que los ingresos medio de hombres mayores de 15 años y el que las mujeres sean casadas influyen en la decisión de que trabajen en Mano de Obra, por lo cual sería conveniente omitir dichas variables para encontrar un mejor modelo con mayor significancia. Además de que se obtiene un R^2igual al 78 % lo cual nos indica que el modelo explica las variaciones de WLFP en ese porcentaje, lo cual es bueno ya se estan ajustando las variables explicativas al modelo.

Analizando la información y obteniendo variables que no tienen significancia con relación al porcentaje de mujeres mayores de 16 años que participan en mano de obra, se procede a omitir dichas variables para mejorar el modelo, con lo cual se obtiene que la siguiente ecuación.

Al realizar el modelo nuevamente sin considerar las variables que presentan bajo nivel de significancia se obtiene que todas las variables son significativas al 90 % con respecto a la variable dependiente, es decir que el el ingreso medio de una mujer mayor de 16 años, el que haya terminado la secundaria, el porcentaje de mujeres divorciadas, el porcentaje de poblacion que vive en zonas urbanas, el que una mujer sea de piel blanca son factores que inducen que las mujeres participen en trabajos de mano de Obra, que si se quisiera ser más estricto y tener un modelo con variables significativas del 99 % tendríamos que omitir la variable de que la mujer mayor de 16 años sea divorciada, pero esto implica que los estimadores mínimos cuadráticos presentaron sesgo por cual se considera conveniente dejar la variable DR que explica a nuestra variable dependiente.

Nuestro modelo presenta una R^2del 77.73 % de relación entre las variables independientes con respecto a la dependiente, por lo cual se considera que es confiable nuestro modelo, y que no cometemos un error al omitir las variables MR y YM ya que no explican a la variable dependientes.

Figura 3.1: Modelo de Regresión Lineal Múltiple en Gretl

Modelo 1: MCO, usando las observaciones 1–50
Variable dependiente: wlfp

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	44.5096	8.97496	4.9593	0.0000
yf	0.987983	0.407583	2.4240	0.0198
ym	−0.174345	0.306207	−0.5694	0.5722
educ	0.285129	0.0931647	3.0605	0.0039
ue	−1.61058	0.313617	−5.1355	0.0000
mr	−0.0782145	0.173139	−0.4517	0.6538
dr	0.437371	0.258336	1.6930	0.0980
urb	−0.0926339	0.0333355	−2.7788	0.0082
wh	−0.0874916	0.0398446	−2.1958	0.0338

Media de la vble. dep.	57.47400	D.T. de la vble. dep.	4.248784
Suma de cuad. residuos	193.9742	D.T. de la regresión	2.175104
R^2	0.780710	R^2 corregido	0.737922
$F(8, 41)$	18.24590	Valor p (de F)	2.90e–11
Log-verosimilitud	−104.8395	Criterio de Akaike	227.6790
Criterio de Schwarz	244.8872	Hannan–Quinn	234.2319

Figura 3.2: Ecuación del Modelo Omitiendo las variables YM y MR

$$\widehat{\text{wlfp}} = \underset{(5.4753)}{41.8336} + \underset{(0.15815)}{0.849264}\,\text{yf} + \underset{(0.069099)}{0.249152}\,\text{educ} - \underset{(0.27686)}{1.67758}\,\text{ue}$$

$$+ \underset{(0.22508)}{0.434104}\,\text{dr} - \underset{(0.029336)}{0.0942172}\,\text{urb} - \underset{(0.035204)}{0.0960861}\,\text{wh}$$

$$T = 50 \quad \bar{R}^2 = 0.7462 \quad F(6, 43) = 25.015 \quad \hat{\sigma} = 2.1404$$

(Desviaciones típicas entre paréntesis)

