OBSERVATORIO PARA EL
DESARROLLO ECONÓMICO
DE MISIONES

SUPLEMENTO ODEM N°3

Estimación de la brecha de ingresos entre la mujer y el hombre en Misiones

UNIVERSIDAD Gastón Dachary

EDITORIAL UGD

SUPLEMENTO ODEM N°3

Estimación de la brecha de ingresos entre la mujer y el hombre en Misiones

Dr. Darío Ezequiel Díaz
Lic. Natalia Ojeda

OBSERVATORIO PARA EL DESARROLLO ECONÓMICO DE MISIONES

UNIVERSIDAD Gastón Dachary

UGD EDITORIAL

1. Introducción

El presente artículo tiene como objetivo medir la brecha de ingresos entre los varones y las mujeres, para el Aglomerado Posadas, utilizando dos tipos de metodologías diferentes, en base a los microdatos de la Encuesta Permanente de Hogares, correspondientes al segundo trimestre del año 2018.

En la literatura académica respecto a medición de brecha salarial por género, existen trabajos de referencia muy interesantes, como ser (Paz, Diferencias de ingresos entre varones y mujeres. Evidencias para Salta (Argentina), 1995); (Paz, 1998); (Esquivel, 2007), (Segura Gómez, 2013), (Broso, 2016), entre otros. Todos ellos resultaron valiosos para la construcción y aplicación de dos metodologías muy diferentes entre sí, que de ahora en más, se denominará a la primera como "Metodología de Regresión Lineal Múltiple", y a la segunda, "Metodología de (Oaxaca, 1973) y (Blinder, 1973) con corrección del sesgo por selección de (Heckman, 1979)"

El presente trabajo es inédito para la provincia de Misiones, y pretende constituir las bases para futuras investigaciones sobre la temática.

2. Desarrollo

2.1 La brecha salarial y la literatura económica

Las diferencias de ingreso entre las personas surgen a partir de los diferentes trabajos y de las aptitudes o habilidades que tienen los individuos.

Según (Di Paola & Berges, 2000, pág. 1), "existen muchas formas en las que los trabajos difieren respecto a su atractivo, lo agradable que resulte, lo riesgoso, las perspectivas que posean, los costos de entrenamiento". Además, "la población difiere en sus gustos, en sus habilidades originales, en su educación y en las oportunidades que se les presentan". Por lo que las desigualdades "reales" surgen entre las personas, especialmente en sus habilidades y oportunidades.

La literatura denomina *diferencia compensatoria* a la diferencia salarial causada por características no monetarias de los distintos puestos de trabajo. Una de esas posibles diferencias salariales es la discriminación.

Según (Paz, 1998) y (Di Paola & Berges, 2000), existe discriminación cuando el mercado ofrece diferentes oportunidades a personas similares que sólo difieren por su sexo, edad, raza, grupo étnico u otras particularidades personales.

Como señala el economista, ganador del premio nobel, Gary Becker, la discriminación puede reflejar el prejuicio de la sociedad contra ciertos grupos, como un cierto "gusto" por la discriminación de los demandantes de trabajo. En (Becker, 1957), a este "gusto por la discriminación", él lo define de la siguiente manera: "si una persona tiene un gusto por la discriminación debe actuar como si fuera a pagar algo, ya sea directa o en una reducción de los ingresos, que se asocia con algunas personas en contra de otras. Cuando la discriminación ocurre, la persona debe, de hecho, pagar o perder ingresos por ese privilegio".

Para medir el grado de discriminación que existe en el mercado de trabajo se suelen evaluar los salarios medios de los diferentes grupos, aunque esto plantea un problema evidente, ya que, por ejemplo, incluso en un mercado de trabajo libre de discriminación, cada persona cobra un salario distinto. Esto se debe a que los individuos se diferencian por la cantidad de capital humano que poseen y por los tipos de trabajo que pueden y quieren realizar.

La literatura reconoce tres fuentes de diferencias salariales entre hombres y mujeres:

1. La proveniente de dotaciones distintas de capital humano entre ambos grupos, por ejemplo, la Hipótesis de Capital Humano (HCH)

2. La derivada de la concentración de uno de los grupos en tipos de actividades determinadas, que podrían tener de acuerdo con sus características de riesgo, o disgusto involucrados, compensaciones diferentes, por ejemplo, la Hipótesis de Segregación (HS)

3. La que proviene de la discriminación en sí misma, que adquiere un carácter residual, en la medida que constituye la parte de la discriminación no explicada por las otras dos razones.

El sesgo de selección: un aspecto metodológico a tener en cuenta

Como señala (Di Paola & Berges, 2000), un aspecto relevante al trabajar con salarios de la población femenina es detectar si existe o no un sesgo en los datos muestrales.

El concepto de oferta de trabajo individual indica lo que, a cada precio del trabajo o salario por hora ofrecida por el mercado, la cantidad de horas que un sujeto está dispuesto a ofrecer en función de sus preferencias entre ocio y trabajo.

En el caso de la mujer, el trabajo involucra un costo de oportunidad dado no solo por el valor de las horas destinadas a su ocio, sino por el valor que la misma atribuye a otras actividades productivas tales como el cuidado del hogar, o la crianza de los niños (sobre todo en los primeros años de vida hasta que alcanzan la edad escolar), entre otras actividades domésticas. Si el valor de estas actividades fuese superior al salario de mercado, no habría incentivo para incorporarse al mismo ya que maximizaría su utilidad no trabajando y realizando otras tareas en su tiempo disponible. El valor mínimo necesario para su incorporación al mercado se denomina "salario de reserva" (Di Paola & Berges, 2000).

Bajo estas consideraciones, la remuneración promedio de las mujeres corresponde al valor calculado en base a la población observada, es decir quienes están efectivamente trabajando y no sobre la población total femenina.

Los datos de la muestra pueden resultan sesgados en un sentido negativo, ya que existe una proporción de la población capaz de percibir mayores salarios y que sin embargo decide no trabajar. Por lo que el promedio de los valores observados será menor que la que resultaría si estas mujeres estuviesen incorporadas en el mercado laboral.

Como afirma (Paz, 1998) y (Di Paola & Berges, 2000), el sesgo de selección resulta relevante porque su tratamiento introducirá cambios en las medidas de la discriminación.

Como se mencionó en la introducción del presente trabajo, se desarrollan dos metodologías diferentes para calcular la brecha existente de ingresos laborales entre varones y mujeres: 1) la "Metodología de Regresión Lineal Múltiple", y la segunda, 2) "Metodología de (Oaxaca, 1973) y (Blinder, 1973) con corrección del sesgo por selección de (Heckman, 1979)"

2.1 Modelos econométricos de cálculo de la brecha de ingresos entre el hombre y la mujer

2.2.1 Modelo 1. Regresión Lineal Múltiple con variables dummy

El modelo de regresión múltiple tiene como objetivo explicar el comportamiento de una variable (llamada endógena, explicada o dependiente), "Y", a través de la información brindada por los valores que tomas otras variables denominadas explicativas (exógenas o independientes), designadas como X_1, X_2, ..., $X_{k.}$ (Perez, 2006).

El modelo lineal queda expresado de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Los coeficientes o parámetros $\beta_{1,}\ \beta_{2,}$ … β_{κ} denotan el efecto que las variables explicativas X_1, X_2,… X_k tienen sobre la variable explicada *Y*. El coeficiente b_0 es denominado *termino constante* o *independiente* en el modelo y, el termino ε es el termino de error del modelo.

Se dispone de T observaciones para cada una de las variables endógena y exógena. Ampliando el modelo:

$$Y = \beta_0 + \beta_1 X_1 t + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_{kt} \quad donde: t = 1, 2, 3,... T$$

La aparición de un término independiente β_0 en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor es siempre 1.

Supongamos que la relación entre la variable *Y* y el conjunto de variables X_1, X_2, …, X_k es como describe el modelo enunciado, y que se dispone de un conjunto de *T* observaciones tanto para las variables endógenas como exógenas. Por otra parte, se asignan valores a los parámetros $\beta_{1,}\ \beta_{2,}$ … β_{κ} a partir de información muestral (Perez, 2006).

El modelo lineal queda formulado bajo las siguientes hipótesis clásicas:

- Las variables X_1, X_2, …, X_k son deterministas, es decir, no son variables aleatorias, ya que su valor es constante y proveniente de una muestra tomada, y además, no están correlacionas con el termino de error ε, es decir, $E(\varepsilon|X_1, X_2, \ldots, X_k) = 0$ (hipótesis de *exogeneidad*)
- El termino de error, ε, es una variable aleatoria con esperanza igual a cero y matriz de covarianzas constantes y diagonal (llamada matriz escalar). Entonces, para todo *t*, la variable e_t tiene media cero y varianza σ^2 que no depende de *t*, y además *Cov* $(\varepsilon_{i,}\ \varepsilon_j) = 0$ para todo *i* y para todo *j* distintos entre sí. El hecho de que la varianza de ε_t sea constante para todo *t* o bien, que no dependa de *t*, se denomina hipótesis de *homescedasticidad* y puede también expresarse como $V(\varepsilon|X_1, X_2, \ldots, X_k) = \sigma^2$ y $V(Y|X_1, X_2, \ldots, X_k) = \sigma^2$. El hecho de que *Cov* $(e_{i,}\ e_j) = 0$ para todo *i* distinto de *j* se denomina hipótesis de *no autocorrelación*.
- Y es una variable aleatoria ya que depende de otra variable aleatoria, el termino error ε. Además la relación entre *Y* y X_1, X_2, …, X_k es efectivamente lineal. Esto conforma la hipótesis de *linealidad*.
- Se supone además la ausencia de errores de especificación, es decir, se supone que todas las variables *X* relevantes para la explicación de la variable *Y*, están incluidas en la definición del modelo lineal.
- Las variables X_1, X_2, …, X_k son linealmente independientes entre sí, no existiendo relación lineal exacta entre ellas. Esta hipótesis se denomina hipótesis de *independencia* y si no se cumple, se dice que el modelo presenta *multicolinealidad*.
- También se considera la hipótesis de *normalidad* de los residuos, el cual consiste en que las variables ε_t sean normales para todo *t*.

Suponiendo que una muestra tiene el siguiente modelo (modelo tipo *nivel-nivel*):

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_{ki} \quad donde: i = 1, 2, 3, \ldots n$$

Con:

$$E(Y_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

$$V(Y_i|X_{1i}, X_{2i}, \ldots, X_{ki}) = \sigma^2$$

En el caso de que todas las variables permanezcan constantes, a excepción de X_j, se tiene:

$$\Delta E(Y|X_1, X_2, \ldots, X_k) = \beta_j \Delta X_j$$

$$\Rightarrow \frac{\Delta E(Y|X_1, X_2, \ldots, X_k)}{\Delta X_j}$$

Se puede interpretar el coeficiente β_j como el número de unidades que varía la media Y cuando X_j varía en una unidad (permaneciendo el resto de las variables constantes). El termino constante β_0 se interpreta como el pronóstico de Y cuando las X_j se anulan.

Modelo	***Variable dependiente***	***Variable independiente***	***Interpretación de*** β_1
Nivel-nivel	*y*	x	$\Delta y = \beta_1 \Delta x$
Nivel-Log	y	log(*x*)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-nivel	log(y)	*x*	$\% \Delta y = (100\beta_1)\, \Delta x$
Log - log	log(y)	log(*x*)	$\% \Delta y = \beta_1 \% \Delta x$

<u>Estimación mediante el modelo de Mínimos Cuadrados Ordinarios (MCO)</u>

Suponiendo nuevamente que se quiere ajustar el modelo de regresión lineal múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_k$$

y se tiene un conjunto de *T* observaciones para cada una de las variables endógenas y exógenas. Se puede escribir el modelo de la forma:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_t \quad donde: t = 1, 2, 3, \ldots T$$

Como se mencionó anteriormente, la aparición (no necesaria) de un término independiente en el modelo puede interpretarse como la presencia de una primera variable X_0 cuyo valor sea siempre 1.

