PS1 - Equipo 03

Catalina Leal Rojas, Lucas Daniel Carrillo Aguirre, Lucas Eduardo Veras Costa, Maria Paula Basto Lozano

6 de septiembre de 2025

1. Introducción

Poner acá la introducción

2. Datos

Poners cosas de los datos

3. Perfil salario-edad

4. La brecha género-salario

En esta sección analizaremos la relación entre ingresos y género femenino. En primer lugar, es importante discutir qué medida de ingreso utilizar. La base de datos de la GNIH incluye diversas medidas, como ingreso por primas monetarias, ingreso mensual, ingreso usual en el mes e ingreso efectivo en el mes, entre otras.

En los estudios que analizan la brecha salarial entre hombres y mujeres, es común emplear el ingreso por hora, ya que permite aislar el efecto de las horas trabajadas en el mes, lo cual podría sesgar los resultados. Asimismo, suele utilizarse la variable de ingreso en logaritmos, lo que mejora la interpretabilidad del coeficiente estimado, que pasaría a representar cuánto impacta, en términos porcentuales, una unidad adicional de la variable independiente sobre el ingreso.

También es importante señalar que nos interesa comparar únicamente a hombres y mujeres que trabajan. Por lo tanto, el análisis se centra en los individuos ocupados de la muestra.

Inicialmente estimamos el siguiente modelo:

$$\log(\omega) = \beta_0 + \beta_1 Female + u \tag{1}$$

Cuadro 1. Resultados del modelo incondicional

	Variable dependiente:				
	log(ysalarym)	log(ysalarymha)	log(yingLabm)	log(ytotalm)	log(ytotalmha)
	(1)	(2)	(3)	(4)	(5)
female	−0.149***	−0.045***	−0.147***	−0.238***	−0.090***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)
Constant	13.977***	8.641***	14.088***	13.981***	8.667***
	(0.011)	(0.010)	(0.011)	(0.010)	(0.009)
Observaciones	9,892	9,892	9,892	14,764	14,764
R^2	0.010	0.001	0.009	0.018	0.003
R^2 ajustado	0.010	0.001	0.009	0.017	0.003
Error estándar residual	0.751 (df = 9890)	0.721 (df = 9890)	0.762 (df = 9890)	0.889 (df = 14762)	0.832 (df = 14762)
F Statistic	97.364*** (df = 1; 9890)	9.559*** (df = 1; 9890)	91.422*** (df = 1; 9890)	263.841*** (df = 1; 14762)	43.342*** (df = 1; 14762)
Nota:					*p<0.1; **p<0.05; ***p<0.01

El Cuadro 1 presenta los resultados de la estimación por mínimos cuadrados ordinarios utilizando diversas medidas de ingreso. *ysalarym* corresponde al ingreso nominal de la ocupación principal, *ysalarymha* al salario por hora de la ocupación principal, *yingLabm* al ingreso proveniente de todas las ocupaciones, *ytotalm* al ingreso total proveniente de ocupaciones e ingresos independientes, y *ytotalmha* al ingreso total de ocupaciones e independientes medido por hora.

Los resultados muestran evidencia de que las mujeres ganan menos que los hombres. Aunque existe cierta variación en los coeficientes de la variable $female$, todos son negativos y estadísticamente significativos, lo que indica pérdidas salariales para las mujeres.

Cabe señalar que en los resultados anteriores todos los coeficientes fueron estimados mediante mínimos cuadrados ordinarios. No obstante, es posible obtenerlos aplicando el Teorema de Frisch-Waugh-Lovell (FWL). Supongamos que queremos estimar el coeficiente de una variable X_1 en una regresión múltiple que también incluye una variable de control X_2, en el siguiente modelo:

$$y = \beta_1 X_1 + \beta_2 X_2 + u$$

Para estimar β_1 usando el Teorema de Frisch-Waugh-Lovell, se siguen los siguientes pasos:

1. Regrese y sobre X_2 y obtenga los residuos.
   Denotemos estos residuos como r_y:

$$r_y = y - \hat{y}_2 \quad \text{donde } \hat{y}_2 = X_2\hat{\beta}_2$$

2. Regrese X_1 sobre X_2 y obtenga los residuos.
   Denotemos estos residuos como r_{X_1}:

$$r_{X_1} = X_1 - \hat{X}_{1,2} \quad \text{donde } \hat{X}_{1,2} = X_2\hat{\gamma}$$

3. Regrese los residuos r_y sobre r_{X_1}.

El coeficiente estimado será exactamente igual a $\hat{\beta}_1$ de la regresión original:

$$\hat{\beta}_1 = \frac{r'_{X_1} r_y}{r'_{X_1} r_{X_1}}$$

Este procedimiento permite estimar el efecto de X_1 sobre y, controlando por X_2, sin necesidad de realizar directamente la regresión múltiple completa. Además, resulta útil para reducir el costo computacional en aplicaciones con gran cantidad de variables.

