

Problem set 2

Max Brando Serna Leyva

Econometría aplicada

Agosto - Diciembre 2024

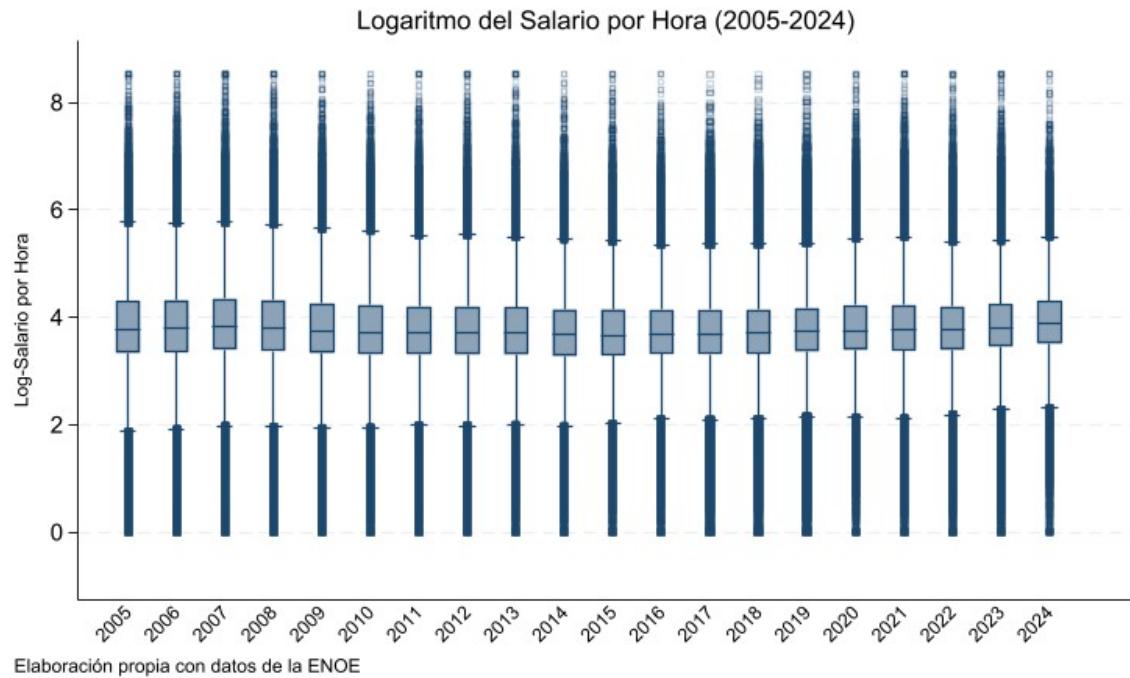
1 Problema 1: ENIGH y ENOE

El objetivo es tener características de individuos y salarios para diferentes puntos en el tiempo.

1. Las bases que utilizaremos son todas las ENIGH hasta 2022, y todas las ENOE de todos los trimestres hasta 2024 (sin ETOE, si tienes problema con el peso, puedes quedarte con el 2º trimestre de cada año). Es muy importante que las bases de datos queden limpias y uniformes en el número de observaciones con todas las variables utilizadas, pues las bases construidas aquí serán utilizadas en los siguientes problemas.
2. **ENIGH:** De la tabla de ingresos obtén el ingreso del mes pasado por trabajo y el ingreso trimestral normalizado y pégalo a la base de población. En la base de ingreso tienes que escoger el ingreso del mes pasado (fíjate en la codificación de las variables), y luego hacer un `collapse` para el folio del individuo. Después tienes que hacer un `merge, by(folio)` con la base de población. Esto lo tienes que hacer para cada año, y después juntar los años con el comando `append` (no se te olvide crear una variable `gen year=1996` para el año 1996, y otros años). El ingreso por trabajo debe de ser homogéneo a través de los años. Para ayuda leer el artículo de Campos, Lustig y Santillan (2014) en *Estudios Económicos*.
3. Al final de este ejercicio, debes tener una base a nivel individual con las muestras de todos los años, similar para la ENOE. **ENIGH:** 1992-2022. **ENOE:** 2005-2024 (no usar ETOE).
4. **ENIGH y ENOE:** Las variables importantes son edad, educación (años y por nivel terminado), ingreso del mes pasado, ingreso trimestral normalizado, horas trabajadas, sexo, rural. Asegúrate de tener el ingreso en términos reales, digamos enero de 2024 (busca en el Banco de México el INPC para que puedas tener el ingreso constante). Restringe las observaciones a individuos entre 25-65 años de edad (ENOE solo se tiene un ingreso laboral). Calcula un `sum, detail` de cada una de las variables para cada año o bien una tabla con diferentes estadísticos para que estés seguro que la limpieza de los datos es correcta, y que no tienes missing values. Es decir, son las mismas observaciones para todas las variables. Asegúrate que eso se cumple, y comenta los pasos de limpieza así como las estadísticas finales. También

puedes checar el comando `tabstat` o `table`. Averigua comandos `putexcel` o `estout` para exportar a Excel.

5. Crea 8 grupos de observaciones. Crea grupos para sexo-edad-educación. Donde grupo de edad es definida como 25-45 y 46-65, y grupo de educación es definida como menos de preparatoria o más o igual que preparatoria.
6. Si el salario es igual a 999999 quiere decir que ese salario no es válido, asegúrate de que ese salario sea convertido a “missing value” (antes de ser cambiado a real, los ceros tampoco los usamos). Define una variable llamada `trabajo` que sea `dummy variable` para los trabajadores, donde trabajador es definido como aquel con un salario válido. Cambia el salario a pesos reales de enero 2024.
7. Crea una variable de salario por hora. Es decir, divide el salario por horas trabajadas en la semana multiplicadas por 4.33 para asegurarte que el salario es por hora. Haz un censoring de los datos, es decir, todo salario por hora menor a 1 cámbialo a valor 1, mientras que todos los salarios por hora mayores a \$5000, restríngelos a \$5000. Asegúrate de no cambiar missing values o valores cero. ¿Cuántas observaciones cambias, a qué se debe?
8. Analiza la dispersión de los datos usando figuras de boxplot. Usa el logaritmo del salario por hora. Pon en eje x los años. Entender significancia y valores en el boxplot.

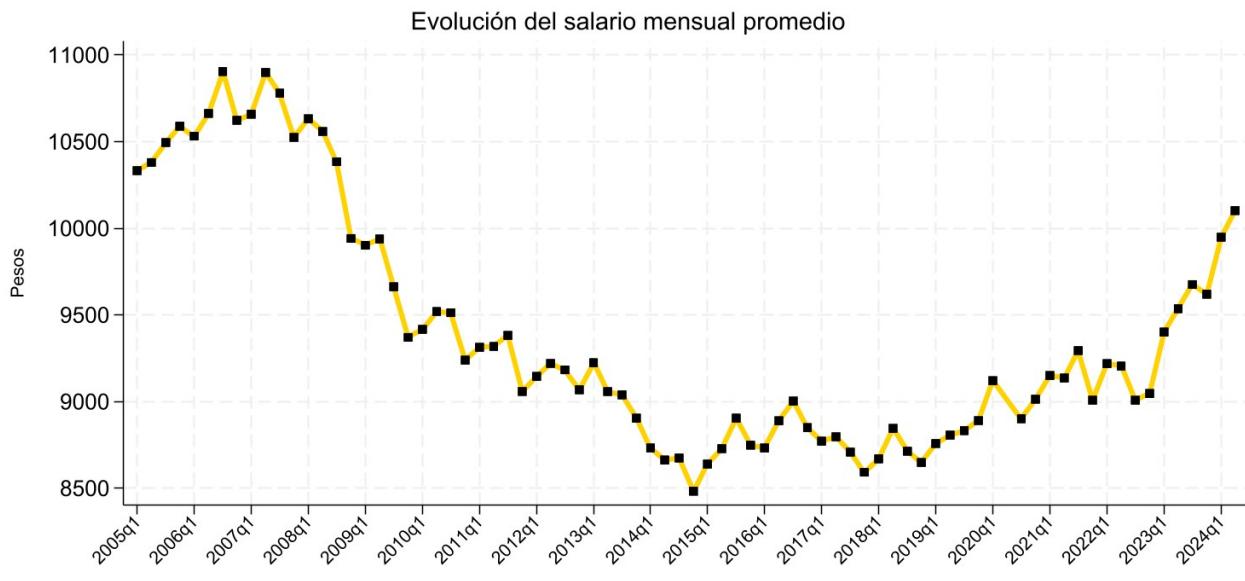


9. Analiza cómo cambia el salario y el salario por hora para la población y para cada grupo (asegúrate de utilizar los pesos factor que incluye la ENIGH y ENOE, averigua en el `help` las propiedades de `fweight` y `aweight`) **únicamente para la población con salario por hora válido**. De ahora en adelante cuando hablamos de cálculos de salario o salario por hora me refiero únicamente a aquellos con ingreso válido (es decir, ingreso positivo, todos los ingresos ceros, missing, de aquellos que no trabajan o no reciben pago no se incluyen). Utiliza el comando `table` y realiza gráficas al respecto. Discute qué grupo ha sido más afectado o beneficiado.

Comencemos analizando la evolución del ingreso mensual y el salario por hora a través del tiempo, considerando la media nacional.

Las figuras de la siguiente página muestran que, para el salario mensual promedio, se observa un claro aumento en los salarios entre 2005 y 2008, alcanzando un máximo superior a los 10,500 pesos. Posteriormente, hay una caída pronunciada que llega a su punto más bajo en 2015, con salarios cercanos a los 8,500 pesos. A partir de 2012, los salarios muestran una estabilidad relativa, con ligeras fluctuaciones, hasta que en 2020 comienza una tendencia alcista significativa que se intensifica en 2022 y se proyecta hacia 2024, alcanzando los 11,000 pesos. Este aumento reciente podría estar relacionado con los ajustes al salario mínimo y otras regulaciones en materia laboral.

El segundo gráfico muestra la evolución del salario por hora promedio en el mismo periodo. Al igual que en el salario mensual, se observa un aumento hasta 2008, cuando el salario por hora alcanza su máximo de 65 pesos aprox. Sin embargo, este indicador también experimenta una fuerte caída entre 2009 y 2012, estabilizándose alrededor de los 50-55 pesos por hora durante varios años. A partir de 2020, el salario por hora empieza a crecer nuevamente de manera sostenida, superando los 60 pesos por hora en 2024. La tendencia sugiere un incremento en los ingresos por hora trabajada, lo que podría reflejar mejoras en la productividad o en las condiciones del mercado laboral en los últimos años.



Elaboración propia con datos de la ENOE. Valores en pesos constantes de enero 2024



Elaboración propia con datos de la ENOE. Valores en pesos constantes de enero 2024

Ahora bien, para entender mejor la evolución de estos indicadores por grupo, podemos construir tablas donde diferenciamos entre cada uno de los 8 grupos. Para el caso del ingreso mensual tenemos lo siguiente, año por año:

Año	Ingreso mensual promedio por grupo								Total
	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	
2005	7,608	13,429	7,667	17,757	4,785	10,366	4,674	13,087	10,450
2006	7,683	13,697	7,679	18,091	4,956	10,503	4,866	13,132	10,681
2007	7,861	13,626	7,735	17,812	4,963	10,452	4,837	12,855	10,716
2008	7,674	13,240	7,472	16,745	4,983	9,998	4,781	12,439	10,383
2009	7,137	12,183	7,191	15,156	4,640	9,480	4,697	11,736	9,719
2010	6,915	11,694	6,917	14,653	4,593	9,133	4,551	10,891	9,423
2011	7,007	11,234	6,971	14,193	4,657	8,969	4,566	10,482	9,268
2012	6,742	11,173	6,845	13,773	4,570	8,768	4,571	10,297	9,155
2013	6,927	11,002	6,705	13,391	4,563	8,655	4,384	10,194	9,057
2014	6,652	10,508	6,479	12,441	4,471	8,252	4,260	9,409	8,639
2015	6,757	10,654	6,582	12,528	4,555	8,319	4,316	9,174	8,756
2016	6,924	10,907	6,708	12,131	4,608	8,356	4,402	9,148	8,868
2017	6,976	10,475	6,698	11,885	4,660	8,134	4,521	8,927	8,719
2018	7,084	10,445	6,840	11,583	4,697	8,117	4,526	8,648	8,720
2019	7,271	10,550	6,881	11,622	4,821	8,201	4,621	8,486	8,823
2020	7,217	10,588	6,894	11,632	4,928	8,418	4,669	9,076	9,016
2021	7,113	10,825	7,114	11,696	5,018	8,483	4,783	8,829	9,147
2022	7,502	10,790	7,159	11,425	5,106	8,583	4,739	8,566	9,120
2023	7,852	11,507	7,463	12,030	5,266	8,875	4,897	8,773	9,558
2024	8,210	11,995	7,839	12,705	5,440	9,256	5,124	9,121	10,025
Total	7,232	11,496	7,094	13,256	4,777	8,920	4,623	9,748	9,392

Elaboración propia con datos de la ENOE.

Valores a pesos constantes de 2024

La evolución del ingreso mensual promedio por grupo muestra disparidades significativas según la edad, género y nivel educativo. Los hombres de 46 a 65 años con mayor educación (Grupo 4) han mantenido consistentemente los ingresos más altos a lo largo de los años, superando los 13,000 pesos en 2024, mientras que las mujeres del mismo rango de edad y nivel educativo (Grupo 8) tienen ingresos más bajos, alcanzando alrededor de 10,941 pesos. En contraste, los grupos con menos educación, especialmente las mujeres jóvenes sin preparatoria (Grupo 5) y las mujeres mayores sin preparatoria (Grupo 7), presentan los ingresos más bajos, con una tendencia estancada alrededor de los 4,000-5,000 pesos durante el periodo. Los hombres jóvenes con menor educación (Grupo 1) también muestran ingresos relativamente bajos, aunque con una ligera mejora hacia 2024. La brecha salarial entre hombres y mujeres, así como entre niveles educativos, es evidente, reflejando una recuperación más lenta para las mujeres y aquellos con menor formación académica. Por último, podemos notar la misma tendencia general en todos los grupos, en la que el salario tiene una caída y una posterior recuperación.

Para el caso del salario por hora:¹

Año	Salario por hora promedio por grupo								Total
	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	
2005	41	71	43	99	35	70	36	91	61
2006	41	73	44	102	37	72	38	93	63
2007	43	73	45	101	38	72	39	90	64
2008	42	71	43	95	38	69	39	88	62
2009	40	66	42	87	36	65	37	83	59
2010	38	63	41	85	34	63	37	78	57
2011	39	60	41	80	35	62	36	74	56
2012	38	60	41	79	35	61	36	75	56
2013	39	59	40	76	35	60	37	73	55
2014	36	57	39	71	33	57	35	67	52
2015	37	57	39	71	35	58	35	67	53
2016	38	58	39	69	35	58	36	65	53
2017	38	56	39	68	35	57	36	64	53
2018	38	56	40	65	35	56	37	62	52
2019	39	57	40	67	36	57	38	61	53
2020	41	60	42	70	37	61	39	69	57
2021	40	61	43	72	37	62	39	68	58
2022	41	59	41	67	37	59	38	62	56
2023	43	63	44	70	39	62	39	64	59
2024	45	66	46	74	40	64	42	68	62
Total	40	62	41	76	36	62	37	70	57

Elaboración propia con datos de la ENOE.

Valores a pesos constantes de 2024

Los hombres de 46 a 65 años con prepa o más (Grupo 4) han mantenido los salarios más altos durante todo el periodo, aunque también experimentaron una disminución en los últimos años, cerrando en 80 pesos por hora en 2024. Las mujeres en este mismo grupo (Grupo 8) también presentan salarios relativamente altos, aunque siempre por debajo de los hombres, con un promedio de 73 pesos en 2024. En contraste, las mujeres jóvenes y mayores sin prepa (Grupos 5 y 7) tienen los salarios más bajos, con promedios que apenas superan los 35 pesos por hora.

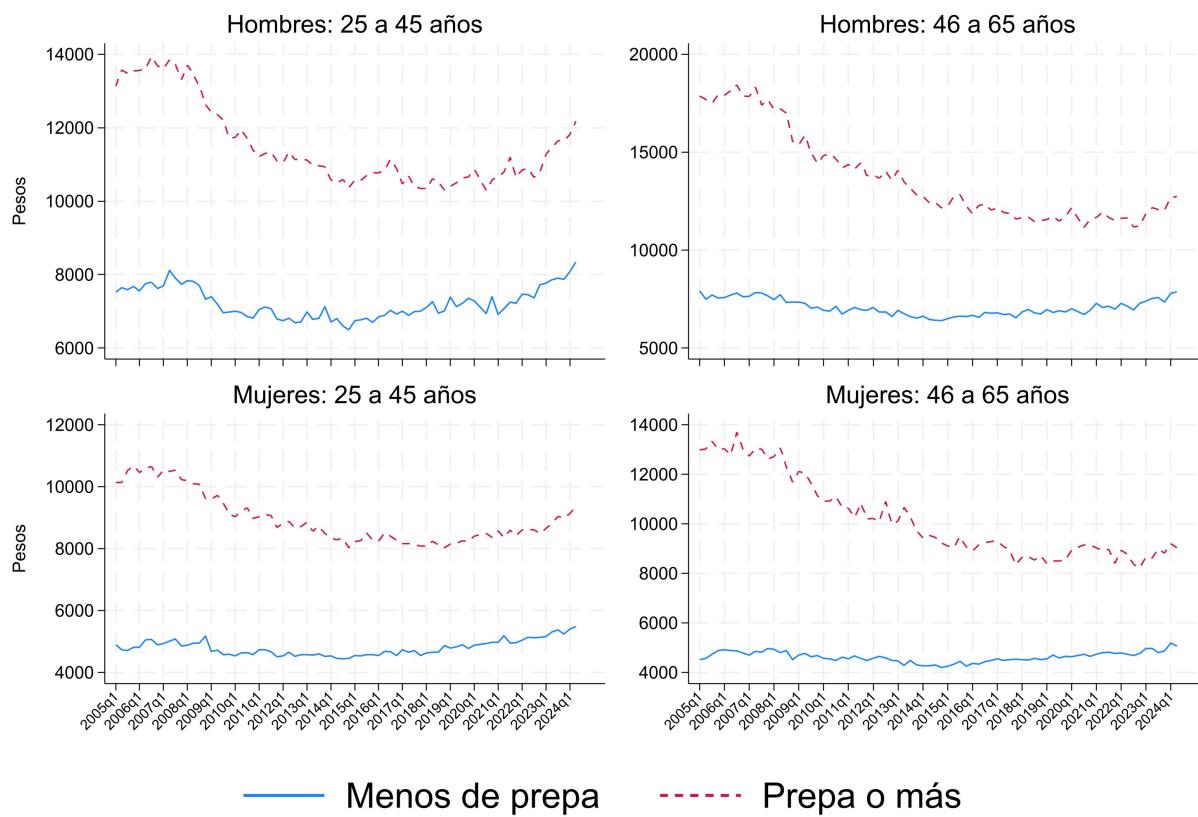
Podemos visualizar la misma información pero a nivel trimestral, y comparando por grupos educativos y por sexo.

(Siguiente página)

¹Para el cálculo de estos indicadores, se excluyeron a las personas que reportaron haber trabajado cero horas a la semana.

Salario mensual promedio por sexo, grupo educativo y etario

Comparación entre grupos educativos

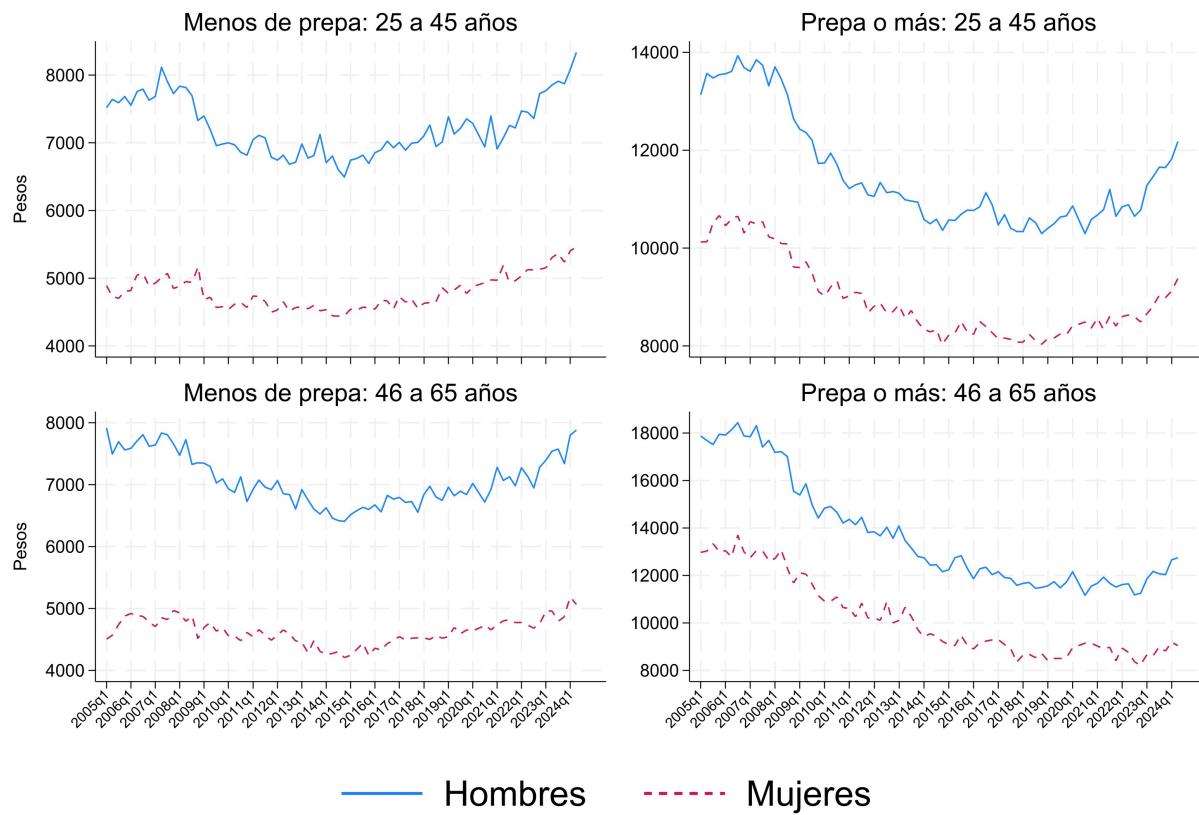


Elaboración propia con datos de la ENOE. Valores en pesos constantes de enero 2024

La gráfica muestra la evolución del salario mensual promedio por sexo, grupo educativo y etario desde 2005 hasta 2024. Los hombres de 25 a 45 años con mayor educación (preparatoria o más) presentan salarios más altos en comparación con aquellos con menos educación, aunque ambos grupos experimentaron una caída pronunciada hasta 2015, seguida de una recuperación significativa hacia 2024. Una tendencia similar se observa en los hombres de 46 a 65 años, donde la brecha salarial entre niveles educativos es más amplia, y aunque ambos grupos se recuperan hacia 2024, los hombres con mayor educación siempre tienen salarios significativamente más altos.

En el caso de las mujeres, tanto jóvenes como mayores, los salarios son consistentemente más bajos en comparación con los hombres. Las mujeres de 25 a 45 años con mayor educación experimentan una ligera mejora hacia 2024, aunque la diferencia con las mujeres de menor educación es menos marcada. Para las mujeres de 46 a 65 años, la caída es más pronunciada hasta 2020, seguida de una leve recuperación. Finalmente, la brecha entre las mujeres con y sin prepa también persiste a lo largo del tiempo.

Salario mensual promedio por sexo, grupo educativo y etario
Comparación entre sexos

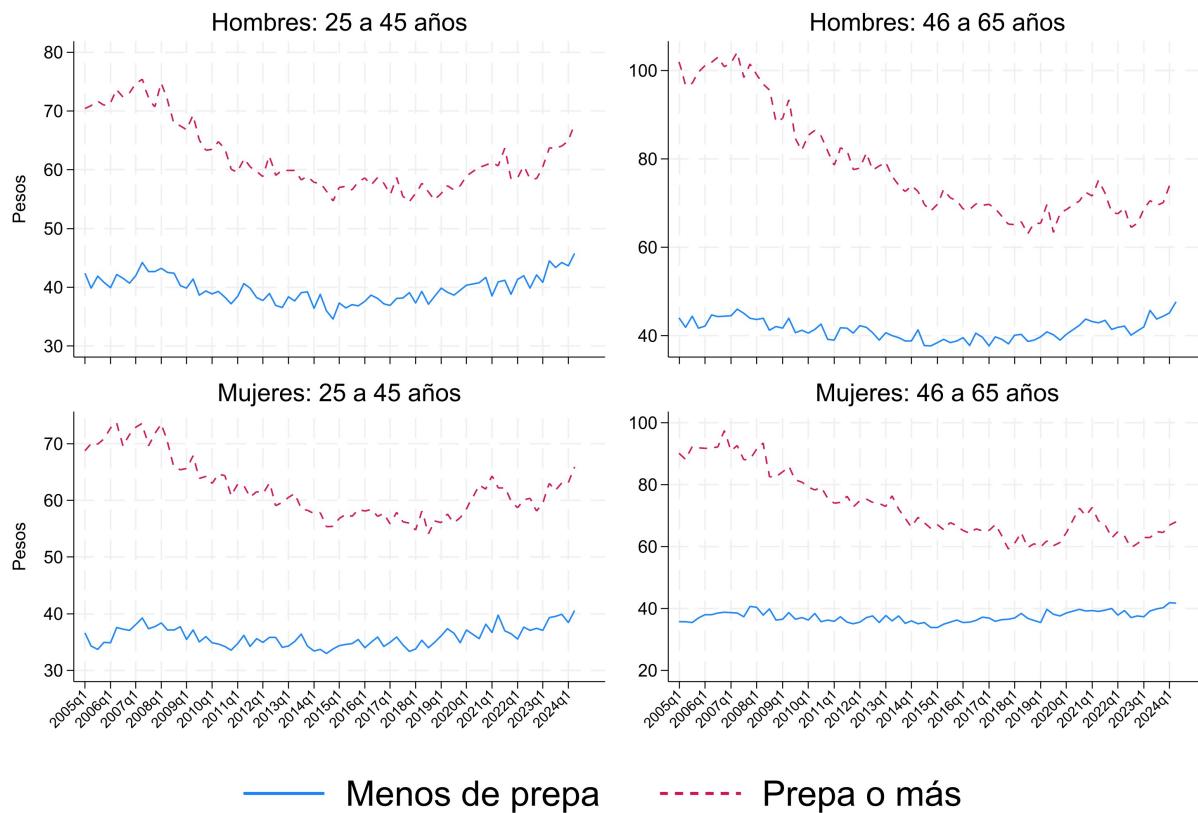


Elaboración propia con datos de la ENOE. Valores en pesos constantes de enero 2024

En todos los grupos, los hombres ganan significativamente más que las mujeres. Entre las personas con menos de preparatoria, tanto en el grupo de 25 a 45 años como en el de 46 a 65 años, los hombres muestran un aumento gradual de los salarios a partir de 2015, mientras que las mujeres experimentan un incremento más lento, manteniendo siempre una brecha considerable. Para los grupos con preparatoria o más, la diferencia de género también es notable. Los hombres de 25 a 45 años con prepa o más muestran una tendencia decreciente hasta 2020, seguida de una recuperación que los lleva a más de 12,000 pesos en 2024. Las mujeres, aunque con una mejora en la misma fecha, siguen muy por debajo de sus pares masculinos. En el grupo de 46 a 65 años con más educación, la tendencia es similar para los hombres, aunque con una recuperación mucho más lenta; para las mujeres, se observa una caída con estancamiento: los hombres alcanzan cerca de 12,000 pesos en 2024, mientras que las mujeres permanecen por debajo de los 10,000 pesos, lo que refleja una persistente brecha salarial de género en todos los niveles educativos y edades.

Salario por hora medio por sexo, grupo educativo y etario

Comparación entre grupos educativos

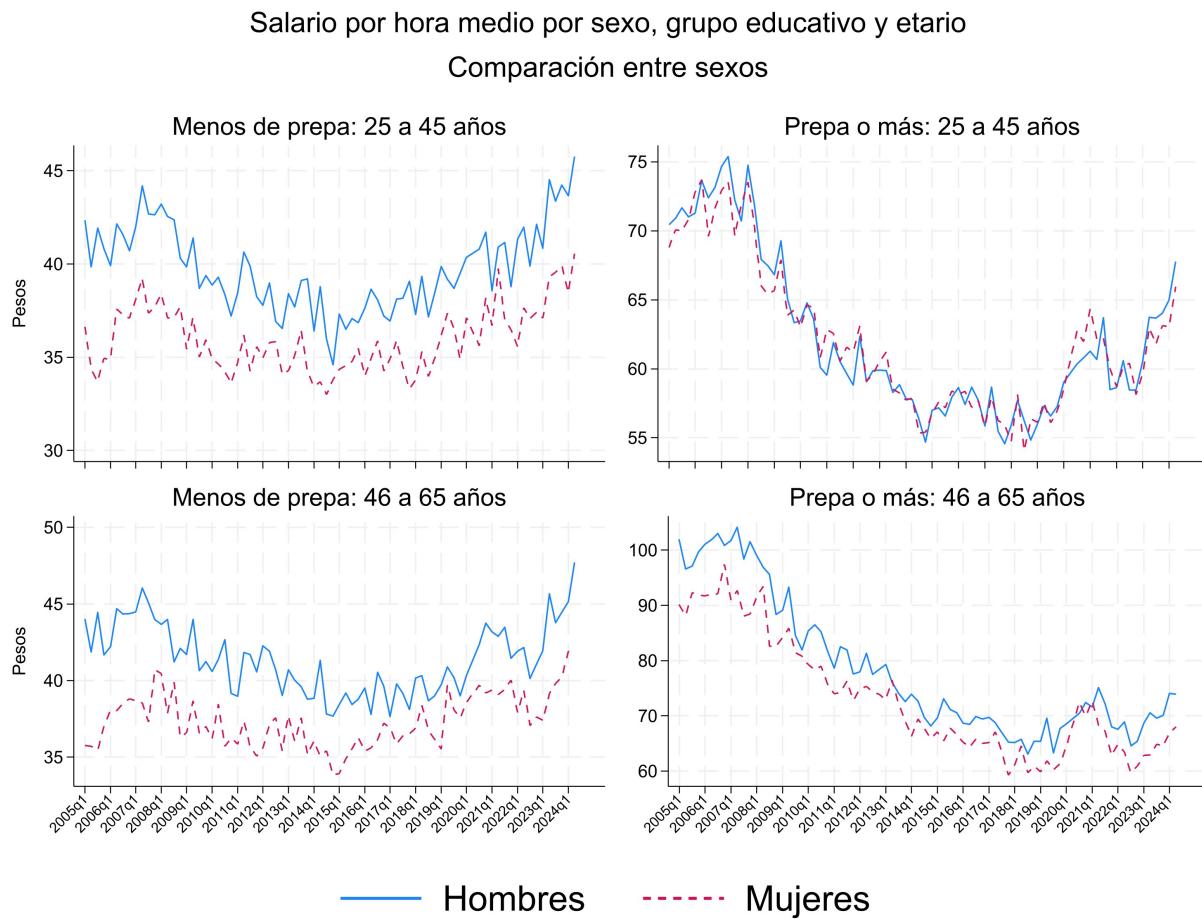


Elaboración propia con datos de la ENOE. Valores en pesos constantes de enero 2024

La gráfica muestra la evolución del salario por hora medio con comparaciones entre aquellos con menor educación (menos de prepa) y aquellos con mayor educación (prepa o más). En los hombres de 25 a 45 años, se observa una clara diferencia en salarios según el nivel educativo. Aquellos con prepa o más mantienen salarios superiores, con una tendencia de crecimiento en los últimos años, mientras que los hombres con menor educación presentan salarios más bajos, aunque con la misma tendencia al alza hacia 2024. En el caso de los hombres de 46 a 65 años, la brecha entre niveles educativos es más evidente, con los más educados alcanzando salarios que oscilan entre 60 y 100 pesos, mientras que aquellos con menos educación permanecen por debajo de los 50 pesos por hora.

En las mujeres, la tendencia es similar. Las mujeres de 25 a 45 años con prepa o más también presentan salarios más altos que aquellas con menos de prepa, aunque en ambos grupos los salarios son consistentemente más bajos en comparación con los hombres. Las mujeres de mayor edad (46 a 65 años) muestran una disminución de sus salarios a lo largo del tiempo, especialmente las más educadas, aunque se observa una ligera recuperación hacia 2024. Las mujeres con menos de prepa en este grupo etario mantienen los salarios más bajos, con pocas

fluctuaciones, permaneciendo por debajo de los 40 pesos por hora en la mayor parte del periodo.



La gráfica muestra que, en el caso de las personas con menos de prepa, los hombres de 25 a 45 años muestran una tendencia al alza más pronunciada que las mujeres, con una clara brecha salarial que se amplía hacia 2024, cuando los hombres alcanzan más de 45 pesos por hora, mientras que las mujeres se quedan alrededor de los 40 pesos. Para el grupo de 46 a 65 años, la disparidad entre hombres y mujeres con menos educación es también evidente, aunque menos acentuada, con los hombres ganando consistentemente más que las mujeres.

Entre las personas con prepa o más, la brecha salarial es más estrecha, especialmente en el grupo de 25 a 45 años, donde los hombres y mujeres presentan salarios relativamente cercanos, aunque los hombres continúan ganando un poco más. En el grupo de 46 a 65 años con mayor educación, los hombres continúan ganando considerablemente más que las mujeres, con diferencias que oscilan entre 10 y 20 pesos por hora. Esto refuerza la persistencia de las brechas salariales de género, aunque estas son menores en los grupos con mayor nivel

educativo, y aparentemente nula en el caso de las personas jóvenes y educadas.

10. Analiza cómo cambia la proporción de trabajadores para la población y para cada grupo (asegúrate de utilizar los pesos factor que incluye la ENIGH y ENOE). Utiliza el comando **table**. Discute qué grupo ha sido más afectado o beneficiado en términos de participación laboral.