Figura 3.3: Modelo MCO con Omisión de Variables YM y MR

Modelo 5: MCO, usando las observaciones 1–50
Variable dependiente: wlfp

	Coeficiente	Desv. Típica	Estadístico t	Valor p
const	41.8336	5.47528	7.6405	0.0000
yf	0.849264	0.158152	5.3699	0.0000
educ	0.249152	0.0690987	3.6057	0.0008
ue	−1.67758	0.276859	−6.0593	0.0000
dr	0.434104	0.225080	1.9287	0.0604
urb	−0.0942172	0.0293363	−3.2116	0.0025
wh	−0.0960861	0.0352037	−2.7294	0.0092

Media de la vble. dep.	57.47400	D.T. de la vble. dep.	4.248784
Suma de cuad. residuos	196.9882	D.T. de la regresión	2.140355
R^2	0.777303	R^2 corregido	0.746229
$F(6, 43)$	25.01455	Valor p (de F)	1.55e–12
Log-verosimilitud	−105.2249	Criterio de Akaike	224.4499
Criterio de Schwarz	237.8341	Hannan–Quinn	229.5467

3.0.1. Contrastes de Regresión Lineal

La regresión múltiple tiene contrastes importantes, los cuales se revisaron para tener certeza de que el análisis del modelo es preciso y no sesgado.

1. **Linealidad.** En el contexto del análisis de regresión, permiten examinar la relación existente entre la variable dependiente y cada una de las variables independientes por separado, tras eliminar de ellas el efecto del resto de las variables independientes incluidas en el análisis. Se contrasta la hipótesis de linealidad, completando la especificación del modelo propuesto con los logaritmos o los cuadrados de las variables explicativas y contrastando la nulidad de sus coeficientes.

Suponiendo que se tienen:

H_o: La relación es lineal bajo los coefientes de los términos cuadráticos que son nulos.

H_1;La relación no es lineal bajo los coeficientes de los términos cuadráticos que son nulos.

ya que se obtiene p mayor que 0.05 entonces no se rechaza la hipotesis nula por lo tanto el modelo presenta relación Lineal.

2. **Normalidad.** Se refiere a que los datos, las variables independientes y la variable dependiente, tienen que representar valores que esten distribuidos normalmente, es decir, los resiudos, es importante ya que cuando los errores de las variables tienen distribución no normal, pueden afectar las relaciones y la significancia. Si se consideran las siguientes Hipótesis

H_o: El error se distribuye normalmente

H_1;El error no se distribuye normalmente

Ya que los datos presentan una media cercana a cero entonces se distribuyen normalmente y no se rechaza la hipótesis nula.

Figura 3.4: Linealidad

```
Regresión auxiliar para el contraste de no linealidad (términos logarítmicos)
MCO, usando las observaciones 1-50
Variable dependiente: uhat

             Coeficiente   Desv. Típica   Estadístico t   Valor p
  ---------------------------------------------------------------
  const       357.805       396.024          0.9035       0.3721
  yf           -1.52769       1.78448       -0.8561       0.3975
  educ          1.84334       1.63605        1.127        0.2671
  ue           -1.53941       1.70794       -0.9013       0.3732
  dr            3.10384       1.70740        1.818        0.0772  *
  urb          -0.292388      0.158823      -1.841        0.0737  *
  wh           -0.148150      0.179316      -0.8262       0.4140
  l_yf         31.0336       34.5785         0.8975       0.3753
  l_educ     -137.464       123.043         -1.117        0.2711
  l_ue          9.99555      10.5668         0.9459       0.3503
  l_dr        -32.3688       16.9520        -1.909        0.0640  *
  l_urb        17.6360        9.37277        1.882        0.0678  *
  l_wh          6.05403       9.53110        0.6352       0.5292

  R-cuadrado = 0.200245

  Estadístico de contraste: TR^2 = 10.0123,
  con valor p  = P(Chi-cuadrado(6) > 10.0123) = 0.124137
```