Como ilustración de este procedimiento, estimamos el siguiente modelo mediante el método clásico de MCO y también aplicando el teorema de FWL:

$$\log(\omega) = \beta_1 + \beta_2 female + \beta_3 age + \beta_4 age^2 \tag{2}$$

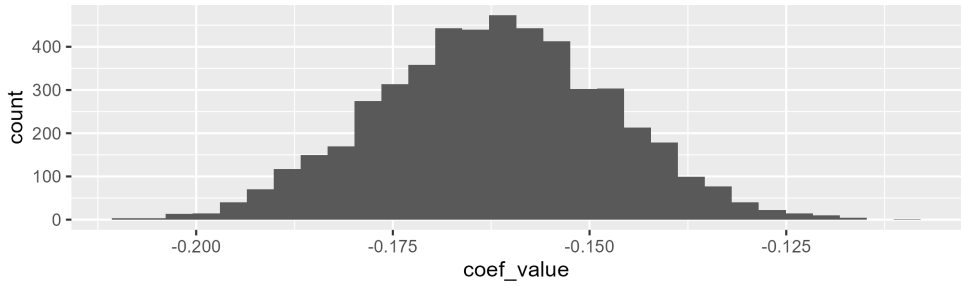
Cuadro 2. Comparación de las estimaciones por MCO y FWL

	Variable dependiente:	
	log(y_ingLab_m)	resid_ing
	(1)	(2)
female	-0.163^{***}	
	(0.015)	
age	0.091^{***}	
	(0.004)	
age_sqr	-0.001^{***}	
	(0.00005)	
resid_fem		-0.163^{***}
		(0.017)
Constant	12.338^{***}	-0.000
	(0.071)	(0.007)
Observations	9,892	9,892
R^2	0.069	0.012
Adjusted R^2	0.069	0.012
Residual Std. Error	0.739 (df = 9888)	0.739 (df = 9890)
F Statistic	245.548^{***} (df = 3; 9888)	119.495^{***} (df = 1; 9890)
Note:		*p<0.1; **p<0.05; ***p<0.01

El Cuadro 2 muestra que los coeficientes de $female$ y $resid\_fem$ son idénticos. Sin embargo, los errores estándar no lo son. Esto ocurre debido a la forma en que se realiza la estimación en dos etapas: al reducir el número de variables en la segunda ecuación, los grados de libertad utilizados también disminuyen.

Otra manera de estimar el error estándar es a través del método bootstrap. Este procedimiento consiste en realizar un remuestreo de la muestra n veces y reestimar el coeficiente en cada repetición. El error estándar se obtiene a partir de la dispersión de esta distribución de estimaciones.

Figura 1. Coeficiente de $female$



Nota: La figura muestra el histograma de los coeficientes de $female$ obtenidos a través del bootstrap con 5000 remuestreos.

La Figura 1 muestra la distribución del coeficiente de $female$. Al calcular el error estándar, obtenemos un valor de 0.01472, prácticamente idéntico al obtenido mediante la estimación por MCO.

Al estimarmos el modelo 1, no hemos considerado ningún control. Es posible que existan variables omitidas influenciando nuestros resultados. De esta manera, a fin de verificar la robustez agregamos los siguientes controles a nuestra estimación:

- **Capital humano**: Incluimos edad y su cuadrado como aproximación a la experiencia laboral, así como el nivel educativo más alto alcanzado. Estas variables capturan diferencias en productividad potencial.
- **Intensidad laboral**: En las especificaciones con salarios mensuales, se consideran las horas usuales de trabajo y la existencia de un segundo empleo. En las especificaciones con salarios por hora, estas variables se omiten deliberadamente para evitar un ajuste redundante.
- **Características del empleo**: Incorporamos ocupación, tamaño de la firma, tipo de relación laboral, formalidad del empleo y la antigüedad en el puesto. Estos factores permiten aproximar la comparabilidad entre trabajos, tal como lo exige la noción de "igual trabajo".

La estrategia empírica consiste en estimar primero la brecha salarial sin controles y, posteriormente, añadir secuencialmente los bloques de covariables. De este modo, se puede observar cómo evoluciona el coeficiente asociado a la variable de género, lo que permite interpretar en qué medida el diferencial incondicional se explica por diferencias observables en características de los trabajadores y de sus empleos. En esta parte estimamos apenas como variable dependente la variable de ingreso de la ocupación principal por hora.
Nota: En la ecuación (5) se omitieron las variables relacionadas con el oficio y el número de personas que trabajan en la empresa.