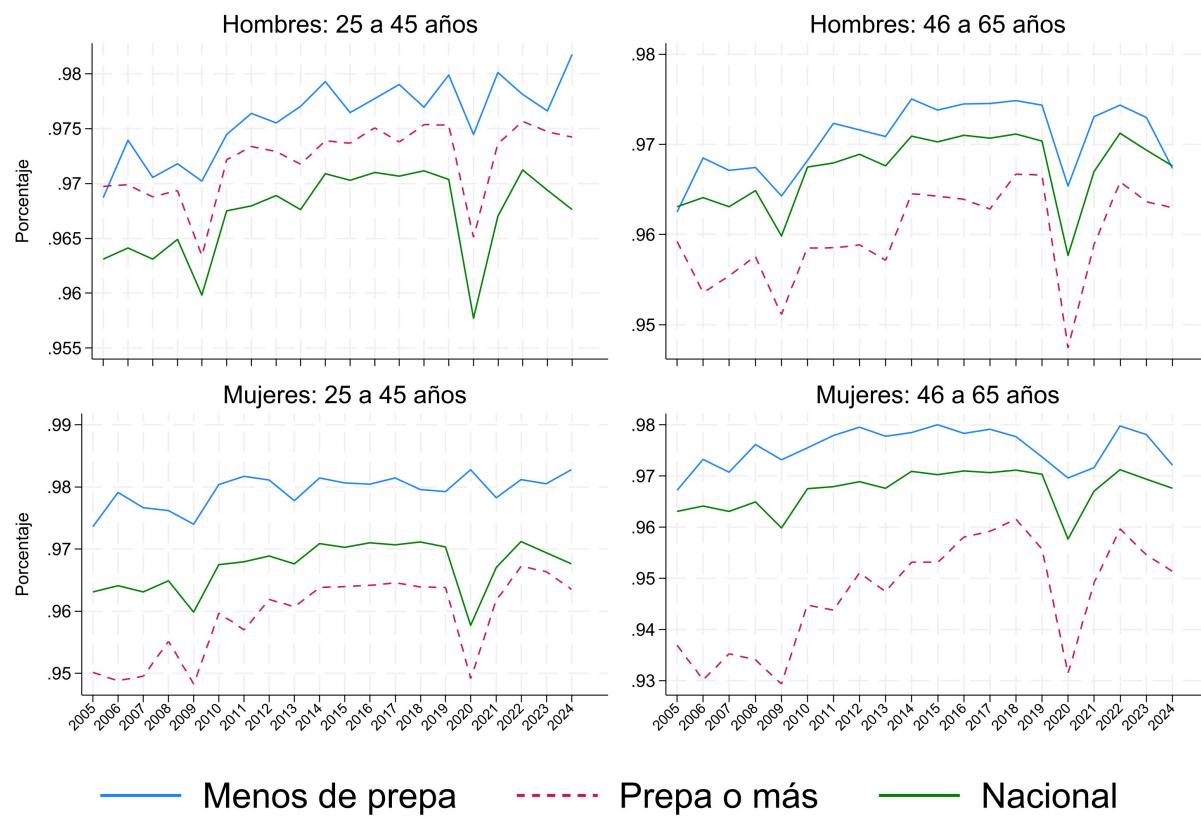
Utilizando el comando **table** obtenemos:

Año	Participación laboral por grupos de edad, sexo y educación								Total
	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	
2005	0.97	0.97	0.96	0.96	0.97	0.95	0.97	0.94	0.96
2006	0.97	0.97	0.97	0.95	0.98	0.95	0.97	0.93	0.96
2007	0.97	0.97	0.97	0.96	0.98	0.95	0.97	0.94	0.96
2008	0.97	0.97	0.97	0.96	0.98	0.96	0.98	0.93	0.96
2009	0.97	0.96	0.96	0.95	0.97	0.95	0.97	0.93	0.96
2010	0.97	0.97	0.97	0.96	0.98	0.96	0.98	0.94	0.97
2011	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.94	0.97
2012	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.95	0.97
2013	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.95	0.97
2014	0.98	0.97	0.98	0.96	0.98	0.96	0.98	0.95	0.97
2015	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.95	0.97
2016	0.98	0.98	0.97	0.96	0.98	0.96	0.98	0.96	0.97
2017	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.96	0.97
2018	0.98	0.98	0.97	0.97	0.98	0.96	0.98	0.96	0.97
2019	0.98	0.98	0.97	0.97	0.98	0.96	0.97	0.96	0.97
2020	0.97	0.97	0.97	0.95	0.98	0.95	0.97	0.93	0.96
2021	0.98	0.97	0.97	0.96	0.98	0.96	0.97	0.95	0.97
2022	0.98	0.98	0.97	0.97	0.98	0.97	0.98	0.96	0.97
2023	0.98	0.97	0.97	0.96	0.98	0.97	0.98	0.95	0.97
2024	0.98	0.97	0.97	0.96	0.98	0.96	0.97	0.95	0.97
Total	0.98	0.97	0.97	0.96	0.98	0.96	0.98	0.95	0.97

Es difícil analizar la tabla por sí misma, por lo que la siguiente página contiene la información en cuatro gráficas.

(Siguiente página

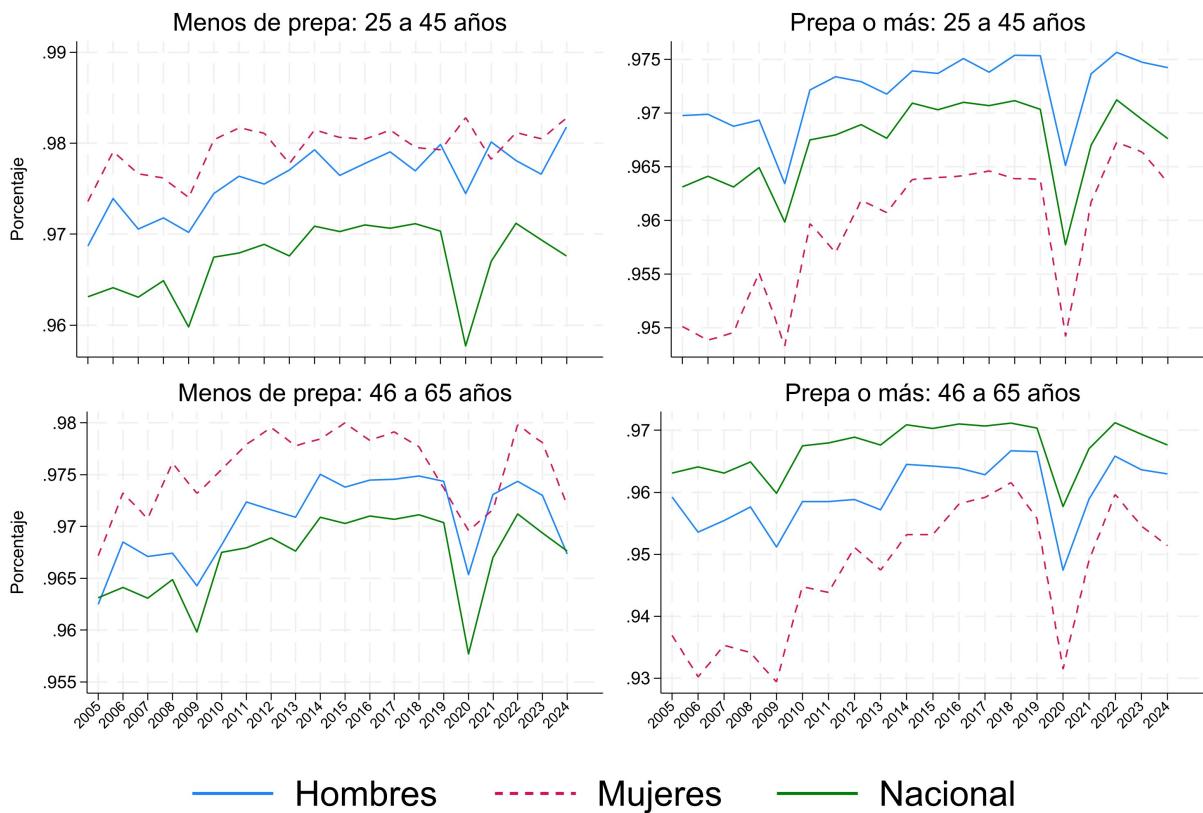
Participación laboral por sexo, grupo educativo y etario
Comparación entre grupos educativos



Elaboración propia con datos de la ENOE.

La gráfica anterior muestra la evolución de la participación laboral por grupo. Si diferenciamos entre aquellos que tienen menos de prepa y los que tienen prepa o más, observamos que en todos los casos, la población con menor educación tiene una participación laboral mayor en todo el periodo. Con excepción del grupo de hombres jóvenes, todos los grupos con mayor educación participan en la fuerza laboral en una proporción menor al promedio nacional. Ahora bien, la caída más abrupta por la pandemia del 2020 se observa en la población mayor, lo cual tiene sentido si consideramos que son el grupo más vulnerable de los dos que analizamos. Además, es visible la mayor afectación en la población menos educada con respecto a su participación laboral en el año de la pandemia.

Participación laboral por sexo, grupo educativo y etario
Comparación entre sexos



Elaboración propia con datos de la ENOE.

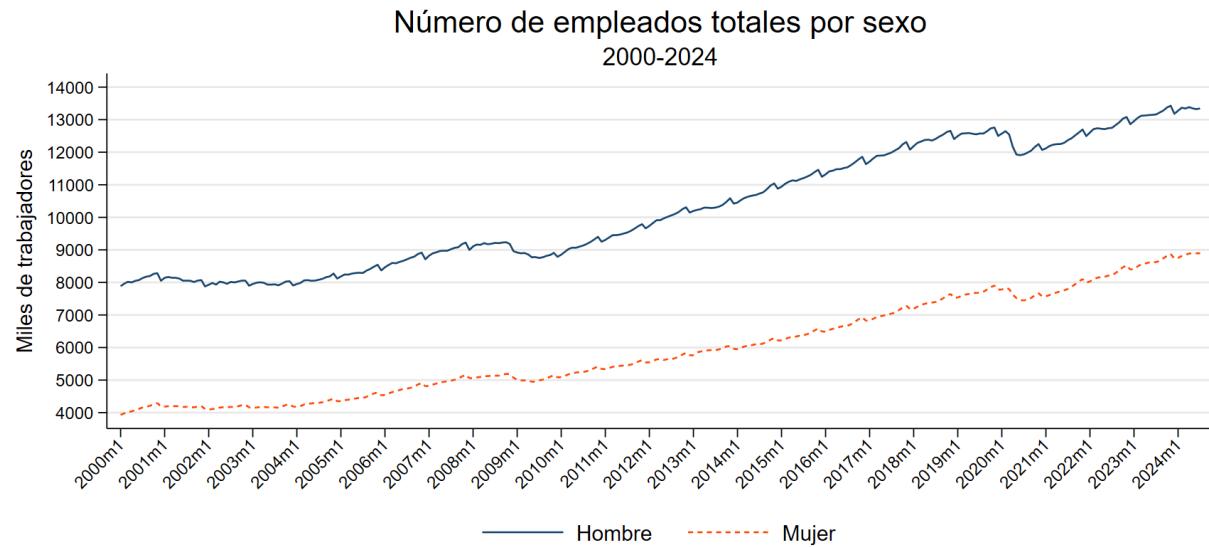
La gráfica anterior muestra la evolución de la participación laboral por grupos, diferenciando entre sexos. Al presentar la información de esta manera, podemos detectar varios elementos importantes: 1) la participación laboral femenina es relativamente mayor que la de los hombres si consideramos a la población menos educada. Esto no implica que hay más mujeres que hombres trabajando (de hecho, la proporción de la fuerza laboral compuesta por hombres es mayor que la de mujeres en todos los grupos), sino que, al interior de cada uno de los dos grupos, las mujeres tienden a participar más en la fuerza laboral que sus pares masculinos. Lo contrario sucede con la población más educada. 2) La participación laboral de las mujeres educadas se vio más afectada que la de sus pares masculinos en el año de pandemia. En el caso de las mujeres jóvenes con menor educación, la participación aumentó, quizás reflejo del trabajo extra que debieron realizar para cuidar de la población vulnerable.

11. ¿Qué tan comparable es el ingreso laboral en ENIGH 2018 a los anteriores? ¿Qué grupos fueron más afectados? Leer de internet la discusión de la comparabilidad de ingreso con la nueva ENIGH.

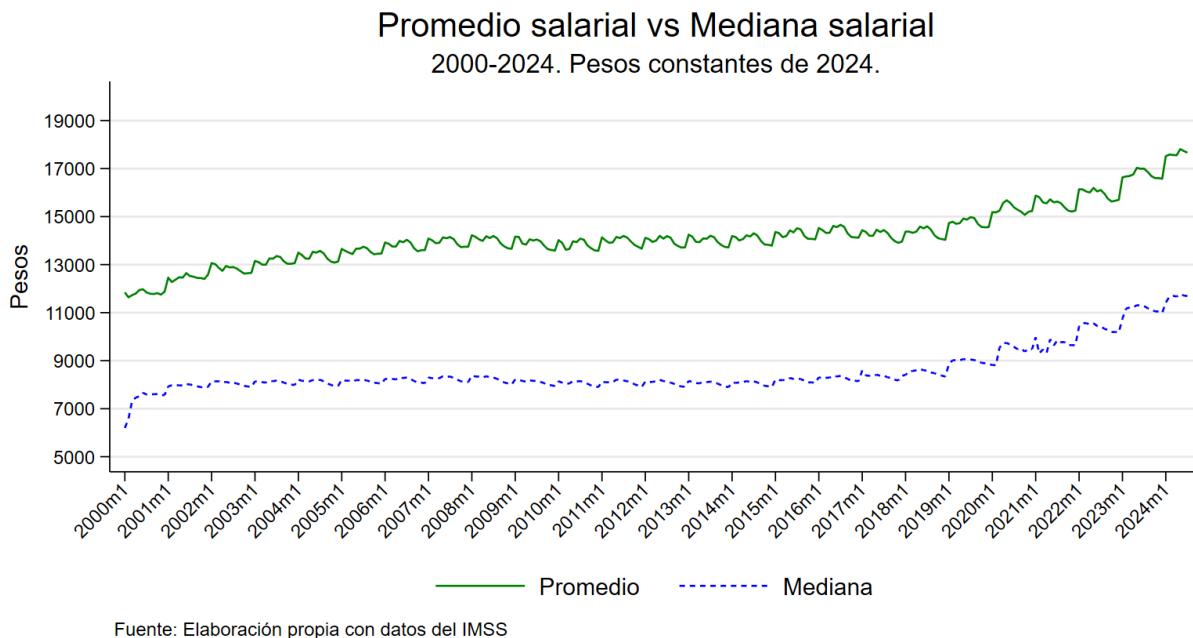
2 Problema 2: IMSS

El objetivo es tener características de individuos y salarios para diferentes puntos en el tiempo.

1. Usando los microdatos del IMSS para todos los meses, grafica el número de empleados totales en el IMSS (solo celdas con masa salarial positiva), por hombre y por mujer

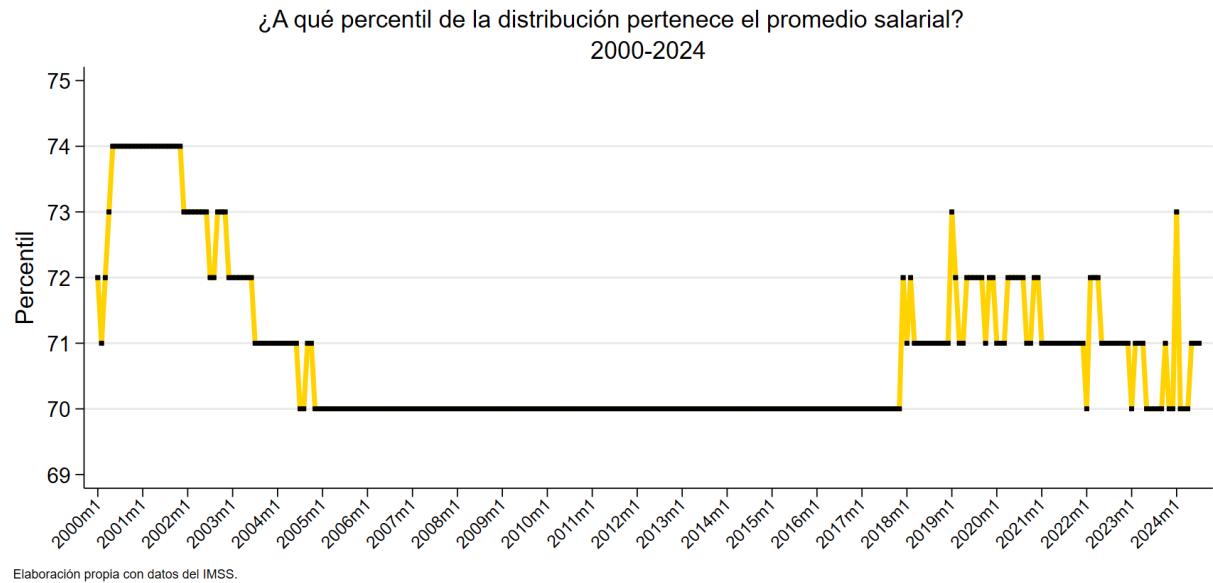


2. Pon los ingresos reales en el mismo periodo que el problema 1. Lleva los ingresos a ingresos mensuales. Grafica el promedio de los ingresos y la mediana de los ingresos para todos los trabajadores en todo el periodo. Razona cómo realizar esto pues tienes celdas a nivel de unidades (diferentes combinaciones de celdas). Hint: calcula el salario promedio (masa sobre número de trabajadores), y luego usa el número de trabajadores como ponderador.



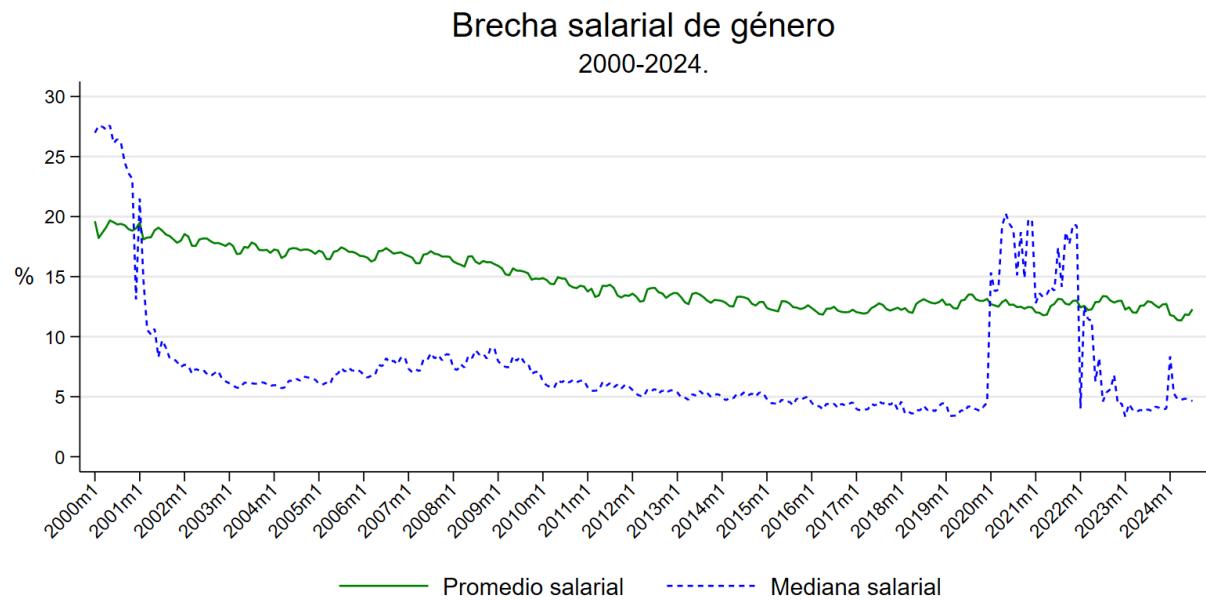
El promedio salarial se ubica por encima de la mediana salarial entre 4,000 y 6,000 pesos dependiendo del mes y año. Este hecho es indicativo de la gran desigualdad en los salarios.

3. Grafica el promedio y la mediana del salario al mes para todos los trabajadores. ¿En qué percentil se encuentra el promedio?



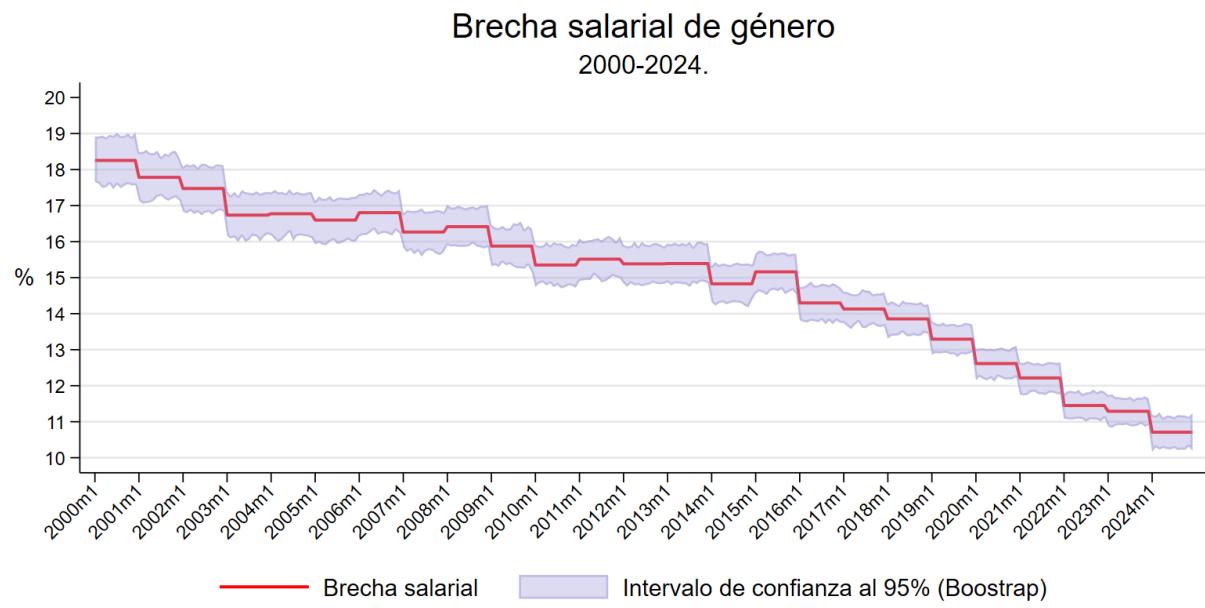
La gráfica nos muestra que el promedio salarial se ha ubicado entre los percentiles 74 y 70 a lo largo del periodo. La tendencia es a la baja: antes del 2005, se ubicaba normalmente entre los percentiles 71 y 74. Después de esa fecha, ha fluctuado entre los percentiles 72 y 70, con un par de excepciones.

4. Grafica la brecha salarial de género para todo el periodo.



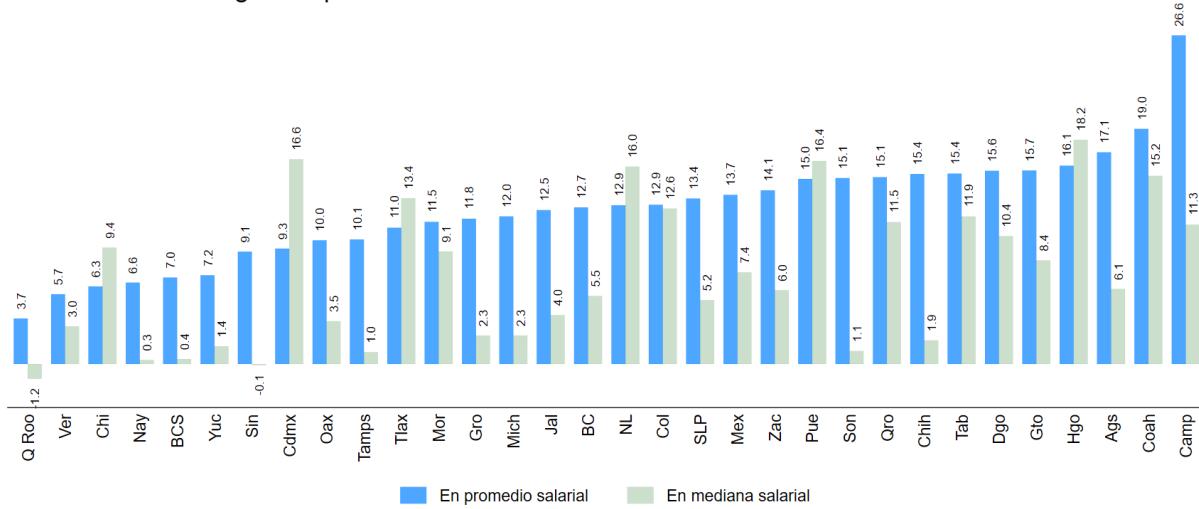
Fuente: Elaboración propia con datos del IMSS

5. Realiza un bootstrap para obtener intervalos de confianza con el método del percentil 95%, e inclúyelos en tu figura de brecha de género.



6. ¿Qué entidad federativa tiene la menor y mayor brecha de género en 2024? ¿Qué entidades son las más afectadas por la COVID (comparar Febrero 2020 con Julio 2024)?

Brecha salarial de género por entidad federativa 2024

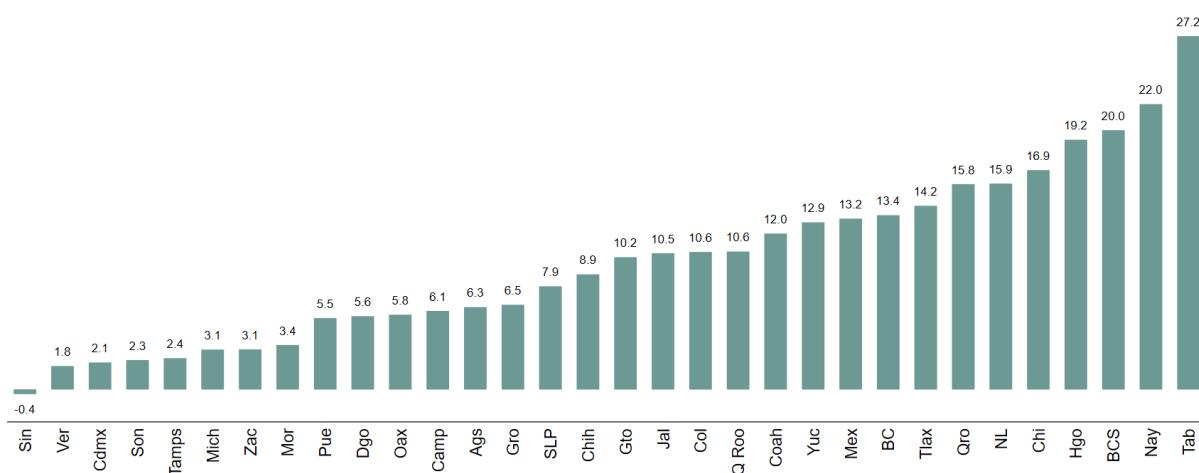


Fuente: Elaboración propia con datos del IMSS

En 2024, considerando la brecha en salarios promedio, la entidad con el mayor nivel de este indicador es Campeche, donde los hombres en promedio ganan más de un cuarto de veces lo que gana una mujer. Por su parte, la entidad con la menor brecha en salarios medios es Quintana Roo, que de hecho tiene una brecha negativa si consideramos salarios medianos.

Variación porcentual en el empleo por entidad federativa

Febrero 2020 - Julio 2024



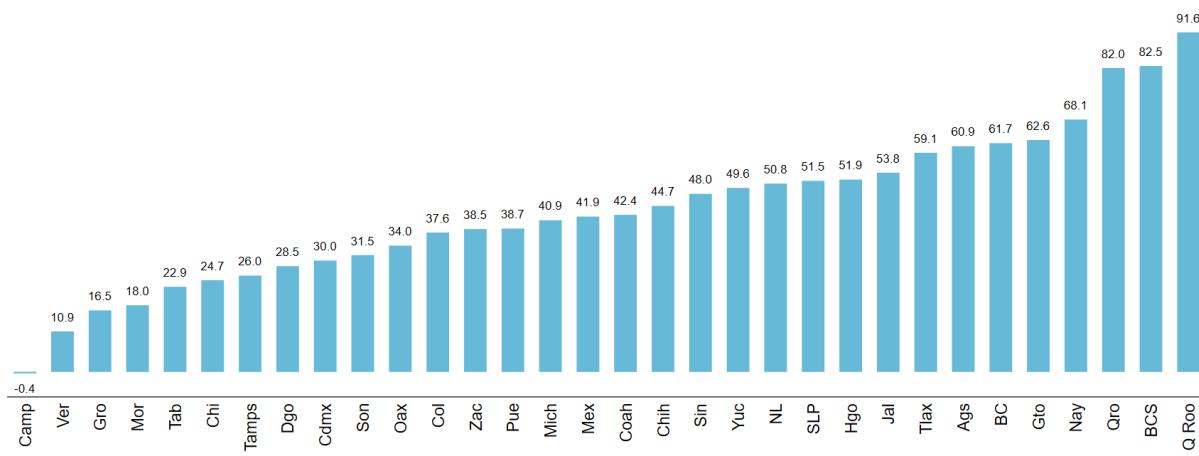
Fuente: Elaboración propia con datos del IMSS

Sinaloa es la entidad más afectada por la pandemia en términos de crecimiento de empleo.

7. ¿Qué entidad federativa ha incrementado más su empleo en términos porcentuales entre 2012 y 2024 (usa el mismo mes o trimestre)?

Variación porcentual en el nivel de empleo por entidad federativa

Mayo 2012 - Mayo 2024



Fuente: Elaboración propia con datos del IMSS

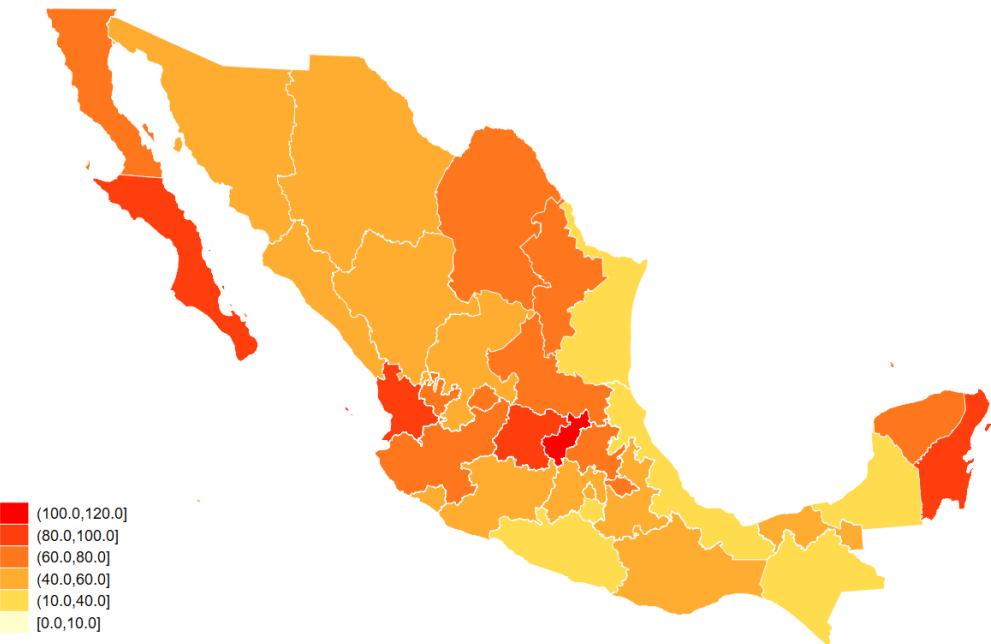
La entidad federativa que más incrementó su empleo en términos porcentuales entre 2012 y 2024 es Quintana Roo, que vio un aumento de casi el doble. La que menos ha incrementado es Campeche, que, de hecho, ha tenido una disminución de casi medio punto porcentual.

- Realiza un mapa en Stata y en R sobre los cambios en empleo y en salario a nivel entidad federativa. Usa dos años base: 2010 y 2017, y el año final 2024. Queremos saber la distribución de cambios en empleo y salario en ese periodo.

Mapas en la siguientes páginas.