El criterio de MCO considera que la función que mejor se ajusta a los datos es la que minimiza la varianza del error ε, lo que es equivalente a minimizar:

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t = (\beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt})2$$

Derivando respecto de los parámetros $\beta_0, \beta_1, \ldots \beta_\kappa$ e igualando a cero tenemos:

La *suma residual,* ya introducida previamente, es un indicador del nivel de error del modelo en su intento de explicar la evolución temporal de la variable Y_t,

Sabiendo que:

$$SR = \sum_{i=1}^{T} (Y_t - \hat{Y})^2 = \hat{\varepsilon}'\hat{\varepsilon} = \left(Y - X\hat{B}\right)'\left(Y - X\hat{B}\right) = Y'Y - \hat{B}'X'Y = Y'Y - \hat{Y}'Y$$

Luego podemos escribir la igualdad $Y\hat{}'Y = Y\hat{}'Y + \hat{\varepsilon}'\hat{\varepsilon}$, y si a los dos miembros de esta igualdad, se resta $T\bar{Y}^2$, tenemos que:

$$(Y'Y - T\bar{Y}^2) = (\hat{Y}'Y - T\bar{Y}^2) + \hat{\varepsilon}'\hat{\varepsilon}, \text{ o sea, } ST = SE + SR$$

Siendo:

- ST = suma total
- SE = suma explicada
- SR = suma residual

A estos tres términos se les llama *Suma de Cuadrado.*

A cada suma de cuadrados dividida por sus grados de libertad se le llama cuadrado medio. Bajo la hipótesis de normalidad de los residuos, *SE* se distribuye según una Chi-cuadrado con *k* grados de libertad, SR según una Chi-cuadrado con *T-k-1* grados de libertad, y ST según una Chi-cuadrado con *n-1* grados de libertad. Por tanto, el *Cuadrado Medio explicado por el modelo* será *CME=SE/k,* y el *Cuadrado Medio residual* será *CM(R)=SR/(T-k-1).*

Se define el *coeficiente de determinación (R^2)* como una medida descriptiva del ajuste global del modelo cuyo valor es el cociente entre la variabilidad explicada (o *suma explicada*) y la variabilidad total (o *suma total),* es decir, R^2=SE/ST=1-SR/ST

Un modelo será mejor cuanto mayor sea su R^2, aunque esta afirmación no sea demasiado severa, ya que este coeficiente depende mucho de nuevas variables introducidas en el modelo, aunque estas empeoren la calidad de la regresión. Este problema se soluciona sustituyendo R^2 por el coeficiente de determinación corregido, que para muestras grandes ya no dependerá del número de variables del modelo (Perez, 2006).

El *coeficiente de correlación múltiple* es la raíz cuadrada del coeficiente de determinación, su valor es R.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{T-1}{T-k-1}$$

Cuando $T \to \infty$, es decir, en muestras grandes, $(T\text{-}1)/(T\text{-}k\text{-}1) \to 1$ y no depende de *k,* que es el número de variables del modelo. Además, $T \to \infty \Rightarrow \bar{R}^2 \to R^2$.

Ahora ya se puede considerar a $\bar{R}^2$ como una buena medida de la calidad de la regresión. El modelo será tanto mejor cuanto mayor sea el coeficiente de determinación corregido $\bar{R}^2$.

El estadístico $\frac{(\hat{B}-B)'X'X(\hat{B}-B)}{k\hat{\sigma}^2}$ sigue una distribución *F(k,T-k-1)*.

Este estadístico permitirá hallar *regiones de confianza* α *un nivel de significación a para el conjunto de parámetros* β_j *del modelo.* Este estadístico también nos permitirá contrastar la hipótesis nula $\beta_1 = \beta_2 = \ldots = \beta_\kappa = 0$

El cuadro del análisis de la varianza quedará como sigue:

Fuente de variación	**Suma de cuadrados**	**Grados de Libertad**	**Cuadrados medios**	**F**
Modelo residual	SE	*K*	CM(E)=SE/k	$\frac{CM(E)}{CM(R)}$
	SR	*T-k*-1	CM(R)=SR/(T-*k*-1)	
Total	**ST**	***T*-1**		

El estadístico más general:

$$T = \frac{(D\hat{B} - DB)'[D(X'X)^{-1}D']^{-1}(D\hat{B} - DB)}{k\hat{\sigma}^2}$$

también sigue una distribución *F(k, T-k-1*) para utilizar una matriz adecuada D.

Consistencia de los estimadores MCO

El teorema de Gauss-Markov asegura que el modelo de regresión lineal bajo sus supuestos típicos, los estimadores MCO de los parámetros $\beta_o, \beta_1, \beta_2 \ldots \beta_\kappa$ son los de menor varianza entre los estimadores lineales e insesgados. Además, los estimadores MCO, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_k$ son estimadores consistentes de $\beta_o, \beta_1, \beta_2 \ldots \beta_\kappa$:

$$p\lim \hat{\beta}_j = \beta_j \ \ donde\ j=0, 1, \ldots k$$

Entonces, los estimadores MCO son consistentes de los parámetros porque su sesgo tiene a cero cuando el tamaño de la muestra tiende a infinito.

Según (Perez, 2006), en la econometría es muy común encontrarse con ***heteroscedasticidad*** en datos de corte transversal, es decir, con varianza no constantes en los termino de error; también con ***multicolinealidad*** (relación lineal entre las variables explicativas), ***no linealidad*** en la expresión matemática del modelo, errores de especificación o lo que es lo mismo, no inclusión de variables adecuadas ni la forma funcional óptima para definir el modelo. Por otra parte, suele encontrarse ***endogeneidad y regresores estocásticos*** (dependencias entre la perturbación y las variables explicativas y regresores aleatorios). A continuación, una breve referencia de cada uno de estos posibles incumplimientos, que generan problemas en las estimaciones.

El hecho de que la varianza de μ_t sea constante para toda *t*, es decir, que no dependa de *t,* se denomina hipótesis de ***homoscedasticidad***. La importancia del incumplimiento de esta hipótesis se encuentra en que los estimadores obtenidos por MCO no son de varianza mínima, aunque sigan siendo insesgados. Además, para cada variable del modelo se estimará una varianza del error.

En el modelo lineal $Y = XB + \mu$, uno de los supuestos relevantes es que las variables $X_1, X_2, \ldots X_k$ son linealmente independientes, es decir, no existe relación lineal exacta entre ellas. Esta hipótesis se denomina **hipótesis de independencia,** y cuando no se cumple decimos que el modelo presenta **multicolinealidad.**

En caso de que exista, el tener fuerte asociación lineal entre las variables explicativas $X'X$ tendría determinantemente cercano a 0 y no sería calculable $(X'X)^{-1}$, *por lo que* entonces no se podría hallar el vector de estimaciones de los parámetros $(X'X)^{-1} X'Y$ [1].

El problema de la falta de normalidad en los residuos

Una de las hipótesis importantes a cumplir en el modelo de regresión múltiple es que los residuos tengan una distribución normal. Aunque esa hipótesis no es necesaria para obtener los estimadores de los parámetros del modelo por el método de los mínimos cuadrados ordinarios, sí es estrictamente necesaria para la realización de la inferencia del modelo.

Para probar la normalidad de los residuos se puede utilizar cualquier contraste de ajuste a una distribución normal, por ejemplo, el de la Chi-cuadrado o el de Kolmogorov-Smirnov. Además, existen contrastes específicos para comprobar el ajuste de un conjunto de datos a una distribución normal, por ejemplo, el Contraste de normalidad de Shapiro Wilks y los de asimetría, curtosis y Jarque-Bera.

---

[1] Los indicadores más comunes de la presencia de multicolinealidad son los siguientes:
- Valores altos en módulo en la matriz de correlaciones de las variables explicativas.
- Poca significatividad de las variables *X* y a la vez R^2 alto.
- Gran significatividad conjunta del modelo (gran rechazo conjunto de $R^2 = 0$)
- Influencia en las estimaciones de la eliminación de una observación en el conjunto de datos.
- **Factores de inflación** de la varianza $VIF = 1/(1-R_j^2)$ elevados (›10), donde R_j^2 es el R^2 de la regresión auxiliar de la variable explicativa *j* en función de las demás variables explicativas.

Soluciones para la falta de normalidad en los residuos

Habitualmente la falta de normalidad en los residuos suele provenir de la presencia de datos atípicos que generan una distribución más apuntada o no simétrica. Estos problemas en los residuos suelen aparecer cuando se omiten variables relevantes en el modelo o cuando existe falta de linealidad en la especificación de este. Si los problemas citados se arreglan, los problemas de normalidad de los residuos suelen solucionarse.

Sin embargo, cuando los residuos no son normales por la presencia de más de una moda, los datos suelen provenir de varias poblaciones, lo cual se puede arreglar con la introducción de variables ficticias en el modelo para las diferentes poblaciones. En otras ocasiones, la solución para la falta de normalidad es la transformación adecuada de las variables para conseguirla, por ejemplo, la transformación de Box Cox y sus derivados.

Error de especificación en la selección de las variables explicativas

Las especificaciones más importantes del modelo lineal relativa a la matriz X es que sea una matriz no estocástica de rango pleno igual a *k* (ausencia de multicolinealidad). De acuerdo con (Perez, 2006), puede haber posibles problemas adicionales con *X,* entre los que se destacan:

1. ***Exclusión de variables relevantes (variables omitidas).*** La teoría económica enseña que el ingreso y los precios afectan conjuntamente a la demanda, por lo tanto, si aislamos el ingreso de la ecuación de la demanda no esperamos obtener un buen estimador para la elasticidad del precio. Sin embargo, en situaciones más complicadas, no suele ser tan evidente averiguar cuáles son las variables para incorporar en una relación, lo que puede llegar a convertirse en un importante problema de especificación.
2. ***Inclusión de variables irrelevantes (redundantes).*** En este caso, la hipótesis incluye variables que nos deberían estar presentes en la ecuación. De todos modos, las consecuencias sobre los procedimientos de inferencia suelen ser menos graves que en los casos donde se omiten variables relevantes.

Existen contrastes para observar si un modelo cuenta con variables omitidas, uno de ellos es el **test de la razón de verosimilitud para variables omitidas** el cual permite añadir un conjunto de variables a una ecuación existente y contrastar si representan una contribución significativa a la explicación de la variable dependiente. La hipótesis nula de este contraste es que los regresores adicionales no son conjuntamente significativos.

Error de especificación en la forma funcional

Puede ocurrir que a pesar de que las variables incluidas en un modelo sean correctas, la forma funcional lineal que las relaciona sea incorrecta. En este caso se presenta un problema de no linealidad.

Una relación $Y = f(X_2, X_3)$ puede especificarse como $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \mu$ o, como $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_2 X_2^2 + \gamma^3 X_3^2 + \delta(X_2 X_3) + \mu$. La segunda ecuación permite tanto una respuesta cuadrática a los regresores como un efecto de interacción. El efecto de interacción se basa en una nueva variable, que es el producto de los dos regresores. Por lo tanto, el efecto esperado de un cambio unitario en X_2 será $\beta_2 + 2\gamma_2 + \delta X_3$, dependiendo de β_2 y

de los niveles de X_2 y X_3. Del mismo modo, el efecto esperado de un cambio unitario en X_3 dependerá tanto del nivel de X_2 como del de X_3. Cuando el error de especificación consiste en utilizar la primera ecuación en lugar de la segunda, se corrige añadiendo los términos X_2^2, X_3^2, y $(X_2\ X_3)$. En otros casos, será necesaria una especificación intrínsecamente no lineal.

2.2.1.1 El Modelo de regresión lineal múltiple para la determinación de la brecha de ingresos por género. Una aplicación empírica

A continuación, se presenta una tabla con las variables utilizadas (codificación propia, descripción, codificación EPH, y construcción del indicador) en el primer modelo (Modelo 1), de "Regresión lineal múltiple con variables dummy", como también, aquellas pertenecientes al Modelo "Metodología de (Oaxaca, 1973) y (Blinder, 1973) con corrección del sesgo por selección de (Heckman, 1979)" (Modelos 2 y 3). Además, en la tabla 2, se presenta la recodificación propia de las ramas laborales, de acuerdo con la Clasificación de Actividades Económicas para Encuestas Sociodemográficas del MERCOSUR (CAES-Mercosur).