Los resultados en el Cuadro 3 muestran que el coeficiente de $female$ sigue siendo negativo y significativo. Además, la inclusión de controles incrementa la magnitud del coeficiente.

Cuadro 3. Resultados de las regresiones con controles

	Variable dependiente:				
	log(y_ingLab_m_ha)				
	(1)	(2)	(3)	(4)	(5)
female	−0.045***	−0.058***	−0.142***	−0.179***	−0.106***
	(0.015)	(0.014)	(0.012)	(0.012)	(0.012)
age		0.068***	0.062***	0.068***	0.038***
		(0.004)	(0.003)	(0.003)	(0.003)
age_sqr		−0.001***	−0.001***	−0.001***	−0.0003***
		(0.00004)	(0.00004)	(0.00004)	(0.00003)
primarios incompleto			0.199**	0.223**	0.167**
			(0.094)	(0.093)	(0.076)
as.factor(primario completo)4			0.289***	0.303***	0.203***
			(0.091)	(0.090)	(0.073)
as.factor(secundario incompleto)5			0.347***	0.355***	0.229***
			(0.091)	(0.089)	(0.073)
as.factor(secundario completo)6			0.565***	0.575***	0.292***
			(0.090)	(0.088)	(0.072)
as.factor(terciario)7			1.253***	1.238***	0.557***
			(0.089)	(0.088)	(0.073)
totalHoursWorked				−0.009***	−0.010***
				(0.0005)	(0.0004)
cuentaPropia					
microEmpresa					−0.392***
					(0.044)
formal					0.261***
					(0.015)
Constant	8.747***	7.392***	6.583***	6.940***	8.793***
	(0.010)	(0.068)	(0.104)	(0.104)	(0.164)
Observations	9,892	9,892	9,891	9,891	9,891
R^2	0.001	0.046	0.335	0.358	0.576
Adjusted R^2	0.001	0.045	0.335	0.358	0.572
Residual Std. Error	0.727 (df = 9890)	0.711 (df = 9888)	0.593 (df = 9882)	0.583 (df = 9881)	0.476 (df = 9798)
F Statistic	9.317*** (df = 1; 9890)	158.028*** (df = 3; 9888)	623.580*** (df = 8; 9882)	612.986*** (df = 9; 9881)	144.719*** (df = 92; 9798)
Note:					*p<0.1; **p<0.05; ***p<0.01

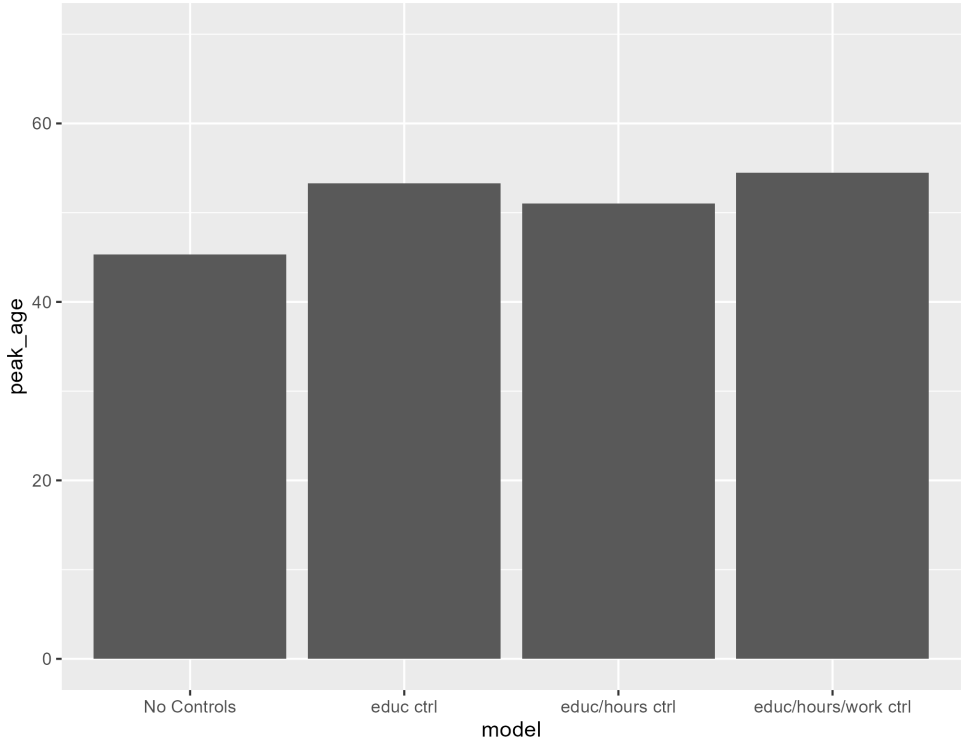
El modelo con todos los controles indica que las mujeres llegan a ganar aproximadamente un 10 % menos que los hombres, dado que ambos tengan la misma experiencia, edad y ejerzan profesiones similares.