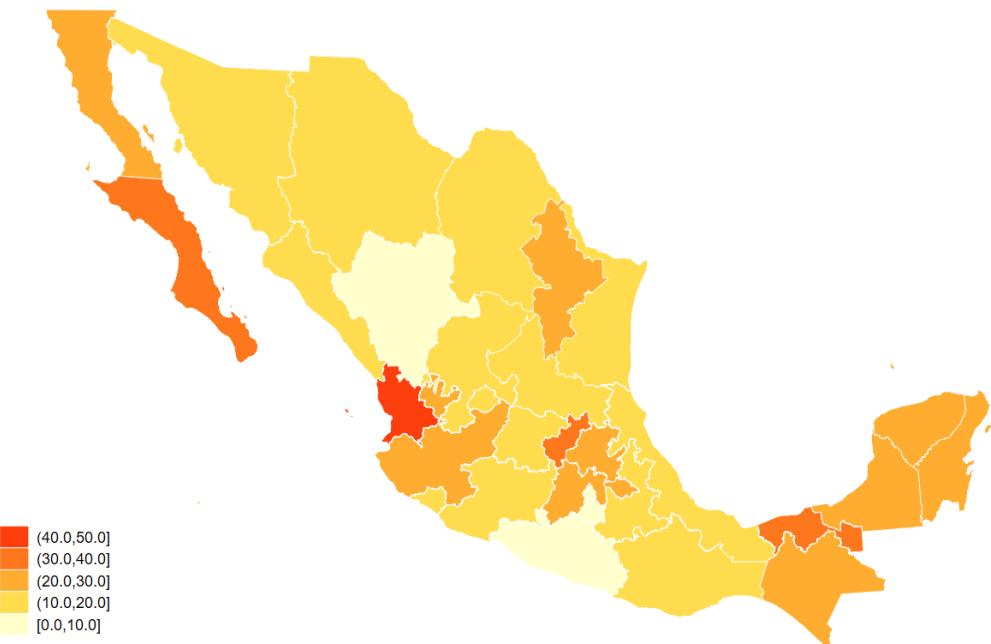
Empleo. Stata:

Variación porcentual en el número de trabajadores por entidad: 2024 vs 2010



Fuente: Elaboración propia con datos del IMSS. Tomamos julio como el mes de referencia

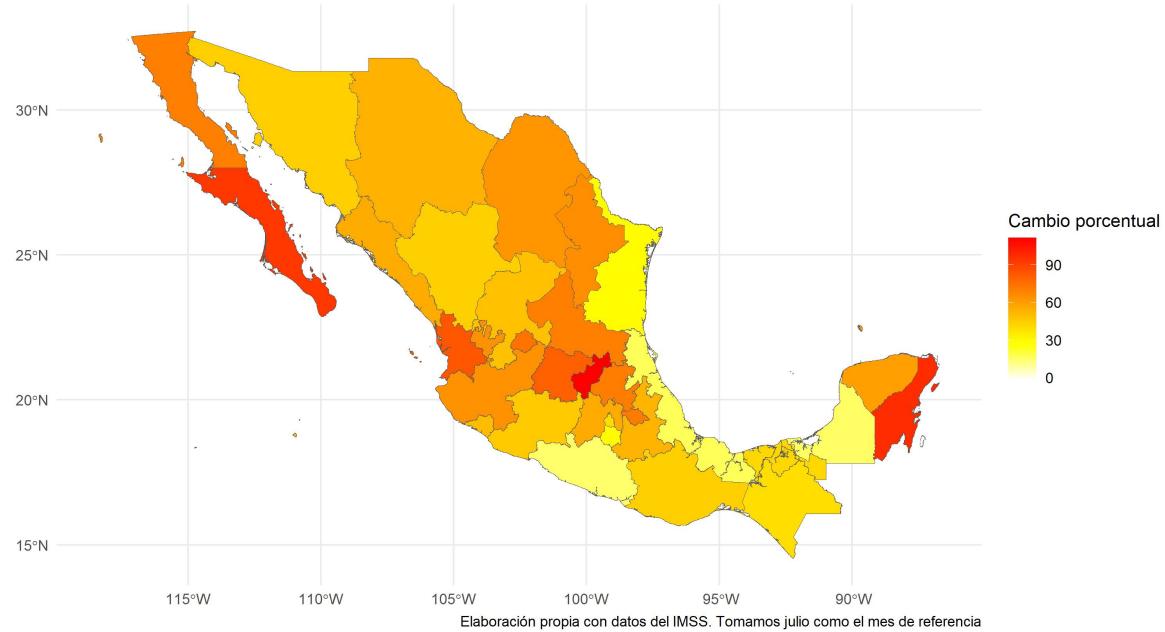
Variación porcentual en el número de trabajadores por entidad: 2024 vs 2017



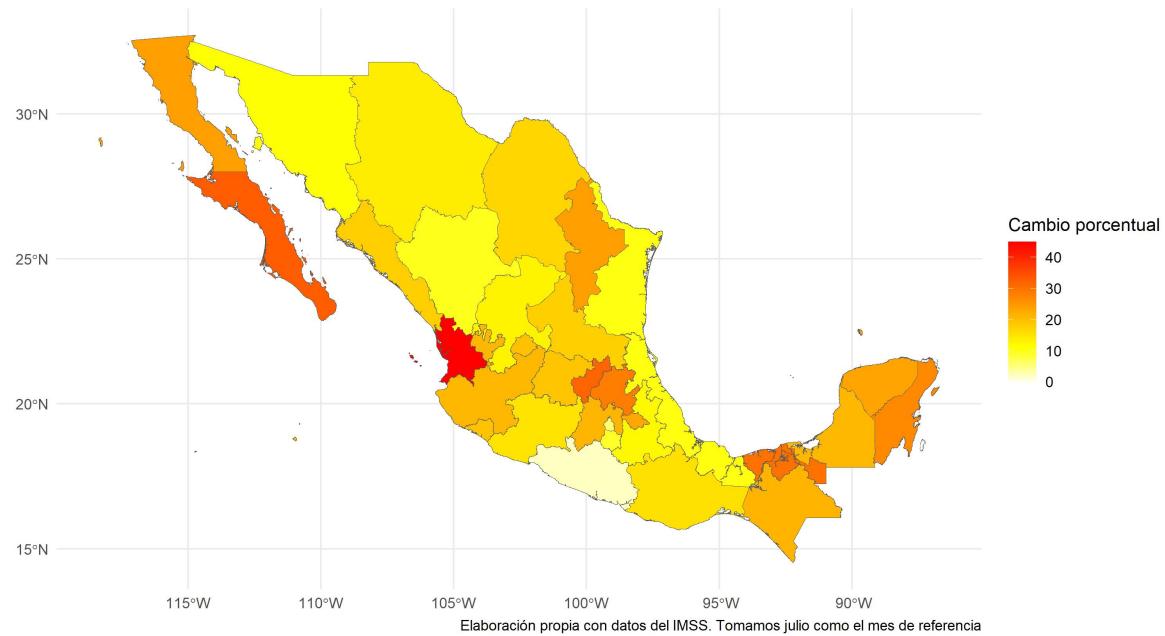
Fuente: Elaboración propia con datos del IMSS. Tomamos julio como el mes de referencia

Empleo. R:

Variación porcentual en el número de trabajadores por entidad
2010 - 2024

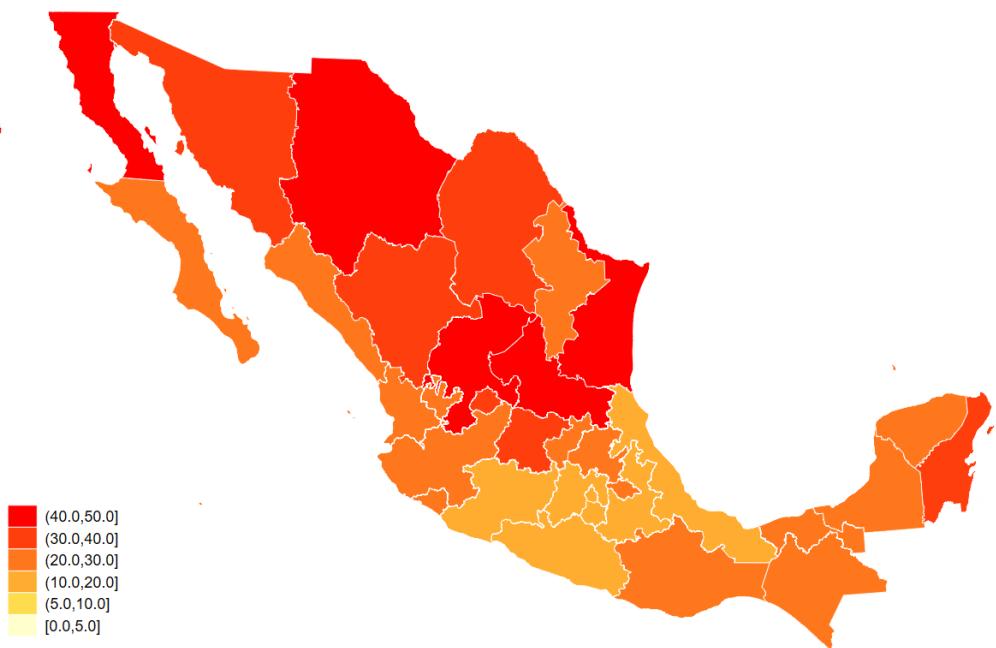


Variación porcentual en el número de trabajadores por entidad
2017- 2024



Salarios. Stata:

Variación porcentual en los salarios promedio por entidad: 2024 vs 2010



Fuente: Elaboración propia con datos del IMSS. Tomamos julio como el mes de referencia

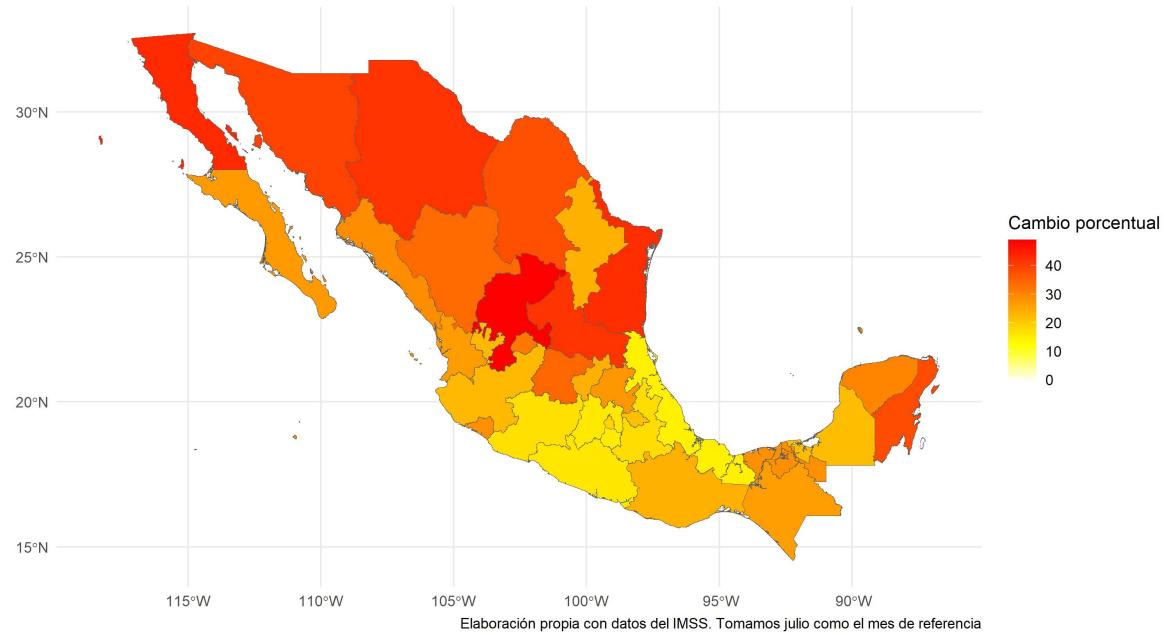
Variación porcentual en los salarios promedio por entidad: 2024 vs 2017



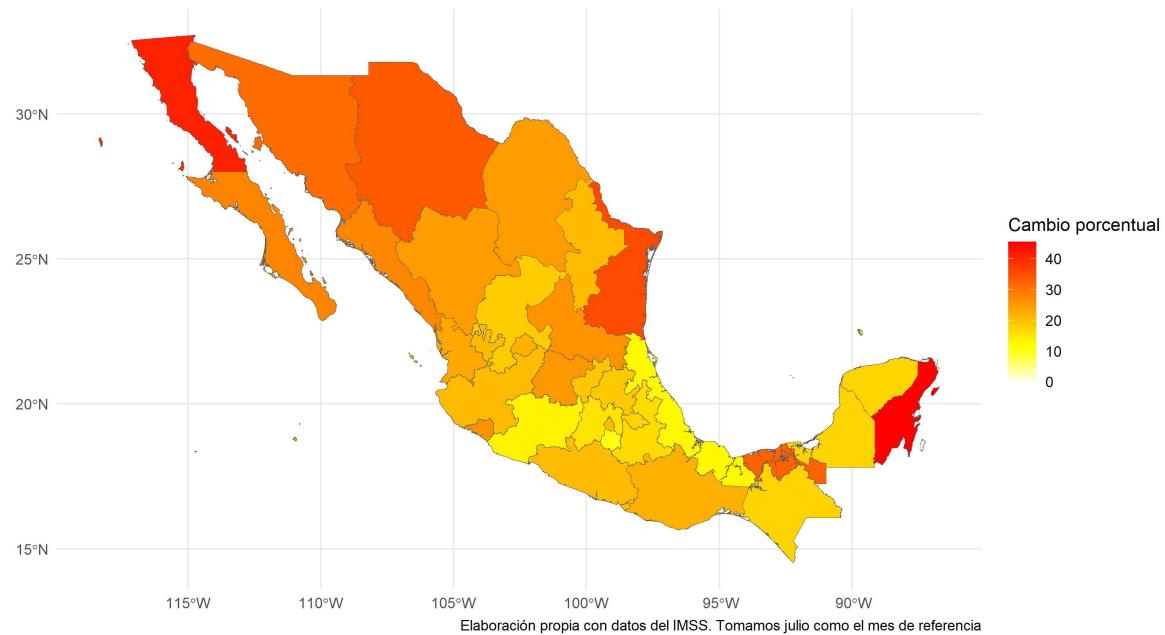
Fuente: Elaboración propia con datos del IMSS. Tomamos julio como el mes de referencia

Salarios. R:

Variación porcentual en los salarios promedio por entidad
2010 - 2024



Variación porcentual en los salarios promedio por entidad
2017- 2024



9. Realiza un mapa en Stata y en R sobre los cambios en empleo y en salario a nivel entidad federativa. Usa Febrero 2020 como base, y el año final Julio 2020 y 2024. Queremos saber la distribución de cambios en empleo y salario en ese periodo.

Mapas en las siguientes páginas.

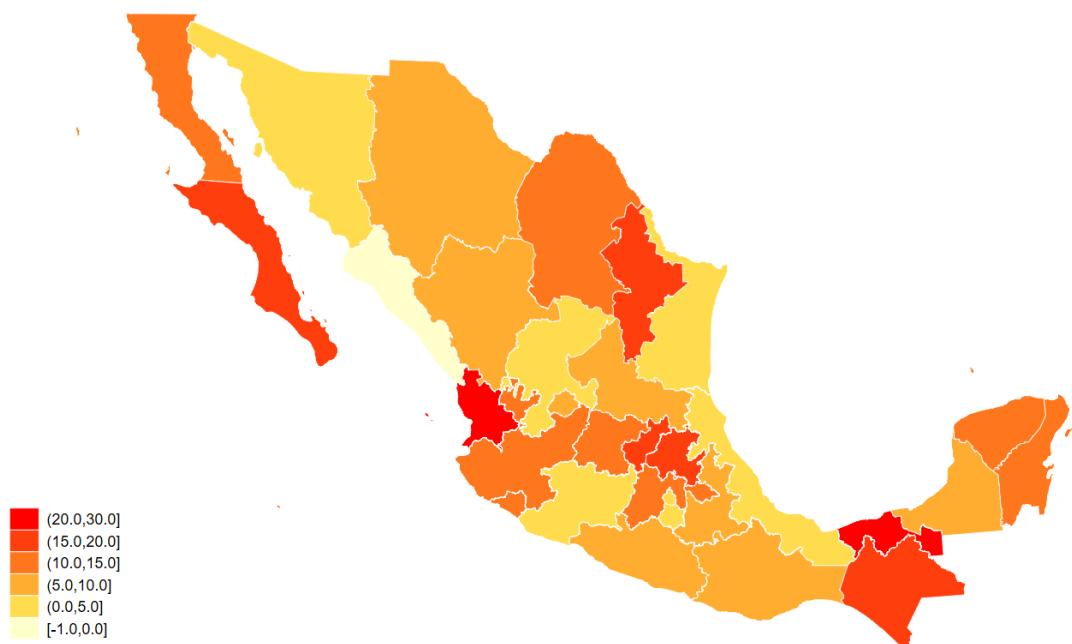
Empleo. Stata:

Variación porcentual en el número de trabajadores por entidad: Julio 2020 vs Febrero 2020



Fuente: Elaboración propia con datos del IMSS.

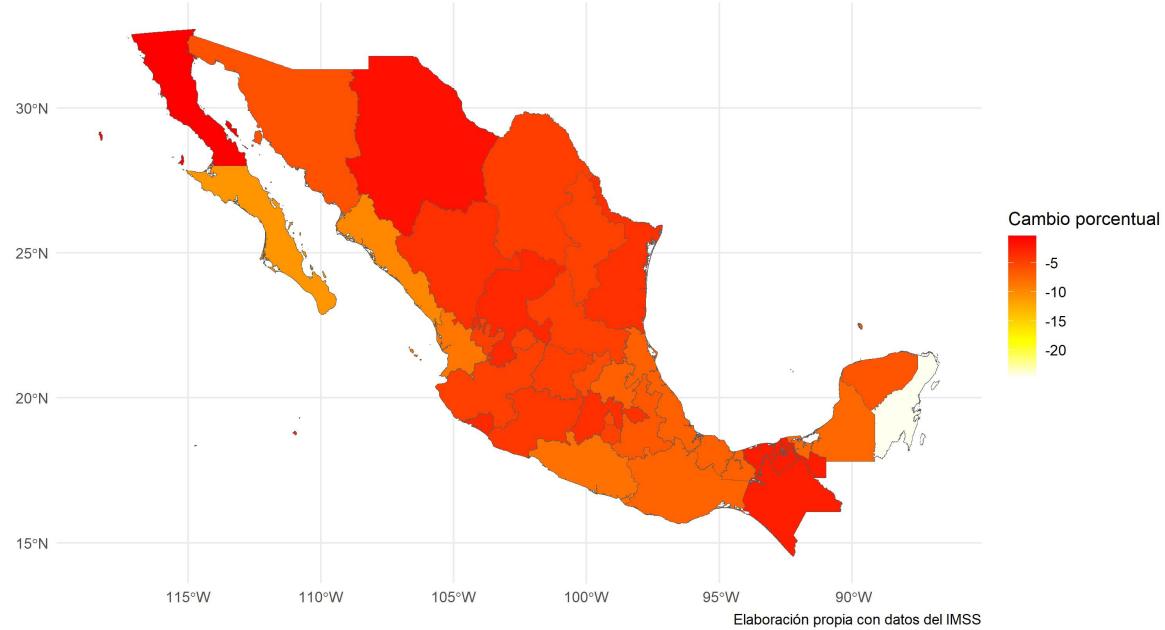
Variación porcentual en el número de trabajadores por entidad: Julio 2024 vs Febrero 2020



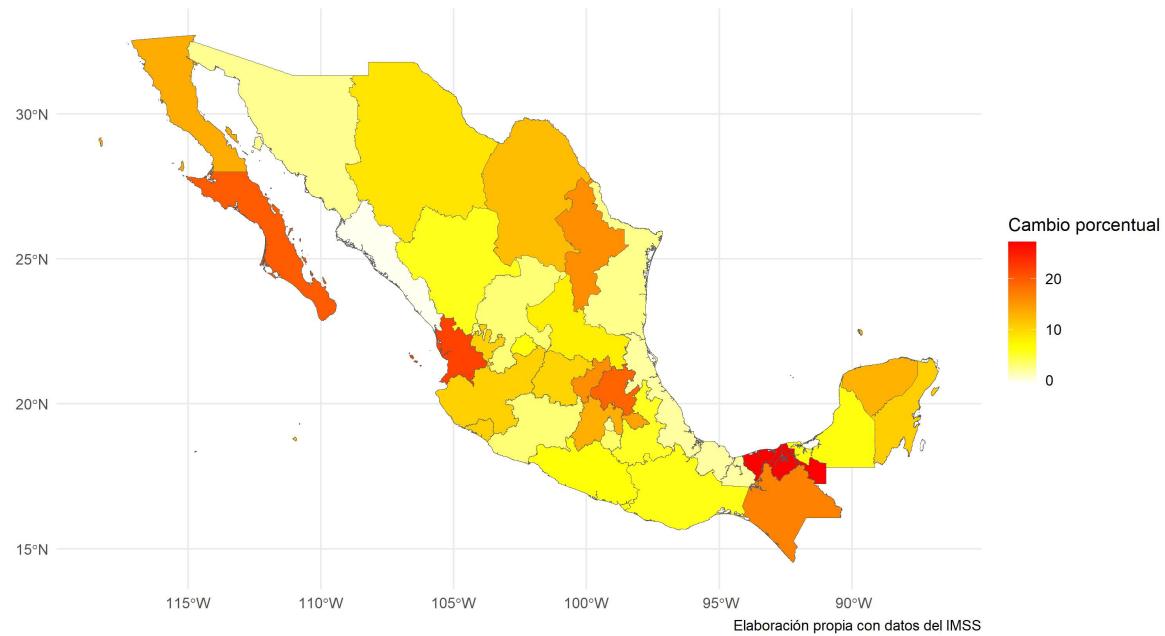
Fuente: Elaboración propia con datos del IMSS.

Empleo. R:

Variación porcentual en el número de trabajadores por entidad
Febrero 2020 - Julio 2020

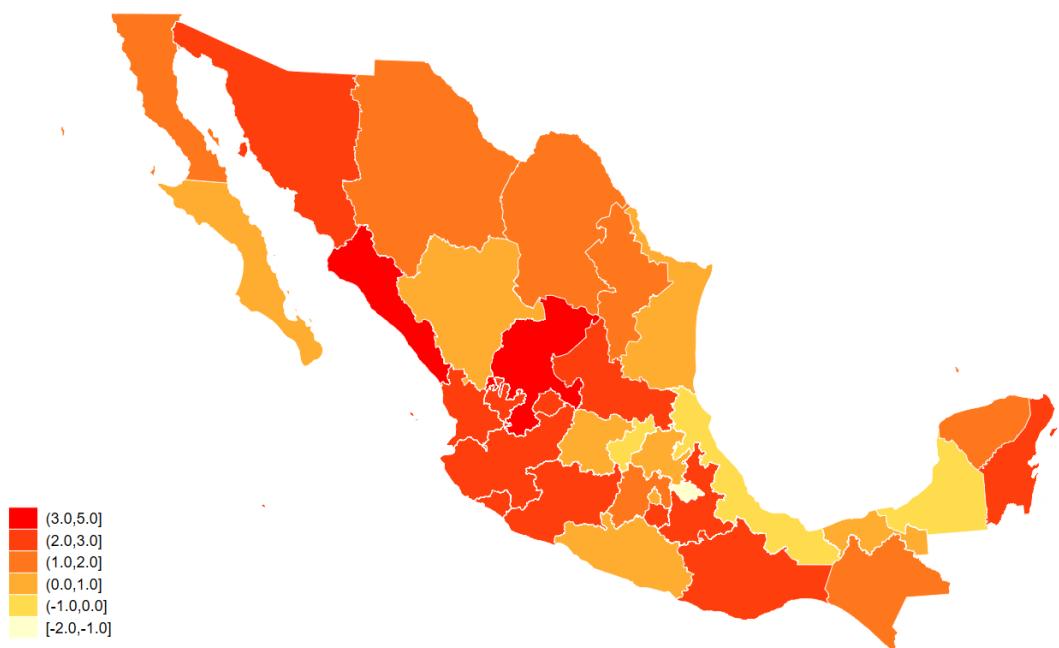


Variación porcentual en el número de trabajadores por entidad
Febrero 2020 - Julio 2024



Salarios. Stata:

Variación porcentual en los salarios promedio por entidad: Julio 2020 vs Febrero 2020



Fuente: Elaboración propia con datos del IMSS.

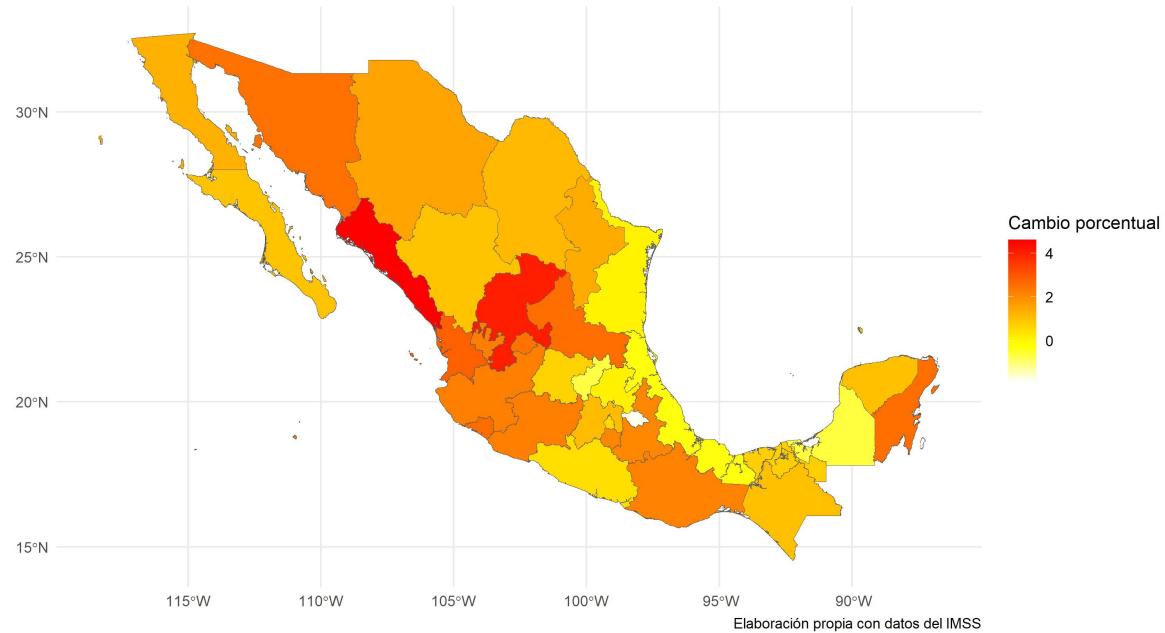
Variación porcentual en los salarios promedio por entidad: Julio 2024 vs Febrero 2020



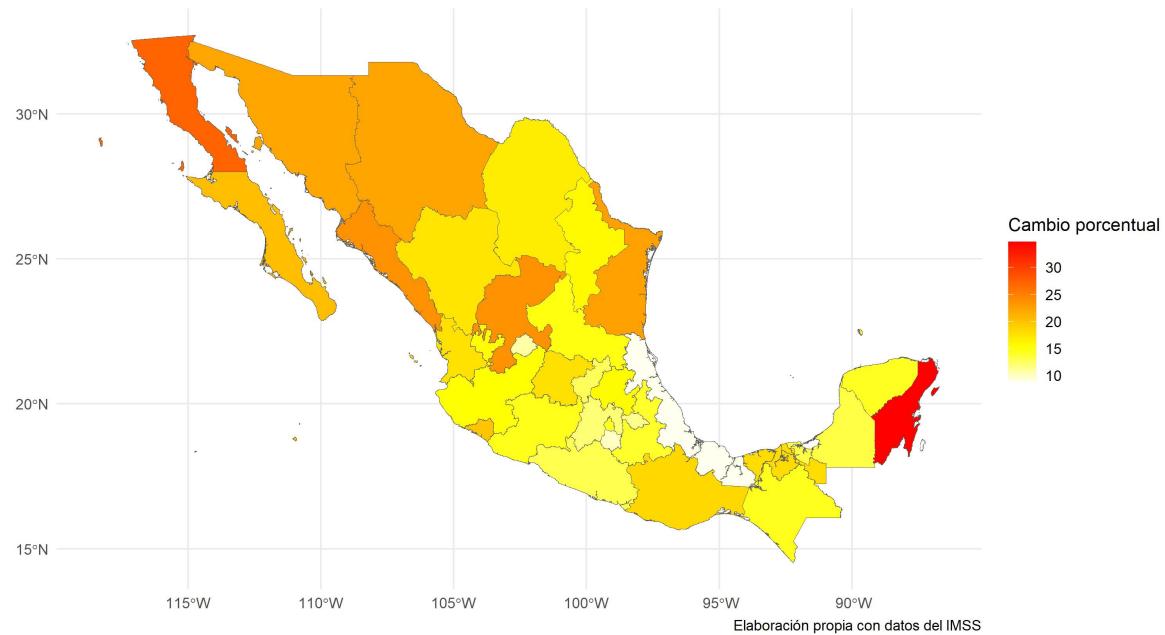
Fuente: Elaboración propia con datos del IMSS.

Salarios. R:

Variación porcentual en los salarios promedio por entidad
Febrero 2020 - Julio 2020



Variación porcentual en los salarios promedio por entidad
Febrero 2020 - Julio 2024

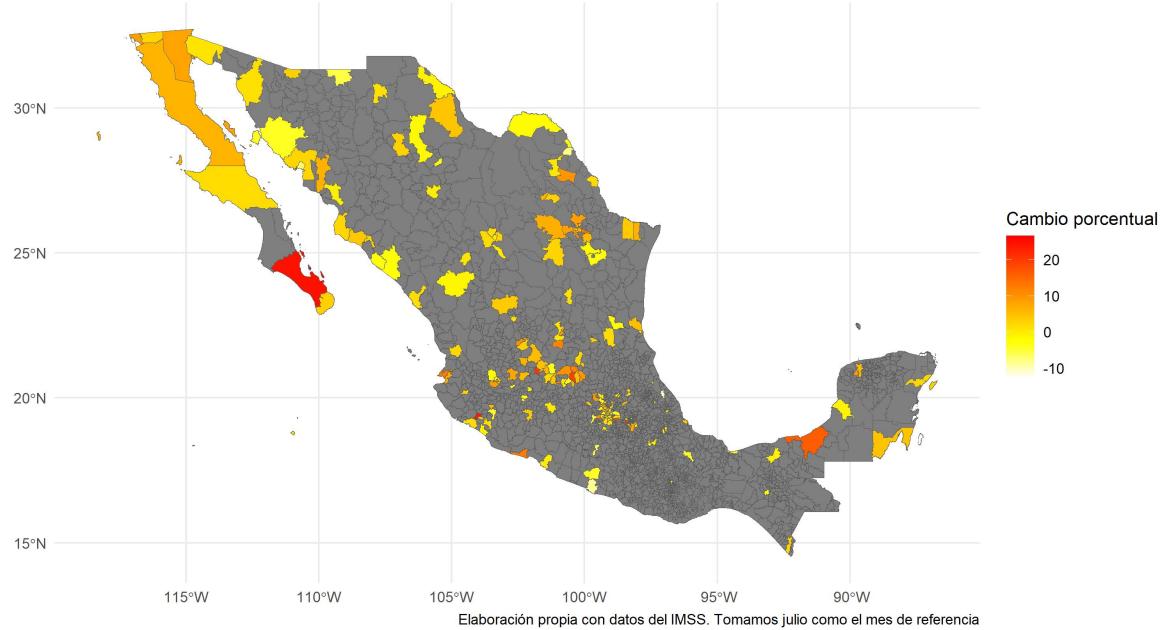


10. Realiza un mapa en R a nivel municipal sobre los cambios en empleo y en salario. Restringe la muestra a aquellos municipios con al menos 10,000 empleados en el año base. Para poder hacer este ejercicio tendrás que hacer un `merge` con la base de municipios de código INEGI, ver el diccionario que presenta el IMSS. Usa año base 2018, y el año final 2019 y 2024. Queremos saber la distribución de cambios en empleo y salario en ese periodo.

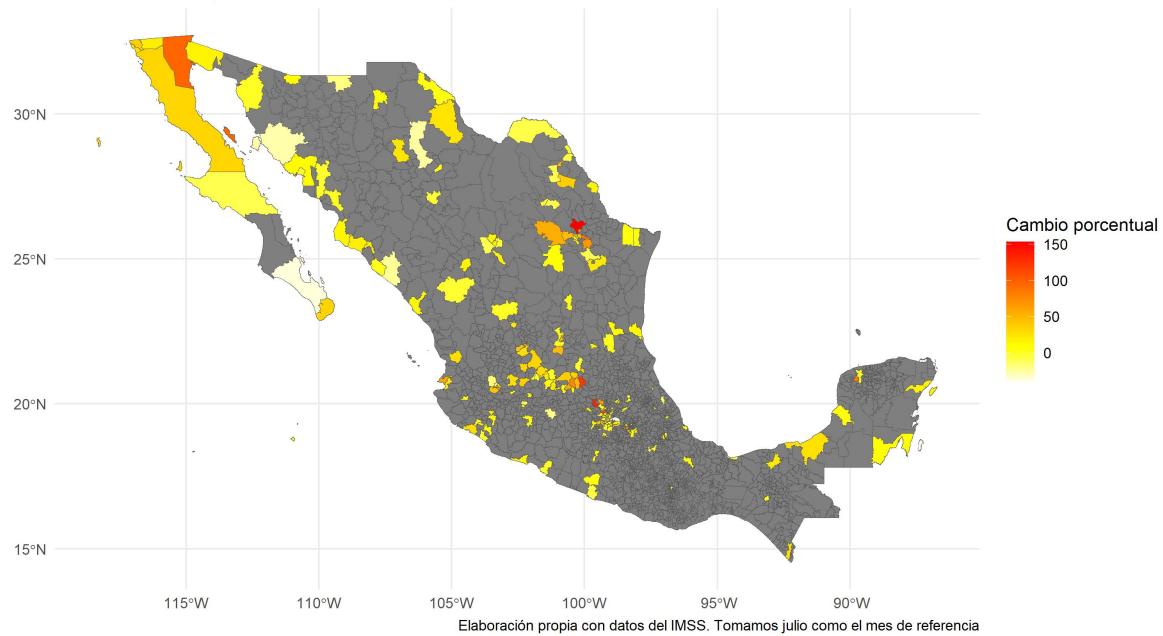
Mapas en las siguientes páginas.

Empleo

Variación porcentual en el número de trabajadores por municipio
2018 - 2019. Municipios con al menos 10,000 empleados en 2018.

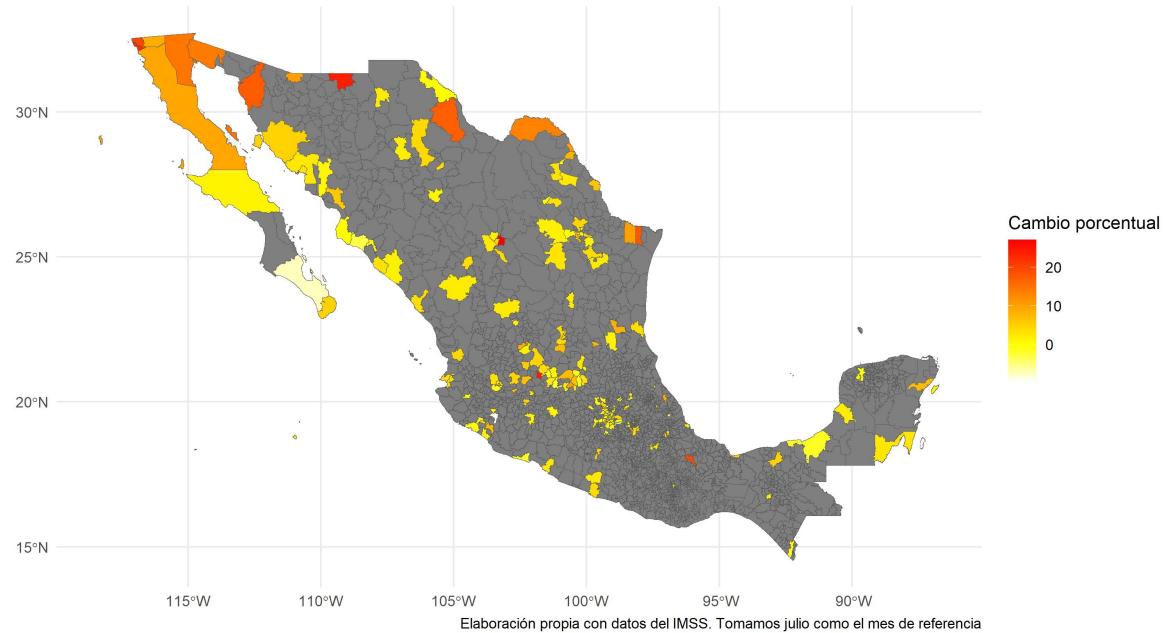


Variación porcentual en el número de trabajadores por municipio
2018 - 2024. Municipios con al menos 10,000 empleados en 2018.

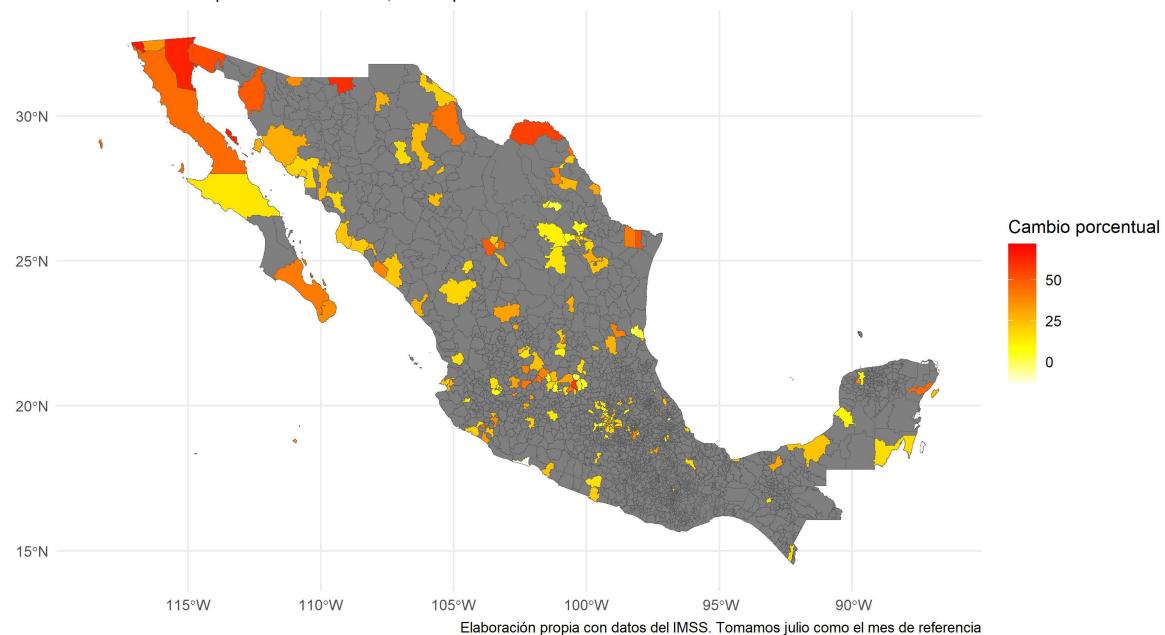


Salarios

Variación porcentual en los salarios promedio por municipio
2018 - 2019. Municipios con al menos 10,000 empleados en 2018.