Tabla Nº 1. Descripción de las variables utilizadas en los métodos econométricos para el cálculo de la brecha de ingresos entre el hombre y la mujer

	Variable	Descripción	Cog. EPH	Construcción
Modelo 1	**Educación (Educ)**	Cantidad de años de educación formal	CH12; CH13; CH14	
	Experiencia (EXPER 6)	Edad menos años de instrucción menos 6 (considerados de no asistencia a la educación formal)	CH06	CH6-EDUC-6
	Mujer	¿es mujer? (dummy)	CH04	CH4 = 2
	Concubinato	¿esta casada o convive con su pareja? (dummy)	CH07	CH07 = 1 y 2
	Horas Trabajadas	Horas trabajadas en la ocupación principal por mes	PP3E_TOT; PP3F_TOT	((PP3E_TOT+PP3F_TOT)/7)*30
	Ing. NO Labor	¿percibe ingresos no laborales? (dummy)	T_Vi	T_Vi > 0
	Jerarquía	Directivo, jefe, trabajador (dummy)	En función al Clasificador Nacional de Ocupaciones (Versión 2001) - INDEC	
	Calificación	Categoria (dummy): profesional, técnico, operativo		
	Rama laboral	Industria, construcción, comercio, educación, servicios domésticos, transporte y comunicaciones, administración pública, otras ramas (Variable dummy)	Readaptación de los códigos de CAES-MERCOSUR	
	Educ*Exper6			
	Mujer*HsTrabajadas		PP3E_TOT	
Modelo 2	Edad		CH06	
	Edad al cuadrado			
	Instrucción	Cantidad de años de educación formal	CH12; CH13; CH14	
	Cantidad de miembros en el hogar (CANTMIEMBROS)	Cantidad total de miembros en el hogar de la mujer	IX_TOT	
	Cantidad de miembros menos de 5 años (CANTMEN5)	Cantidad total de miembros en el hogar de la mujer menores de 5 años	IX_TOT; CH06	IX_TOT<5
	Cantidad de miembros en el hogar de entre 6 y 14 años (CANT6A14)	Cantidad total de miembros en el hogar de la mujer de entre 6 y 14 años	IX_TOT; CH06	6<IX_TOT<14
	Perceptores	Cantidad de perceptores en el hogar		
	Jefatura Femenina (MUJERJEFE)	Indica si la mujer es o no jefe de hogar (variable dummy)	CH03; CH04	CH03=1 y CH04=2
Modelo 3	**Ingreso Laboral (YL)**	**Ingresos obtenidos del trabajo**	p21; TOT_P12	P21+TOT_P12
	Ingreso Laboral por Hora (YLpoHs)	Ingresos obtenidos por horas de trabajo	PP3E_TOT; P21	P21/PP3E_TOT
	Educación (Educ)	Cantidad de años de educación formal		
	Experiencia (EXPER 6)	Edad menos años de instrucción menos 6 (considerados de no asistencia a la educación formal)		
	Horas Trabajadas (HsTrabaj)	Horas trabajadas en la ocupación principal por mes	PP3E_TOT;+PP3F_TOT	(PP3E_TOT+PP3F_TOT)/7)30

Fuente: Elaboración propia en base a EPH y (Di Paola & Berges, 2000); (Oaxaca, 1973) y (Blinder, 1973)

Tabla Nº 2. Recodificación de las ramas de actividades económicas según CAES

Ramas	CAES	CODIGO
Actividades primarias	01 al 09	1
Industria	10 al 33	2
Construcción	40	3
Comercio	45 al 48	4
Educación	85	5
Servicio Financiero e inmobiliario	64;65;66; 68	6
Otros servicios	35 al 39; 77 al 82 ; 86 al 88 ; 90 al 96 -	7
Servicios Doméstico	97;98	8
Transporte y comunicaciones	49 al 53; 58 al 63	9
Administración Pública	83; 84	10
Otras ramas	99	11
Servicios proefesionales, administrativos y de apoyo	69 al 75	12
Hoteles y restaurantes	55 al 56	13

Fuente: Elaboración propia

A continuación, se representa el modelo de regresión lineal múltiple, a partir de la salida del software econométrico Stata.

```
regress LnYL Educ Exper6 Mujer Concubinato HsTrabaj YNoL Direccion Profesional Tecnico
Operativo Industria Comercio FinancInmobil OtrosScios ScioDomest Educacion TranspComunic
AdminPublic MenoresCuatroAños EducXExper6 MujerXHsTrabaj [fweight=Pondiio]
```

Source	SS	df	MS	Number of obs	=	112,555
				F(20, 112534)	=	8526.63
Model	37281.488	20	1864.0744	Prob > F	=	0.0000
Residual	24601.966	112,534	.218618071	R-squared	=	0.6024
				Adj R-squared	=	0.6024
Total	61883.454	112,554	.549811237	Root MSE	=	.46757

LnYL	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Educ	.0585697	.0008605	68.06	0.000	.0568831	.0602562
Exper6	.0180527	.0003639	49.61	0.000	.0173395	.018766
Mujer	-.5834958	.0086123	-67.75	0.000	-.6003757	-.5666159
Concubinato	.1576957	.0032248	48.90	0.000	.1513751	.1640163
HsTrabaj	.0015847	.0000366	43.32	0.000	.0015131	.0016564
YNoL	.1621524	.0041011	39.54	0.000	.1541142	.1701906
Direccion	.150125	.0142915	10.50	0.000	.1221139	.178136
Profesional	.5318079	.0077694	68.45	0.000	.5165799	.5470359
Tecnico	.4427208	.0058853	75.23	0.000	.4311858	.4542559
Operativo	.291168	.0042213	68.98	0.000	.2828943	.2994417
Industria	-.4988367	.0081502	-61.21	0.000	-.514811	-.4828625
Comercio	-.1337729	.0052065	-25.69	0.000	-.1439775	-.1235683
FinancInmobil	.120505	.0102485	11.76	0.000	.1004181	.1405918
OtrosScios	-.184755	.0042113	-43.87	0.000	-.193009	-.176501
ScioDomest	-.3180419	.0063128	-50.38	0.000	-.330415	-.3056689
TranspComunic	-.0222286	.0075162	-2.96	0.003	-.0369603	-.0074969
AdminPublic	.0634179	.0044721	14.18	0.000	.0546526	.0721832
MenoresCuatroAños	.1182738	.0034677	34.11	0.000	.1114772	.1250704
EducXExper6	.0003307	.0000301	11.00	0.000	.0002718	.0003897
MujerXHsTrabaj	.0029243	.0000505	57.89	0.000	.0028253	.0030233
_cons	7.802338	.0127478	612.05	0.000	7.777353	7.827324

A partir de 112.555 observaciones, y con un R-cuadrado ajustado mayor al 60%, y con todos los coeficientes de las variables estadísticamente significativas, a un valor *p,* menor a 0.05, el modelo de regresión lineal múltiple quedaría expresado, de la siguiente manera:

*LnYL=0.0585697*Educ+0.0180527*Exper6-0.5834958 * Mujer + 0.1576957 * Concubinato + 0.015847 * HsTrabaj + 0.1621524 * YNoL + 0.150125 * Direccion + 0.5318079 * Profesional +0.4427208 * Tecnico+0.291168*Operativo-0.4988367*Industria- 0.1337729 * Comercio +0.120505 * FinancInmobil-0.184755OtrosScios-0.3180419*ScioDomest -0.0222286 * TranspComunic + 0.0634179 * AdminPublic + 0.1182738 * MenoresCuatroAños + 0.0003307 EducXExper6 + 0.0029243 *MujerXHsTrabaj [fweight=Pondiio]*

Variable	Obs	Mean	Std. Dev.	Min	Max
YL	112,555	15749.87	10890.32	500	84000
YLporHs	112,555	110.6898	116.4411	11.66667	1866.667
Educ	112,555	12.11189	3.710216	1	18
Exper6	112,555	20.14305	12.85911	-2	63
HsTrabaj	112,555	161.8807	63.23507	12.85714	360

La media de los ingresos laborales, correspondientes al segundo trimestre de 2018, era $15.749,87. El ingreso promedio laboral por hora trabajada, ascendía a $110,69. La cantidad promedio de años formales de educación, ascendía a 12,1 y la experiencia, a 20,1 años. El número promedio de horas trabajadas mensuales de la ocupación principal era de 161,88 (5,4 horas trabajadas diarias).

```
. by Mujer: summarize YL YLporHs Educ Exper6 HsTrabaj[fweight=Pondiio]
```

-> Mujer = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
YL	62,058	16685.02	10332.52	800	60000
YLporHs	62,058	106.6976	135.5682	11.66667	1866.667
Educ	62,058	11.69292	3.637929	1	17
Exper6	62,058	19.52467	12.66068	-2	63
HsTrabaj	62,058	181.2915	58.94349	25.71429	360

-> Mujer = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
YL	50,497	14600.62	11434.77	500	84000
YLporHs	50,497	115.5959	87.1285	13.07639	551.5152
Educ	50,497	12.62679	3.73335	1	18
Exper6	50,497	20.903	13.05893	-1	56
HsTrabaj	50,497	138.0259	60.09181	12.85714	360

De un total de 54.497 mujeres, la media de los ingresos laborales, correspondientes al segundo trimestre de 2018, era $14.600,62. El ingreso promedio laboral por hora trabajada, ascendía a $115,60. La cantidad promedio de años formales de educación, ascendía a 12,6 y la experiencia, a 20,9. El número promedio de horas trabajadas mensuales de la ocupación principal, era de 138,0 (4,6 horas trabajadas diarias).

De un total de 62.058 hombres, la media de los ingresos laborales, correspondientes al segundo trimestre de 2018, era $16.685,02. El ingreso promedio laboral por hora trabajada ascendía a $106,70. La cantidad promedio de años formales de educación, ascendía a 11,7 y la experiencia, a 19,5. El número promedio de horas trabajadas mensuales de la ocupación principal, era de 181,3 (6,0 horas trabajadas diarias).

A continuación, se plantean los supuestos del modelo de regresión lineal múltiple, con sus respectivos testeos (pruebas)

1. Los valores de las variables independientes han de ser fijos.

2. El número de observaciones debe ser mayor que el número de variaciones independientes: $n > k$

3. Debe haber suficiente variabilidad en los valores de las variables independientes: $Var\ (x_i) > L$

4. El termino de perturbación está normalmente distribuido

$$\varepsilon_i \sim N(0,\sigma)$$

5. Para cada conjunto de casos con una x_i dada, el valor medio de la perturbación (ε_i) es cero

$$\forall x_i \quad E(\varepsilon_i) = 0$$

6. En el caso de que las sean estocásticas, no existe correlación entre estas y los términos de perturbación.

$$Cov\ (x_i, \varepsilon_i) = 0$$

7. Para cada conjunto de casos con una dada, la varianza de es constante y homocedástica.

$$\forall x_i \quad Var(\varepsilon_i) = \sigma^2$$

8. No hay relación exacta (no hay multicolinealidad) en los regresores.

$$Cov\ (z_{xi}, z_{xj}) < 1; \ (i \neq j)$$

9. No existe autocorrelación entre las perturbaciones

$$Cov\ (\varepsilon_i, \varepsilon_j) = 0; \ (i \neq j)$$

10. El modelo de regresión es lineal en sus parámetros.

11. El modelo de la regresión está correctamente especificado.

Pruebas:

1. Depende que se cumpla el (6)

2. Se cuenta con N=112.555; n=405 y k=20, se recomienda una proporción de cinco veces superior del número de casos sobre el de parámetros

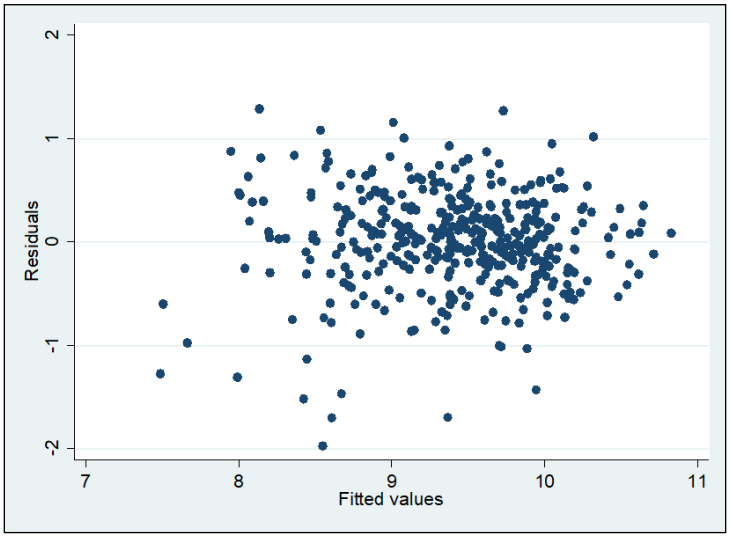
3. Obteniendo la desviación típica de las variables independientes. O bien, con el coeficiente de variación.