Figura 3.5: Estadístico para el contraste de Normalidad

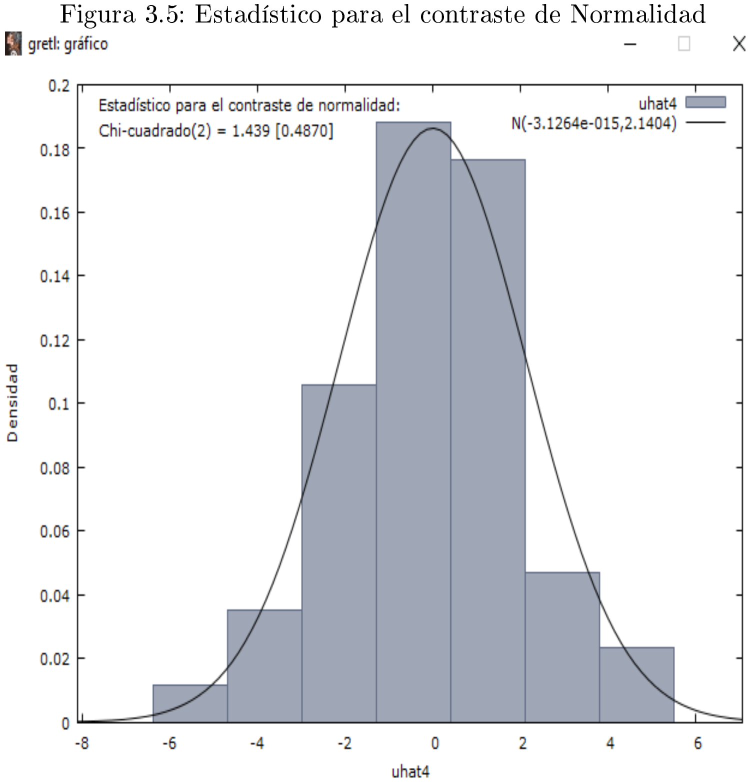


Figura 3.6: Contraste de Heterocedasticidad

Contraste de heterocedasticidad de White
MCO, usando las observaciones 1-50
Variable dependiente: uhat^2

	Coeficiente	Desv. Típica	Estadístico t	Valor p	
const	333.864	255.735	1.306	0.2052	
yf	-3.22749	11.1293	-0.2900	0.7745	
educ	-4.51590	6.00558	-0.7520	0.4601	
ue	-0.635927	17.3708	-0.03661	0.9711	
dr	-26.3046	12.7654	-2.061	0.0514	*
urb	2.96772	2.04387	1.452	0.1606	
wh	-3.15321	2.50979	-1.256	0.2222	
sq_yf	0.179799	0.220939	0.8138	0.4245	
X2_X3	-0.00570935	0.108358	-0.05269	0.9585	
X2_X4	0.0329672	0.408640	0.08068	0.9364	
X2_X5	0.496423	0.377763	1.314	0.2023	
X2_X6	-0.0933231	0.0627530	-1.487	0.1512	
X2_X7	-0.0205301	0.0952325	-0.2156	0.8313	
sq_educ	0.00164282	0.0423328	0.03881	0.9694	
X3_X4	-0.282204	0.210547	-1.340	0.1938	
X3_X5	0.441408	0.158014	2.793	0.0106	**
X3_X6	-0.0297516	0.0191651	-1.552	0.1348	
X3_X7	0.0504961	0.0340743	1.482	0.1525	
sq_ue	0.523690	0.460996	1.136	0.2682	
X4_X5	-0.548527	0.715060	-0.7671	0.4512	
X4_X6	-0.0445001	0.104027	-0.4278	0.6730	
X4_X7	0.352856	0.101659	3.471	0.0022	***
sq_dr	0.303270	0.232627	1.304	0.2058	
X5_X6	-0.0411331	0.0867299	-0.4743	0.6400	
X5_X7	-0.242820	0.109288	-2.222	0.0369	**
sq_urb	0.00811383	0.00564328	1.438	0.1646	
X6_X7	0.0115429	0.0113184	1.020	0.3189	
sq_wh	-0.00411299	0.00747579	-0.5502	0.5877	

R-cuadrado = 0.833224

Estadístico de contraste: TR^2 = 41.661215,
con valor p = P(Chi-cuadrado(27) > 41.661215) = 0.035535

3. **Heterocedasticidad.** En estadística se dice que un modelo de regresión lineal presenta heterocedasticidad cuando la varianza de las perturbaciones no es constante a lo largo de las observaciones. De ella se deriva que los datos con los que se trabaja son heterogéneos, ya que provienen de distribuciones de probabilidad con distinta varianza.