Un aspecto interesante sería estimar la edad pico en la que las mujeres alcanzarían su máximo salarial. Esto puede calcularse mediante un problema sencillo de maximización:

$$pico = -\frac{\beta_2}{2\beta_3},$$

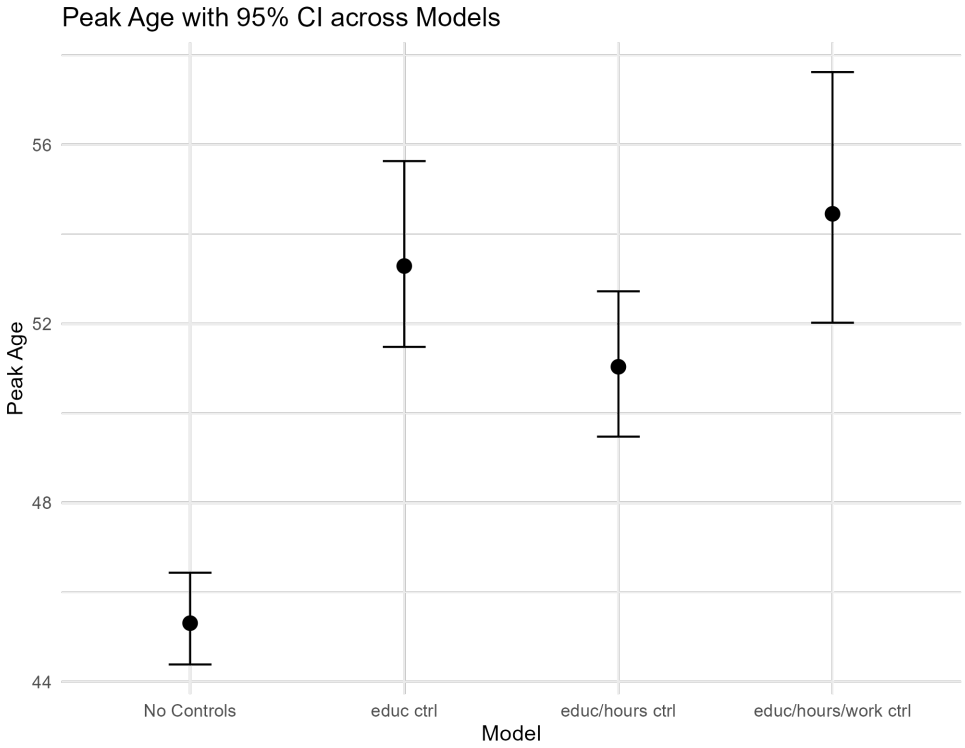
donde β_2 es el coeficiente de age y β_3 es el coeficiente de age^2.

Figura 2. Edad pico según nuestros modelos



La Figura 2 muestra los resultados de la edad pico. En general, los modelos indican que la edad pico estaría entre los 45 y 54 años. La inclusión de controles parece incrementar dicha edad. Cabe señalar que, al estimar la edad pico a partir de los coeficientes de MCO, no es posible calcular sus errores estándar ni los intervalos de confianza. Para generar dichos intervalos, empleamos el método bootstrap. En particular, remuestreamos los datos 2000 veces y registramos la distribución de la edad pico. Con esta distribución construimos el intervalo de confianza al 5 % para nuestros modelos.

Figura 3. Intervalos de confianza de la edad pico



La Figura 3 muestra los resultados de los intervalos de confianza. Como era de esperar, la inclusión de más controles amplía dichos intervalos; sin embargo, el incremento no supera los 2 años con respecto al valor central de la estimación.

5. Prediciendo Salarios

Referencing tables is very easy you can do the following: In Table **??** ...

Citing papers is easier, for example Albouy et al. (2020) shows that crime around parks. For many papers in parenthesis Albouy et al. (2020), McMillen et al. (2019)

6. ?

7. Conclusion

Referencias

Albouy, D., Christensen, P., and Sarmiento-Barbieri, I. (2020). Unlocking amenities: Estimating public good complementarity. *Journal of Public Economics*, 182:104110.

McMillen, D., Sarmiento-Barbieri, I., and Singh, R. (2019). Do more eyes on the street reduce crime? evidence from chicago's safe passage program. *Journal of urban economics*, 110:1–25.

Nikolov, P. (2020). Writing tips for economics research papers.