Variable	Obs	Mean	Std. Dev.	Min	Max
Educ	112,555	12.11189	3.710216	1	18
Exper6	112,555	20.14305	12.85911	-2	63
Mujer	112,555	.4486429	.4973577	0	1
Concubinato	112,555	.5666918	.4955344	0	1
HsTrabaj	112,555	161.8807	63.23507	12.85714	360
YNoL	112,555	.1867976	.3897507	0	1
Direccion	112,555	.0126516	.111766	0	1
Profesional	112,555	.0641109	.2449514	0	1
Tecnico	112,555	.1345476	.3412412	0	1
Operativo	112,555	.5121585	.4998544	0	1
Industria	112,555	.0351828	.1842424	0	1
Comercio	112,555	.1187864	.3235384	0	1
FinancInmo~l	112,555	.0208343	.1428299	0	1
OtrosScios	112,555	.2115588	.4084154	0	1
ScioDomest	112,555	.1052286	.3068491	0	1
TranspComu~c	112,555	.047479	.2126621	0	1
AdminPublic	112,555	.193612	.3951301	0	1
MenoresCua~s	112,555	.2710941	.4445266	0	1
EducXExper6	112,555	230.202	153.5984	-34	952
MujerXHsTr~j	112,555	61.92432	79.57777	0	360

```
. summarize LnYL?*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
LnYLresidual	405	.0049294	.4711638	-1.976293	1.288882
LnYLrstand~d	405	.0105401	1.007792	-4.227184	2.756882
LnYLrstudent	405	.0105376	1.007808	-4.227501	2.756963

A continuación, se grafican los residuos con los valores predichos de la variable dependiente. Con esto podemos comprobar si la media = 0 de los residuos es constante a lo largo de los distintos valores de X (supuesto 5), en cuyo caso también se cumpliría la no correlación entre ε_i y los X_i (supuesto 6).

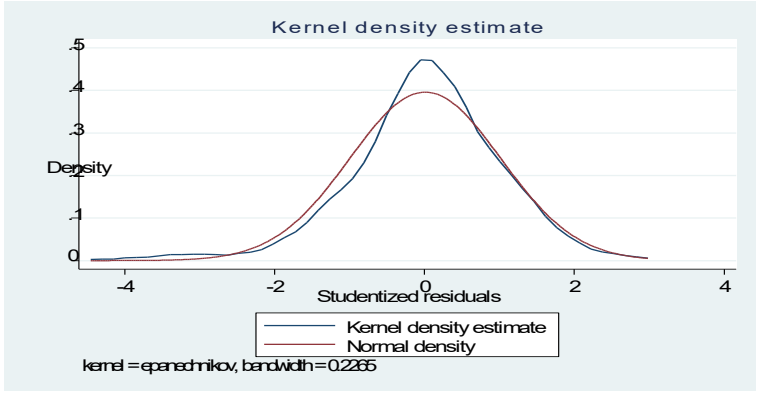


Otro requisito es evaluar la normalidad en la distribución de los residuos (supuesto 4). Para comprobarlo, usamos las pruebas de Shapiro-Wilk y Shapiro Francia.

```
. swilk LnYLresidual- LnYLrstudent

                    Shapiro-Wilk W test for normal data

    Variable |       Obs        W           V          z       Prob>z
-------------+--------------------------------------------------------
LnYLresidual |       405    0.97552      6.815      4.569    0.00000
LnYLrstand~d |       405    0.97552      6.815      4.569    0.00000
LnYLrstudent |       405    0.97551      6.817      4.570    0.00000

. sfrancia LnYLresidual- LnYLrstudent

                   Shapiro-Francia W' test for normal data

    Variable |       Obs        W'          V'         z       Prob>z
-------------+--------------------------------------------------------
LnYLresidual |       405    0.97441      7.673      4.410    0.00001
LnYLrstand~d |       405    0.97441      7.673      4.410    0.00001
LnYLrstudent |       405    0.97441      7.674      4.411    0.00001
```

Cada variable puede rechazarse con un nivel de significatividad inferior al 0,05 la hipótesis nula de que la distribución es normal.

De un modo gráfico se puede comprobar utilizando un gráfico de probabilidades “pnom”, de cuantiles “qnorm”o el de superposición de las dos distribuciones, seguida de la opción normal (Mercado, Macías, & Bernardi, 2009).




```
. sktest LnYLresidual LnYLrstandard LnYLrstudent

                    Skewness/Kurtosis tests for Normality
                                                              ------ joint ------
    Variable |       Obs  Pr(Skewness)  Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+--------------------------------------------------------------------
LnYLresidual |       405        0.0000        0.0000        29.81        0.0000
LnYLrstand~d |       405        0.0000        0.0000        29.81        0.0000
LnYLrstudent |       405        0.0000        0.0000        29.81        0.0000
```

En el examen estadístico de los coeficientes de asimetría y curtosis se observa que existen estos dos problemas.

Otro de los diagnósticos es el de homocedasticidad (supuesto 7). Para probarlo, se utiliza la prueba de Cook-Weisberg (Mercado, Macías, & Bernardi, 2009):

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of LnYL

         chi2(1)      = 10436.42
         Prob > chi2  =   0.0000
```

A partir del resultado de esta prueba, se rechaza la hipótesis nula, es decir, hay presencia de heterocedasticidad.

Otro supuesto, es la ausencia de multicolinealidad (supuesto 8). El criterio más utilizado es el de la tolerancia, conocido como el *factor de inflación de la varianza (VIF)*.

```
. vif

    Variable |      VIF      1/VIF
-------------+----------------------
       Exper6 |    11.27    0.088702
   EducXExper6 |    10.99    0.091029
         Mujer |     9.45    0.105865
  MujerXHsTr~j |     8.32    0.120179
         Educ |     5.25    0.190556
     HsTrabaj |     2.75    0.363035
    Operativo |     2.29    0.436260
      Tecnico |     2.08    0.481580
   ScioDomest |     1.93    0.517640
  Profesional |     1.86    0.536272
  AdminPublic |     1.61    0.622038
   OtrosScios |     1.52    0.656593
     Comercio |     1.46    0.684522
         YNoL |     1.32    0.760223
  TranspComu~c |     1.32    0.760232
  Concubinato |     1.31    0.760608
    Direccion |     1.31    0.761293
  MenoresCua~s |     1.22    0.817438
    Industria |     1.16    0.861412
  FinancInmo~l |     1.10    0.906502
-------------+----------------------
    Mean VIF |     3.48
```

Como regla se recomienda que el factor no supere el valor de 10, lo que equivale al 0,10 de su inverso. Cuando una variable de la ecuación tiene un coeficiente de correlación múltiple con el resto de las variables superior a 0,95, los problemas de eficiencia de los estimadores serán altos.

Para verificar el supuesto 11, se utilizará el test de Ramsey:

```
. ovtest

Ramsey RESET test using powers of the fitted values of LnYL
       Ho:  model has no omitted variables
                F(3, 112531) =     828.22
                    Prob > F =      0.0000
```

El test nos muestra que el modelo ha omitido variables importantes.

A continuación, se presentan las soluciones a los problemas planteados (no cumplimiento de algunos supuestos):

Heterocedasticidad

- Siguiendo los trabajos de Huber (1967) y White (1982), se obtiene una regresión con errores típicos robustos, que conducirán a ser más exigentes a la hora de rechazar sus respectivas hipótesis nulas.

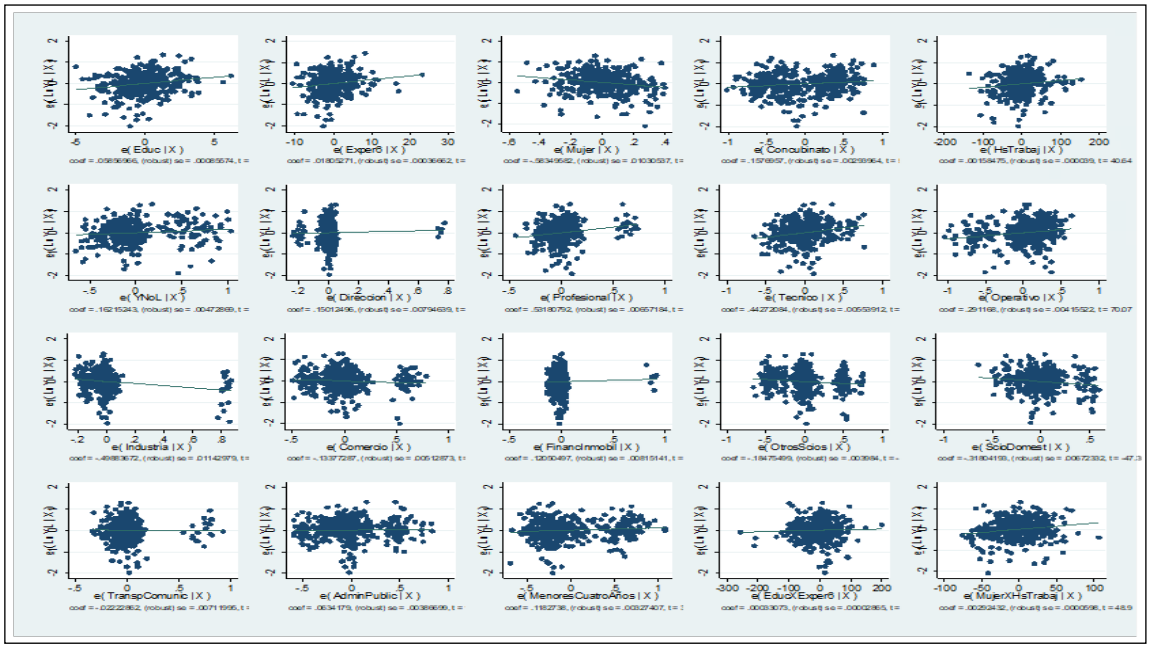
```
. regress LnYL Educ Exper6 Mujer Concubinato HsTrabaj YNoL Direccion Profesional Tecnico Operativo Industria  Comercio  FinancInmo
> bil OtrosScios ScioDomest TranspComunic AdminPublic MenoresCuatroAños EducXExper6 MujerXHsTrabaj [fweight=Pondiio], robust

Linear regression                               Number of obs     =    112,555
                                                F(20, 112534)     =    7987.64
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6024
                                                Root MSE          =     .46757

------------------------------------------------------------------------------
             |               Robust
        LnYL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Educ |   .0585697   .0008557    68.44   0.000     .0568924    .0602469
      Exper6 |   .0180527   .0003666    49.24   0.000     .0173341    .0187713
       Mujer |  -.5834958   .0103054   -56.62   0.000    -.6036942   -.5632974
  Concubinato |   .1576957   .0029396    53.64   0.000      .151934    .1634574
    HsTrabaj |   .0015847    .000039    40.64   0.000     .0015083    .0016612
        YNoL |   .1621524   .0047287    34.29   0.000     .1528843    .1714206
   Direccion |    .150125   .0079464    18.89   0.000     .1345502    .1656998
 Profesional |   .5318079   .0065718    80.92   0.000     .5189272    .5446886
     Tecnico |   .4427208   .0055391    79.93   0.000     .4318642    .4535774
  Operativo |    .291168   .0041552    70.07   0.000     .2830238    .2993122
   Industria |  -.4988367   .0114298   -43.64   0.000    -.5212389   -.4764345
   Comercio |  -.1337729   .0051287   -26.08   0.000    -.1438251   -.1237206
FinancInmobil |    .120505   .0081514    14.78   0.000     .1045283    .1364816
  OtrosScios |   -.184755    .003984   -46.37   0.000    -.1925636   -.1769464
  ScioDomest |  -.3180419   .0067233   -47.30   0.000    -.3312195   -.3048643
TranspComunic |  -.0222286   .0071199    -3.12   0.002    -.0361836   -.0082736
  AdminPublic |   .0634179    .003867    16.40   0.000     .0558386    .0709971
MenoresCuatroAños |   .1182738   .0032741    36.12   0.000     .1118567    .1246909
  EducXExper6 |   .0003307   .0000287    11.54   0.000     .0002746    .0003869
MujerXHsTrabaj |   .0029243   .0000598    48.90   0.000     .0028071    .0030415
       _cons |   7.802338   .0147613   528.57   0.000     7.773407     7.83127
------------------------------------------------------------------------------
```

Tratamiento de la normalidad

Para el tratamiento de la normalidad, según (Mercado, Macías, & Bernardi, 2009),

- Si la distribución es conocida, aunque no sea normal, se aplica estimación MV.
- Si la distribución es desconocida, se puede utilizar:
    - Transformaciones buscando normalidad
    - Regresión robusta
- Si la normalidad es debida a valor atípicos:
- Se utilizan variables ficticias
- Se eliminan si hay suficientes datos

Si no se verifica la normalidad del modelo, entonces los estimadores MCO dejan de ser MV y por tanto pierden la eficacia dentro de los estimadores insesgados, sin embargo, siguen siendo Estimadores Lineales Insesgados y Óptimos (ELIO). Mantienen la consistencia y la normalidad asintótica, pero también pierden la eficiencia asintótica (Mercado, Macías, & Bernardi, 2009).