Existen diferentes razones o situaciones en las que cabe encontrarse con perturbaciones heteroscedásticas. La situación más frecuente es en el análisis de datos de corte transversal, ya que los individuos no suelen tener un comportamiento homogéneo. La Heterocedasticidad puede arruinar nuestros resultados y hacernos caer en un error, podríamos asumir que algo está relacionado cuando en realidad no lo está.

Considerando el contrate de White se supone que:

H_o: No hay Heterocedasticidad

H_1;Existe heterocedasticidad

Con lo cual se rechaza la hipótesis nula y existe heterocedasticidad, lo cual se refirma con una $R^2 = 83.32\,\%$ ya que para que no exista heterocedasticidad tendría que ser suficientemente pequeño.

4. **Colinealidad.** Existe colinealidad perfecta cuando una de las variables independientes se relaciona de forma perfectamente lineal con una o más del resto de variables independientes de la ecuación. La colinealidad es un problema en el caso de la colinealidad perfecta no es posible estimar los coeficientes de la ecuación de regresión y en el caso de la colinealidad parcial, aumenta el tamaño de los residuos tipificados y esto produce coeficientes de regresión muy inestables: pequeños cambios en los datos.

Figura 3.7: Colinealidad

```
Factores de inflación de varianza (VIF)
Mínimo valor posible = 1.0
Valores mayores que 10.0 pueden indicar un problema de colinealidad

         yf    1.955
       educ    1.680
         ue    1.525
         dr    1.350
        urb    1.981
         wh    1.166

VIF(j) = 1/(1 - R(j)^2), donde R(j) es el coeficiente de correlación múltiple
entre la variable j y las demás variables independientes

Diagnósticos de colinealidad de Belsley-Kuh-Welsch:

                         --- proporciones de la varianza ---
    lambda      cond    const       yf     educ       ue       dr      urb       wh
     6.869     1.000    0.000    0.000    0.000    0.001    0.000    0.000    0.000
     0.057    10.949    0.000    0.014    0.001    0.272    0.001    0.140    0.006
     0.037    13.718    0.003    0.002    0.008    0.172    0.028    0.095    0.207
     0.019    19.087    0.001    0.093    0.000    0.147    0.754    0.009    0.000
     0.010    25.802    0.004    0.314    0.033    0.053    0.048    0.654    0.364
     0.006    33.209    0.151    0.484    0.102    0.001    0.130    0.050    0.422
     0.002    62.018    0.840    0.093    0.857    0.354    0.038    0.051    0.000

  lambda = valores propios de X'X, del más grande al más pequeño
  cond   = índice de condición
  nota: las columnas de proporciones de la varianza suman 1.0
```

Ya que los valores obtenidos del contraste de colinealidad son menores a 10.0, entonces no existe relacion entre las variables independientes.

- Realizando la corrección de la heterocedasticidad se obtiene que la variable de URB presenta una significancia de 95 % y las variables YF, EDUC, UE, WH, presentan una significancia del 99 % por lo tanto la relación de las variables presentan un R^2del 90.3 % .

Figura 3.8: Corrección de Heterocedasticidad

$$\widehat{\text{wlfp}} = \underset{(3.3453)}{49.9443} + \underset{(0.14888)}{0.817620}\,\text{yf} + \underset{(0.039766)}{0.279675}\,\text{educ} - \underset{(0.16201)}{1.77531}\,\text{ue}$$
$$- \underset{(0.025285)}{0.0680404}\,\text{urb} - \underset{(0.030417)}{0.197106}\,\text{wh}$$
$$T = 50 \quad \bar{R}^2 = 0.8923 \quad F(5,44) = 82.194 \quad \hat{\sigma} = 2.2113$$