Variación porcentual en los salarios promedio por municipio
2018 - 2024. Municipios con al menos 10,000 empleados en 2018.



11. Con base en tu análisis anterior, ¿qué se podría decir sobre el cambio del salario mínimo en la zona de la frontera norte? Describe los cambios al respecto.

Respuesta: Los incrementos al salario mínimo que tuvieron comienzo alrededor de 2019, no parecen haber tenido un efecto negativo sobre el empleo de los municipios de la frontera norte. De hecho, existieron incrementos importantes en el número de trabajadores y por supuesto, en el nivel salarial (efecto faro). A pesar de que el análisis visual no es suficiente para establecer una causalidad entre aumentos al salario mínimo y cambios en el empleo, es evidencia previa de que la relación negativa entre salarios y empleo no es estricta al menos en el caso de México en los últimos años. Se requerirá un análisis más completo para evaluar los datos y establecer causalidad.

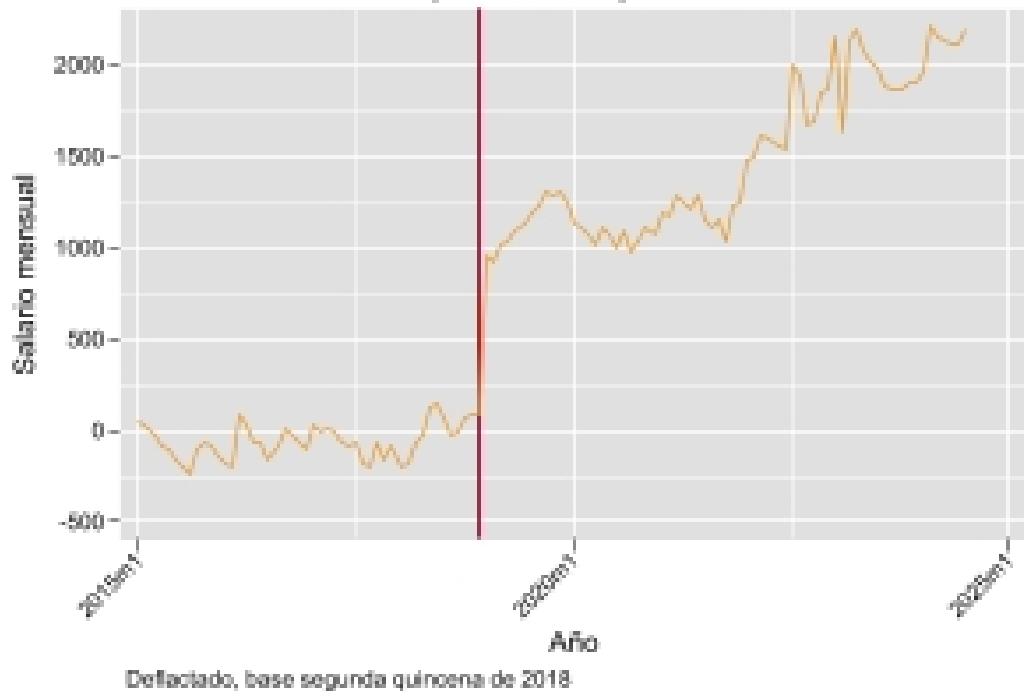
12. **Control sintético:** Efecto del programa de salario mínimo y reducción de IVA en la ZLFN a nivel entidad (programa empieza en 2019). Quédate con meses de enero 2015 a julio 2024. Realiza control sintético a nivel entidad. Es decir, tendrás 1 observación tratada que es la ZLFN (todos los municipios que la conforman), y 31 observaciones a nivel entidad federativa (realizar un `collapse` sum de empleo y de masa salarial y después se calcula el promedio). Tendrás por ejemplo Tamaulipas sin incluir la ZLFN como control potencial, BC no entra como control porque todos los municipios son parte de la ZLFN. Se pide:

- Estimar control sintético para 10 modelos (los mejores en RMSPE), discute cuál se eligió.
- Estimar valor p de cada modelo para empleo y salario promedio.
- Discutir similitudes o diferencias al artículo sobre efectos del salario mínimo en la ZLFN de Campos-Vazquez, Delgado y Rodas (2020). Enlace al artículo.

A continuación, los resultados de los modelos de salario y empleo.

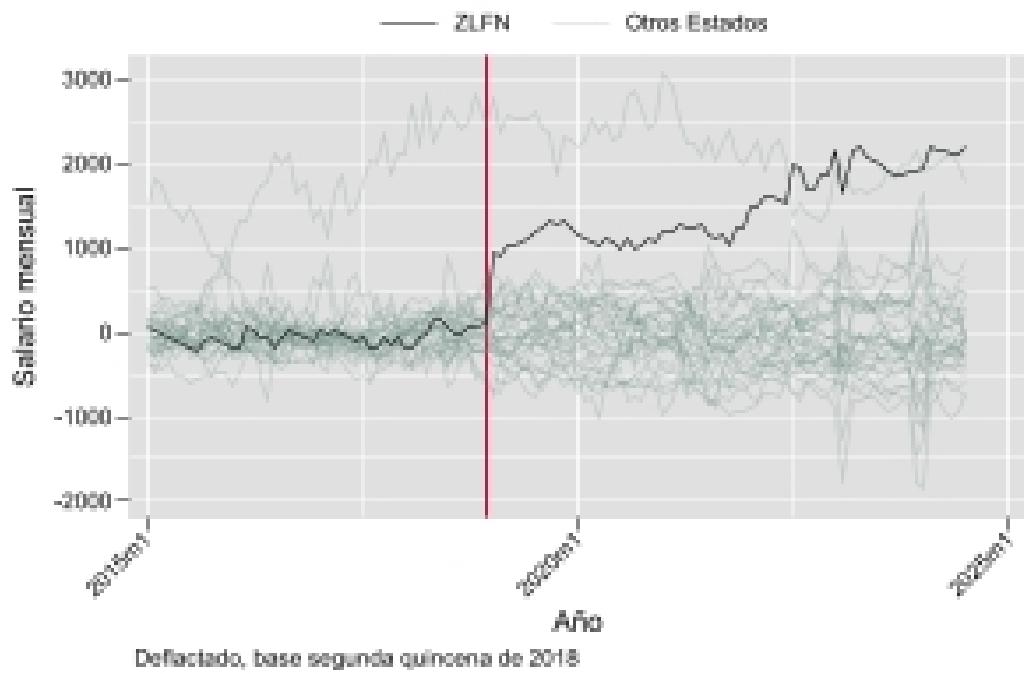
Efecto en la ZLFN, salario mensual

Grupo de control por Estado

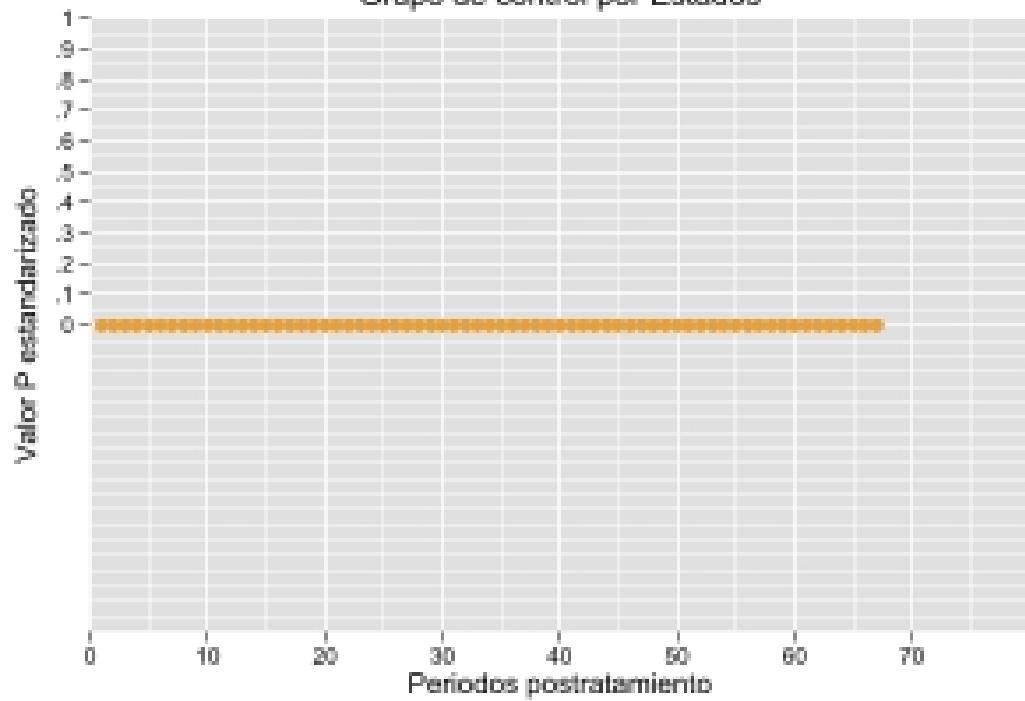


Diferencia entre placebos, salario promedio

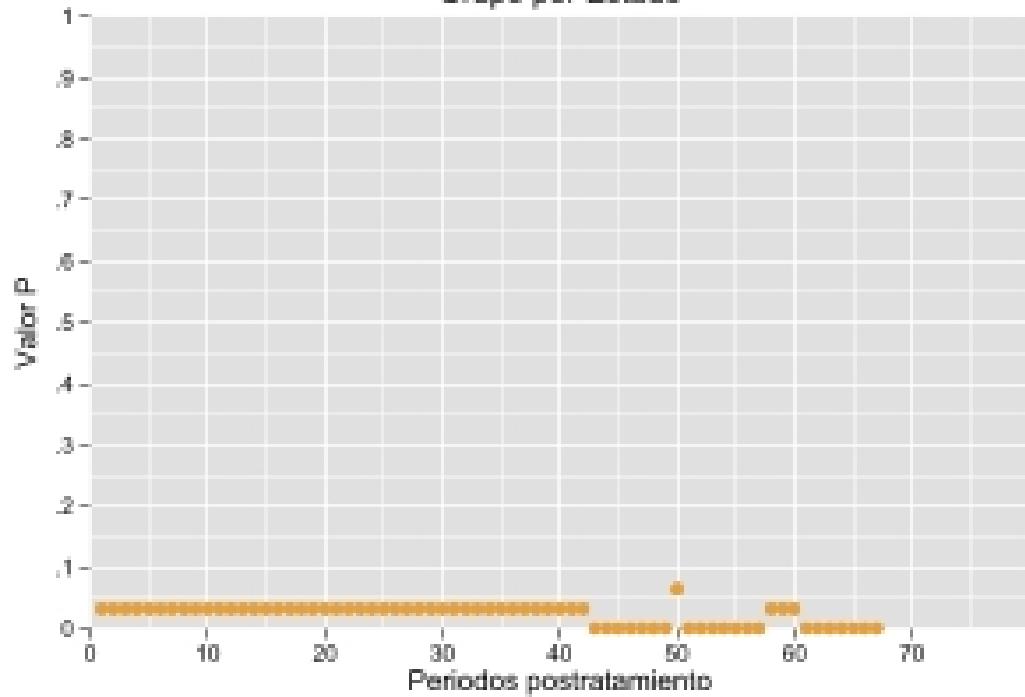
Grupo de control por Estado

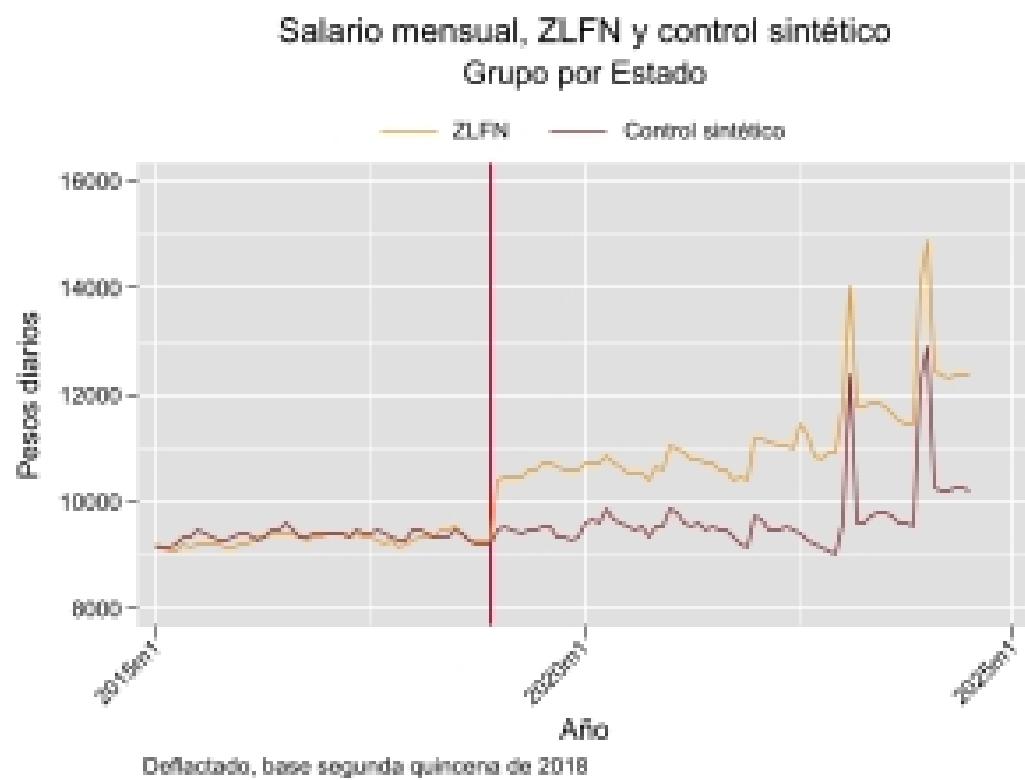
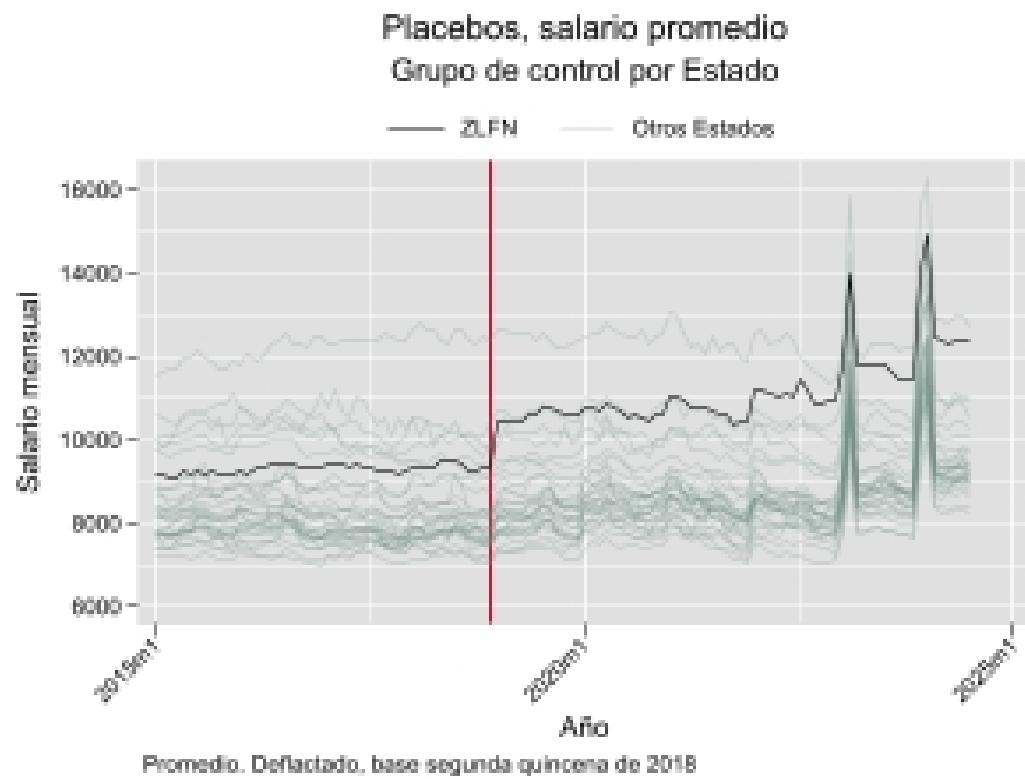


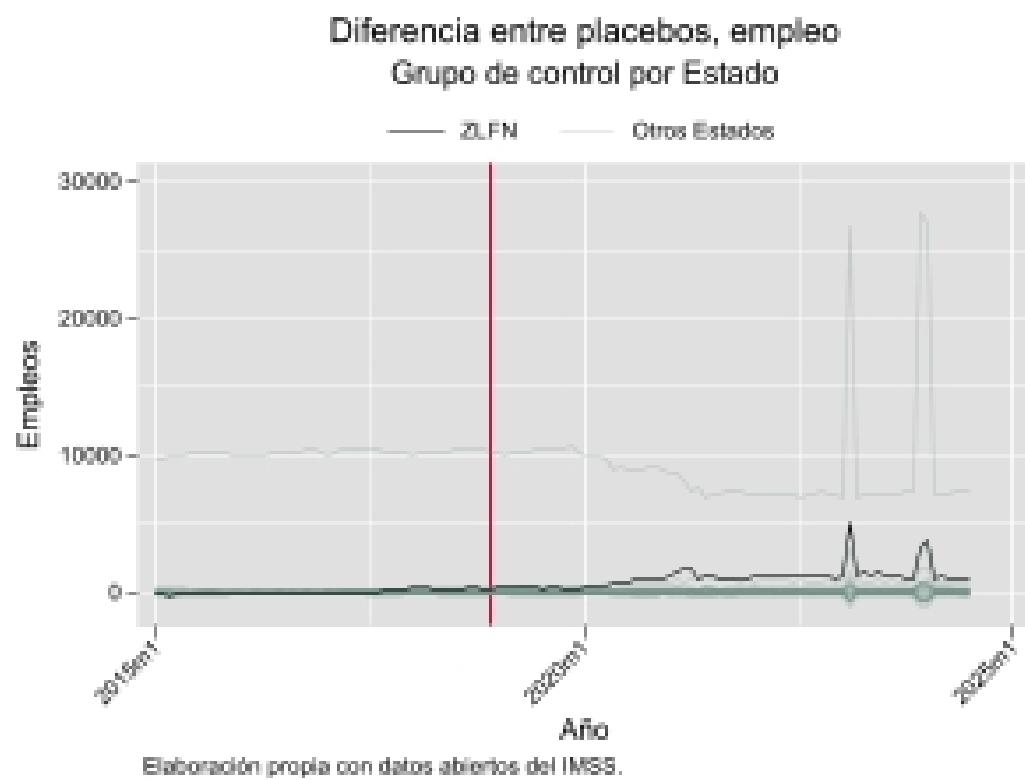
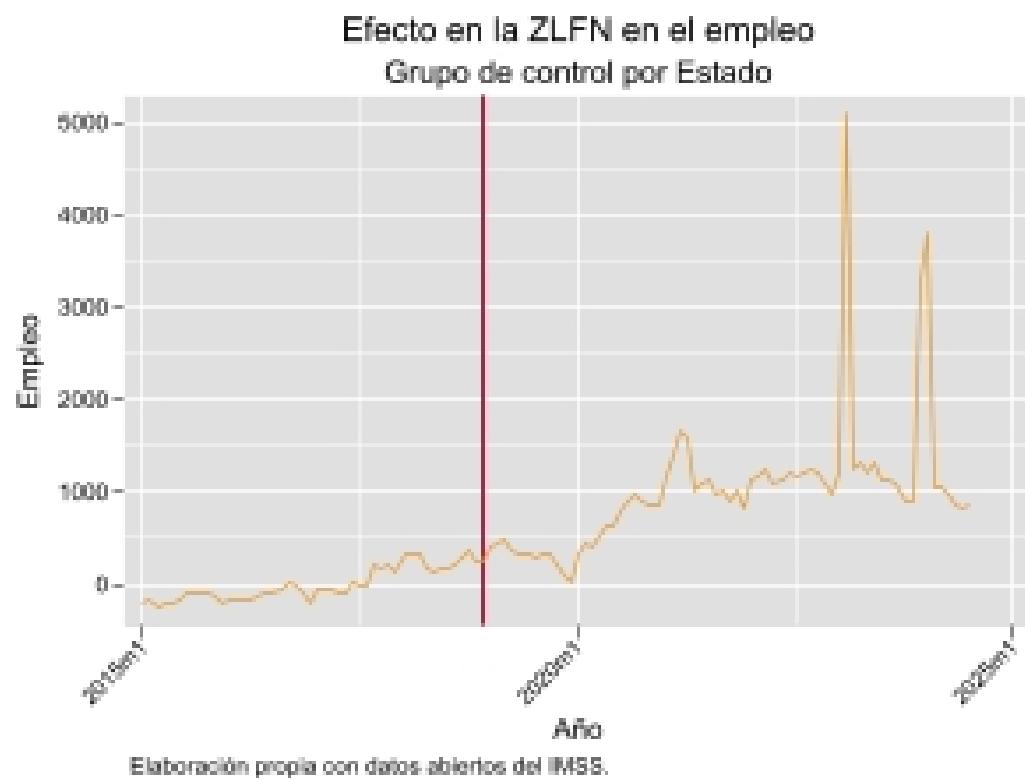
Valores P estandarizados
Grupo de control por Estados



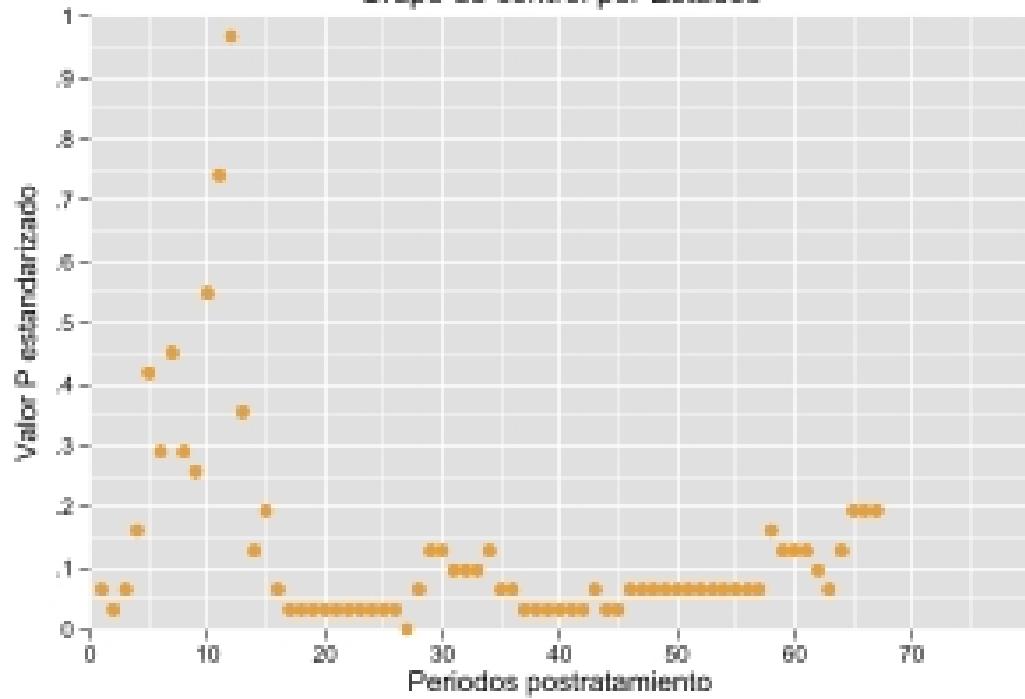
Valores P
Grupo por Estado



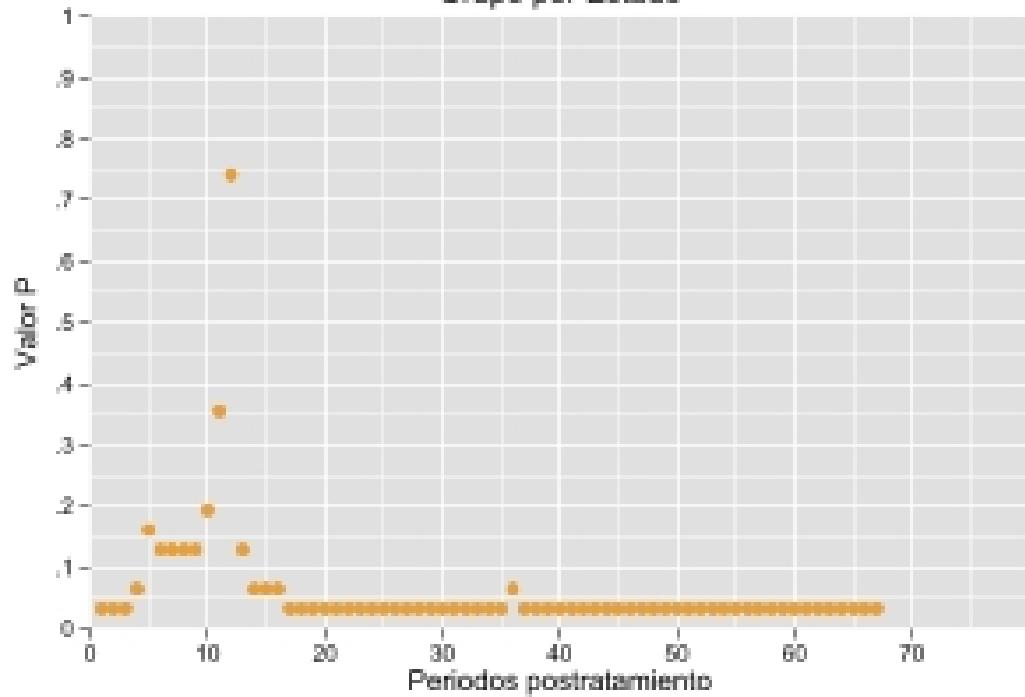




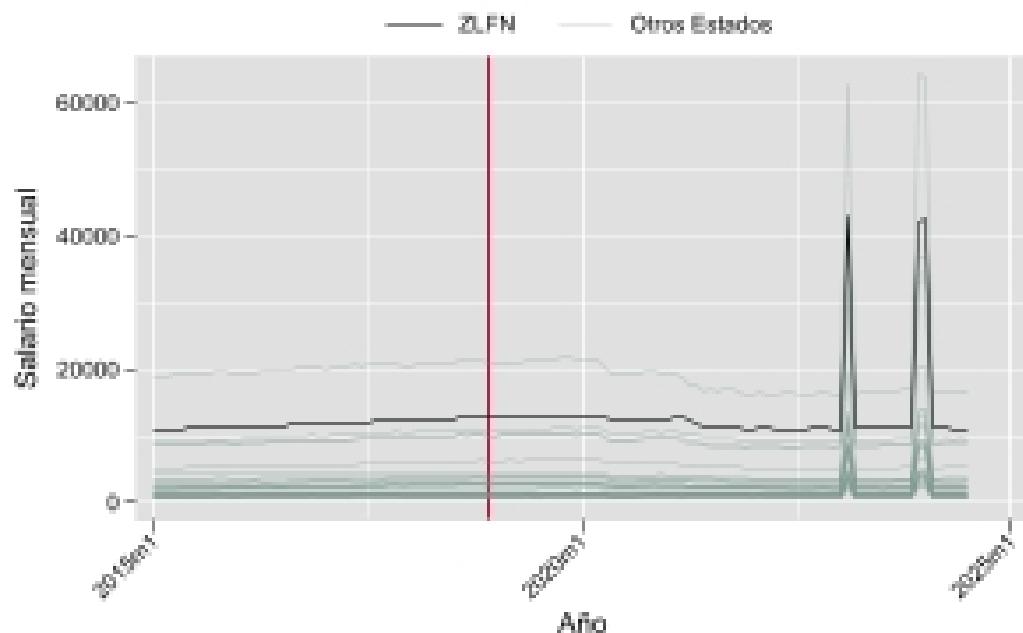
Valores P estandarizados
Grupo de control por Estados



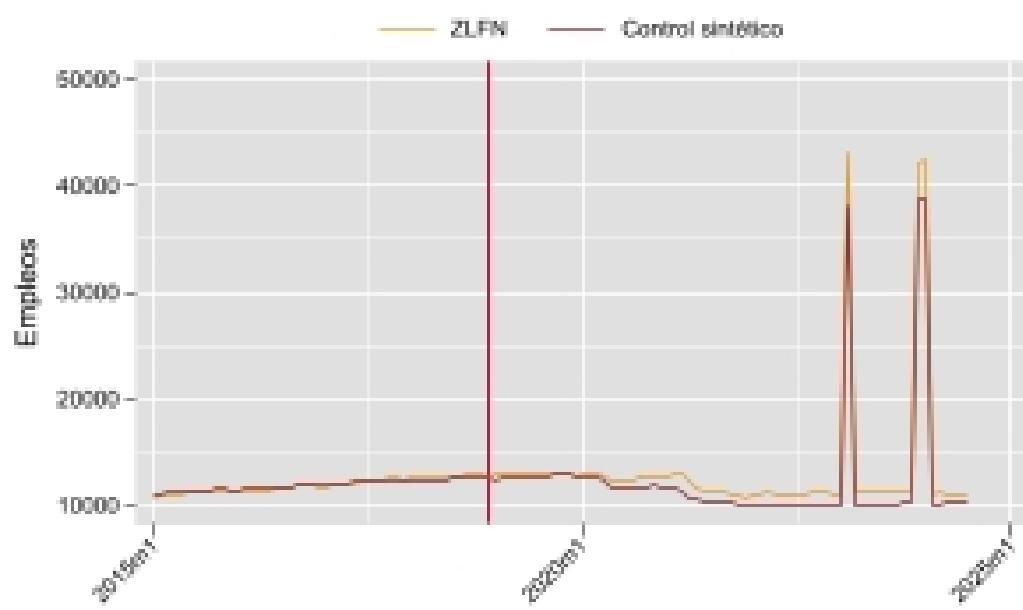
Valores P
Grupo por Estado



Placebos, cambio en el empleo Grupo de control por Estado



Empleos ZLFN y control sintético Grupo por Estado



3 Problema 3: Regresión Cuantil

Quantile Regression: `qreg`.

1. Explica el método de *quantile regression* y para qué sirve principalmente.

Respuesta: El modelo de regresión de cuantiles busca estimar los efectos condicionales de las variables independientes en diferentes cuantiles de la distribución de la variable dependiente, lo que nos permite analizar la heterogeneidad en la distribución.

En contraste, MCO estima el efecto promedio de X sobre Y al calcular la esperanza condicional de Y en X : $E[Y|X] = \beta X$. Sin embargo, esto puede omitir variaciones importantes en los estimadores β que tienen lugar cuando se consideran distintos niveles de la distribución del vector (Y, X) .

2. ¿Cómo se interpretan los parámetros de regresión cuantil? Explica.

Respuesta: Los estimadores de la regresión cuantil tienen una interpretación similar los de MCO, pero en lugar de describir el efecto de las variables explicativas sobre y , describe su efecto en el cuantil τ -ésimo de y . Esto significa que $\hat{\beta}_\tau$ nos dice cómo las variables X influyen en el cuantil τ de la distribución de y .

Para un cuantil τ -ésimo dado (por ejemplo, $\tau = 0.25$ o $\tau = 0.75$), el parámetro $\hat{\beta}_\tau$ en el modelo de regresión cuantil:

$$Q_\tau(Y|X) = X\hat{\beta}_\tau$$

es el cambio en el cuantil τ -ésimo de la variable dependiente Y , condicional en X , debido a un cambio unitario en la variable independiente X . Es decir, mide el **efecto marginal de X sobre el cuantil τ -ésimo** de la distribución de Y , en lugar de sobre su media, como ocurre en OLS.

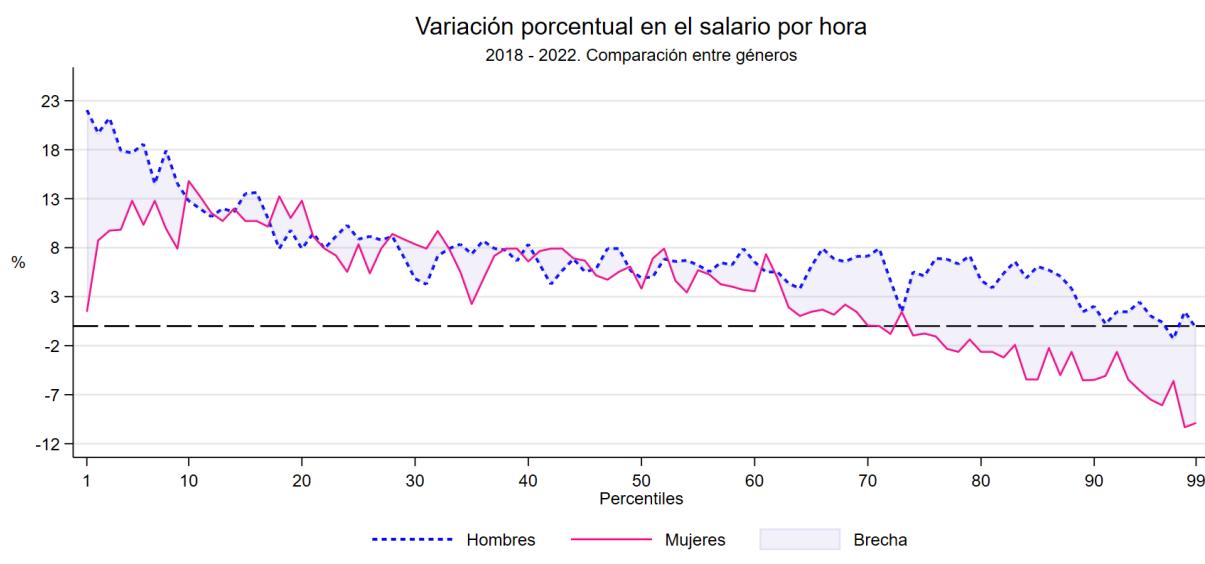
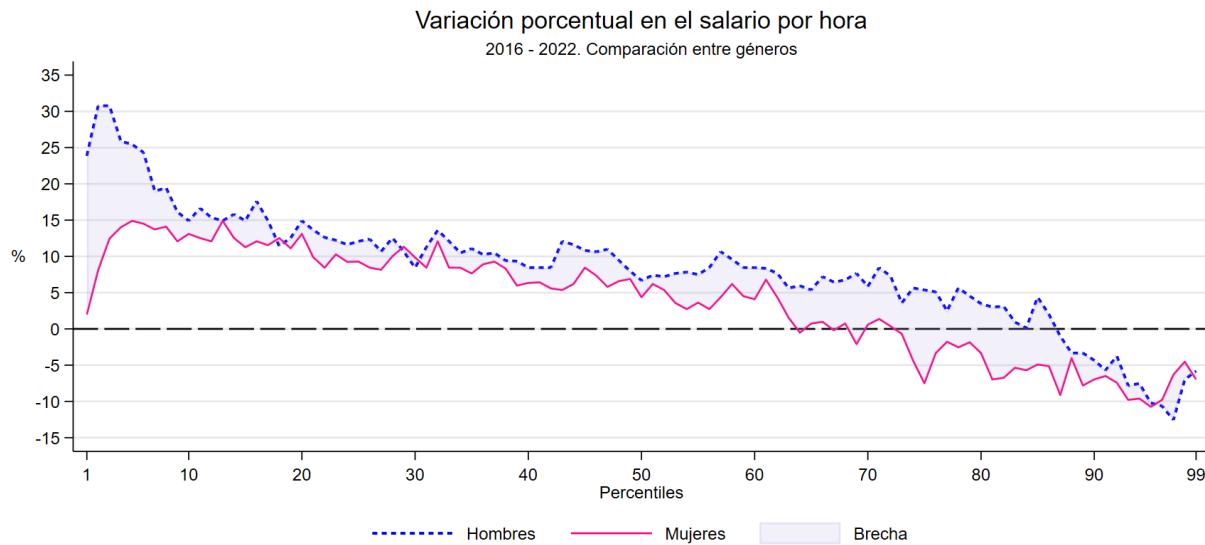
A manera de ejemplo, consideremos un modelo que estima los salarios (Y) en función de los años de educación (X) usando regresión cuantil. Supongamos que estamos interesados en el cuantil $\tau = 0.50$ (el percentil 50). Si el parámetro $\hat{\beta}_{0.50}$ para la variable "años de educación" es 1,500, entonces:

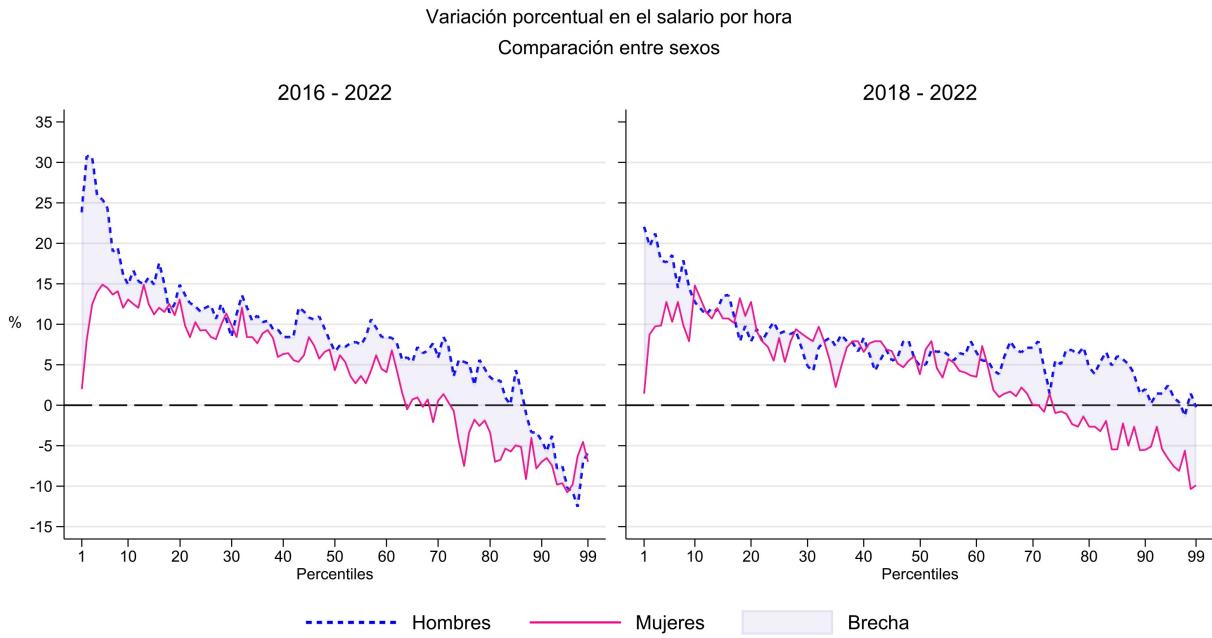
Un año adicional de educación está asociado con un aumento de 1,500 unidades en el salario para las personas en el percentil 50 de la distribución salarial, manteniendo todo lo demás constante.