Source	SS	df	MS	Number of obs	=	112,555
				F(20, 112534)	=	8526.63
Model	37281.488	20	1864.0744	Prob > F	=	0.0000
Residual	24601.966	112,534	.218618071	R-squared	=	0.6024
				Adj R-squared	=	0.6024
Total	61883.454	112,554	.549811237	Root MSE	=	.46757

LnYL	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Educ	.0585697	.0008605	68.06	0.000	.0568831	.0602562
Exper6	.0180527	.0003639	49.61	0.000	.0173395	.018766
Mujer	-.5834958	.0086123	-67.75	0.000	-.6003757	-.5666159
Concubinato	.1576957	.0032248	48.90	0.000	.1513751	.1640163
HsTrabaj	.0015847	.0000366	43.32	0.000	.0015131	.0016564
YNoL	.1621524	.0041011	39.54	0.000	.1541142	.1701906
Direccion	.150125	.0142915	10.50	0.000	.1221139	.178136
Profesional	.5318079	.0077694	68.45	0.000	.5165799	.5470359
Tecnico	.4427208	.0058853	75.23	0.000	.4311858	.4542559
Operativo	.291168	.0042213	68.98	0.000	.2828943	.2994417
Industria	-.4988367	.0081502	-61.21	0.000	-.514811	-.4828625
Comercio	-.1337729	.0052065	-25.69	0.000	-.1439775	-.1235683
FinancInmobil	.120505	.0102485	11.76	0.000	.1004181	.1405918
OtrosScios	-.184755	.0042113	-43.87	0.000	-.193009	-.176501
ScioDomest	-.3180419	.0063128	-50.38	0.000	-.330415	-.3056689
TranspComunic	-.0222286	.0075162	-2.96	0.003	-.0369603	-.0074969
AdminPublic	.0634179	.0044721	14.18	0.000	.0546526	.0721832
MenoresCuatroAños	.1182738	.0034677	34.11	0.000	.1114772	.1250704
EducXExper6	.0003307	.0000301	11.00	0.000	.0002718	.0003897
MujerXHsTrabaj	.0029243	.0000505	57.89	0.000	.0028253	.0030233
_cons	7.802338	.0127478	612.05	0.000	7.777353	7.827324

Interpretaciones del Modelo 1

En el Aglomerado Posadas, con datos del segundo trimestre de 2018 de la Encuesta Permanente de Hogares, el modelo de regresión lineal múltiple nos indica lo siguiente:

- Por cada año de educación formal, los ingresos laborales se incrementan un 5,9%.
- Por cada año de experiencia laboral, los ingresos aumentan un 1,8%.
- Las personas que están en situación de concubinato perciben un 15,8% más de ingresos.
- Aquellos que ejercen tareas de dirección, obtienen un 15,0% más de ingresos laborales.
- De acuerdo con la calificación profesional, aquellos que son profesionales ganan un 53,2% más de aquellos que no lo son; los técnicos, un 44,3% más, y lo de conocimiento operativo, un 29,1%.
- Respecto a las actividades económicas, las personas que trabajan en la industria obtienen un 49,9% menos de ingresos; en el comercio (-13,4%); en la administración

pública (6,3%); sector financiero e inmobiliaria (12,1%); transporte y comunicaciones (-2,2%); y en el servicio doméstico (-31,8%).
- Aquellas personas que tienen hijos menores de cuatro años perciben un 11,8% más de ingresos.
- La interacción entre experiencia y educación genera un mínimo efecto en el adicional de ingresos; sólo un 0,03% más.

Y, por último, con respecto al objetivo del presente trabajo, el modelo 1 estima que las mujeres reciben un 58,3% menos de ingresos que los varones.

2.0.2. Metodología de (Oaxaca, 1973) y (Blinder, 1973) con corrección del sesgo por selección de (Heckman, 1979)

El cálculo de la brecha de ingresos entre géneros se efectúa a partir de funciones de ingreso Mincerianas estimadas para ambos sexos y aplicando la técnica de A. Blinder (1973) y R. Oaxaca (1973).

Según (Di Paola & Berges, 2000), esta supone que en ausencia de discriminación los efectos generados por las dotaciones de capital humano deberían ser idénticos para ambos grupos; de manera que, si se equipararan las diferencias en dotaciones y segregación en ocupaciones determinadas, y aún se registraran diferencias, éstas podrían atribuirse a discriminación.

Es decir, se estima con un carácter residual, una vez que se identifican las diferencias del primer tipo o diferencias explicadas.

(1) $LN\ Y_v = X_v \beta_v + U_v$;
(2) $LN\ Y_m = X_m \beta_m + U_m$

En las anteriores ecuaciones de ingresos, los subíndices indican el sexo de los individuos, Y_i es el vector columna de los ingresos, X_i es la matriz de variables independientes y β el vector columna de coeficientes a estimar. Los *U* son los términos de perturbación.

Suponiendo que los errores se distribuyen normalmente con media cero ($U_v = U_m = 0$) y varianza constante y evaluando las funciones en los valores promedio de las variables de la muestra, se cumple para una regresión estimada por OLS que:

(3) $\overline{LnY_v} = \overline{X_v}\beta_v + U_v$

$\overline{LnY_m} = \overline{X_m}\beta_m + U_m$

De tal forma la diferencia en las medias de los logaritmos de ingreso estimados para ambos grupos es:

(4) $\overline{LnY_v} - \overline{LnY_m} = \overline{X_v}\beta_v - \overline{X_m}\beta_m$

Siendo la diferencia entre los vectores de coeficientes de ambos grupos:

(5) $\Delta\beta = \beta_v - \beta_m \therefore \beta_v = \beta_m - \Delta\beta$

Sustituyendo (4) en (3), la ecuación queda finalmente:

(6) $\overline{LnY_v} - \overline{LnY_m} = \beta_v (\overline{X_v} - \overline{X_m}) + \overline{X_m}\Delta\beta$

Esta ecuación expresa que el promedio entre los ingresos de ambos grupos se puede desagregar en los efectos de las diferencias en sus correspondientes dotaciones de capital humano y en los efectos de la discriminación, manifiestos por las diferencias en los coeficientes estimados.

Cuando se estiman funciones de ingresos pertinentes sobre todo a la población femenina ocupada, es decir sobre la base de aquellas que reciben ingresos, hay que tomar en cuenta que se incurre en un problema de sesgo de selección cuando la población femenina excluida del análisis no tiene las mismas características que la observada.

Para la corrección de este, se utiliza la técnica propuesta por Heckman que consiste primeramente en estimar una función Probit de participación laboral para el total de las mujeres de la muestra (de entre 15 y 65 años de edad), es decir, sumando las mujeres de la muestra, económicamente activas y no activas.

Una vez que se estima esta ecuación y considerando los residuos de esta se calcula el inverso de la ratio de Mills (), que se añade como un regresor más en las funciones de ingreso.

Si la nueva variable lambda resulta significativa se puede concluir que existe sesgo de selección y en este caso, los coeficientes que intervendrán en el cálculo de la discriminación serán los "corregidos" por sesgo, es decir los que resultan de la última ecuación. En caso de sesgo negativo, se "sobreestima" la brecha; ya que, si se incorporara al mercado laboral, el grupo autoseleccionado, el salario promedio sería mayor (Di Paola & Berges, 2000).

2.2.2.1 Modelos de variable dependiente limitada

La expresión funcional del modelo de análisis de la regresión múltiple es $y = F(x_1, x_2,..., x_n)$. La regresión múltiple admite la posibilidad de trabajar con variables dependientes cuyo rango de valores está restringido (variables binarias con valores o y 1, variables con valores enteros positivos, etc.). En general los modelos que admiten variables dependientes con rango restringido se denominan ***modelos de variables dependientes limitada.***

La mayoría de las variables económicas que se analizan presentan valores que están limitados de alguna manera, en muchas ocasiones porque deben ser positivos. Por ejemplo, el salario por hora, los precios de las viviendas, y los tipos de interés nominales deben ser mayores que cero. Pero no todas esas variables requieren un trato especial. No suele ser necesario ningún modelo econométrico especial para tratar las variables que son estrictamente positivas pero que toman muchos valores diferentes. Cuando la variable dependiente es discreta ***(modelos de elección discreta)*** y toma un reducido número de valores, no tiene sentido que la tratemos como si fuera una variable aproximadamente continua. El hecho de que la variable dependiente sea discreta no implica necesariamente que los modelos lineales no sean apropiados. Sin embargo, para respuestas binarias ***(modelos de elección binaria)***, suelen utilizarse modelos Logit y Probit y en ciertos casos el ***modelo lineal de probabilidad.*** También para respuestas múltiples ***(modelos de elección múltiple)*** se utilizan los modelos mencionados (Perez, 2006).

Modelos Logit y Probit

Se puede considerar estos modelos como de respuesta binaria:

$$P(Y = 1|X_1, X_2,..., X_k) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Que, para evitar los problemas del modelo lineal de probabilidad, se especifican como $Y=G(X\beta)$, donde G es una función que toma valores estrictamente entre 0 y 1 (0<G(Z)<1) para todos los números reales *z*. según las diferentes definiciones de G tenemos los distintos modelos de elección binaria.

En el caso Probit tenemos:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \emptyset(v)\boldsymbol{d}v$$

Donde $\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}}$ es la función de densidad normal (0,1).

La expresión del modelo Probit será:

$$Y = G(z) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \int_{-\infty}^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \mathbf{d}v$$

Los modelos Probit y Logit, como son modelos no lineales, no podremos estimar por MCO y tendremos que emplear métodos de máxima verosimilitud.

Supongamos que tenemos *n* observaciones idénticas e independientemente distribuidas (muestra aleatoria) que siguen el modelo:

$$P(Y = 1|\mathbf{X}) = G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

Para obtener el estimador de máxima verosimilitud (MV), condicionado a las variables explicativas necesitamos la función de verosimilitud:

$$L(\beta) = \prod_{Y_t=1} P_i \prod_{Y_t=0} (1 - P_i) = \prod_{i=1}^{n} G(X_i'\beta)^{Y_i}(1 - G(X_i'\beta)^{1-Y_i}$$

Con:

$$P(Y = 1|X_{1i}, X_{2i}, \dots, X_{ki}) = G(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) = G(X_i'\beta)$$

El estimador de MV de β es el que maximiza el logaritmo de la función de verosimilitud:

$$l(\beta) = \ln L\,(\beta) = lnL(\beta) = \sum_{i=1}^{n} [Y_i \ln G(X_i'\beta) + (1 - Y_i)\ln\,(1 - G(X_i'\beta))]$$

Que será un estimador consistente, asintóticamente normal y asintóticamente eficiente.

Las condiciones de primer orden serán:

$$= \sum_{i=1}^{n} \left[\frac{Y_i}{G(X_i'\beta)} - \frac{(1 - Y_i)}{(1 - G(X_i'\beta))} \right] X_i g(X_i'\beta) =$$

$$= \sum_{i=1}^{n} \left[\frac{Y_i - G(X_i'\beta)}{G(X_i'\beta)(1 - G(X_i'\beta)} \right] X_i g(X_i'\beta) = 0$$

Donde g(.) es la función de densidad de la normal o la logística (derivada de la función de distribución).