(Desviaciones típicas entre paréntesis)

Figura 4.1: Ecuación Modelo de Regresión Lineal sobre el Porcentaje de Mujeres mayores de 16 años que trabajan en Mano de Obra

$$\begin{aligned}\widehat{\text{wlfp}} = \underset{(5.4753)}{41.8336} + \underset{(0.15815)}{0.849264}\,\text{yf} + \underset{(0.069099)}{0.249152}\,\text{educ} - \underset{(0.27686)}{1.67758}\,\text{ue} \\ + \underset{(0.22508)}{0.434104}\,\text{dr} - \underset{(0.029336)}{0.0942172}\,\text{urb} - \underset{(0.035204)}{0.0960861}\,\text{wh}\end{aligned}$$

$$T = 50 \quad \bar{R}^2 = 0.7462 \quad F(6, 43) = 25.015 \quad \hat{\sigma} = 2.1404$$

(Desviaciones típicas entre paréntesis)

4. Conclusiones

Realizando un Modelo de Regresión Lineal a los datos obtenidos de Gretl, donde se tiene una variable dependiente acerca del Porcentaje de Mujeres Mayores de 16 años que trabajan en Mano de Obra se obtiene la ecuación (Figura 4.1).

Por lo cual podemos notar que las variables de ingreso medio que aporta una mujer mayores de 15 años, que haya terminado la secundaria y que sea sea divorciada explica de manera positiva nuestra variable dependiente, es decir, explican mejor a la variable dependiente, por lo contario el porcentaje de desempleo, el que la persona (en este caso haciendo referencia a la población femenil) viva en una zona urbana y el que sea de piel blanca influyen muy poco o explican muy poco nuestra variable, pero de aún asi influyen o explican el porcentaje de mujeres que toman la decisión de trabajar como mano de obra, la variable que más influye en esta decisión es que solo haya estudiado hasta la secundaria, lo que nos muestra que actualmente es muy importante el nivel de educación para poder solicitar un mejor trabajo, asi como que actualmente estudiar hasta la secundaria ya no es tan revelante como años atrás en el que podías solicitar un buen trabajo con este nivel de educación pero es claro, ya que el nivel de población ha incrementado, por lo tanto la competencia laboral también, con lo cual se determina que es necesario seguirse preparando profesionalmente y actualizandose en conocimientos para un mejor empleo.

Pero si además de tener un nivel de educación hasta la secundaria, también cuentas con una media de ingresos bajo y eres una mujer divorciada entonces te encuentras con pocas probabilidades de que seas contratada, recurres a la posibilidad más cercana que es trabajar como mano de obra don de se considera que no requieres mucha experiencia profesional, solo fuerza y habilidades físicas. Así como tenemos variables con mayor influencia explicativa, también contamos con variables que si explican la variable dependiente pero no tiene mucha influencia, que en este caso son que tengas que vivir en alguna zona urbana o el color de piel de la persona que solicita este tipo de trabajo.

Finalmente se puede decir que actualmente la participación de la mujer en el campo laboral es muy notable, no solo en actividades que les corresponden, si no que en actividades que se consideran solo para hombres, esto se debe ante la necesidad de contar con más ingresos al hogar y quizás el hecho de que te soliciten un mejor grado de educación.

Referencias

[1] 18reglin_SPSS.pdf

[2] ocw.ehu.eus/file.php/23/GretlHetero.pdf

[3] www.uam.es/personal_pdi/economicas/anadelsur/pdf/heterocedasticidad.pdf

[4] www.uam.es/personal_pdi/ciencias/horra/AnalisisDatos-Apuntes/Regresion-Simple.pdf

[5] www.dm.uba.ar/materias/estadistica_Q/2011/1/clase %20regresion %20simple.pdf

[6] 9. Modelos de regresión lineal múltiple. Diagnosis y validación -Análisis de resíduos. Gráficos

[7] www.fce.unal.edu.co/uifce/proyectos-de-estudio/pdf/GRETL