3. Utilizando la ENIGH, estima el cambio del log salario por hora para cada cuantil entre 2016, 2018 y 2022 para hombres y para mujeres (cuantiles: 1,...,99). Con esto, crea 2 gráficas (eje x los cuantiles, y eje y el cambio del log salario por hora) y ponlas en el mismo renglón.,

es decir cambio entre 2016-2022 y 2018-2022. ¿Cómo interpretas tus resultados? ¿Bajó la desigualdad? ¿Implicaciones para la distribución de salarios de hombres y mujeres?

Siguiente página: Presento las dos gráficas con la variación porcentual en los salarios por sexo, para los dos períodos mencionados, en una sola columna. En la página que le sigue, las muestro en el mismo renglón; con ello es posible visualizar mejor los resultados.





Elaboración propia con datos de la ENIGH.

En los **percentiles inferiores** de la distribución (1 - 10) se observa una brecha relativamente alta en las tasas de crecimiento del salario por hora entre hombres y mujeres, lo que podría ser un claro indicativo de que la brecha en los salarios medios se ha incrementado en ambos períodos. Por otro lado, para los **percentiles de la parte media** de la distribución (11 - 60), la brecha entre las tasas parece ser menor que en el caso anterior, aunque aún positiva, si observamos únicamente el periodo completo (2016 a 2022); no obstante, para esos mismos percentiles pero en el periodo corto (2018 a 2022), la brecha en tasas de crecimiento es positiva para algunos percentiles y negativa para otros, lo que pudiera indicar que la brecha en salarios por hora medios no aumentó entre esos años pero sí en el periodo completo. Finalmente, para la **parte alta de la distribución** (percentiles 61 a 99), si analizamos el periodo completo, observamos que, aunque las tasas de crecimiento del salario por hora han sido positivas para los hombres entre los percentiles 61 y 87, éste no ha sido el caso para las mujeres, quienes han observado una caída en los salarios en todos los percentiles desde el 61 al 99. El resultado de esto podría ser un incremento en la desigualdad salarial en toda la parte alta de la distribución, pues aunque algunos hombres de ciertos percentiles han experimentado caídas en sus salarios, estas caídas no son tan pronunciadas como en el caso de las mujeres. Es importante notar que la mayor parte de este efecto tuvo lugar entre los años 2018 y 2022, donde los hombres tuvieron crecimientos salariales, a la par que los salarios femeninos mostraban caídas en aquellos percentiles.

4. Utilizando la ENIGH, estima la regresión cuantil de log de salario por hora en función de años de escolaridad, edad, edad², dummy de rural y dummy de female para los cuantiles: 1,2,...,99 para los años de 2016, 2018 y 2022 por separado. También estima las regresiones usando OLS.

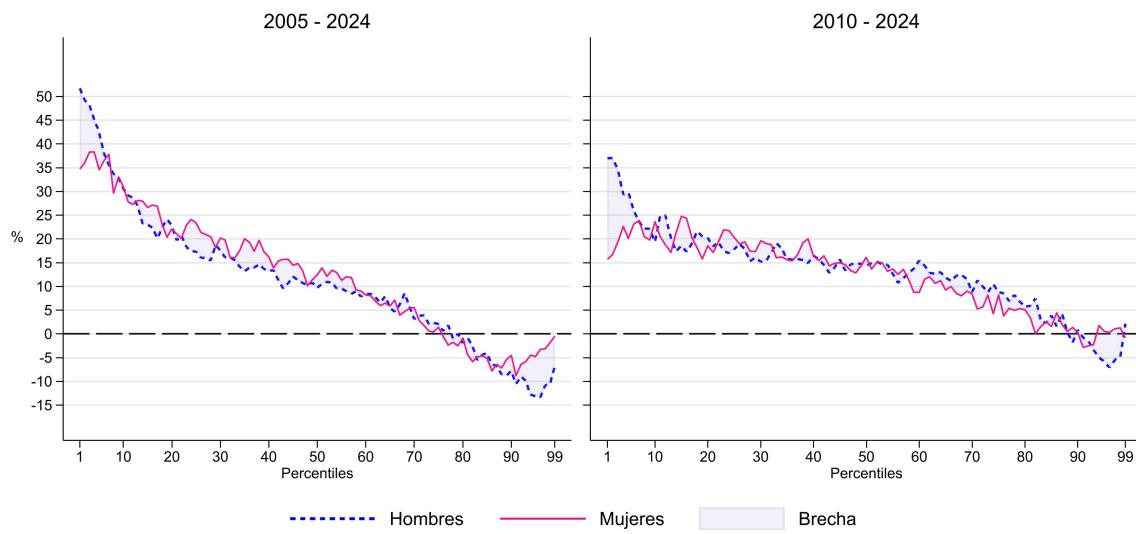
Respuesta: Se realizó en Stata.

5. Realiza los incisos 3 y 4 para la ENOE con los años 2005, 2010, 2018 y 2024, utiliza únicamente los trimestres I y II (conjuntamente).

Siguiente página

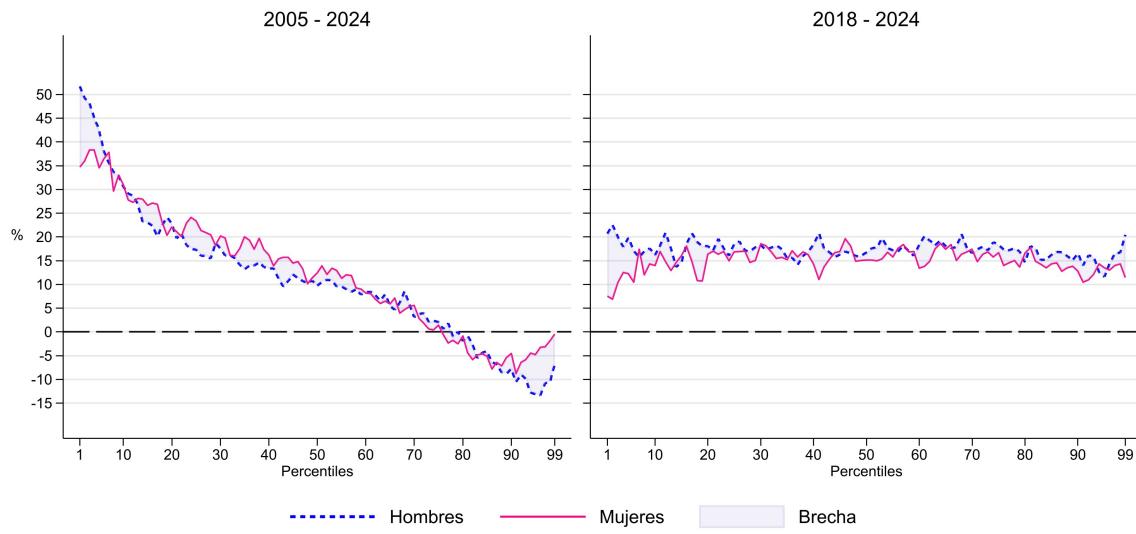
ENOE:

Variación porcentual en el salario por hora
Comparación entre sexos



Fuente: Elaboración propia con datos de la ENOE. Se considera el primer semestre de cada año.

Variación porcentual en el salario por hora
Comparación entre sexos



Fuente: Elaboración propia con datos de la ENOE. Se considera el primer semestre de cada año.

6. Grafica los coeficientes de esas regresiones para la variable escolaridad y género, incluyendo su intervalo de confianza al 95% y el coeficiente obtenido por OLS. En cada renglón pon dos gráficas por variable, del lado izquierdo para 2016 y del lado derecho para 2022 para la ENIGH, y para la ENOE utiliza 2018 vs 2024 (trimestres I y II únicamente). Interpreta y explica tus resultados. Explica las implicaciones de estos resultados para tu respuesta en el inciso 3.

Respuesta: Las gráficas de la siguiente página muestran el efecto estimado que las variables de escolaridad y género tienen sobre el salario por hora. En el caso de los resultados de la **ENIGH**, tenemos lo siguiente:

- **Escolaridad.** Los retornos a la educación son menores en 2022 que en 2016 para todos los percentiles. Además, en ambos períodos podemos observar que algunos percentiles de la parte inferior (los más bajos) y aquellos por encima del decil 6, gozan de un retorno mayor que el promedio (el estimador de MCO) por cada año de educación, lo cual no ocurre con los percentiles ubicados debajo de este decil, con excepción de los más bajos. Es decir, la parte baja y media de los trabajadores tienen en general retornos educativos menores que la parte alta de la distribución, lo que podría acrecentar la desigualdad.
- **Género.** Los estimadores de regresión cuantil y de MCO representan la brecha de género. Es decir, en ambos períodos las gráficas muestran que ser mujer trabajadora implica un salario menor que el que reciben sus pares masculinos, en alrededor de 10 y 15%; este valor es similar en ambos períodos, así como su magnitud a lo largo de la distribución. En cuanto a ello, los resultados muestran que la brecha de género es superior en la parte más baja de la distribución, mientras que para la parte media y alta, con algunas excepciones, la brecha es inferior al promedio (aunque aún existe). Ahora bien, esto no es inconsistente con lo hallado en el inciso 3, donde vimos que la brecha en tasas de crecimiento es alta en la parte baja, lo que podría explicar las mayores brechas en el salario por hora en esos mismos percentiles y su profundización entre 2016 y 2022. Aunado a esto, vimos en el inciso 3 que la brecha en tasas de crecimiento para la parte media y alta eran positivas en ambos casos, aunque mayor en la parte alta que en la media. Esto se refleja en brechas más altas entre 2016 y 2022 para ambos grupos.

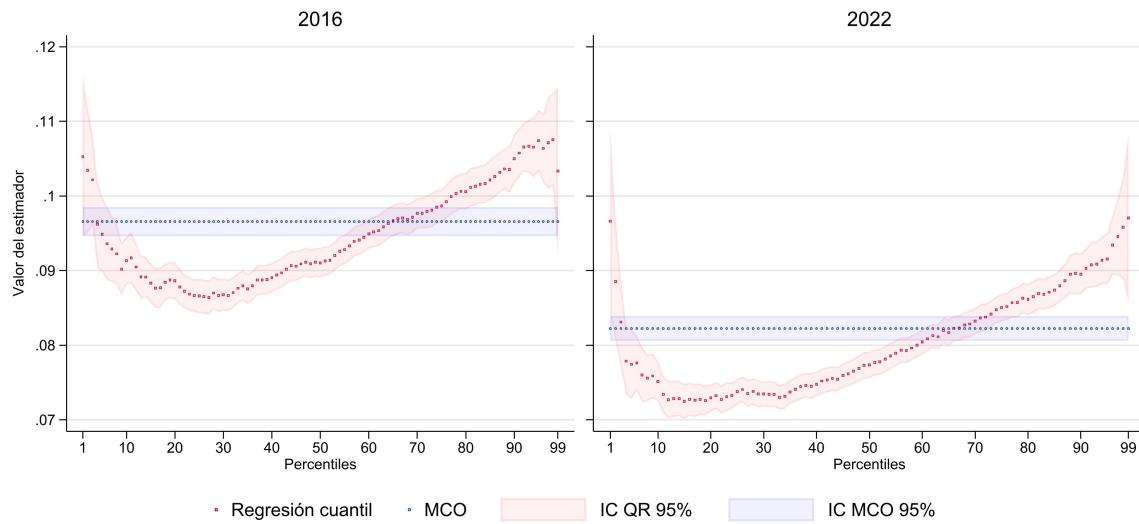
Por su parte, la **ENOE**, que aunque muestra magnitudes ligeramente menores en comparación con la ENIGH, revela un escenario parecido en el caso de la educación, con excepción quizás de que los percentiles más bajos ya no gozan de un retorno educativo mayor al promedio. Es interesante también que la ENOE no indica una reducción en los retornos educativos en toda la distribución como sí apunta lo hallado en la ENIGH; esto nos dice que, entre 2022 y 2024, los retornos se recuperaron a los niveles prepandemia de 2018, aunque un análisis mayor es necesario. En el caso del género, los resultados son muy similares, aunque la distancia entre la brecha en la parte alta y la media es mayor en la ENOE que en la

ENIGH. Todo lo cual resulta consistente con el inciso 3. A continuación, presento las figuras elaboradas.

ENIGH:

Efecto de la educación sobre el salario por percentiles

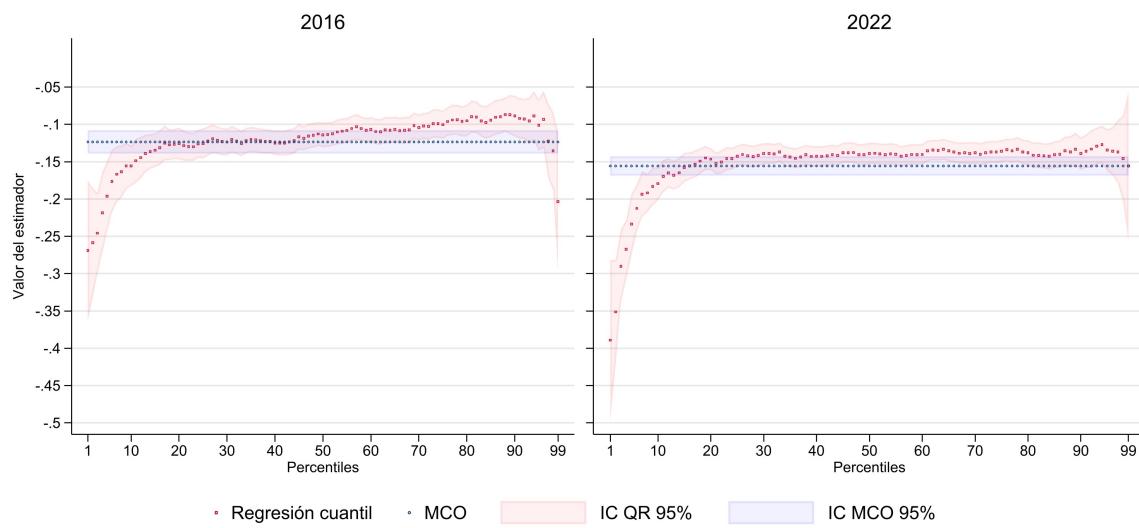
Estimadores de regresión cuantil (QR) vs MCO



Elaboración propia con datos de la ENIGH.

Brecha de género en el salario por percentiles

Estimadores de regresión cuantil (QR) vs MCO

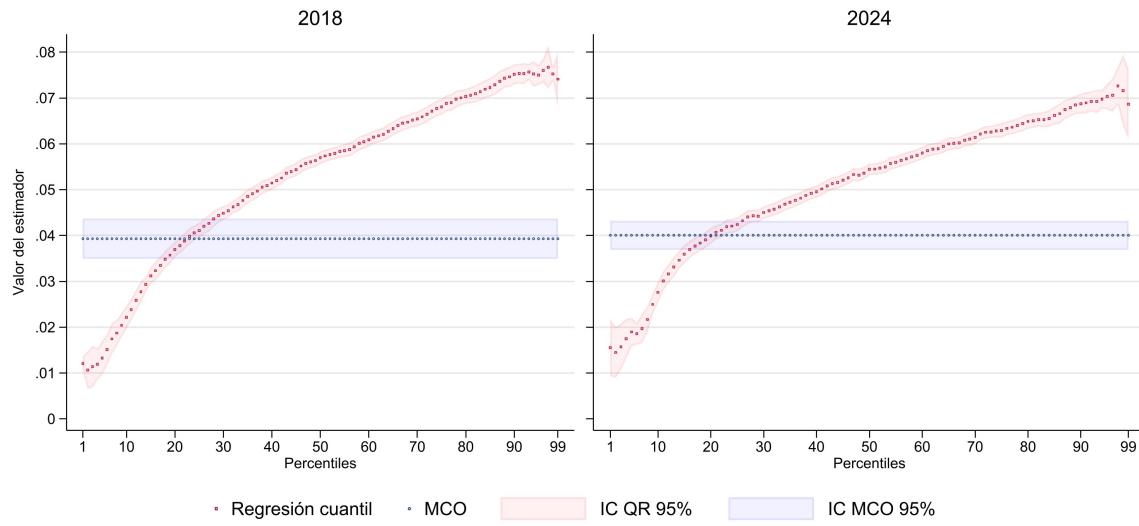


Elaboración propia con datos de la ENIGH.

ENOE:

Efecto de la educación sobre el salario por percentiles

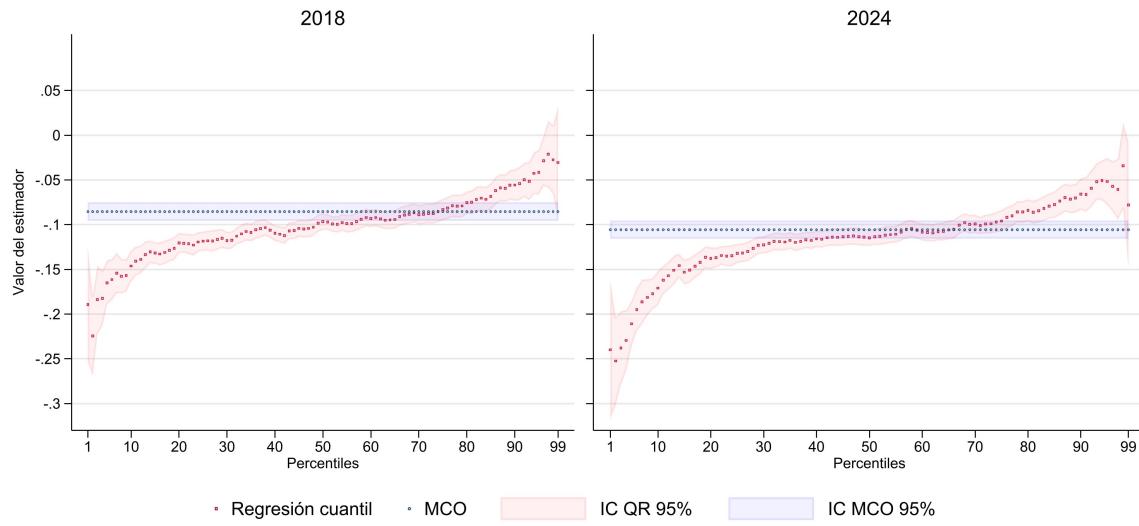
Estimadores de regresión cuantil (QR) vs MCO



Elaboración propia con datos de la ENOE. Primer semestre de cada año.

Brecha de género en el salario por percentiles

Estimadores de regresión cuantil (QR) vs MCO



Elaboración propia con datos de la ENOE. Primer semestre de cada año.

4 Problema 4: Bootstrap

Bootstrap methods

1. Utiliza el comando `bootstrap` para obtener errores estándar de una regresión simple de la ENIGH:
 - (a) Estima la regresión de log de salario por hora en función de años de escolaridad, edad, edad², `dummy` de rural y `dummy` de female con errores estándar robustos para el año 2022.
 - (b) Bootstrap no paramétrico con 100 repeticiones. Organiza en una tabla tus resultados: error estándar robusto de OLS, error estándar del bootstrap, el intervalo de confianza del estimador de bootstrap con el método del percentil, el método percentil-t al 95% de confianza y compáralo con la t de OLS para rechazar o no la hipótesis nula de un efecto cero.

Tabla 1. La primera columna presenta los resultados de la regresión por MCO descrita en el inciso a). Los valores son:

- i. Valor del estimador $\hat{\beta}$.
- ii. Error estándar robusto.
- iii. Intervalo de confianza del estimador al 95%.
- iv. Estadístico t -student.

La segunda y tercera columna presentan los resultados del bootstrap. Para la columna dos tenemos los siguientes valores:

- i. Valor del estimador $\hat{\beta}$ con bootstrap.
- ii. Error estándar robusto con bootstrap.
- iii. Intervalo de confianza del estimador al 95%.
- iv. Estadístico z .

La columna tres muestra el error estándar robusto, que difiere del mostrado en la columna uno, debido a que no utiliza factores de expansión. Además de su valor, muestra también:

- i. Valor del error estándar robusto
- ii. Error estándar robusto del error estándar robusto.
- iii. Intervalo de confianza del error al 95%.
- iv. Estadístico z

Tabla 1: Resultados de MCO y del bootstrap no paramétrico de 100 repeticiones.

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ BS	(3) $\hat{\sigma}_{Robusto}$ BS
years_ed	0.0823*** (0.000832)	0.0756*** (0.000592)	0.000611*** (4.13e-06)
	0.0806 - 0.0839 (98.81)	0.0744 - 0.0767 (127.7)	0.000602 - 0.000619 (147.8)
edad	0.0364*** (0.00242)	0.0333*** (0.00189)	0.00177*** (1.28e-05)
	0.0316 - 0.0411 (15.01)	0.0296 - 0.0370 (17.60)	0.00174 - 0.00179 (138.3)
edad2	-0.000342*** (2.85e-05)	-0.000319*** (2.26e-05)	2.08e-05*** (1.70e-07)
	-0.000398 - -0.000286 (-12.01)	-0.000363 - -0.000275 (-14.15)	2.05e-05 - 2.11e-05 (122.5)
rural	-0.245*** (0.00715)	-0.138*** (0.00567)	0.00524*** (2.76e-05)
	-0.259 - -0.231 (-34.22)	-0.149 - -0.127 (-24.34)	0.00518 - 0.00529 (190.0)
female	-0.156*** (0.00647)	-0.157*** (0.00476)	0.00471*** (2.03e-05)
	-0.168 - -0.143 (-24.06)	-0.166 - -0.148 (-33.00)	0.00467 - 0.00475 (232.0)
constante	1.820*** (0.0506)	1.988*** (0.0373)	0.0368*** (0.000239)
	1.721 - 1.919 (35.96)	1.915 - 2.061 (53.24)	0.0363 - 0.0373 (153.7)
Observations	85,250	85,250	85,250
R-squared	0.260		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENIGH

Tabla 2: Resumen de estadísticas descriptivas de medias Bootstrap 100

	Bootstrap 100 mean
_b[years_ed]	.0756282
_b[edad]	.0334306
_b[edad2]	-.0003209
_b[rural]	-.1379643
_b[female]	-.1572827
_b[_cons]	1.983675
_se[years_ed]	.0006105
_se[edad]	.0017659
_se[edad2]	.0000208
_se[rural]	.0052339
_se[female]	.0047097
_se[_cons]	.0367774
hip	0
t_years_ed	123.8936
icl_years_ed_bt	.0745431
icu_years_ed_bt	.0768219
ictl_years_ed_bt	121.1264
ictu_years_ed_bt	125.4929
t_edad	18.93136
icl_edad_bt	.0302041
icu_edad_bt	.037039
ictl_edad_bt	17.20455
ictu_edad_bt	20.83804
t_edad2	-15.42407
icl_edad2_bt	-.0003781
icu_edad2_bt	-.0002839
ictl_edad2_bt	-17.97245
ictu_edad2_bt	-13.72167
t_rural	-26.36008
icl_rural_bt	-.1499227
icu_rural_bt	-.1274296
ictl_rural_bt	-28.57246
ictu_rural_bt	-24.26656
t_female	-33.39559
icl_female_bt	-.1675493
icu_female_bt	-.14897
ictl_female_bt	-35.4574
ictu_female_bt	-31.58925
Observations	100

Elaboración propia con datos de la ENIGH

- (c) Bootstrap no paramétrico con 1000 repeticiones. Organiza en una tabla tus resultados: error estándar robusto de OLS, error estándar del bootstrap, el intervalo de confianza del estimador de bootstrap con el método del percentil, el método percentil-t al 95% de confianza y compáralo con la t de OLS para rechazar o no la hipótesis nula de un efecto cero.

Resultados en las siguientes dos páginas.

Tabla 3: Resultados de MCO y del bootstrap no paramétrico de 1000 repeticiones.

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ BS	(3) $\hat{\sigma}_{Robusto}$ BS
years_ed	0.0823*** (0.000832)	0.0756*** (0.000622)	0.000611*** (4.07e-06)
	0.0806 - 0.0839 (98.81)	0.0744 - 0.0768 (121.6)	0.000603 - 0.000619 (149.9)
edad	0.0364*** (0.00242)	0.0333*** (0.00177)	0.00177*** (1.19e-05)
	0.0316 - 0.0411 (15.01)	0.0298 - 0.0367 (18.79)	0.00174 - 0.00179 (148.9)
edad2	-0.000342*** (2.85e-05)	-0.000319*** (2.09e-05)	2.08e-05*** (1.55e-07)
	-0.000398 - -0.000286 (-12.01)	-0.000360 - -0.000278 (-15.24)	2.05e-05 - 2.11e-05 (133.9)
rural	-0.245*** (0.00715)	-0.138*** (0.00518)	0.00524*** (3.06e-05)
	-0.259 - -0.231 (-34.22)	-0.148 - -0.128 (-26.69)	0.00518 - 0.00530 (171.0)
female	-0.156*** (0.00647)	-0.157*** (0.00471)	0.00471*** (2.11e-05)
	-0.168 - -0.143 (-24.06)	-0.166 - -0.148 (-33.33)	0.00467 - 0.00475 (223.5)
constante	1.820*** (0.0506)	1.988*** (0.0366)	0.0368*** (0.000235)
	1.721 - 1.919 (35.96)	1.916 - 2.060 (54.27)	0.0363 - 0.0373 (156.7)
Observations	85,250	85,250	85,250
R-squared	0.260		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENIGH

Tabla 4: Resumen de estadísticas descriptivas de medias Bootstrap 1000

	Bootstrap 1000
	mean
byears_ed	.0755928
edad	.0331805
edad2	-.0003182
brural	-.138333
bfemale	-.1568777
_b_cons	1.989628
se_years_ed	.0006106
se_edad	.0017668
se_edad2	.0000208
se_rural	.0052351
se_female	.0047086
_se_cons	.0368036
hip	0
t_years_ed	123.7982
icl_years_ed_bt	.0738211
icu_years_ed_bt	.0748079
ictl_years_ed_bt	120.3188
ictu_years_ed_bt	122.1927
t_edad	18.78029
icl_edad_bt	.0284882
icu_edad_bt	.0308301
ictl_edad_bt	16.192
ictu_edad_bt	17.49502
t_edad2	-15.2904
icl_edad2_bt	-.0003755
icu_edad2_bt	-.0003466
ictl_edad2_bt	-18.08636
ictu_edad2_bt	-16.6131
t_rural	-26.42436
icl_rural_bt	-.1535547
icu_rural_bt	-.145153
ictl_rural_bt	-29.20029
ictu_rural_bt	-27.72459
t_female	-33.31719
icl_female_bt	-.1711192
icu_female_bt	-.1628795
ictl_female_bt	-36.25423
ictu_female_bt	-34.61593
Observations	1000

Elaboración propia con datos de la ENIGH

(d) Utiliza el método `jackknife` y calcula los errores estándar.

Tabla 5: MCO y errores estándar: método Jackknife

Variables	(1) $\hat{\beta}$	(2) σ Jackknife
years_ed	0.0756*** (0.000611)	0.000611*** (4.00e-06)
edad	0.0744 - 0.0768 (123.8)	0.000603 - 0.000618 (152.6)
edad2	0.0333*** (0.00177)	0.00177*** (1.24e-05)
rural	0.0298 - 0.0367 (18.82)	0.00174 - 0.00179 (142.4)
female	-0.000319*** (2.08e-05)	2.08e-05*** (1.63e-07)
constante	-0.000360 - -0.000278 (-15.33)	2.05e-05 - 2.11e-05 (128.0)
Observations	85,250	85,250

Errores estándar y estadísticos t en paréntesis

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENIGH

- (e) Realiza nuevamente (a) pero en lugar de tomar un bootstrap de toda la muestra N, realiza un bootstrap con $0.25 * N$. Es decir, el tamaño de muestra es menor en cada bootstrap.

Tabla 6: Bootstrap del 25% de la muestra y con 100 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ BS	(3) $\hat{\sigma}_{Robusto}$ BS
years_ed	0.0803*** (0.00162)	0.0743*** (0.00127)	0.00121*** (1.71e-05)
	0.0771 - 0.0835 (49.48)	0.0718 - 0.0768 (58.70)	0.00118 - 0.00124 (70.65)
edad	0.0330*** (0.00469)	0.0355*** (0.00381)	0.00348*** (4.44e-05)
	0.0238 - 0.0422 (7.027)	0.0280 - 0.0430 (9.321)	0.00339 - 0.00356 (78.33)
edad2	-0.000298*** (5.50e-05)	-0.000339*** (4.40e-05)	4.08e-05*** (6.02e-07)
	-0.000405 - -0.000190 (-5.413)	-0.000425 - -0.000253 (-7.714)	3.96e-05 - 4.20e-05 (67.83)
rural	-0.268*** (0.0147)	-0.146*** (0.0106)	0.0107*** (0.000137)
	-0.297 - -0.240 (-18.24)	-0.167 - -0.126 (-13.80)	0.0104 - 0.0109 (78.13)
female	-0.144*** (0.0132)	-0.157*** (0.00995)	0.00938*** (7.90e-05)
	-0.170 - -0.118 (-10.96)	-0.176 - -0.137 (-15.76)	0.00922 - 0.00953 (118.6)
constante	1.893*** (0.0980)	1.943*** (0.0819)	0.0726*** (0.000828)
	1.701 - 2.086 (19.32)	1.783 - 2.104 (23.73)	0.0710 - 0.0743 (87.76)
Observations	21,506	21,506	21,506
R-squared	0.250		

En paréntesis: errores estándar robustos y estadísticos t

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENIGH

Tabla 7: Resumen de medias BS 100 con 25% de muestra

	Bootstrap 100 mean
_b[years_ed]	.0743575
_b[edad]	.0356656
_b[edad2]	-.0003407
_b[rural]	-.1463274
_b[female]	-.1575059
_b[_cons]	1.938467
_se[years_ed]	.001209
_se[edad]	.0034776
_se[edad2]	.0000408
_se[rural]	.0106868
_se[female]	.0093797
_se[_cons]	.0726406
hip	0
t_years_ed	61.51879
icl_years_ed_bt	.0718103
icu_years_ed_bt	.0770281
ictl_years_ed_bt	58.90289
ictu_years_ed_bt	64.28022
t_edad	10.25636
icl_edad_bt	.0260789
icu_edad_bt	.0419203
ictl_edad_bt	7.44483
ictu_edad_bt	12.09427
t_edad2	-8.34268
icl_edad2_bt	-.0004253
icu_edad2_bt	-.0002377
ictl_edad2_bt	-10.51557
ictu_edad2_bt	-5.878258
t_rural	-13.69189
icl_rural_bt	-.1716143
icu_rural_bt	-.1276815
ictl_rural_bt	-15.85859
ictu_rural_bt	-12.04106
t_female	-16.79153
icl_female_bt	-.1740443
icu_female_bt	-.1374089
ictl_female_bt	-18.63147
ictu_female_bt	-14.67399
Observations	100

Elaboración propia con datos de la ENIGH

(f) Realiza nuevamente (b) pero en lugar de tomar un bootstrap de toda la muestra N, realiza un bootstrap con $0.25 * N$. Es decir, el tamaño de muestra es menor en cada bootstrap.