Cuando se interpretan las estimaciones en los modelos Probit y Logit, generalmente, lo que interesa es conocer el efecto de variaciones en una variable X_j sobre la probabilidad de respuestas, que si la variable es continua será:

$$\Delta \hat{P}(Y = 1|\mathbf{X}) \approx \left[g\ (\boldsymbol{X}\ \hat{\beta}) \hat{\beta}_j \right] \Delta X_j$$

Como $g\ (\boldsymbol{X}\ \hat{\boldsymbol{\beta}})$ depende de X habrá que calcular los efectos parciales para valores interesantes de X (las medias muestrales, los valores máximos y mínimos de la variable de interés, etc.) también se puede calcular el efecto parcial para cada individuo y después calcular su media.

El efecto parcial de una variable continúa X_j sobre la probabilidad de respuesta $P(Y=1|\mathbf{X})$

$$\frac{\partial P(Y = 1|\mathbf{X})}{\partial X_j} = \ g\ (\boldsymbol{X}\ \hat{\beta}) \hat{\beta}_j$$

Donde g(.) es la función de densidad de la logística *(logit)* o de la normal estándar *(probit)*. Este efecto varía de individuo a individuo. Como en el caso del Probit y del Logit, $g(z)>0$ para todo z, **el signo del efecto parcial de** X_j es el mismo que el de β_j.

El **efecto relativo de dos variables continúas** X_j y X_h no depende de X. Notese que el cociente de los efectos parciales es β_j/β_h.

Si X_1, por ejemplo, es una variable explicativa ficticia, el efecto parcial de que varíe de 1 a o vendrá dado por:

$$G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) - G(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

que también varía de un individuo a otro, pues depende de los valores de todas las X_j.

2.2.2.1.2 Una aplicación empírica del Modelo Probit de participación en el mercado de Trabajo.

El modelo planteado para la corrección del sesgo por selección, siguiendo los trabajos de (Esquivel, 2007), (Di Paola & Berges, 2000) y (Paz, 1998), entre otros, es:

$$P_i = \beta_0 + \beta_1 edad + \beta_2 edadcuadrado + \beta_3 instrucción + \beta_4 cantmiembro + \beta_5 cantmen5 + \beta_6 cant6a14 + \beta_7 perceptores + \beta_8 mujerjefe + \mu$$

Siendo μ el término residual

La edad de la mujer se incorporó porque se supone que, a mayor edad, mayor será la probabilidad de que la mujer participe en el mercado laboral, alcanzando un valor máximo de

probabilidad, a partir de la cual comenzaría a disminuir. La edad al cuadrado se incorporó para captar posibles efectos no lineales de sus variaciones.

La variable instrucción se construye a partir de la cantidad de años de la educación formal al momento de efectuada la encuesta. Es de esperar que el signo del parámetro sea positivo, por lo que estaría señalando que a mayor cantidad de años de educación formal o de instrucción, mayor es la probabilidad que las mujeres participen en el mercado laboral.

También se agregó la cantidad de miembros o integrantes del hogar, como representación del tamaño del hogar. Se parte del supuesto que a medida que se incrementa el número de miembros de la familia, disminuye la probabilidad de participación de la mujer en el mercado laboral, su costo de oportunidad es mayor, por ende, su salario de reserva.

Se incorporó la cantidad de menores de 5 años del hogar, suponiendo que, a mayor cantidad de estos, menor es la probabilidad de participación en el mercado de trabajo. Asimismo, otra variable que representa la cantidad de menores entre 6 y 14 años.

Se añadió una variable que mide la cantidad de perceptores de ingresos, con el supuesto que a mayor cantidad de integrantes que reciban alguna fuente de ingreso, menos es la probabilidad de que la mujer ingrese en el mercado laboral.

Por último, una variable dummy que mide si la jefatura del hogar corresponde a una mujer.

A partir de la utilización de los microdatos de la Encuesta Permanente de Hogares, correspondiente al segundo de trimestre de 2018, se aplica el modelo Probit, con el programa Eviews, que arroja las siguientes salidas:

Dependent Variable: PART
Method: ML - Binary Probit (Newton-Raphson / Marquardt steps)
Date: 03/18/19 Time: 00:44
Sample: 1 435
Included observations: 435
Convergence achieved after 5 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-5.308480	0.645607	-8.222461	0.0000
EDAD	0.258918	0.033316	7.771593	0.0000
EDADCUADRADO	-0.003146	0.000423	-7.430811	0.0000
INSTRUCCION	0.035823	0.013000	2.755569	0.0059
CANTMIEMBROS	-0.363772	0.057683	-6.306368	0.0000
CANTMEN5	0.291618	0.118799	2.454724	0.0141
CANT6A14	0.873970	0.274025	3.189387	0.0014
PERCEPTORES	0.698266	0.102049	6.842450	0.0000
MUJERJEFE	0.574163	0.192132	2.988371	0.0028

McFadden R-squared	0.278847	Mean dependent var	0.526437
S.D. dependent var	0.499876	S.E. of regression	0.410490
Akaike info criterion	1.039093	Sum squared resid	71.78190
Schwarz criterion	1.123410	Log likelihood	-217.0027
Hannan-Quinn criter.	1.072372	Deviance	434.0054
Restr. deviance	601.8214	Restr. log likelihood	-300.9107
LR statistic	167.8160	Avg. log likelihood	-0.498857
Prob(LR statistic)	0.000000		

Obs with Dep=0	206	Total obs	435
Obs with Dep=1	229		

Estimation Equation:
==========
I_PART = C(1) + C(2)*EDAD + C(3)*EDADCUADRADO + C(4)*INSTRUCCION + C(5)*CANTMIEMBROS + C(6)*CANTMEN5 + C(7)*CANT6A14 + C(8)*PERCEPTORES + C(9)*MUJERJEFE

Forecasting Equation:
==========
PART = 1-@CNORM(-(C(1) + C(2)*EDAD + C(3)*EDADCUADRADO + C(4)*INSTRUCCION + C(5)*CANTMIEMBROS + C(6)*CANTMEN5 + C(7)*CANT6A14 + C(8)*PERCEPTORES + C(9)*MUJERJEFE))

Substituted Coefficients:
==========
PART = 1-@CNORM(-(-5.30848035395 + 0.258917902401*EDAD - 0.00314573013722*EDADCUADRADO + 0.0358234296705*INSTRUCCION - 0.363772474798*CANTMIEMBROS + 0.291618315468*CANTMEN5 + 0.873970194277*CANT6A14 + 0.698265782852*PERCEPTORES + 0.574162728987*MUJERJEFE))

Todos los parámetros estimados son significativos individualmente. La significatividad conjunta es muy alta porque el p-valor del estadístico de la razón de verosimilitud es muy pequeño. El Pseudo R^2 de McFadden no se acerca demasiado a la unidad (0.2788). Los valores de los criterios de información (Akaike, Schwarz y Hannan-Quinn) son adecuados porque son bajos y muy parecidos.

Otro criterio para medir la bondad del ajuste del modelo Probit es el criterio del porcentaje de predicciones correctas que consiste en observar el porcentaje de veces en que el valor de Yi observado coincide con su predicción.

Expectation-Prediction Evaluation for Binary Specification
Equation: UNTITLED
Date: 03/18/19 Time: 00:48
Success cutoff: C = 0.5

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	154	46	200	0	0	0
P(Dep=1)>C	52	183	235	206	229	435
Total	206	229	435	206	229	435
Correct	154	183	337	0	229	229
% Correct	74.76	79.91	77.47	0.00	100.00	52.64
% Incorrect	25.24	20.09	22.53	100.00	0.00	47.36
Total Gain*	74.76	-20.09	24.83			
Percent Gai...	74.76	NA	52.43			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(# of Dep=0)	133.57	72.00	205.57	97.55	108.45	206.00
E(# of Dep=1)	72.43	157.00	229.43	108.45	120.55	229.00
Total	206.00	229.00	435.00	206.00	229.00	435.00
Correct	133.57	157.00	290.56	97.55	120.55	218.11
% Correct	64.84	68.56	66.80	47.36	52.64	50.14
% Incorrect	35.16	31.44	33.20	52.64	47.36	49.86
Total Gain*	17.48	15.91	16.66			
Percent Gai...	33.21	33.60	33.41			

*Change in "% Correct" from default (constant probability) specification
**Percent of incorrect (default) prediction corrected by equation

Se puede observar que el modelo predice adecuadamente el 77,47% de las observaciones. Se predicen mejor los unos (la participación de la mujer en el mercado de trabajo) con un 79,91% de aciertos frente a un 74,76% de la no participación.

Para cuantificar e interpretar los efectos de las variables explicativas sobre la probabilidad de la mujer de participar en el mercado de trabajo se tendrá presente que el efecto parcial de una variable explicativa continua X_j sobre la probabilidad de respuesta $P(Y=1\backslash \mathbf{X})$ es

$$\frac{\partial P(Y=1 \setminus \boldsymbol{X}}{\partial X_j} = g(\boldsymbol{X}\beta)\beta_j$$

Como para cada observación, se tiene un efecto, se evita el cálculo de todas las observaciones, calculando los efectos para las observaciones medias.

A continuación, se obtienen los estadísticos descriptivos para todas las variables del modelo, incluida la media.

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

	CANT6A14	CANTMEN5	CANTMIEM...	EDAD	EDADCUA...	INSTRUCCI...	MUJERJEFE	PERCEPT...
	CANT6A14	CANTMEN5	CANTMIEM...	EDAD	EDADCUA...	INSTRUCCI...	MUJERJEFE	PERCEPT...
Mean	0.064368	0.420690	4.052874	36.42069	1526.117	9.075862	0.222989	2.234483
Median	0.000000	0.000000	4.000000	35.00000	1225.000	12.00000	0.000000	2.000000
Maximum	3.000000	3.000000	13.00000	65.00000	4225.000	17.00000	1.000000	5.000000
Minimum	0.000000	0.000000	1.000000	15.00000	225.0000	0.000000	0.000000	1.000000
Std. Dev.	0.346801	0.695631	2.133658	14.14605	1102.081	5.899406	0.416730	1.011549
Skewness	6.294000	1.559446	1.239084	0.279523	0.754686	-0.505014	1.330984	0.815705
Kurtosis	45.88424	4.649607	5.522618	1.943148	2.501882	1.737700	2.771518	3.204772
Jarque-Bera	36204.97	225.6326	226.6517	25.90911	45.78965	47.37070	129.3813	48.99962
Probability	0.000000	0.000000	0.000000	0.000002	0.000000	0.000000	0.000000	0.000000
Sum	28.00000	183.0000	1763.000	15843.00	663861.0	3948.000	97.00000	972.0000
Sum Sq. Dev.	52.19770	210.0138	1975.784	86848.01	5.27E+08	15104.50	75.37011	444.0828
Observations	435	435	435	435	435	435	435	435

Con lo que

$$g(\overline{\boldsymbol{X}}\hat{\beta}) = \varphi(\hat{\beta}_0 + \hat{\beta}_1\overline{edad} + \hat{\beta}_2\overline{edadcuadrado} + \hat{\beta}_3\overline{instrucción} + \hat{\beta}_4\overline{cantmiembros} + \hat{\beta}_5\overline{cantmen5} + \hat{\beta}_6\overline{cant6a14} + \hat{\beta}_7\overline{perceptores} + \hat{\beta}_8\overline{mujerjefe} =$$

$$\varphi = Función\ de\ densidad\ de\ la\ normal\ (0{,}1) = 0.038772181$$

Luego se puede obtener el efecto parcial estimados de las variables continuas, para los valores medios de las X, multiplicando los coeficientes estimados del modelo Probit por 0,038772181.

B_0	-5,3084804	C	Media	Probabilidad
B_1	0,2589179	Edad	36,421	59,9%
B_2	-0,0031457	Edad²	1.526,117	-
B_3	0,0358234	Instrucción	9,076	50,8%
B_4	-0,3637725	CantMiembros	4,053	42,5%
B_5	0,2916183	CantMen5	0,0421	56,0%
B_6	0,8739702	Cant6a14	0,064	67,4%
B_7	0,6982658	Perceptores	2,234	14,2%
B_8	0,5741627	MujerJefe	0,223	22,2%

Es posible visualizar en este modelo Probit que la probabilidad de participación en el mercado laboral para la mujer se incrementa con la edad (un 59,9%, con un promedio de 36 años), la instrucción (50,8%, con un promedio de 9 años de estudios formales), la cantidad de hijos menores de 5 años (56,0%), como, asimismo, la cantidad de menores entre 6 a 14 años (67,4%); la cantidad de perceptores de ingresos en el hogar (14,2%, con un promedio de más de dos perceptores) y si la mujer es jefa del hogar (22,2%). Sólo cuando aumenta la cantidad de miembros a más de cuatro integrantes, se reduce la probabilidad de participación de la mujer en un 42,5%.