Tabla 8: Bootstrap del 25% de la muestra y con 100 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ BS	(3) $\hat{\sigma}_{Robusto}$ BS
years_ed	0.0803*** (0.00162)	0.0743*** (0.00120)	0.00121*** (1.59e-05)
	0.0771 - 0.0835 (49.48)	0.0720 - 0.0766 (62.05)	0.00118 - 0.00124 (75.96)
edad	0.0330*** (0.00469)	0.0355*** (0.00350)	0.00348*** (4.62e-05)
	0.0238 - 0.0422 (7.027)	0.0286 - 0.0424 (10.14)	0.00338 - 0.00357 (75.22)
edad2	-0.000298*** (5.50e-05)	-0.000339*** (4.08e-05)	4.08e-05*** (6.06e-07)
	-0.000405 - -0.000190 (-5.413)	-0.000419 - -0.000259 (-8.308)	3.96e-05 - 4.20e-05 (67.32)
rural	-0.268*** (0.0147)	-0.146*** (0.0107)	0.0107*** (0.000125)
	-0.297 - -0.240 (-18.24)	-0.167 - -0.125 (-13.66)	0.0104 - 0.0109 (85.74)
female	-0.144*** (0.0132)	-0.157*** (0.00956)	0.00938*** (8.36e-05)
	-0.170 - -0.118 (-10.96)	-0.176 - -0.138 (-16.40)	0.00921 - 0.00954 (112.2)
constante	1.893*** (0.0980)	1.943*** (0.0736)	0.0726*** (0.000897)
	1.701 - 2.086 (19.32)	1.799 - 2.088 (26.39)	0.0709 - 0.0744 (80.94)
Observations	21,506	21,506	21,506
R-squared	0.250		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENIGH

Tabla 9: Resumen de medias BS 1000 con 25% de muestra

	Bootstrap 1000 mean
_b[years_ed]	.0742663
_b[edad]	.035589
_b[edad2]	-.0003403
_b[rural]	-.1464761
_b[female]	-.1574106
_b[_cons]	1.941987
_se[years_ed]	.0012098
_se[edad]	.003475
_se[edad2]	.0000408
_se[rural]	.0106769
_se[female]	.0093749
_se[_cons]	.0726158
hip	0
t_years_ed	61.39699
icl_years_ed_bt	.0708106
icu_years_ed_bt	.0727849
ictl_years_ed_bt	57.49363
ictu_years_ed_bt	59.73086
t_edad	10.24248
icl_edad_bt	.0253825
icu_edad_bt	.0310269
ictl_edad_bt	7.214668
ictu_edad_bt	8.920622
t_edad2	-8.340889
icl_edad2_bt	-.0004661
icu_edad2_bt	-.0003896
ictl_edad2_bt	-11.38481
ictu_edad2_bt	-9.593479
t_rural	-13.71803
icl_rural_bt	-.1783822
icu_rural_bt	-.1606119
ictl_rural_bt	-16.56397
ictu_rural_bt	-14.9673
t_female	-16.79042
icl_female_bt	-.1867836
icu_female_bt	-.1698192
ictl_female_bt	-19.94081
ictu_female_bt	-18.11187
Observations	1000

Elaboración propia con datos de la ENIGH

2. Compara respuestas de a-f. Explica detalladamente.

Respuesta. Los resultados son muy similares entre los incisos a) hasta d); es decir, entre los modelos estimados mediante bootstrap y jackknife. Para el caso de los incisos e) y f), encuentro diferencias en el tamaño de los errores estándar robustos, y por lo tanto en los intervalos de confianza de los mismos: los errores robustos son mayores en el caso de los incisos e) y f), quizás debido a que el tamaño de la muestra es del 25% solamente. Eso demuestra que la precisión de los estimadores aumenta con el tamaño de la muestra.

3. Utiliza el comando `bsample` ahora y haz las repeticiones por ti mismo. Es decir, repite a, b, d y e. En cada `bsample`, asegúrate de guardar el coeficiente, error y la t, para luego poder encontrar el bootstrap error estándar y el percentil-t. Al final, tendrás una base de datos de 100 observaciones o 1000 observaciones, y puedes obtener la parte a y b fácilmente.

Resultados en las siguientes páginas.

Tabla 10: Resumen de medias bsample con 100 repeticiones

	bsample 100 mean
bsample	50.5
byears_ed	.0822556
byears_ed_se	.0008325
t_years_ed	98.80866
bedad	.0363576
bedad_se	.0024227
t_edad	15.00736
bedad2	-.0003423
bedad2_se	.0000285
t_edad2	-12.0131
brural	-.2446887
brural_se	.0071501
t_rural	-34.22181
bfemale	-.1556211
bfemale_se	.0064687
t_female	-24.0576
iclyears_ed_boot	.0822556
icuyears_ed_boot	.0822556
ictlyears_ed_boot	98.80866
ictuyears_ed_boot	98.80866
icledad_boot	.0363576
icuedad_boot	.0363576
ictledad_boot	15.00736
ictuedad_boot	15.00736
icledad2_boot	-.0003423
icuedad2_boot	-.0003423
ictledad2_boot	-12.0131
ictuedad2_boot	-12.0131
iclrural_boot	-.2446887
icurural_boot	-.2446887
ictlrural_boot	-34.22181
icturural_boot	-34.22181
iclfemale_boot	-.1556211
icufemale_boot	-.1556211
iclfemale_boot	-24.0576
ictufemale_boot	-24.0576
Observations	100

Elaboración propia con datos de la ENIGH

Tabla 11: Resumen de medias bsample con 1000 repeticiones

	bsample 1000
	mean
bsample	500.5
byears_ed	.0822556
byears_ed_se	.0008325
t_years_ed	98.80866
bedad	.0363576
bedad_se	.0024227
t_edad	15.00736
bedad2	-.0003423
bedad2_se	.0000285
t_edad2	-12.0131
brural	-.2446887
brural_se	.0071501
t_rural	-34.22181
bfemale	-.1556211
bfemale_se	.0064687
t_female	-24.0576
iclyears_ed_boot	.0822556
icuyears_ed_boot	.0822556
ictlyears_ed_boot	98.80866
ictuyears_ed_boot	98.80866
icledad_boot	.0363576
icuedad_boot	.0363576
ictledad_boot	15.00736
ictuedad_boot	15.00736
icledad2_boot	-.0003423
icuedad2_boot	-.0003423
ictledad2_boot	-12.0131
ictuedad2_boot	-12.0131
iclrural_boot	-.2446887
icurural_boot	-.2446887
ictlrural_boot	-34.22181
icturural_boot	-34.22181
iclfemale_boot	-.1556211
icufemale_boot	-.1556211
iclfemale_boot	-24.0576
ictufemale_boot	-24.0576
Observations	1000

Elaboración propia con datos de la ENIGH

Tabla 12: Resumen de medias bsample 100 con 25% de muestra

	bsample 100 mean
bsample	50.5
byears_ed	.0822556
byears_ed_se	.0008325
t_years_ed	98.80866
bedad	.0363576
bedad_se	.0024227
t_edad	15.00736
bedad2	-.0003423
bedad2_se	.0000285
t_edad2	-12.0131
brural	-.2446887
brural_se	.0071501
t_rural	-34.22181
bfemale	-.1556211
bfemale_se	.0064687
t_female	-24.0576
iclyears_ed_boot	.0822556
icuyears_ed_boot	.0822556
ictlyears_ed_boot	98.80866
ictuyears_ed_boot	98.80866
icledad_boot	.0363576
icuedad_boot	.0363576
ictledad_boot	15.00736
ictuedad_boot	15.00736
icledad2_boot	-.0003423
icuedad2_boot	-.0003423
ictledad2_boot	-12.0131
ictuedad2_boot	-12.0131
iclrural_boot	-.2446887
icurural_boot	-.2446887
ictlrural_boot	-34.22181
icturural_boot	-34.22181
iclfemale_boot	-.1556211
icufemale_boot	-.1556211
iclfemale_boot	-24.0576
ictufemale_boot	-24.0576
Observations	100

Elaboración propia con datos de la ENIGH

Tabla 13: Resumen de medias bsample 1000 con 25% de muestra

	bsample 1000
	mean
bsample	500.5
byears_ed	.0822556
byears_ed_se	.0008325
t_years_ed	98.80866
bedad	.0363576
bedad_se	.0024227
t_edad	15.00736
bedad2	-.0003423
bedad2_se	.0000285
t_edad2	-12.0131
brural	-.2446887
brural_se	.0071501
t_rural	-34.22181
bfemale	-.1556211
bfemale_se	.0064687
t_female	-24.0576
iclyears_ed_boot	.0822556
icuyears_ed_boot	.0822556
ictlyears_ed_boot	98.80866
ictuyears_ed_boot	98.80866
icledad_boot	.0363576
icuedad_boot	.0363576
ictledad_boot	15.00736
ictuedad_boot	15.00736
icledad2_boot	-.0003423
icuedad2_boot	-.0003423
ictledad2_boot	-12.0131
ictuedad2_boot	-12.0131
iclrural_boot	-.2446887
icurural_boot	-.2446887
ictlrural_boot	-34.22181
icturural_boot	-34.22181
iclfemale_boot	-.1556211
icufemale_boot	-.1556211
iclfemale_boot	-24.0576
ictufemale_boot	-24.0576
Observations	1000

Elaboración propia con datos de la ENIGH

- Utiliza la ENOE II trimestre y realiza el problema 1 y 2, solo los incisos a, b, d y e. Año 2024.

Tabla 14: Resultados de MCO y del bootstrap no paramétrico de 100 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ Bootstrap 100	(3) se Bootstrap 100
years_ed	0.0455*** (0.00224) 0.0411 - 0.0499 (20.34)	0.0411*** (0.00157) 0.0380 - 0.0442 (26.11)	0.000643*** (1.11e-05) 0.000621 - 0.000664 (57.77)
edad	0.0305*** (0.00418) 0.0223 - 0.0386 (7.293)	0.0325*** (0.00281) 0.0270 - 0.0381 (11.58)	0.00252*** (1.22e-05) 0.00250 - 0.00254 (207.1)
edad2	-0.000309*** (4.81e-05) -0.000404 - -0.000215 (-6.436)	-0.000335*** (3.28e-05) -0.000400 - -0.000271 (-10.22)	2.88e-05*** (1.38e-07) 2.86e-05 - 2.91e-05 (209.3)
rural	-0.296*** (0.0144) -0.325 - -0.268 (-20.52)	-0.292*** (0.00917) -0.310 - -0.274 (-31.85)	0.00955*** (5.94e-05) 0.00943 - 0.00967 (160.8)
female	-0.0662*** (0.0112) -0.0881 - -0.0443 (-5.917)	-0.0841*** (0.00664) -0.0971 - -0.0711 (-12.67)	0.00666*** (3.24e-05) 0.00659 - 0.00672 (205.1)
constante	2.921*** (0.0923) 2.740 - 3.101 (31.66)	2.968*** (0.0620) 2.847 - 3.090 (47.88)	0.0535*** (0.000265) 0.0529 - 0.0540 (201.6)
Observations	105,164	105,164	105,164
R-squared	0.070		

Robust se tstat in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENOE

Tabla 15: Resumen de estadísticas descriptivas de medias BS 100

	Bootstrap 100
	mean
_b[anios_esc]	.0412598
_b[eda]	.0324675
_b[edad2]	-.0003333
_b[rural]	-.2905724
_b[female]	-.0834025
_b[_cons]	2.965641
_se[anios_esc]	.0006435
_se[eda]	.002522
_se[edad2]	.0000288
_se[rural]	.0095569
_se[female]	.0066588
_se[_cons]	.0534882
hip	0
t_anios_esc	64.09698
icl_anios_esc_bt	.0381696
icu_anios_esc_bt	.0441168
ictl_anios_esc_bt	61.32117
ictu_anios_esc_bt	66.93938
t_edad	12.87401
icl_edad_bt	.0273861
icu_edad_bt	.0387745
ictl_edad_bt	10.85099
ictu_edad_bt	15.33494
t_edad2	-11.55555
icl_edad2_bt	-.0004118
icu_edad2_bt	-.0002795
ictl_edad2_bt	-14.22749
ictu_edad2_bt	-9.67851
t_rural	-30.40529
icl_rural_bt	-.3084263
icu_rural_bt	-.2741624
ictl_rural_bt	-32.36472
ictu_rural_bt	-28.87455
t_female	-12.52536
icl_female_bt	-.0958127
icu_female_bt	-.0710728
ictl_female_bt	-14.4582
ictu_female_bt	-10.69906
Observations	100

Elaboración propia con datos de la ENOE

Tabla 16: Resultados de MCO y del bootstrap no paramétrico de 1000 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ Bootstrap 1000	(3) se Bootstrap 1000
years_ed	0.0455*** (0.00224)	0.0411*** (0.00147)	0.00150*** (3.96e-05)
	0.0411 - 0.0499 (20.34)	0.0382 - 0.0440 (27.88)	0.00142 - 0.00158 (37.83)
edad	0.0305*** (0.00418)	0.0325*** (0.00258)	0.00255*** (2.17e-05)
	0.0223 - 0.0386 (7.293)	0.0275 - 0.0376 (12.60)	0.00251 - 0.00259 (117.6)
edad2	-0.000309*** (4.81e-05)	-0.000335*** (3.00e-05)	2.97e-05*** (2.70e-07)
	-0.000404 - -0.000215 (-6.436)	-0.000394 - -0.000277 (-11.18)	2.91e-05 - 3.02e-05 (109.9)
rural	-0.296*** (0.0144)	-0.292*** (0.0104)	0.0101*** (0.000114)
	-0.325 - -0.268 (-20.52)	-0.312 - -0.271 (-28.03)	0.00989 - 0.0103 (88.79)
female	-0.0662*** (0.0112)	-0.0841*** (0.00692)	0.00674*** (3.48e-05)
	-0.0881 - -0.0443 (-5.917)	-0.0976 - -0.0705 (-12.16)	0.00667 - 0.00681 (193.9)
constante	2.921*** (0.0923)	2.968*** (0.0571)	0.0567*** (0.000511)
	2.740 - 3.101 (31.66)	2.856 - 3.080 (52.00)	0.0557 - 0.0577 (111.1)
Observations	105,164	105,164	105,164
R-squared	0.070		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENOE

Tabla 17: Resumen de estadísticas descriptivas de medias BS 1000

	Bootstrap 1000 mean
_b[anios_esc]	.0411009
_b[eda]	.0325187
_b[edad2]	-.0003351
_b[rural]	-.2918096
_b[female]	-.0842028
_b[_cons]	2.968575
_se[anios_esc]	.0014959
_se[eda]	.0025525
_se[edad2]	.0000297
_se[rural]	.0101204
_se[female]	.0067432
_se[_cons]	.0567307
hip	0
t_anios_esc	27.47377
icl_anios_esc_bt	.0373183
icu_anios_esc_bt	.039224
ictl_anios_esc_bt	25.70048
ictu_anios_esc_bt	26.69503
t_edad	12.74358
icl_edad_bt	.0256223
icu_edad_bt	.0291166
ictl_edad_bt	9.981412
ictu_edad_bt	11.37884
t_edad2	-11.29694
icl_edad2_bt	-.0004293
icu_edad2_bt	-.0003738
ictl_edad2_bt	-14.67201
ictu_edad2_bt	-12.69161
t_rural	-28.84213
icl_rural_bt	-.3215677
icu_rural_bt	-.305834
ictl_rural_bt	-32.05272
ictu_rural_bt	-30.45849
t_female	-12.48868
icl_female_bt	-.1054044
icu_female_bt	-.0933555
ictl_female_bt	-15.59842
ictu_female_bt	-13.85565
Observations	1000

Elaboración propia con datos de la ENOE

Tabla 18: Bootstrap con 25% de muestra con 100 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ Bootstrap 100	(3) se Bootstrap 100
years_ed	0.0457*** (0.00449)	0.0418*** (0.00267)	0.00306*** (0.000152)
	0.0369 - 0.0545 (10.18)	0.0366 - 0.0471 (15.68)	0.00276 - 0.00336 (20.10)
edad	0.0265*** (0.00849)	0.0319*** (0.00495)	0.00515*** (7.42e-05)
	0.00983 - 0.0431 (3.118)	0.0222 - 0.0416 (6.453)	0.00500 - 0.00529 (69.38)
edad2	-0.000237** (9.67e-05)	-0.000312*** (5.88e-05)	5.99e-05*** (9.40e-07)
	-0.000427 - -4.77e-05 (-2.453)	-0.000427 - -0.000197 (-5.306)	5.81e-05 - 6.17e-05 (63.75)
rural	-0.310*** (0.0305)	-0.296*** (0.0183)	0.0208*** (0.000385)
	-0.370 - -0.250 (-10.16)	-0.332 - -0.260 (-16.23)	0.0201 - 0.0216 (53.97)
female	-0.0558** (0.0240)	-0.0827*** (0.0120)	0.0136*** (0.000147)
	-0.103 - -0.00878 (-2.326)	-0.106 - -0.0592 (-6.879)	0.0133 - 0.0139 (92.46)
constante	2.944*** (0.188)	2.942*** (0.104)	0.114*** (0.00167)
	2.577 - 3.312 (15.69)	2.739 - 3.146 (28.33)	0.111 - 0.118 (68.29)
Observations	26,567	26,567	26,567
R-squared	0.072		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENOE

Tabla 19: Resumen de medias BS 100 con 25% de muestra

	Bootstrap 100
	mean
_b[anios_esc]	.042123
_b[eda]	.0313825
_b[edad2]	-.0003042
_b[rural]	-.2969606
_b[female]	-.0820324
_b[_cons]	2.94652
_se[anios_esc]	.0030582
_se[eda]	.0051571
_se[edad2]	.00006
_se[rural]	.0208149
_se[female]	.0135605
_se[_cons]	.1142775
hip	0
t_anios_esc	13.77625
icl_anios_esc_bt	.037155
icu_anios_esc_bt	.0472627
ictl_anios_esc_bt	12.70848
ictu_anios_esc_bt	15.19523
t_edad	6.091496
icl_edad_bt	.0194289
icu_edad_bt	.0400503
ictl_edad_bt	3.684549
ictu_edad_bt	7.816027
t_edad2	-5.072913
icl_edad2_bt	-.0004083
icu_edad2_bt	-.0001921
ictl_edad2_bt	-6.836483
ictu_edad2_bt	-3.16362
t_rural	-14.27793
icl_rural_bt	-.3327831
icu_rural_bt	-.2591735
ictl_rural_bt	-16.31994
ictu_rural_bt	-12.23296
t_female	-6.052744
icl_female_bt	-.1102998
icu_female_bt	-.060661
ictl_female_bt	-8.242102
ictu_female_bt	-4.440203
Observations	100

Elaboración propia con datos de la ENOE

Tabla 20: Bootstrap con 25% de muestra con 1000 repeticiones

Variables	(1) $\hat{\beta}$ MCO	(2) $\hat{\beta}$ Bootstrap 1000	(3) se Bootstrap 1000
years_ed	0.0457*** (0.00449)	0.0418*** (0.00305)	0.00306*** (0.000164)
	0.0369 - 0.0545 (10.18)	0.0358 - 0.0478 (13.69)	0.00274 - 0.00338 (18.62)
edad	0.0265*** (0.00849)	0.0319*** (0.00514)	0.00515*** (8.57e-05)
	0.00983 - 0.0431 (3.118)	0.0218 - 0.0420 (6.208)	0.00498 - 0.00532 (60.10)
edad2	-0.000237** (9.67e-05)	-0.000312*** (5.99e-05)	5.99e-05*** (1.07e-06)
	-0.000427 - -4.77e-05 (-2.453)	-0.000429 - -0.000195 (-5.208)	5.78e-05 - 6.20e-05 (55.80)
rural	-0.310*** (0.0305)	-0.296*** (0.0207)	0.0208*** (0.000454)
	-0.370 - -0.250 (-10.16)	-0.337 - -0.256 (-14.33)	0.0199 - 0.0217 (45.82)
female	-0.0558** (0.0240)	-0.0827*** (0.0134)	0.0136*** (0.000136)
	-0.103 - -0.00878 (-2.326)	-0.109 - -0.0565 (-6.178)	0.0133 - 0.0138 (100.0)
constante	2.944*** (0.188)	2.942*** (0.115)	0.114*** (0.00189)
	2.577 - 3.312 (15.69)	2.717 - 3.168 (25.57)	0.111 - 0.118 (60.65)
Observations	26,567	26,567	26,567
R-squared	0.072		

En paréntesis: errores estándar robustos y estadísticos t y z

*** p<0.01, ** p<0.05, * p<0.1

Elaboración propia con datos de la ENOE

Tabla 21: Resumen de medias BS 1000 con 25% de muestra

	Bootstrap 1000 mean
_b[anios_esc]	.042006
_b[eda]	.0318614
_b[edad2]	-.000311
_b[rural]	-.2963644
_b[female]	-.0827654
_b[_cons]	2.941238
_se[anios_esc]	.0030587
_se[eda]	.0051478
_se[edad2]	.0000599
_se[rural]	.0208085
_se[female]	.0135687
_se[_cons]	.1143677
hip	0
t_anios_esc	13.73107
icl_anios_esc_bt	.033223
icu_anios_esc_bt	.0382009
ictl_anios_esc_bt	12.17038
ictu_anios_esc_bt	12.95216
t_edad	6.196698
icl_edad_bt	.016422
icu_edad_bt	.0248668
ictl_edad_bt	3.159328
ictu_edad_bt	4.816911
t_edad2	-5.201621
icl_edad2_bt	-.0004767
icu_edad2_bt	-.000388
ictl_edad2_bt	-8.02268
ictu_edad2_bt	-6.534361
t_rural	-14.25988
icl_rural_bt	-.3559835
icu_rural_bt	-.3236739
ictl_rural_bt	-18.16866
ictu_rural_bt	-15.80083
t_female	-6.102906
icl_female_bt	-.1202835
icu_female_bt	-.1008195
ictl_female_bt	-8.984376
ictu_female_bt	-7.443954
Observations	1000

Elaboración propia con datos de la ENOE

Tabla 22: Resumen de medias bsample con 100 repeticiones

	bsample 100 mean
bsample	50.5
banios_esc	.0454986
banios_esc_se	.0022365
t_anios_esc	20.34401
beda	.030453
beda_se	.0041757
t_eda	7.292947
bedad2	-.0003094
bedad2_se	.0000481
t_edad2	-6.436038
brural	-.2963682
brural_se	.0144442
t_rural	-20.51815
bfemale	-.0661834
bfemale_se	.0111861
t_female	-5.916576
iclanios_esc_boot	.0454986
icuanios_esc_boot	.0454986
ictlanios_esc_boot	20.34401
ictuanios_esc_boot	20.34401
icleda_boot	.030453
icueda_boot	.030453
ictleda_boot	7.292947
ictueda_boot	7.292947
icledad2_boot	-.0003094
icuedad2_boot	-.0003094
ictledad2_boot	-6.436038
ictuedad2_boot	-6.436038
iclrural_boot	-.2963682
icurural_boot	-.2963682
ictlrural_boot	-20.51815
icturural_boot	-20.51815
iclfemale_boot	-.0661834
icufemale_boot	-.0661834
ictlfemale_boot	-5.916576
ictufemale_boot	-5.916576
Observations	100

Elaboración propia con datos de la ENOE

Tabla 23: Resumen de medias bsample con 1000 repeticiones

	bsample 1000
	mean
bsample	500.5
banios_esc	.0454986
banios_esc_se	.0022365
t_anios_esc	20.34401
beda	.030453
beda_se	.0041757
t_eda	7.292947
bedad2	-.0003094
bedad2_se	.0000481
t_edad2	-6.436038
brural	-.2963682
brural_se	.0144442
t_rural	-20.51815
bfemale	-.0661834
bfemale_se	.0111861
t_female	-5.916576
iclanios_esc_boot	.0454986
icuanios_esc_boot	.0454986
ictlanios_esc_boot	20.34401
ictuanios_esc_boot	20.34401
icleda_boot	.030453
icueda_boot	.030453
ictleda_boot	7.292947
ictueda_boot	7.292947
icledad2_boot	-.0003094
icuedad2_boot	-.0003094
ictledad2_boot	-6.436038
ictuedad2_boot	-6.436038
iclrural_boot	-.2963682
icurural_boot	-.2963682
ictlrural_boot	-20.51815
icturural_boot	-20.51815
iclfemale_boot	-.0661834
icufemale_boot	-.0661834
ictlfemale_boot	-5.916576
ictufemale_boot	-5.916576
Observations	1000

Elaboración propia con datos de la ENOE

Tabla 24: Resumen de medias bsample 100 con 25% de muestra

	bsample 100 mean
bsample	50.5
banios_esc	.0454986
banios_esc_se	.0022365
t_anios_esc	20.34401
beda	.030453
beda_se	.0041757
t_eda	7.292947
bedad2	-.0003094
bedad2_se	.0000481
t_edad2	-6.436038
brural	-.2963682
brural_se	.0144442
t_rural	-20.51815
bfemale	-.0661834
bfemale_se	.0111861
t_female	-5.916576
iclanios_esc_boot	.0454986
icuanios_esc_boot	.0454986
ictlanios_esc_boot	20.34401
ictuanios_esc_boot	20.34401
icleda_boot	.030453
icueda_boot	.030453
ictleda_boot	7.292947
ictueda_boot	7.292947
icledad2_boot	-.0003094
icuedad2_boot	-.0003094
ictledad2_boot	-6.436038
ictuedad2_boot	-6.436038
iclrural_boot	-.2963682
icurural_boot	-.2963682
ictlrural_boot	-20.51815
icturural_boot	-20.51815
iclfemale_boot	-.0661834
icufemale_boot	-.0661834
ictlfemale_boot	-5.916576
ictufemale_boot	-5.916576
Observations	100

Elaboración propia con datos de la ENOE

Tabla 25: Resumen de medias bsample 1000 con 25% de muestra

	bsample 1000
	mean
bsample	500.5
banios_esc	.0454986
banios_esc_se	.0022365
t_anios_esc	20.34401
beda	.030453
beda_se	.0041757
t_eda	7.292947
bedad2	-.0003094
bedad2_se	.0000481
t_edad2	-6.436038
brural	-.2963682
brural_se	.0144442
t_rural	-20.51815
bfemale	-.0661834
bfemale_se	.0111861
t_female	-5.916576
iclanios_esc_boot	.0454986
icuanios_esc_boot	.0454986
ictlanios_esc_boot	20.34401
ictuanios_esc_boot	20.34401
icleda_boot	.030453
icueda_boot	.030453
ictleda_boot	7.292947
ictueda_boot	7.292947
icledad2_boot	-.0003094
icuedad2_boot	-.0003094
ictledad2_boot	-6.436038
ictuedad2_boot	-6.436038
iclrural_boot	-.2963682
icurural_boot	-.2963682
ictlrural_boot	-20.51815
icturural_boot	-20.51815
iclfemale_boot	-.0661834
icufemale_boot	-.0661834
ictlfemale_boot	-5.916576
ictufemale_boot	-5.916576
Observations	1000

Elaboración propia con datos de la ENOE

5 Problema 5: Métodos No Paramétricos

1. Discute el trade-off que existe entre varianza y sesgo al momento de escoger un kernel o bandwidth en estimación no paramétrica.

Respuesta. Antes de discutir ese trade-off, conviene primero recordar la idea detrás de la densidad de kernel, la cual tiene como objetivo calcular estimar la función de densidad de probabilidad (PDF) de una variable aleatoria a partir de las observaciones que tenemos. Dicho esto, la función de densidad de probabilidad estimada $\hat{f}()$ en un punto específico w se define como:

$$\hat{f}(w) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} K\left(\frac{w - W_i}{h}\right)$$

Donde:

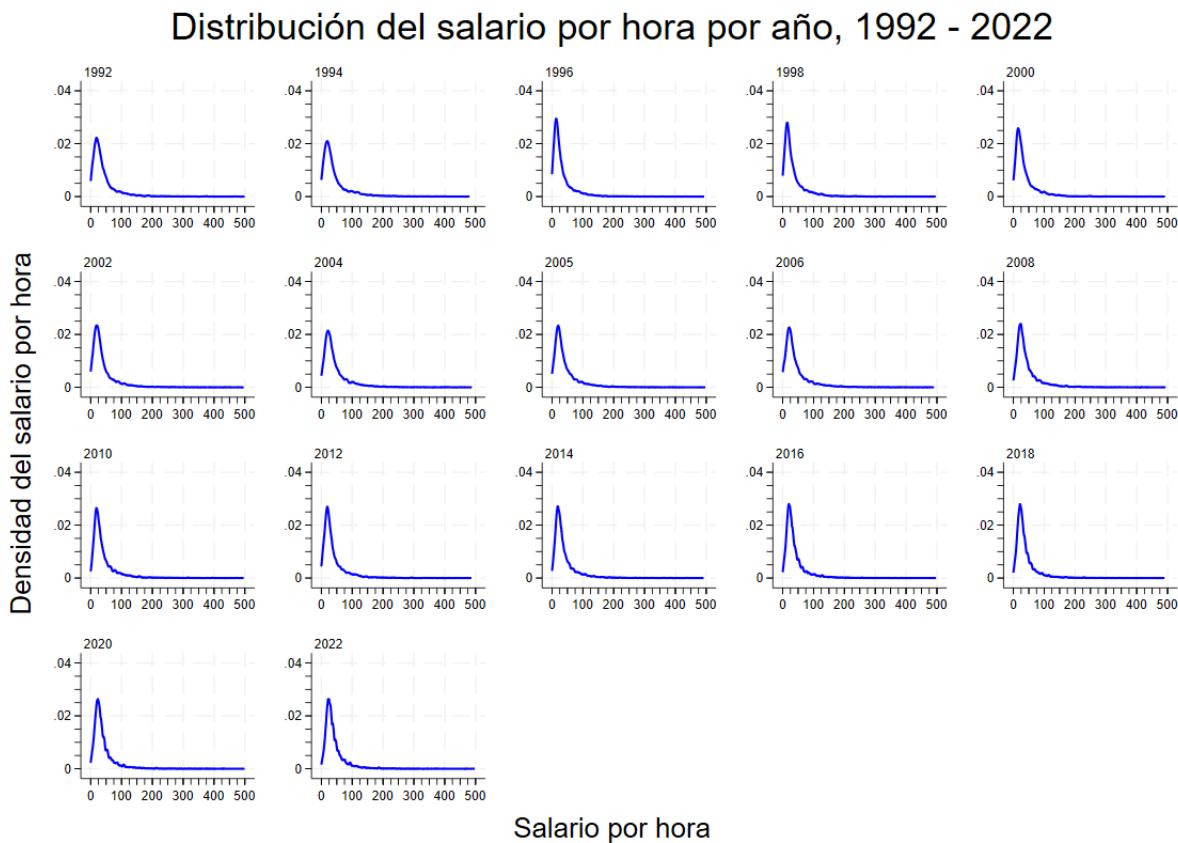
- N es el tamaño de la muestra
- W_i son las observaciones, donde $i = 1, 2, \dots, N$.
- h es conocido como el ancho de banda (bandwidth) y controla la "suavidad" de la densidad estimada.
- $K()$ es la función de núcleo (kernel), que evalúa la distancia entre el punto de interés w y el dato observado W_i , escalada por el ancho de banda h . La función de núcleo asigna pesos a cada observación según qué tan cerca esté W_i de w . Algunos núcleos comunes son el Gaussiano o el Epanechnikov, como vimos en clase.

Ahora bien, en el cálculo de $\hat{f}(w)$, existe un importante **trade-off** entre varianza y sesgo al momento de elegir el kernel $K()$ y, sobre todo, el bandwidth h . Este trade-off es crucial porque afecta la precisión y de la estimación de la función de densidad subyacente. En este sentido, el **sesgo** mide cuánto se aleja la estimación promedio de la verdadera función de densidad, y éste suele aumentar cuando h es grande, lo que suaviza demasiado la estimación, resultando en una pérdida de detalle o estructura en los datos. Por otro lado, la **varianza** mide la dispersión de las estimaciones alrededor del valor estimado promedio: una varianza alta implica que pequeñas variaciones en los datos pueden llevar a grandes diferencias en la estimación de la densidad según la muestra, y ésta es mayor cuando se utiliza un h pequeño.

En resumen:

- h pequeño:
 - **Varianza alta:** El estimador sigue demasiado de cerca los datos, capturando el ruido y las fluctuaciones aleatorias.

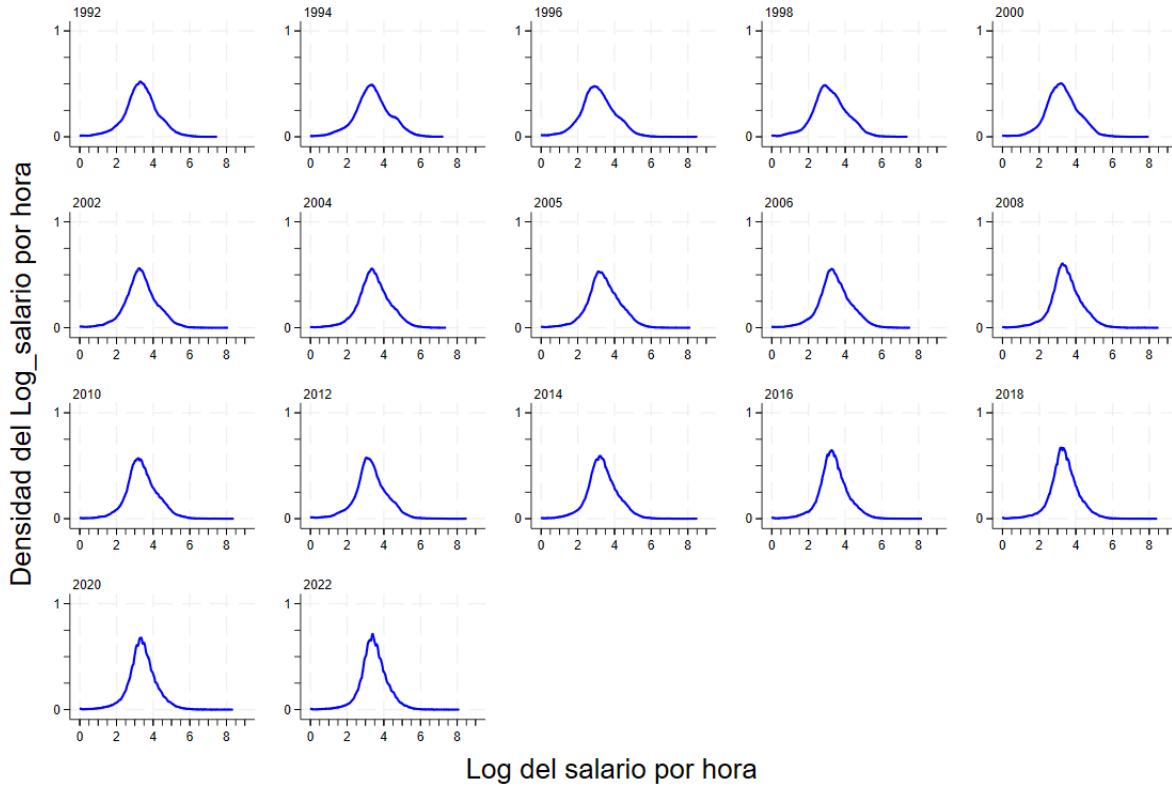
- **Sesgo bajo:** Dado que cada punto de datos tiene un gran peso, la estimación se ajusta estrechamente a los datos y refleja con precisión la estructura local de la densidad.
 - h grande:
 - **Varianza baja:** El estimador es más estable y la densidad estimada será suave y menos sensible a las fluctuaciones aleatorias de los datos.
 - **Sesgo alto:** El exceso de suavidad hace que se pierdan características locales importantes de la función de densidad, lo que lleva a una subestimación de los picos y la verdadera variabilidad de los datos.
2. Grafica una distribución no paramétrica (`kdensity`) del salario por hora para cada año. ¿Qué opinas? Ahora genera una variable que sea el logaritmo natural del salario por hora. Grafica una distribución no paramétrica para esta nueva variable. ¿Cómo cambian tus resultados? (Hint: Puedes limitar tus gráficas a cierto dominio, si esto las hace ver mejor... pero utiliza el mismo criterio en todas las gráficas, gráficas que no se entiendan y no se vean bonitas son castigadas).



Fuente: Elaboración propia con datos de la ENIGH

Las densidades estimadas indican que la distribución se ha concentrado sistemáticamente alrededor de los 30 y 40 pesos por hora, con colas derechas muy largas a pesar de que acotamos el eje x. Esto es indicativo de una desigualdad salarial enorme; además, es interesante notar que existen concentraciones en salarios bajos mayores en los años de la encuesta posteriores a los años de crisis en México, lo que demuestra la afectación que reciben los trabajadores tras esos periodos de shock.

Distribución del log_salario por hora por año, 1992 - 2022



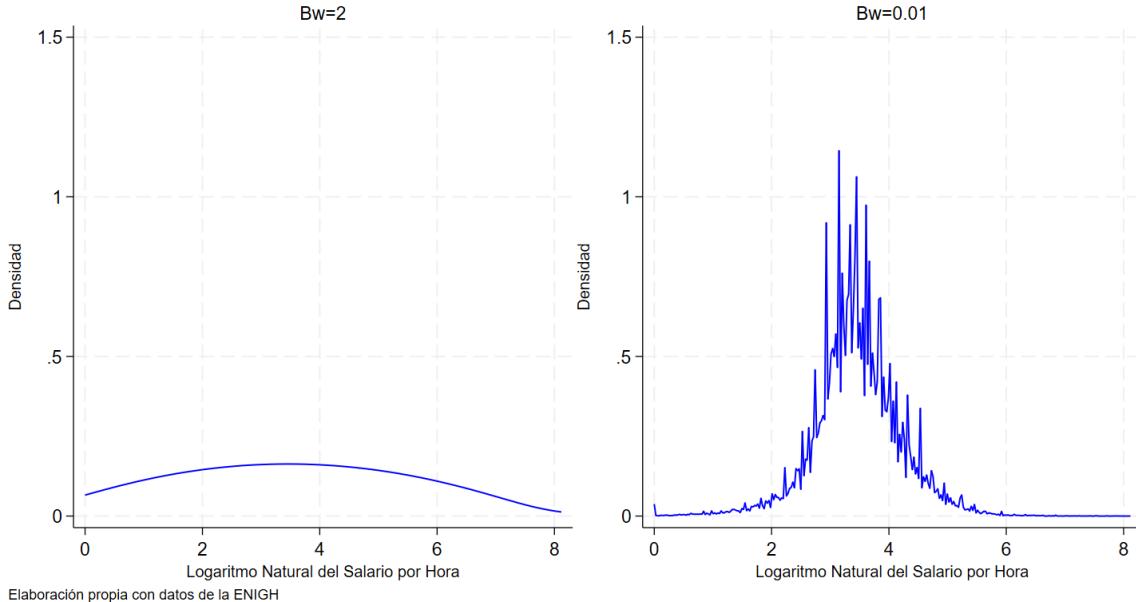
Fuente: Elaboración propia con datos de la ENIGH

La adición que esta visualización de los salarios nos brinda, es que la distribución salarial se ha tornado más apretada en un cierto valor, es decir, la densidad se vuelve cada vez más delgada, manteniendo colas largas (sobre todo la derecha), indicando una profundización de la desigualdad; esto debido a que, mientras que los salarios en logaritmos se concentran cada vez más en un punto, las colas han crecido en ambos lados.

3. Utilizando el logaritmo natural del salario por hora, crea dos gráficas con distribuciones no paramétricas una aumentando el *bandwidth* sustancialmente, y otra disminuyéndolo (utiliza un año únicamente). Haz lo mismo para dos diferentes kernels (manteniendo el *bandwidth* constante). ¿Cómo afecta tu gráfica el *bandwidth* y el kernel? Explica.

Distribución no paramétrica del salario por hora, 2022

Para diferentes bandwidth



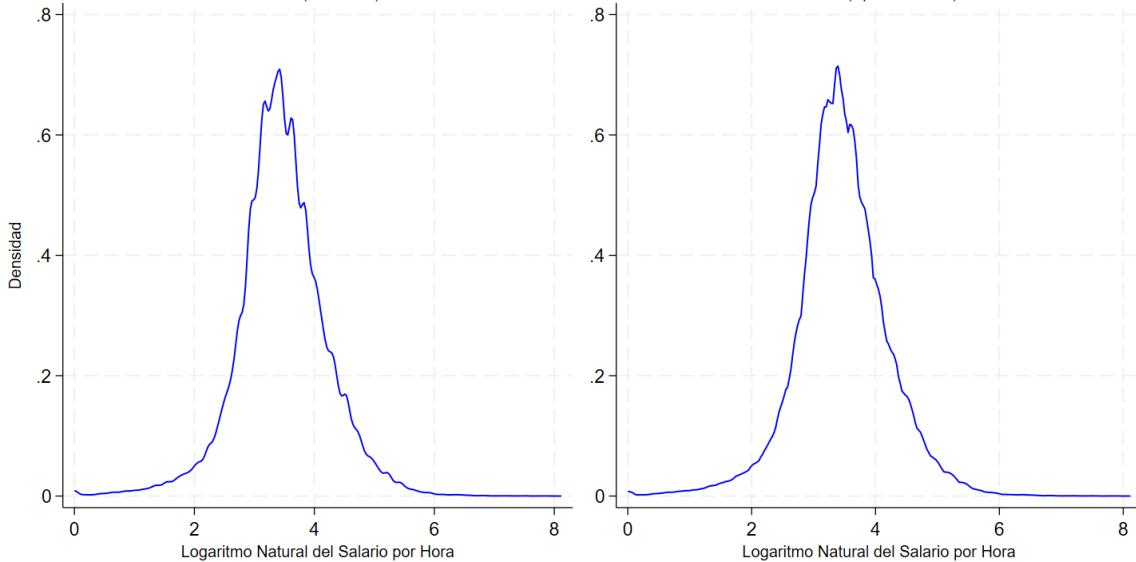
Elaboración propia con datos de la ENIGH

Distribución no paramétrica del salario por hora, 2022

Para diferentes Kernels considerando Bw óptimo (default de Stata)

Kernel(Gaussian)

Kernel(Epanechnikov)

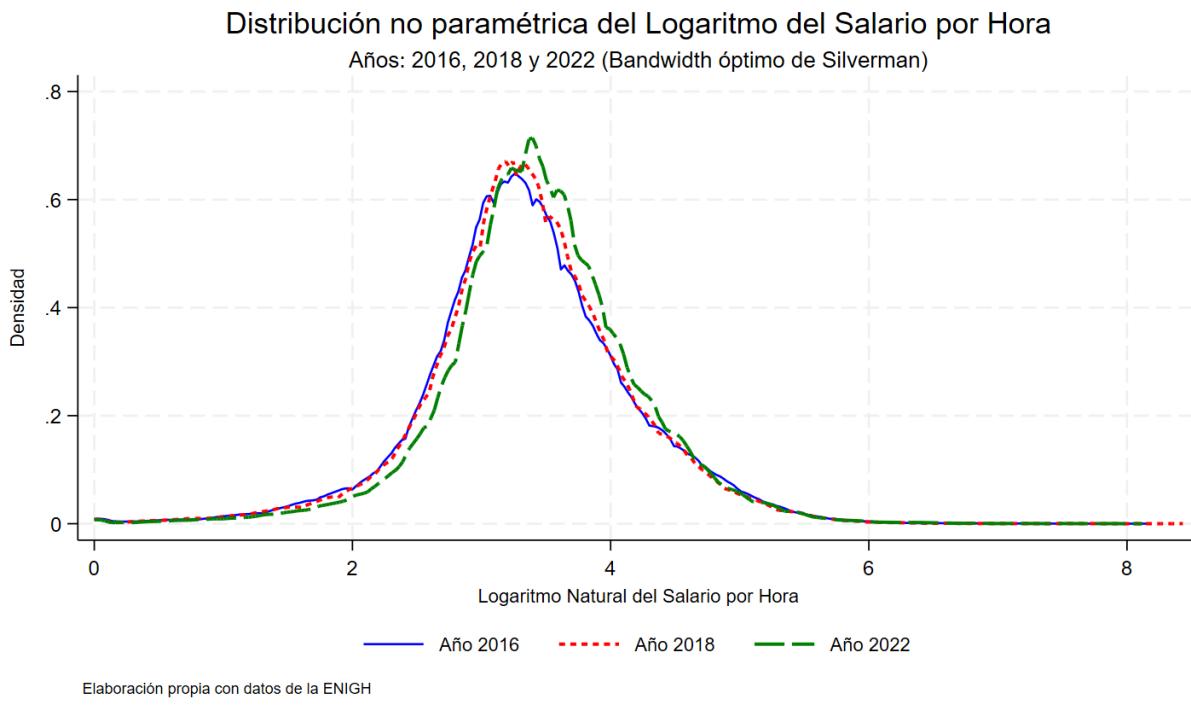


Elaboración propia con datos de la ENIGH

Las diferencia entre las dos figuras con bandwidths distintos, es consistente con la discusión del primer inciso: un h relativamente alto como el de la figura superior izquierda nos arroja una densidad con una varianza muy baja, pero con un sesgo muy alto. La figura se vuelve plana entre mayor sea el bandwidth, y esconde las características locales de la densidad verdadera. En contraste, un h relativamente bajo crea el efecto contrario mostrado en la figura superior derecha: la varianza es muy alta, y el sesgo muy bajo.

Por otro lado, consistente con lo que vimos en clase, la decisión entre usar un u otro kernel es de menor importancia, como observamos en las dos figuras anteriores ubicadas en la parte inferior.

4. ENIGH: Grafica la distribución no paramétrica del logaritmo del salario por hora del año 2016 con la de los años 2018 y 2022 en el mismo gráfico utilizando el *bandwidth* óptimo de Silverman (1986). ¿Qué puedes concluir al respecto?



Lo primero que podemos ver es que, entre 2016 y 2022, la densidad se ha desplazado hacia la derecha ligeramente, lo que indica un aumento salarial para casi todos niveles salariales. Entre 2016 y 2018, el mecanismo actúa en la misma dirección, pero en menor medida, indicando que la mayor parte del aumento vino en el periodo después de 2018, lo que coincide con la rondas de aumentos al salario mínimo. Además, el valor del log-salario de concentración se movió hacia la derecha y aumentó la densidad de ese punto, lo que indica además de un salario medio mayor, una mayor concentración en ese nivel entre 2016 y 2022.

5. Grafica regresiones no paramétricas:

- Explica las diferencias entre `lowess` y `lpoly`.

Recordemos que en una regresión no paramétrica tenemos:

$$y = m(x) + \epsilon$$

Donde $m(x)$ es una función de x a estimar, y que, en el contexto de las regresiones tipo "Local Linear Regressions" (LLR), es:

$$m(x) = a_0 + b_0(x - x_0)$$

Es decir, $m(x)$ es una función lineal en la vecindad de x_0 . De este modo, el estimador de la LLR $\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$ en la vecindad de x_0 , minimiza:

$$\Sigma_{i=1}^N = K\left(\frac{x_i - x_0}{h}\right)(y_i - a_0 - b_0(x_i - x_0))^2,$$

con respecto a a_0 y b_0 , y donde K es una función kernel ponderadora. Entonces, el estimador en exactamente x_0 es $\hat{m}(x_0) = \hat{a}_0$. Más generalmente, una "**Local Polynomial estimator**" (`lpoly`) de grado p minimiza:

$$\Sigma_{i=1}^N = K\left(\frac{x_i - x_0}{h}\right)(y_i - a_{0,0} - a_{0,1}(x_i - x_0) - \dots - a_{0,p}\frac{(x_i - x_0)^p}{p!})^2,$$

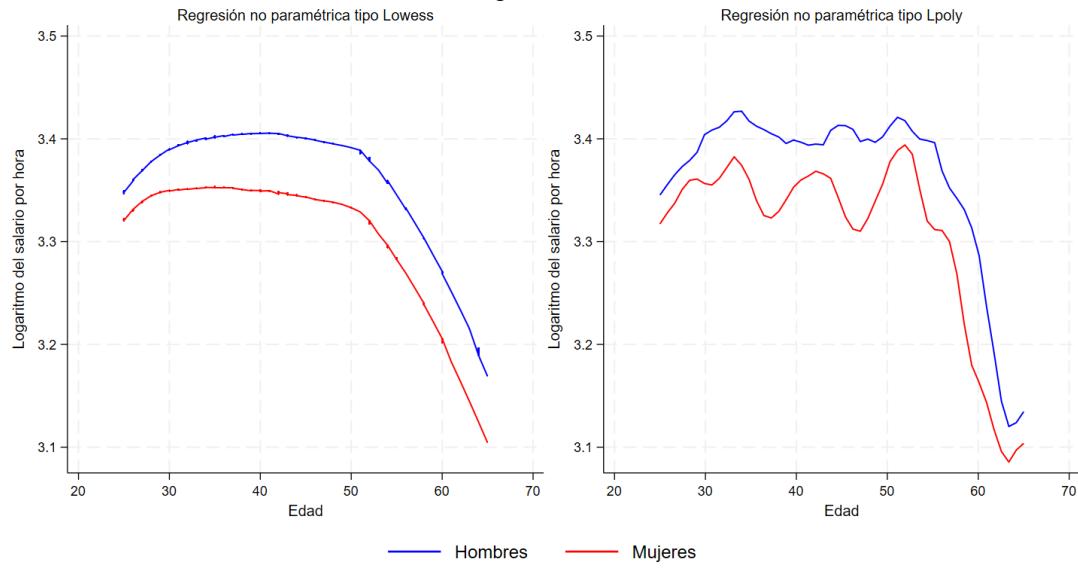
por lo que $\hat{m}^{(s)}(x_0) = \hat{a}_{0,s}$. Por otro lado, el estimador de la "**Locally Weighted Scatterplot Smoothing**" (`lowess`), es una variante del estimador de `lpoly`, donde las principales diferencias son:

- `lowess` usa un bandwidth variable $h_{0,k}$ en vez de uno fijo h ; determinado por la distancia entre x_0 a su vecino k -ésimo más cercano ("*kth nearest neighbor*").
- Utiliza el kernel tricúbico $K(z) = \frac{70}{81}(1 - |z|^3)^3 \mathbf{1}(|z| < 1)$
- Requiere una computación mayor

- (b) Escoge el año de 2018. Grafica no paramétricamente utilizando `lowess` y `lpoly` el logaritmo del salario por hora separadamente para hombres y mujeres. En el eje y es el salario y realiza dos figuras, una donde el eje x tenemos la edad y otra donde tenemos años de escolaridad. Sé claro en tus supuestos.

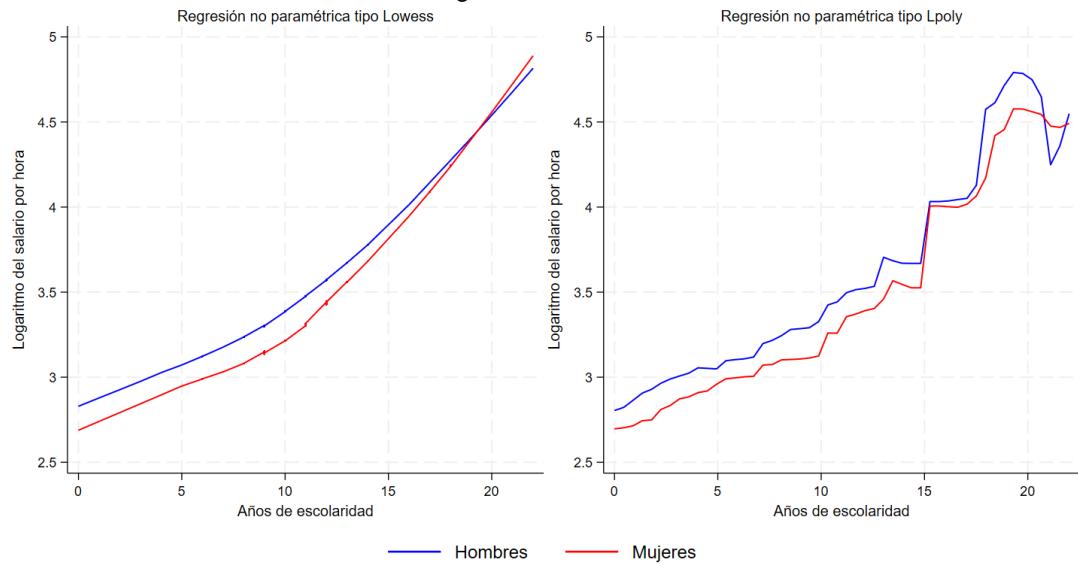
Gráficas en la siguiente página. No se realizó ningún supuesto adicional a los que los comandos `lowess` y `lpoly` de Stata realizan.

2018: Log-salario vs Edad



Elaboración propia con datos de la ENIGH

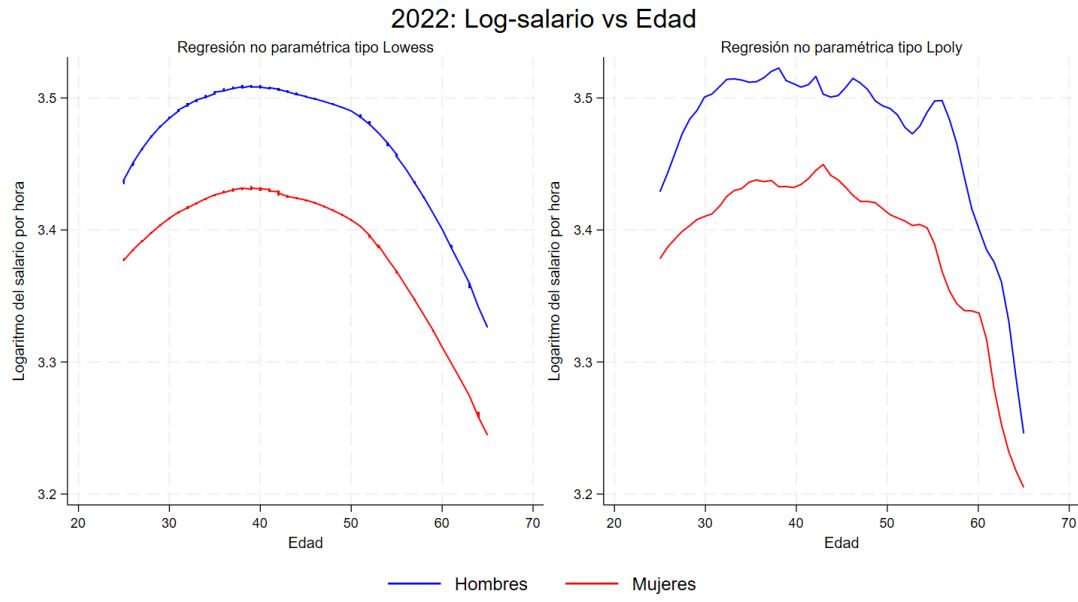
2018: Log-salario vs Escolaridad



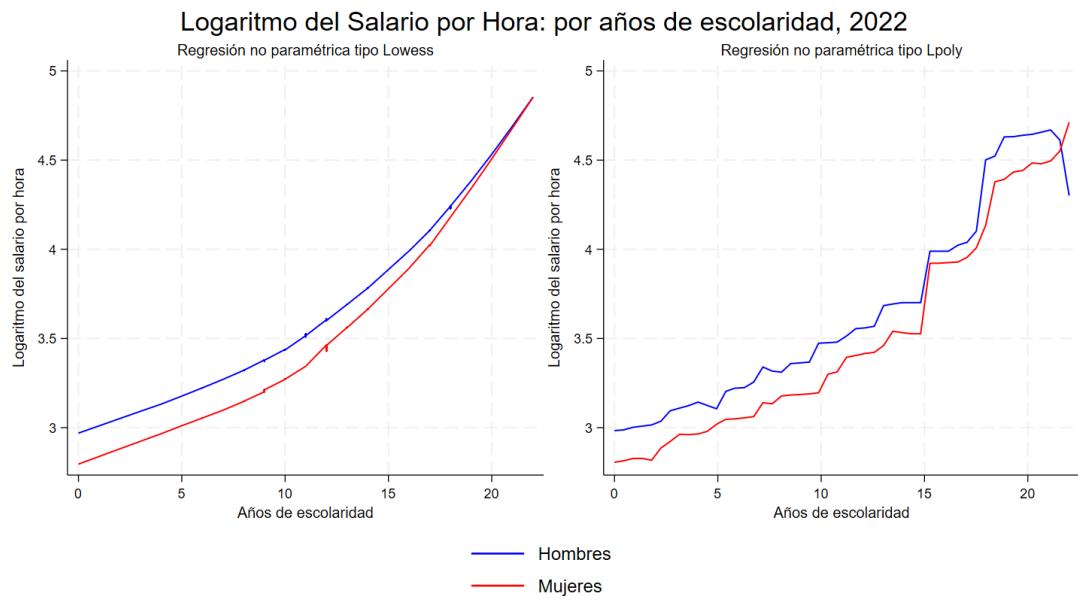
Elaboración propia con datos de la ENIGH

- (c) Escoge el año de 2022. Repite el ejercicio b.

Gráficas en la siguiente página. No se realizó ningún supuesto adicional a los que los comandos `lowess` y `lpoly` de Stata realizan.



Elaboración propia con datos de la ENIGH



Elaboración propia con datos de la ENIGH

- (d) Explica diferencias y similitudes. Explica posibles razones. ¿Crees que esto nos ayude a explicar qué pasó con el ingreso laboral en esos dos años en términos de comparabilidad?
- En el caso de **los salarios y la edad**, la principal diferencia entre 2018 y 2022 es que aparece que la brecha salarial entre géneros se ha acentuado para todas las edades, a pesar de que para ambos géneros el retorno a la edad parece ser mayor de un periodo a otro. Parece también que el salario para las personas de mayor edad aumentó más que para el resto, quizás reflejo de los programas de asistencia puestos entre ambos periodos; esto se aprecia en la reducción de la cola derecha de las gráficas de la parte

superior. También parece que las diferencias en el log-salario por edades se ha reducido en el caso de las mujeres de los 20 a los 50 años: esto se visualiza en la reducción del ruido en las regresiones `lpoly`.

Por otro lado, para **los salarios y la educación**, podríamos decir que el retorno educativo es más alto para todos los niveles de escolaridad en 2022 en comparación con 2018, y pareciera que la brecha de género se ha incrementado ligeramente; no obstante, los resultados visuales son muy similares entre ambos años. Algo adicional es que pareciera haber un cambio estructural a partir de los 15 años de educación, pues el log-salario da un salto y tiene una pendiente más inclinada, revelando retornos educativos más que proporcionalmente mayores que los del resto.

Finalmente, como veremos en el siguiente inciso, es difícil concluir la relación exacta entre los salarios y las variables utilizadas, dado que no es posible controlar por otras variables explicativas, lo que podría ocultar los efectos que otras variables pueden estar teniendo sobre los salarios y que, además de no ser capturados por las variables edad y educación, podrían incluso estar correlacionados con la edad o la educación.

- (e) Un problema de las gráficas `lowess` y `lpoly` es que no controlan por ninguna variable explicativa. Existe análisis semiparamétrico que permite controlar por variables explicativas. (Puedes leer el libro de Cameron y Trivedi para una mayor guía.)

- Explica qué es análisis semiparamétrico.

Respuesta. El análisis semiparamétrico combina componentes paramétricos y componentes no paramétricos. Algunos modelos que implementan este enfoque son:

- Modelo parcialmente lineal: $\mathbb{E}[y|\mathbf{x}, \mathbf{z}] = \mathbf{x}'\beta + \lambda(\mathbf{z})$
- Modelo parcialmente lineal generalizado: $\mathbb{E}[y|\mathbf{x}, \mathbf{z}] = g(\mathbf{x}'\beta + \lambda(\mathbf{z}))$

En ambos modelos, la parte paramétrica es β , mientras que las no paramétricas son $g()$ y $\lambda(\mathbf{z})$

- Asimismo, si tomas una regresión de salario vs variables explicativas, y de edad vs variables explicativas, tomas los residuales y graficas con `lowess` y `lpoly` podrías solucionar ese problema. Haz ese cálculo para 2018 y 2022, explica la intuición del mecanismo. Discute tus resultados. Es decir, grafica la relación entre salario y edad, controlando por diferentes variables explicativas.

Respuesta. Controlando por los años de escolaridad y la variable rural, la idea general del mecanismo es la siguiente:

- Con la regresión de salario Y vs variables explicativas Z ("regresión 1") podemos calcular la parte de Y no explicada por Z, es decir, los residuales de la regresión 1. Después, con la regresión de edad X vs Z ("regresión 2", obtenemos la parte de X no explicada por Z; i.e. los residuales de la regresión 2.

Teniendo esto, podemos graficar el salario (Y) no explicado por Z (los residuales de la regresión 1) contra la edad (X) no explicada por Z, lo que nos permitiría obtener, mediante la regresión `lowess` o `lpoly`, una regresión del salario (Y) vs la edad (X) controlando por variables explicativas (Z), solucionando el problema planteado al principio.

Más formalmente, el mecanismo anterior se basa en el siguiente modelo parcialmente lineal, basado en la propuesta de Robinson (1988):

$$y = x'\beta + \lambda(z) + u,$$

Por lo tanto, tomando esperanzas condicionales en z

$$E[y|z] = E[x|z]'\beta + \lambda(z)$$

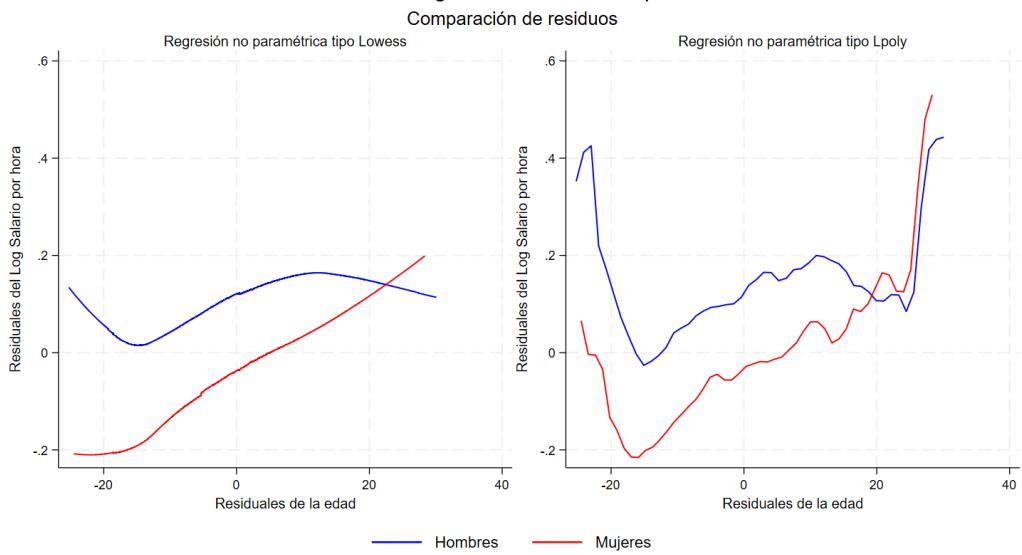
Puesto que si asumimos media condicional cero $E[u|x, z] = 0$ implica que $E[u|z] = 0$. Así, si restamos el resultado anterior del modelo original obtenemos

$$y - E[y|z] = (x - E[x|z])'\beta + u$$

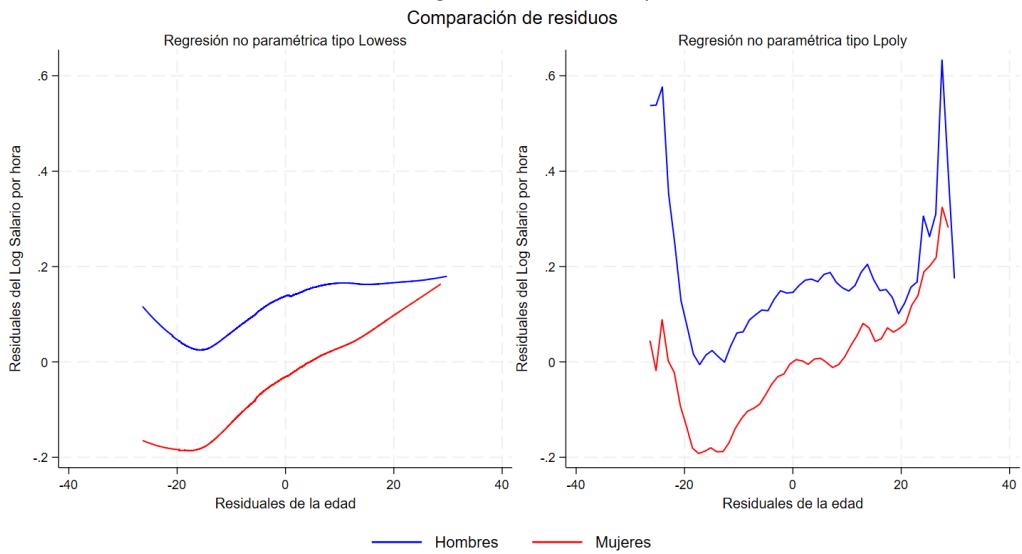
Suponer una forma lineal para $E[y|z]$ y $E[x|z]$ y estimar por MCO, equivale al proceso planteado en este ejercicio de la tarea, si después estimamos β mediante `lowess` o `lpoly`.

Gráficos en la siguiente página.

2018: Residuos del logaritmo del salario por hora vs edad

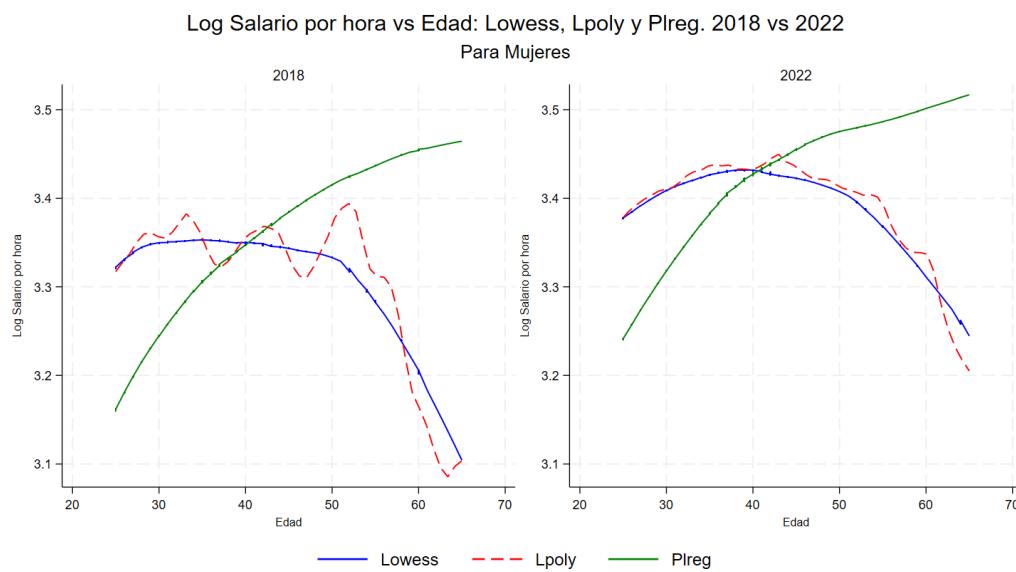
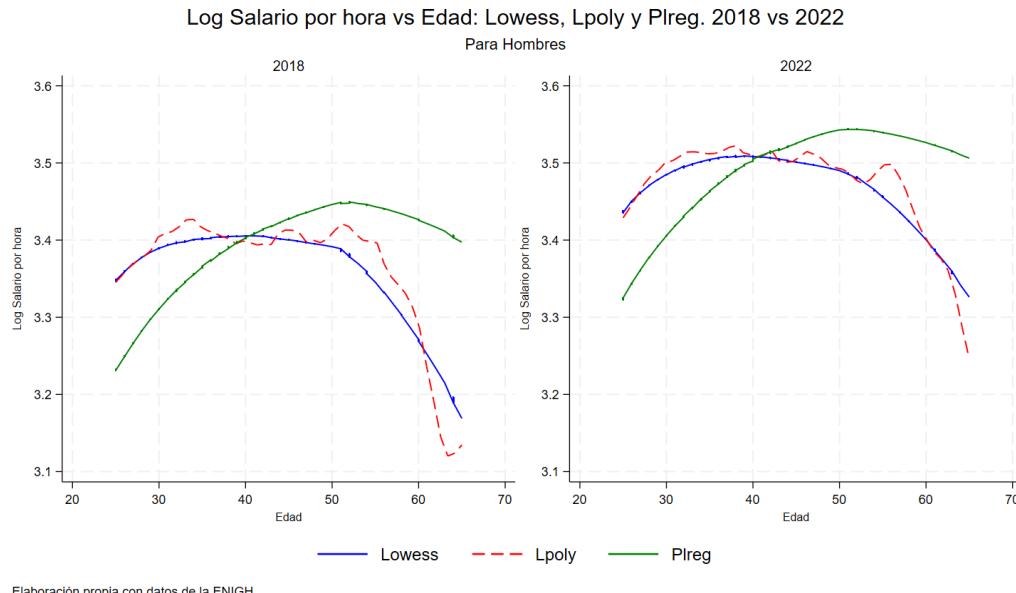


2022: Residuos del logaritmo del salario por hora vs edad



Los resultados son similares a los hallados sin controlar por variables explicativas, aunque en una magnitud menor: las brechas, una vez controlando por escolaridad y una variable rural, siguen manteniéndose, tanto por `lowess` como por `lpoly`. A pesar de que no es posible distinguir la edad en el eje X, pareciera que la brecha es más amplia en los trabajadores más jóvenes, si comparamos con los resultados de los incisos b) y c). Algo interesante es que, una vez que controlamos por los años de educación, vemos que los retornos de la edad ya no tienen una tendencia estrictamente creciente: al nacer y crecer son altos, después se reducen y vuelven a crecer hasta alcanzar un punto máximo. Esto es el caso con `lowess`, pero no con `lpoly`

- iii. Baja el `ado file plreg`, checa el `help`, y compara tus gráficas y resultados anteriores con el resultado de ese comando también. Discute.

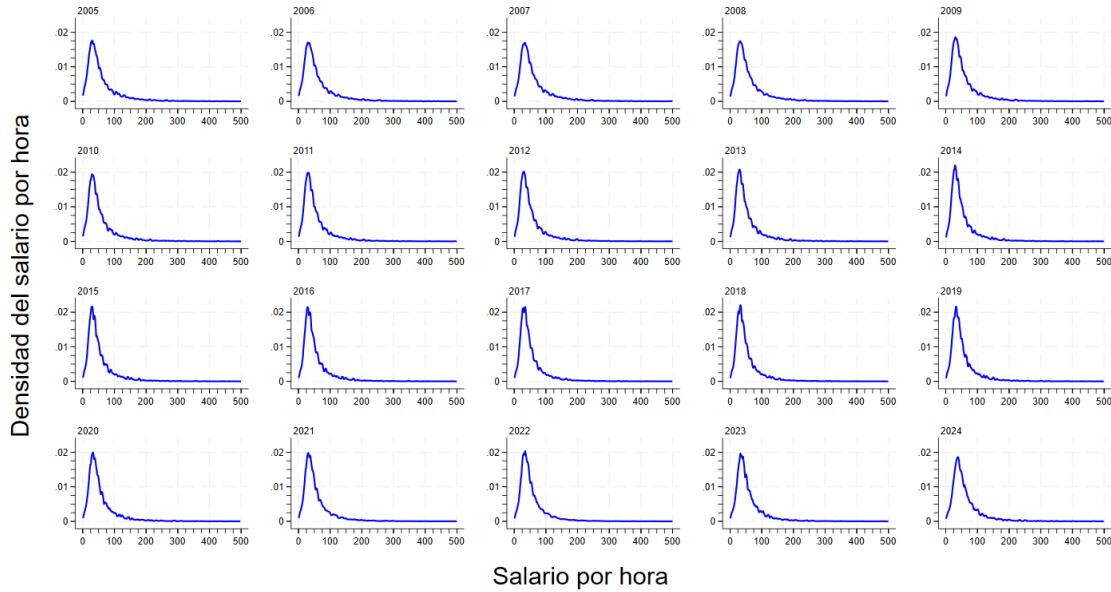


Los resultados de `plreg`, los cuales son estimados mediante un modelo parcialmente lineal de Yatechew², guardan un parecido a los estimados con el método anterior: esto se debe a que ambos parten de la misma familia de modelos, con la diferencia de que Yatechew estima $E[y|z]$ y $E[x|z]$ usando métodos no paramétricos, en vez de MCO como hicimos nosotros.