Aplicación empírica del Modelo 2 (Metodología Oaxaca etc etc)

Luego de aplicado el modelo Probit, se realiza la regresión lineal múltiple para las mujeres, y se aplican los tests de cumplimiento de los supuestos de homocedasticidad, no multicolinealidad, normalidad).

```
. regress LnIngreLaborMensu Instruccion Exper6 lnHsTraba Profesional Tecnico Operativo Lamda Industria Comercio Educacion ScioDome
> stico TransporteComunicac AdminPublica Construccion [fweight= PONDIIO]
```

Source	SS	df	MS	Number of obs	=	33,057
				F(14, 33042)	=	1585.97
Model	12902.3937	14	921.599547	Prob > F	=	0.0000
Residual	19200.5486	33,042	.581095231	R-squared	=	0.4019
				Adj R-squared	=	0.4017
Total	32102.9423	33,056	.97116839	Root MSE	=	.7623

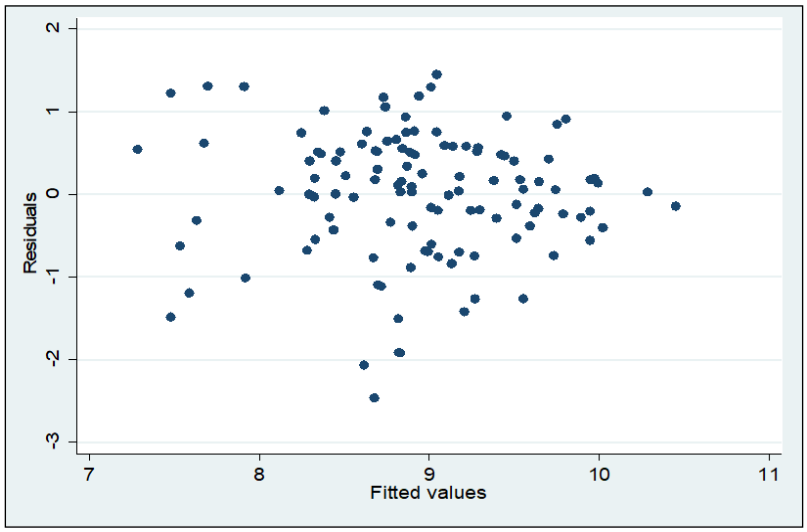
LnIngreLaborMensu	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Instruccion	.021438	.0008378	25.59	0.000	.0197957	.0230802
Exper6	.0072404	.0002886	25.09	0.000	.0066748	.007806
lnHsTraba	.6789928	.008817	77.01	0.000	.6617113	.6962744
Profesional	.6136415	.0216618	28.33	0.000	.5711837	.6560993
Tecnico	.4604761	.0203376	22.64	0.000	.4206137	.5003385
Operativo	.1926378	.0143695	13.41	0.000	.164473	.2208025
Lamda	-.2019782	.0091114	-22.17	0.000	-.2198369	-.1841195
Industria	-.7974452	.0200846	-39.70	0.000	-.8368118	-.7580786
Comercio	-.594055	.0124837	-47.59	0.000	-.6185235	-.5695865
Educacion	.4206605	.0169872	24.76	0.000	.3873649	.453956
ScioDomestico	-.6099389	.0188153	-32.42	0.000	-.6468174	-.5730603
TransporteComunicac	-.1374957	.0234531	-5.86	0.000	-.1834647	-.0915267
AdminPublica	.1920987	.0151396	12.69	0.000	.1624246	.2217728
Construccion	-.6189588	.0176969	-34.98	0.000	-.6536453	-.5842722
_cons	5.297854	.0453028	116.94	0.000	5.209059	5.38665

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of LnIngreLaborMensu

         chi2(1)      =  1937.48
         Prob > chi2  =   0.0000
```

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
   Operativo |      2.82    0.354427
      Tecnico |      2.15    0.464700
ScioDomest~o |      2.04    0.489523
 Profesional |      1.73    0.578931
   Educacion |      1.61    0.621080
    Comercio |      1.38    0.722236
   lnHsTraba |      1.28    0.779586
 Instruccion |      1.26    0.790922
Construccion |      1.23    0.810355
   Industria |      1.21    0.823981
AdminPublica |      1.21    0.828557
Transporte~c |      1.18    0.848417
       Lamda |      1.16    0.858752
      Exper6 |      1.16    0.863562
-------------+----------------------
    Mean VIF |      1.53
```

```
. regress LnIngreLaborMensu Instruccion Exper6 lnHsTraba Profesional Tecnico Operativo Lamda Industria Comercio Educacion ScioDome
> stico TransporteComunicac AdminPublica Construccion [fweight= PONDIIO], robust

Linear regression                               Number of obs     =     33,057
                                                F(14, 33042)      =    2589.76
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4019
                                                Root MSE          =      .7623
```

LnIngreLaborMensu	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
Instruccion	.021438	.0009237	23.21	0.000	.0196274	.0232485
Exper6	.0072404	.0003431	21.11	0.000	.006568	.0079128
lnHsTraba	.6789928	.0103264	65.75	0.000	.6587527	.699233
Profesional	.6136415	.0166854	36.78	0.000	.5809375	.6463455
Tecnico	.4604761	.0185434	24.83	0.000	.4241304	.4968219
Operativo	.1926378	.0153444	12.55	0.000	.1625623	.2227133
Lamda	-.2019782	.0074045	-27.28	0.000	-.2164913	-.1874651
Industria	-.7974452	.0167708	-47.55	0.000	-.8303164	-.7645739
Comercio	-.594055	.0141546	-41.97	0.000	-.6217985	-.5663114
Educacion	.4206605	.0120706	34.85	0.000	.3970016	.4443194
ScioDomestico	-.6099389	.0189545	-32.18	0.000	-.6470904	-.5727873
TransporteComunicac	-.1374957	.0230118	-5.98	0.000	-.1825996	-.0923918
AdminPublica	.1920987	.0131276	14.63	0.000	.1663682	.2178292
Construccion	-.6189588	.0217661	-28.44	0.000	-.6616211	-.5762965
_cons	5.297854	.0532584	99.47	0.000	5.193466	5.402243




```
. swilk LnIngreLaborMensuResidual- LnIngreLaborMensuRstudent

                    Shapiro-Wilk W test for normal data

    Variable |        Obs        W         V         z       Prob>z
-------------+-------------------------------------------------------
LnIngreLab~l |        119    0.96769     3.087     2.524     0.00579
LnIngreLab~d |        119    0.96769     3.087     2.524     0.00580
LnIngreLab~t |        119    0.96768     3.088     2.525     0.00579

. sfrancia LnIngreLaborMensuResidual- LnIngreLaborMensuRstudent

                  Shapiro-Francia W' test for normal data

    Variable |        Obs        W'        V'        z       Prob>z
-------------+-------------------------------------------------------
LnIngreLab~l |        119    0.96815     3.348     2.415     0.00786
LnIngreLab~d |        119    0.96815     3.347     2.415     0.00786
LnIngreLab~t |        119    0.96814     3.348     2.416     0.00785
```

A continuación, se corre la regresión lineal múltiple para los varones, y el testeo del cumplimiento de los supuestos (homocedasticidad, no multicolinealidad, normalidad).

```
. regress LnIngreLaborMensu Instruccion Exper6 LNHorasTrabajadasMensuales Profesional Tecnico Operativo Industria Comercio Educacion
>  ScioDomestico  AdminPublica Construccion [fweight= PONDIIO]
```

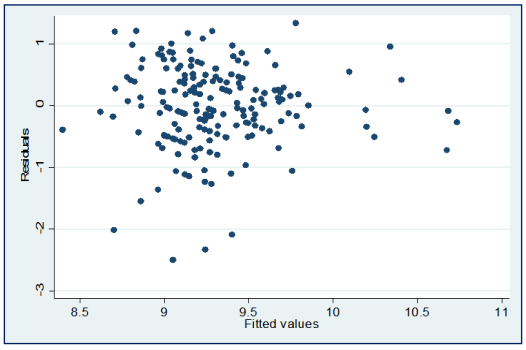
Source	SS	df	MS		
Model	6914.02745	12	576.168954	Number of obs	= 51,034
Residual	22546.3481	51,021	.441903297	F(12, 51021)	= 1303.83
				Prob > F	= 0.0000
				R-squared	= 0.2347
				Adj R-squared	= 0.2345
Total	29460.3756	51,033	.577280889	Root MSE	= .66476

LnIngreLaborMensu	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Instruccion	.0138466	.000587	23.59	0.000	.0126961	.0149972
Exper6	-.0015082	.0002209	-6.83	0.000	-.0019413	-.0010751
LNHorasTrabajadasMensuales	.2630349	.0069794	37.69	0.000	.2493552	.2767146
Profesional	1.051384	.0150507	69.86	0.000	1.021884	1.080883
Tecnico	.1042267	.0107596	9.69	0.000	.0831379	.1253156
Operativo	.1349871	.0085149	15.85	0.000	.1182977	.1516764
Industria	-.5081461	.012821	-39.63	0.000	-.5332754	-.4830169
Comercio	-.2822318	.0085238	-33.11	0.000	-.2989385	-.2655252
Educacion	-.2496993	.0117479	-21.25	0.000	-.2727252	-.2266733
ScioDomestico	.1186733	.0159081	7.46	0.000	.0874932	.1498534
AdminPublica	.2978367	.010035	29.68	0.000	.2781679	.3175054
Construccion	-.1934647	.0103541	-18.68	0.000	-.2137588	-.1731707
_cons	7.826152	.0364716	214.58	0.000	7.754668	7.897637

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of LnIngreLaborMensu

         chi2(1)      =  1884.61
         Prob > chi2  =   0.0000
```

```
. vif
```

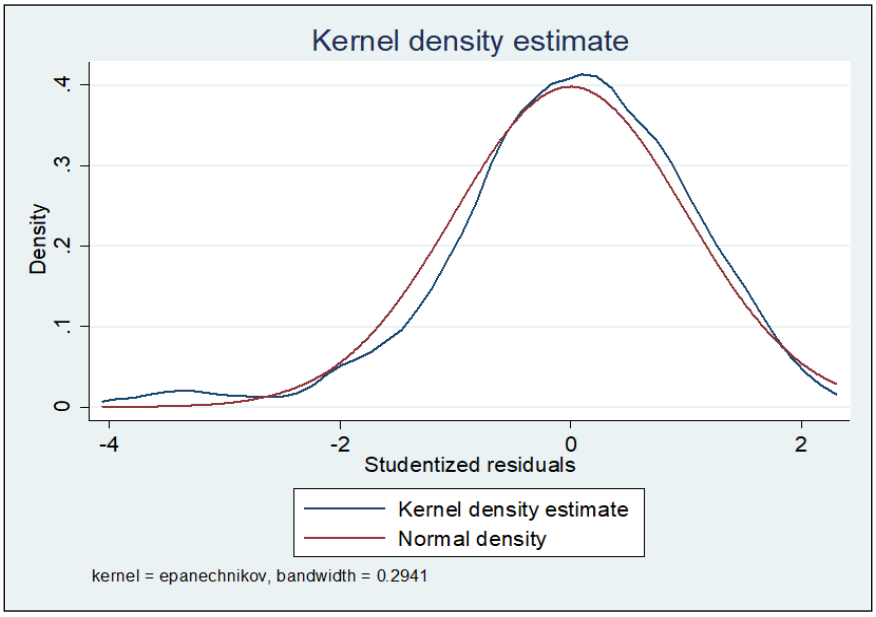
Variable	VIF	1/VIF
Operativo	2.05	0.487249
Tecnico	1.73	0.577002
Comercio	1.44	0.694946
AdminPublica	1.43	0.699473
Profesional	1.43	0.700522
Educacion	1.41	0.710061
ScioDomest~o	1.39	0.717063
Instruccion	1.34	0.748420
Construccion	1.32	0.757195
LNHorasTra~s	1.20	0.836618
Industria	1.17	0.851781
Exper6	1.12	0.892624
Mean VIF	1.42	



```
. regress LnIngreLaborMensu Instruccion Exper6 LNHorasTrabajadasMensuales Profesional Tecnico Operativo Industria Comercio Educacion
>  ScioDomestico  AdminPublica Construccion [fweight= PONDIIO], robust

Linear regression                               Number of obs     =     51,034
                                                F(12, 51021)      =   1783.77
                                                Prob > F          =    0.0000
                                                R-squared         =    0.2347
                                                Root MSE          =    .66476

------------------------------------------------------------------------------
             |               Robust
LnIngreLabo~u|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Instruccion |   .0138466   .0005982    23.15   0.000     .0126741    .0150192
       Exper6 |  -.0015082   .0002335    -6.46   0.000    -.0019658   -.0010506
LNHorasTrab~s |   .2630349   .0067118    39.19   0.000     .2498797     .27619
  Profesional |   1.051384   .0127346    82.56   0.000     1.026424    1.076344
      Tecnico |   .1042267   .0103213    10.10   0.000     .0839969    .1244565
    Operativo |   .1349871   .0072927    18.51   0.000     .1206933    .1492808
    Industria |  -.5081461   .0166126   -30.59   0.000    -.5407071   -.4755852
     Comercio |  -.2822318   .0092194   -30.61   0.000    -.3003019   -.2641618
    Educacion |  -.2496993   .0120432   -20.73   0.000     -.273304   -.2260945
  ScioDomest~o |   .1186733   .0160253     7.41   0.000     .0872635     .150083
  AdminPublica |   .2978367   .0083488    35.67   0.000     .2814729    .3142005
  Construccion |  -.1934647   .0090902   -21.28   0.000    -.2112817   -.1756478
         _cons |   7.826152   .0355393   220.21   0.000     7.756495    7.89581
------------------------------------------------------------------------------
```

Shapiro-Francia W' test for normal data

Variable	Obs	W'	V'	z	Prob>z
Ingresolab~l	183	0.96217	5.710	3.582	0.00017
IngresoLab~d	183	0.96217	5.710	3.581	0.00017
IngresoLab~t	183	0.96216	5.712	3.582	0.00017



A continuación, los principales estadísticos descriptivos (media, desvío estándar, valores mínimo y máximo) de las variables regresoras, para la mujer y el varón.

Mujer

```
. summarize LnIngrelabormensuEstimada  Instruccion Exper6 lnHsTraba Profesional Tecnico Operativo Industria Comercio  Educacion Scio
> Domestico TransporteComunicac AdminPublica Construccion [fweight = PONDIIO]

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
LnIngrelab~a |     33,057    8.936095    .6247554   7.283286   10.45122
 Instruccion |     65,414    10.52272    5.531796          0         17
      Exper6 |     65,414    17.63456    13.78251          0         54
   lnHsTraba |     33,057    4.962206     .538577   3.247047   5.921196
 Profesional |     65,414    .0351454    .1841486          0          1
-------------+---------------------------------------------------------
     Tecnico |     65,414    .0514569    .2209294          0          1
  Operativo  |     65,414    .3027028    .4594312          0          1
  Industria  |     65,414     .028312    .1658639          0          1
   Comercio  |     65,414    .0978842    .2971603          0          1
  Educacion  |     65,414    .0557067    .2293563          0          1
-------------+---------------------------------------------------------
ScioDomest~o |     65,414    .0578928    .2335424          0          1
Transporte~c |     65,414    .0198123    .1393558          0          1
AdminPublica |     65,414    .0521601    .2223515          0          1
Construccion |     65,414    .0378359    .1908008          0          1
```

Varón

```
. summarize IngresLaboralMensPredicho Instruccion Exper6 LNHorasTrabajadasMensuales Profesional Tecnico Operativo Industria Comerci
>  ScioDomestico AdminPublica Construccion

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
IngresLabo~o |        183    9.305365     .368812   8.395237   10.73521
 Instruccion |        304    9.233553    5.755875          0         17
      Exper6 |        304    17.12171    14.14173          0         52
LNHorasTra~s |        183    5.017693    .4537257   3.064725   5.731953
 Profesional |        304    .0328947    .1786552          0          1
-------------+---------------------------------------------------------
     Tecnico |        304    .0921053    .2896513          0          1
  Operativo  |        304    .3552632    .4793821          0          1
  Industria  |        304    .0427632    .2026563          0          1
   Comercio  |        304    .1315789    .3385898          0          1
ScioDomest~o |        304    .0296053    .1697752          0          1
-------------+---------------------------------------------------------
AdminPublica |        304    .0953947    .2942437          0          1
Construccion |        304    .0723684    .2595242          0          1
```

Volviendo a las ecuaciones (1) y (2),

(1) $LN\,Y_v = X_v\,\beta_v + U_v$;
(2) $LN\,Y_m = X_m\,\beta_m + U_m$

Cabe recordar que los subíndices denotan el sexo de los individuos, Y_i es el vector columna de los ingresos, X_i es la matriz de variables independientes y β el vector columna de coeficientes a estimar. Los U son los términos de perturbación.

Como se dijo también anteriormente, teniendo en cuenta la estimación de los residuos, a partir de las diferencias de primer tipo o diferencia explicada, se llegaba a la siguiente ecuación final:

$$(5)\ \overline{LnY_v} - \overline{LnY_m} = \beta_v\,(\overline{X_v} - \overline{X_m}) + \overline{X_m}\Delta\beta$$

El cual mostraba que la media entre los ingresos de ambos grupos puede descomponerse en los efectos de las diferencias de sus respectivas dotaciones de capital humano y en los efectos de la discriminación (diferencia de los coeficientes estimados). A partir de los resultados obtenidos, se observa lo siguiente:

Variables	Hombres		Mujeres	
	Promedio	Coeficientes	Promedio	Coeficientes
Constante		7.826152		5.297854
Instrucción	9.233553	0.138466	10.52272	0.021438
Experiencia	17.12171	-0.015082	17.63456	0.0072404
LNHorastrabajadasMensuales	5.017693	0.2630349	4.962206	0.6789928
Profesional	0.0328947	1.051384	0.0351454	0.6136415
Tecnico	0.0921053	0.1042267	0.0514569	0.4604761
Operativo	0.3552632	0.1349871	0.3027028	0.1926378
Industria	0.0427632	-0.5081461	0.028312	-0.7974452
Comercio	0.1315789	-0.2822318	0.0978842	-0.594055
ScioDomestico	0.0296053	0.1186733	0.0578928	-0.6099389
AdminPublica	0.0953947	0.2978367	0.0521601	0.1920987
Construcción	0.0723684	-0.1934647	0.0378359	-0.6189588
Lamda				-0.2019782
LNIngresoLaboralEstimado	9.305365		8.936095	
R2 ajustado	0.2345		0.4017	

Para la descomposición de la brecha, mediante el método de Blinder – Oaxaca (Blinder, 1973 y Oaxaca, 1973), se obtiene:

$$(7)\ \overline{Y_v} - \overline{Y_m} = (\overline{X_v} - \overline{X_m})B_m + \overline{X_v}\,(B_v - B_m)$$

O bien, si se usa como grupo de comparación al otro sexo:

$$(7')\ \overline{Y_v} - \overline{Y_m} = (\overline{X_v} - \overline{X_m})B_v + \overline{X_m}\,(B_v - B_m)$$

El primer miembro del lado derecho de (7) u (7') es una estimación de la parte de la brecha que se explica por diferencias en las X's (dotaciones de capital humano y posición ocupacional), mientras que el segundo miembro muestra la parte que no puede ser explicada por estos factores; lo que se denomina "el residual".

El logaritmo natural del ingreso laboral estimado de los hombres es de 9,305365, y el de las mujeres, 8,936095, por lo que la diferencia a favor de los primeros es del 36,9%. De esta brecha de ingresos a favor del hombre, únicamente 5,6 puntos porcentuales se explican por las variables consideradas en el modelo (instrucción, experiencia, calificación, rama de actividad, cantidad de horas trabajadas, y la corrección por sesgo de selección); es decir, el 15,0% del total de la brecha. Esto nos dice que del 36,9% que es la brecha a favor del hombre, 31,3% no se explica por ninguna de las variables consideradas en el modelo (es decir, el 85,0%). Aunque pueda suponerse que existan otras variables que expliquen esta diferencia no relacionadas con la discriminación, podemos conjeturar que puede existir un componente discriminatorio hacia la mayor remuneración laboral a favor del varón, en detrimento de la mujer.

	Explicada	
Instrucción	-1.289167	**-0.02763716**
Experiencia	-0.51285	**-0.01099448**
LNHorastrabajadasMensuales	0.055487	**0.00040175**
Profesional	-0.0022507	**-0.00152821**
Tecnico	0.0406484	**0.02494355**
Operativo	0.0525604	**0.02420281**
Industria	0.0144512	**0.00278385**
Comercio	0.0336947	**-0.02686968**
ScioDomestico	-0.0282875	**0.01680433**
AdminPublica	0.0432346	**-0.02637046**
Construcción	0.0345325	**0.00663365**
Lamda	-0.3663	**0.07398461**
Total		**0.05635455**

3. Conclusión

El presente trabajo utilizó dos metodologías diferentes, para medir la brecha de ingresos entre los hombres y las mujeres, para el Aglomerado Posadas, en base a los microdatos de la Encuesta Permanente de Hogares, correspondientes al segundo trimestre del año 2018.

Una de ellas fue la "Metodología de Regresión Lineal Múltiple" y, la segunda, la "Metodología de (Oaxaca, 1973) y (Blinder, 1973) con corrección del sesgo por selección de (Heckman, 1979)".

De acuerdo al primer modelo, la brecha de ingresos a favor del varón respecto a la mujer fue del 58,3%. Con respecto al segundo modelo, la brecha fue del 36,9%.

A partir de dos metodologías econométricas diferentes, podemos inferir que existe discriminación en los ingresos laborales percibidos entre el hombre y la mujer, puesto que, a poner en igualdad de condiciones a ambos géneros, en términos de educación, experiencia, calificación y jerarquía laboral, rama de actividad, entre otras variables, existe una brecha considerable y significativa estadísticamente, que genera nuevos interrogantes para continuar investigando, como identificar nuevas variables, nuevas interacciones entre las mismas y utilizar otros modelos econométricos, como los de datos de panel, no lineales, entre otros.

Referencias bibliográficas

Becker, G. (1957). *The economics of discrimination*. Chicago: Chicago University Press.

Blinder, A. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*(8), 436-455.

Broso, M. (2016). Introducción a la medición de la brecha salarial por género y sus determinantes. *Economistas para qué, 11*.

Di Paola, R., & Berges, M. (2000). *Sesgo de selección y estimación de la brecha por género entre la mujer y el hombre.* Mar del Plata: AAEP.

Esquivel, V. (2007). Género y diferencias de salarios en la Argentina. In M. Novick, & H. Palomino, *Estructura productiva y empleo* (pp. 363-392). Buenos Aires: Ministerio de Trabajo, Empleo y Seguridad Social.

Heckman, J. (1979). Sample bias as a specification error. *Econométrica, 47*(1), 153-161.

Mercado, M., Macías, E., & Bernardi, F. (2009). *Análisis de datos con Stata*. Madrid: CIS.

Oaxaca, R. (1973). Male-Female wage differentials in urban labor markets. *International Economic Review, 14*(3), 693-709.

Paz, J. (1995). *Diferencias de ingresos entre varones y mujeres. Evidencias para Salta (Argentina)*. Salta: AAEP.

Paz, J. (1998). *Brecha de ingresos entre géneros. (Comparación entre el Gran Buenos Aires y el Noroeste Argentino)*. Salta: AAEP.

Perez, C. (2006). *Problemas resueltos de econometría.* Madrid: Ediciones Paraninfo.

Segura Gómez, C. (2013). *Determinantes del diferencial salarial por género en Colombia durante el periodo 2004-2012: Una aplicación de regresión por cuantiles.* La Plata: Tesis de Maestría. Universidad Nacional de la Plata.

OBSERVATORIO PARA EL
DESARROLLO ECONÓMICO
DE MISIONES

STAFF

Rector
Dr. Ricardo Biazzi

Vicerrector
Dr. Alfredo Poenitz

Secretaria de Extensión
Lic. Gabriela Lichowski

Secretaría de Investigación y Desarrollo
Mg. Mario Bortoluzzi
Ing. Héctor Ruidías

Director Técnico
Dr. Darío Díaz

Analista Técnica
Lic. Natalia Ojeda

Corrección
Esp. Paola A. Torres B.

Diseño y Diagramación
Brutal Creativos

UNIVERSIDAD Gastón Dachary

EDITORIAL UGD