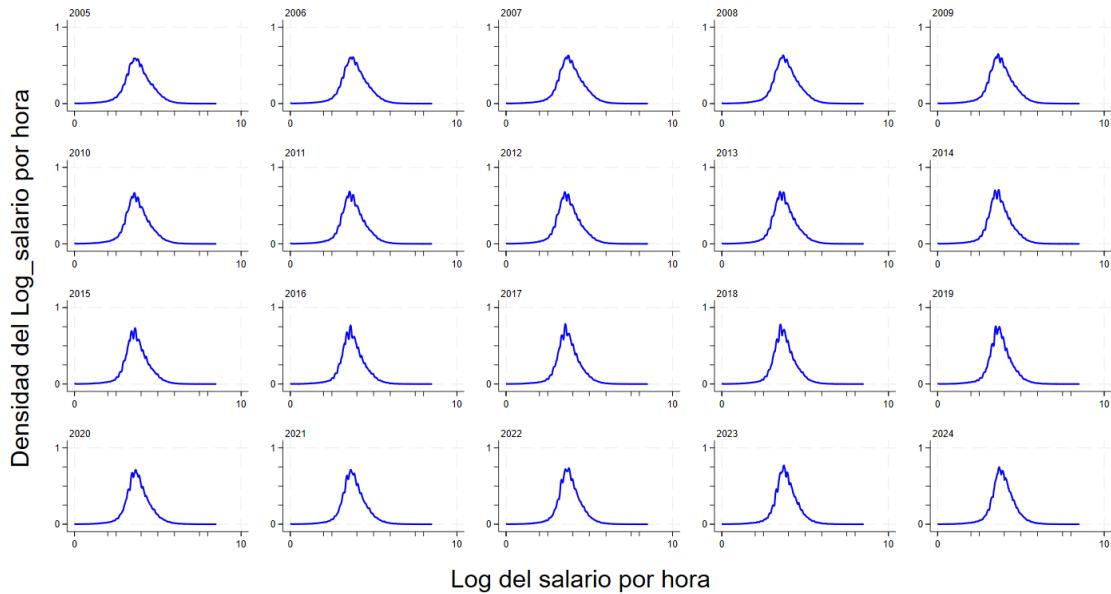
²Véase https://www.stata.com/meeting/uk13/abstracts/materials/uk13_verardi.pdf

6. Repite el ejercicio con ENOE, y discute la comparabilidad entre ENIGH y ENOE.

Distribución del salario por hora por año, 2005 - 2024

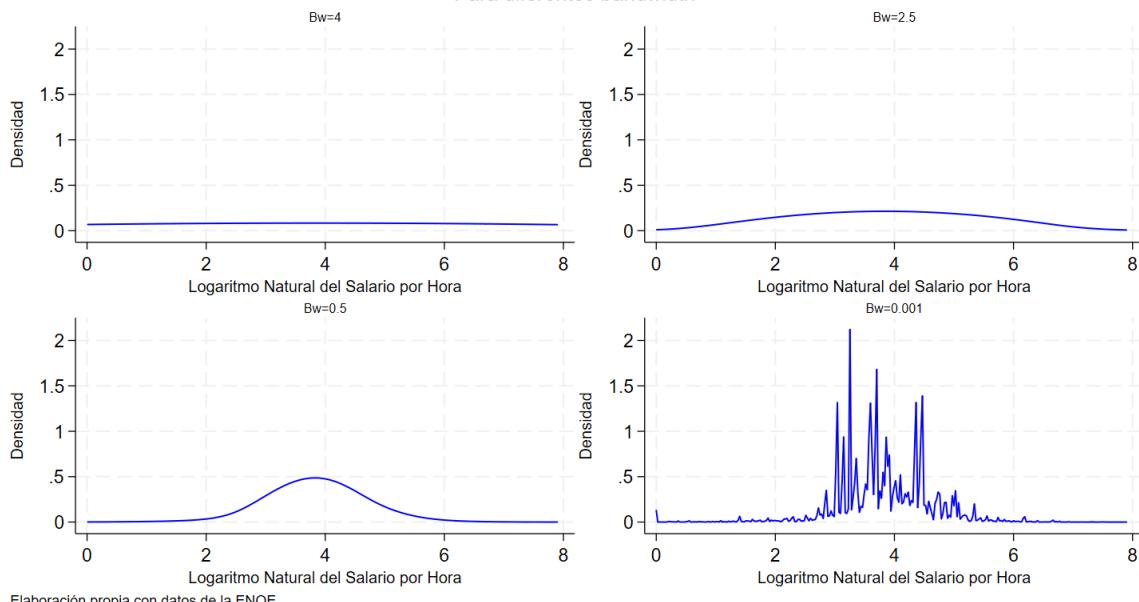


Distribución del log_salario por hora por año, 2005 - 2024



Distribución no paramétrica del salario por hora, 2023

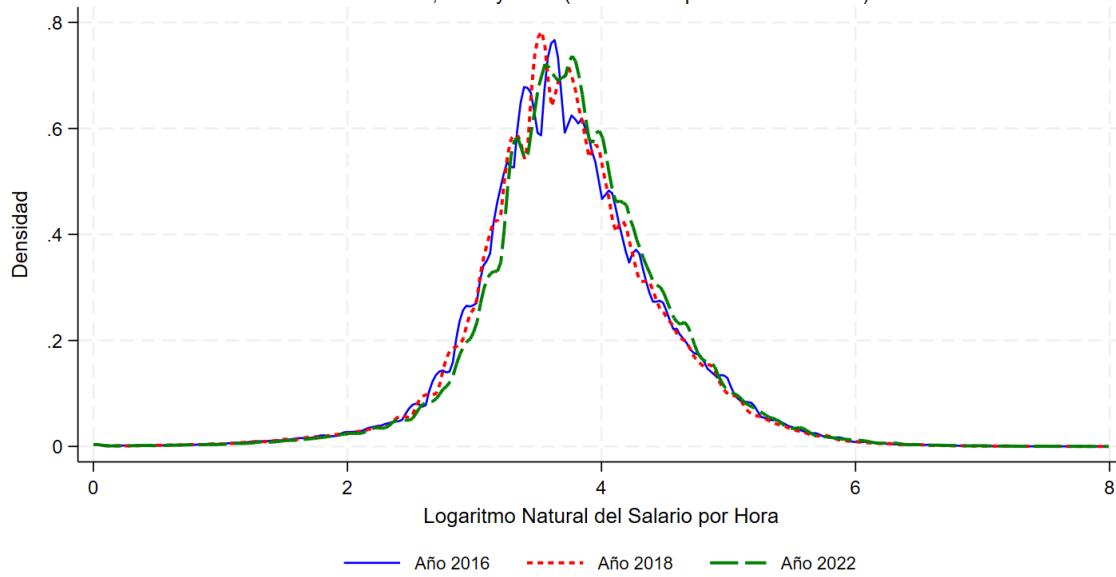
Para diferentes bandwidth



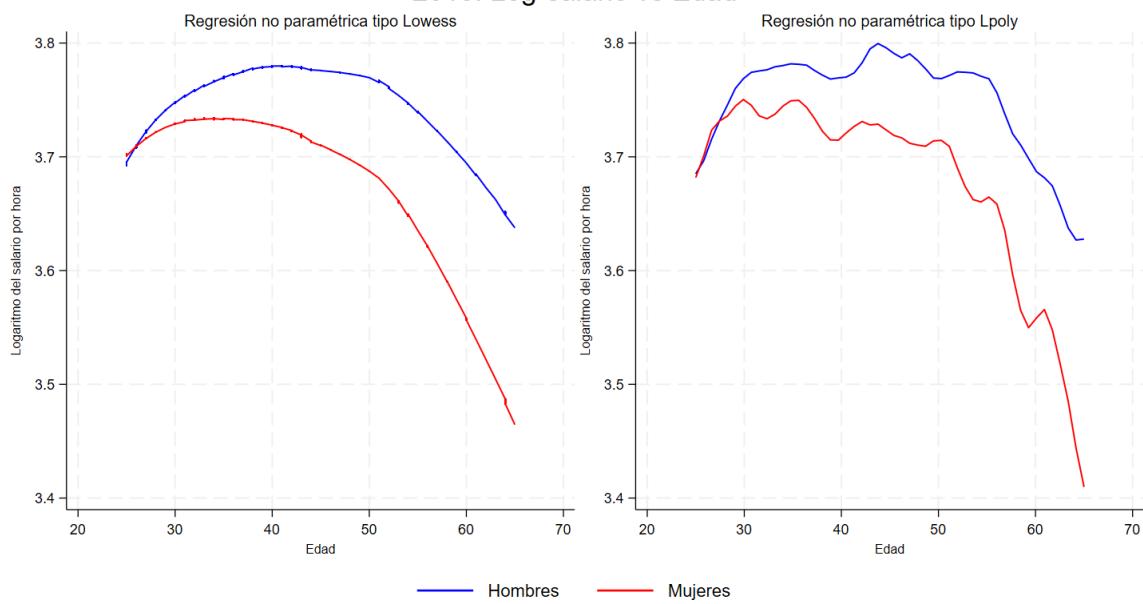
Elaboración propia con datos de la ENOE

Distribución no paramétrica del Logaritmo del Salario por Hora

Años 2016, 2018 y 2022 (Bandwidth óptimo de Silverman)

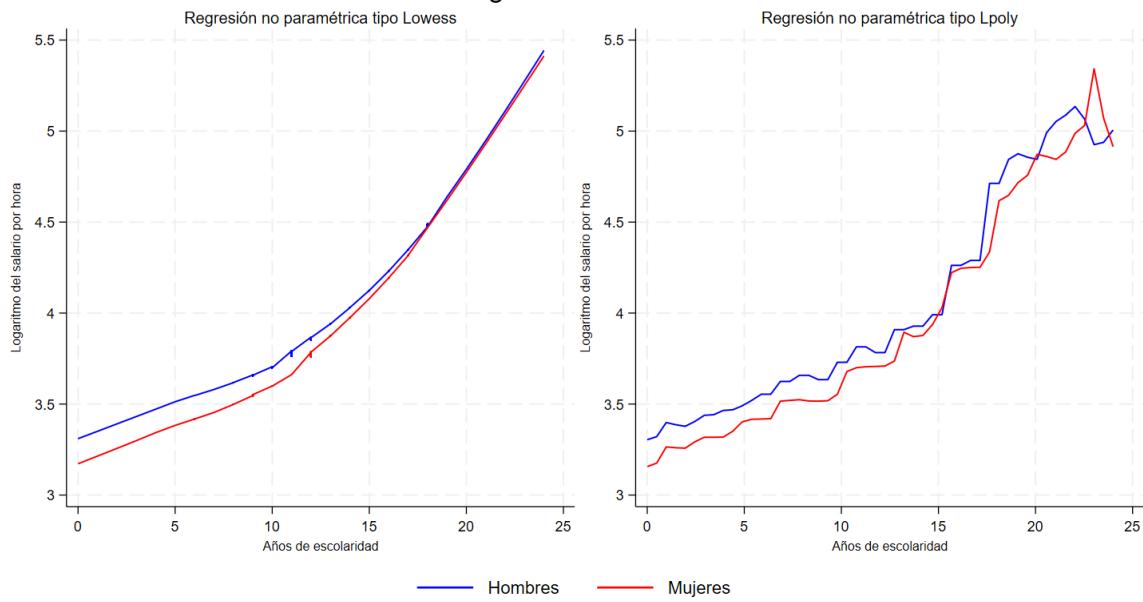


2018: Log-salario vs Edad



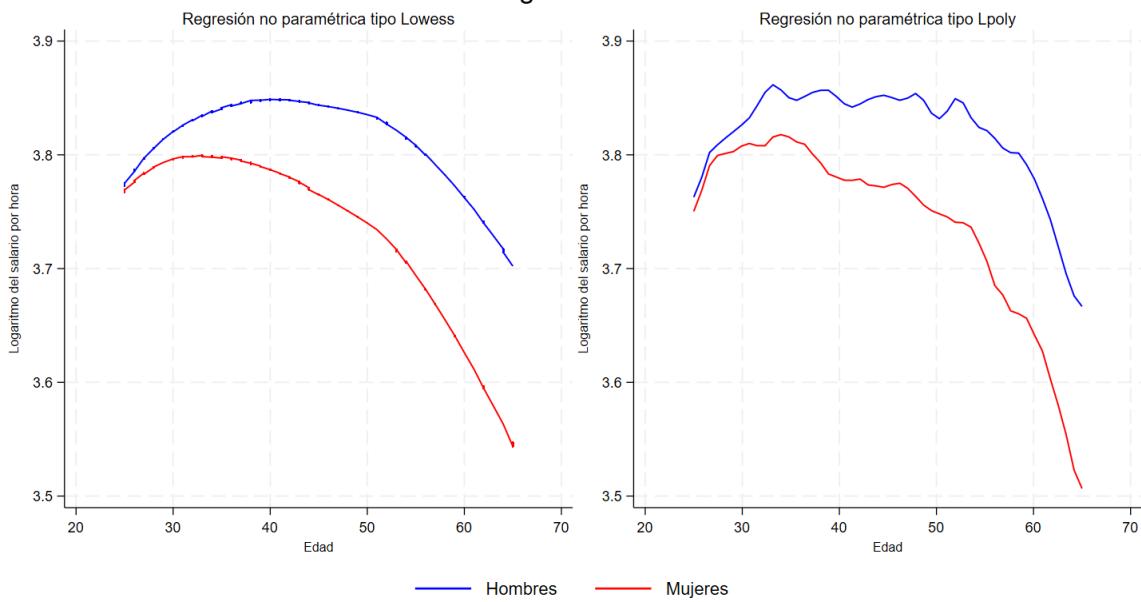
Elaboración propia con datos de la ENOE

2018: Log-salario vs Educación



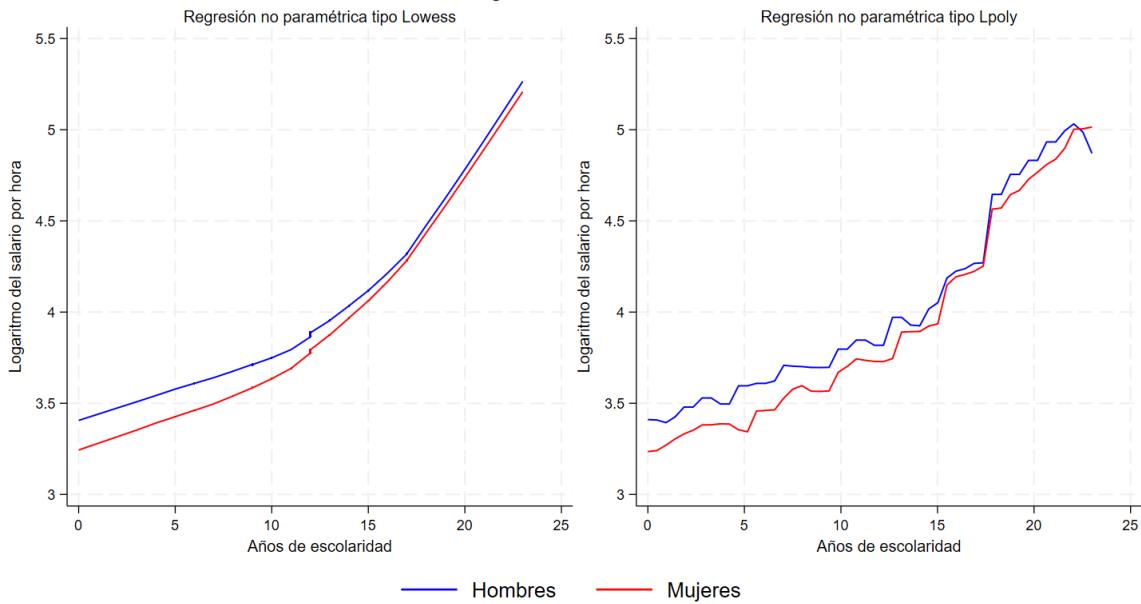
Elaboración propia con datos de la ENOE

2022: Log-salario vs Edad



Elaboración propia con datos de la ENOE

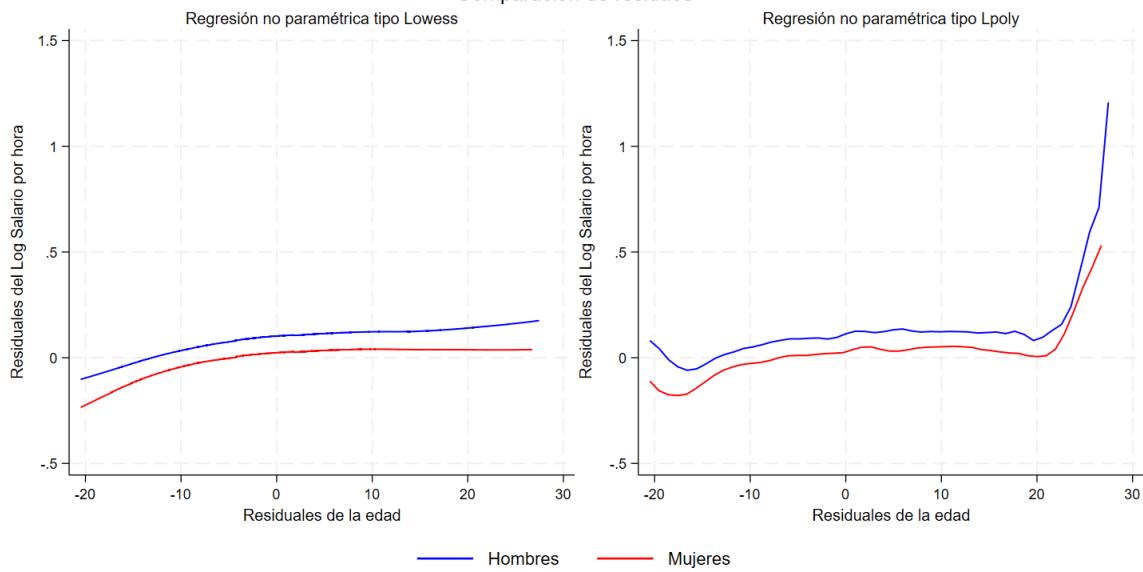
2022: Log-salario vs Escolaridad



Elaboración propia con datos de la ENOE

2018: Residuos del logaritmo del salario por hora vs edad

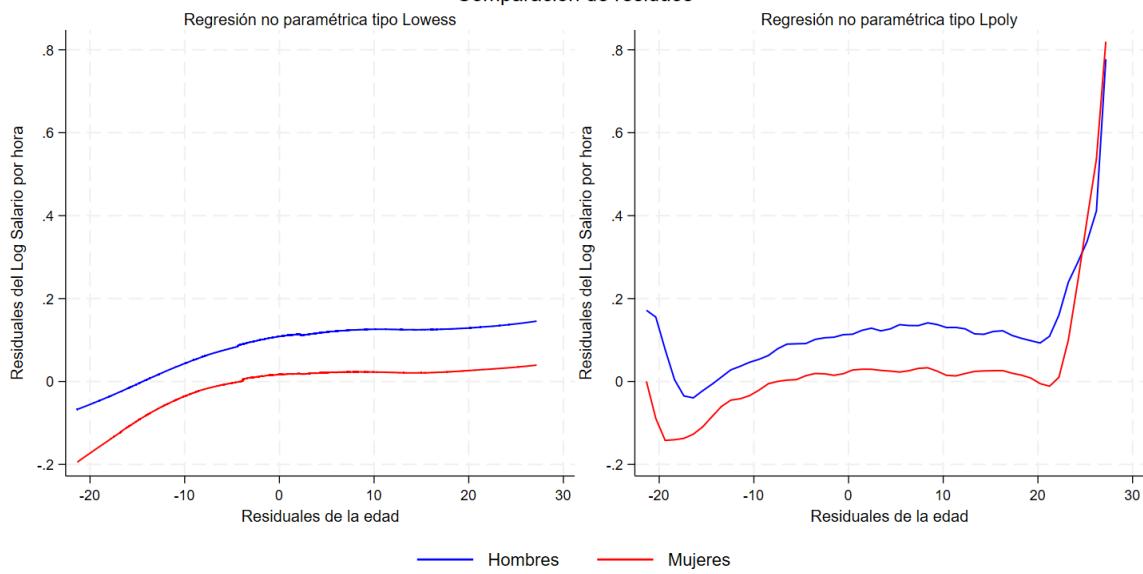
Comparación de residuos



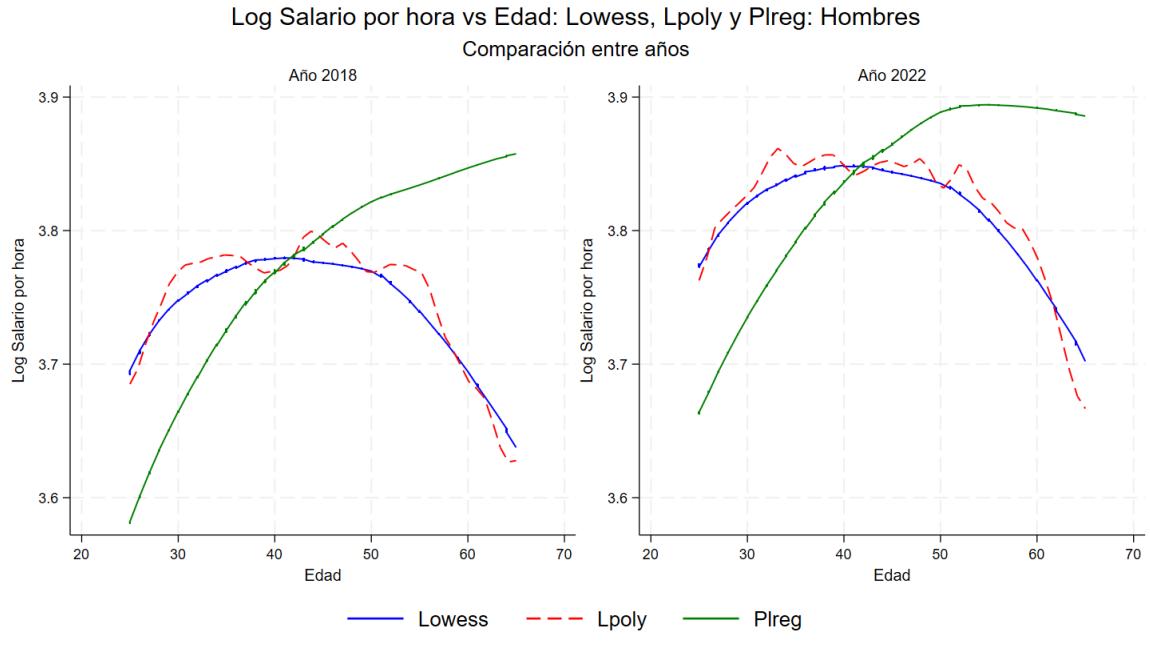
Elaboración propia con datos de la ENOE

2022: Residuos del logaritmo del salario por hora vs edad

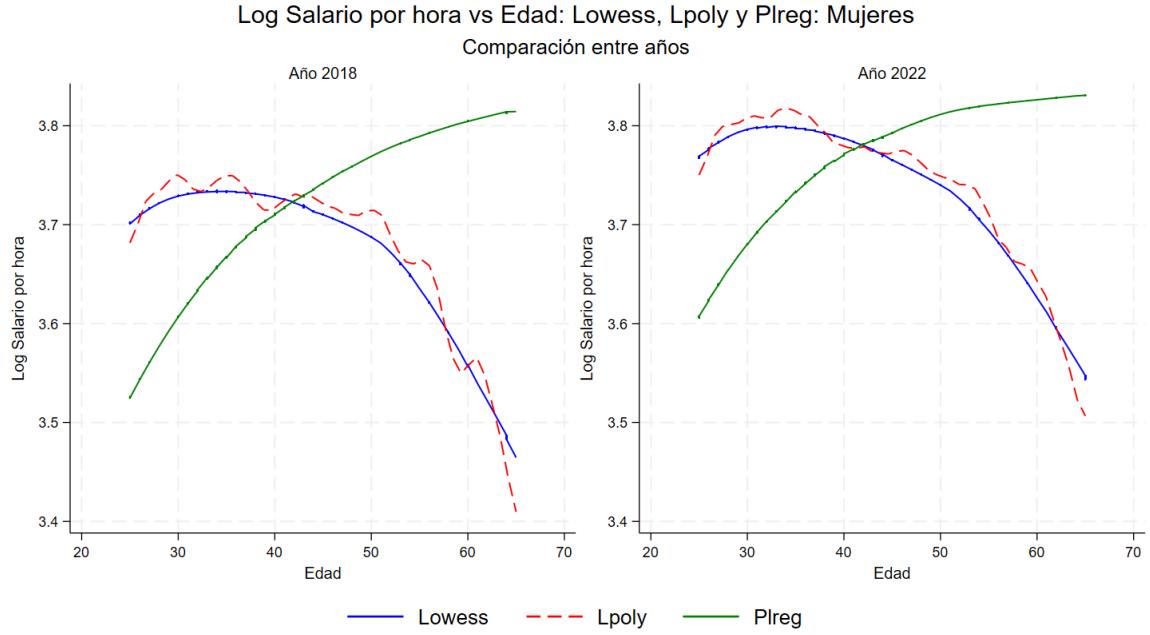
Comparación de residuos



Elaboración propia con datos de la ENOE



Elaboración propia con datos de la ENOE



Elaboración propia con datos de la ENOE

Los resultados son muy similares, con el detalle de que la ENOE estima salarios más altos que la ENIGH, en promedio. Esto se refleja en el valor alrededor del cual se comprimen las densidades salariales, y en los mayores niveles tanto para hombres y mujeres en las gráficas del salario contra la edad y la educación. Una diferencia interesante es que la ENOE no muestra una brecha de género tan amplia en todas las edades como la ENIGH: en la ENOE, el log-salario de los jóvenes es prácticamente igual entre sexos, mientras que la ENIGH sí tiene

una brecha importante. La razón de esto puede ser un problema con los datos, aunque habría que hacer un análisis más profundo. Finalmente, los resultados de los modelos parcialmente lineales, tanto con el método del residuo como con `plreg`, son similares en términos de los hallazgos cualitativos, con la excepción de, nuevamente, un comportamiento distinto en los jóvenes.

6 Problema 6: Imputación

1. Lee el artículo de Campos Vázquez (2013) en la revista de *Ensayos de Economía* de la UANL y el apéndice del artículo de Levy y López Calva (2016). Enlace. Resume rápidamente el primer artículo, los métodos de imputación utilizados, y explica un método de imputación adicional que no sea explicado en el artículo (los artículos de Rubin en la literatura o su libro te pudieran ayudar).

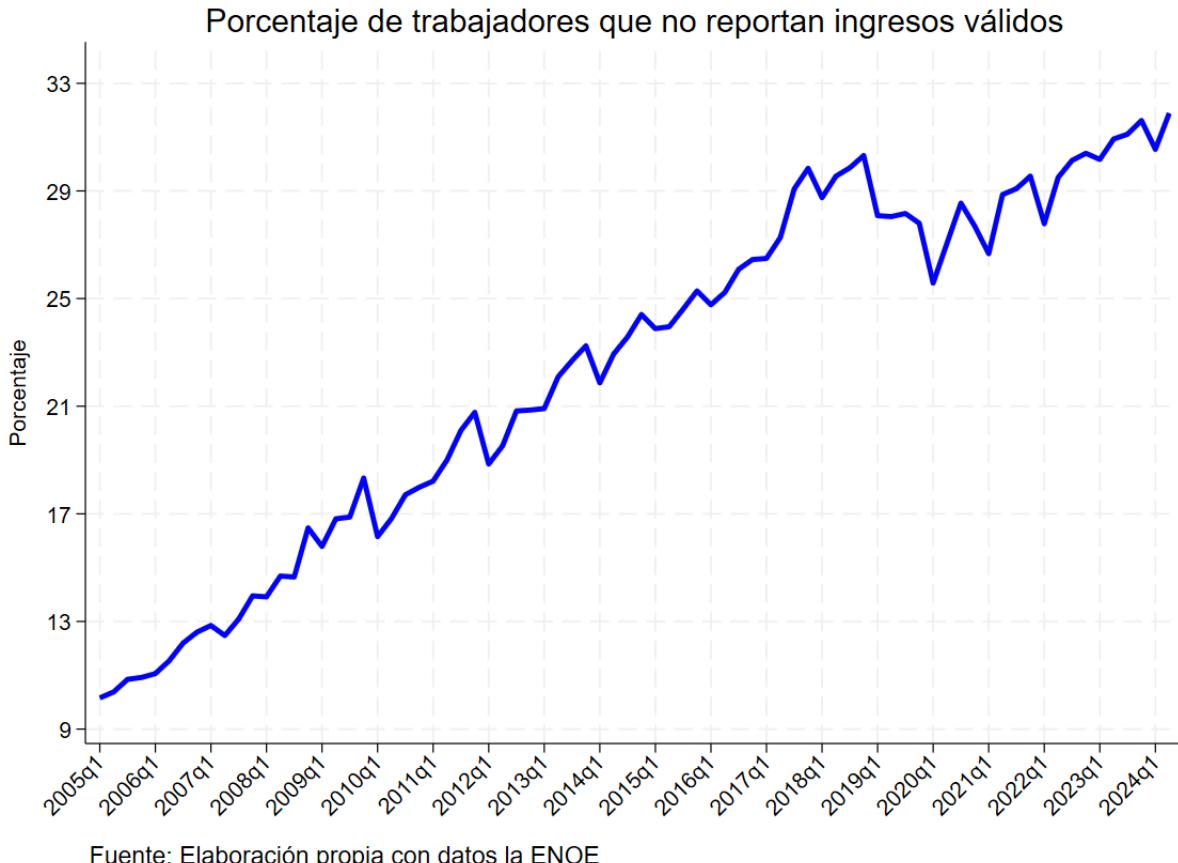
Respuesta. El artículo de Campos Vázquez (2013) examina la corrección de ingresos inválidos por medio de métodos de imputación y analiza los efectos de dichos ingresos laborales no reportados en la medición de la pobreza en México. En los últimos años, la proporción de trabajadores que no declara ingresos ha aumentado considerablemente, lo que puede distorsionar las estadísticas sobre pobreza y desigualdad si no se corrige. El autor utiliza diversos métodos de imputación para corregir estos ingresos no reportados y demostrar que el impacto sobre la medición de la pobreza es significativo, lo que sugiere que las estadísticas deberían incluir una corrección por ingresos no reportados.

Por otro lado, en el artículo se emplean cuatro métodos de imputación:

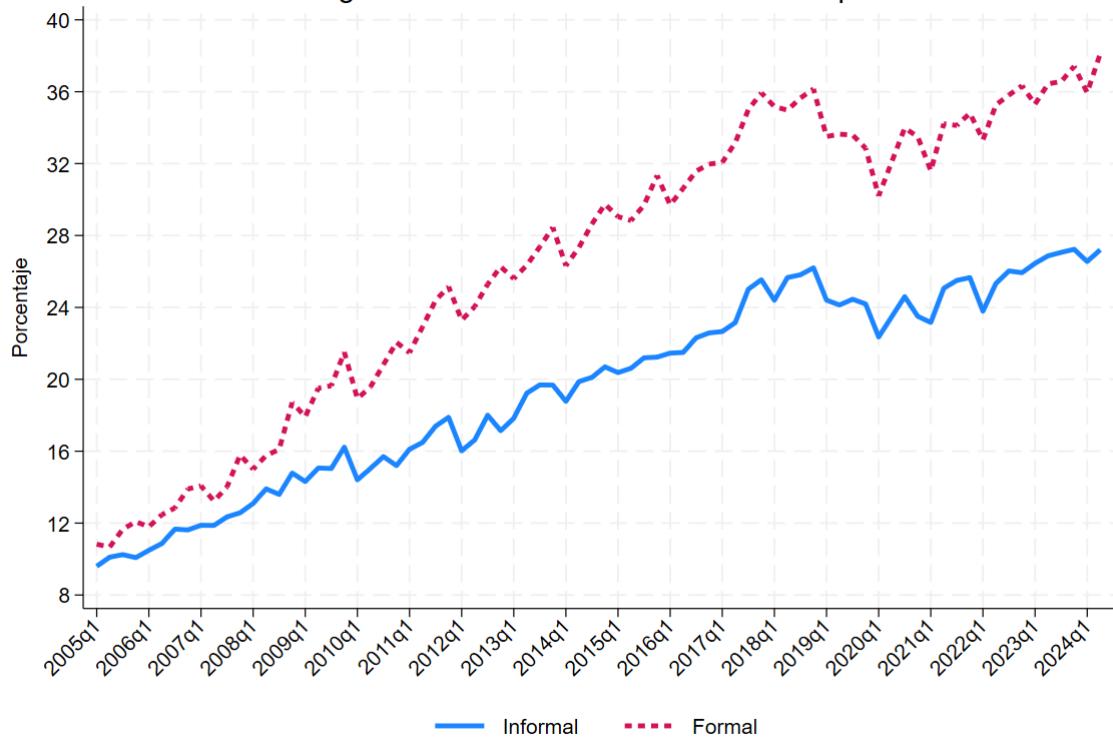
- **Pareamiento por puntajes de propensión.** Este método se realiza en dos pasos: (1) se calcula un puntaje de propensión, es decir, estimamos $\mathbb{P}[Y_i \text{ sea faltante}]$ mediante la estimación de un modelo de probabilidad donde se asumen formas funcionales logísticas o normales; y (2) se calcula la probabilidad predicha o el puntaje de propensión para todo i . Las variables explicativas en el modelo tienen que incluir aquellas que determinan la probabilidad de ingreso faltante.
- **Método Hot-Deck.** Este método consiste en reemplazar datos faltantes con valores observados de forma aleatoria. Se divide la muestra en g grupos según las variables que consideremos adecuado, dentro de cada grupo se divide entre ingresos válidos e inválidos, y después, para cada grupo g , hacemos: (1) remuestreamos mediante bootstrap y obtenemos una muestra de los ingresos válidos con el mismo tamaño N_0 que tiene el g ; (2) de esta muestra N_0 , tomamos una muestra aleatoria de tamaño N_1 donde N_1 es el tamaño de la submuestra con ingresos inválidos de g ; y (3) utilizamos los ingresos de la muestra N_1 para imputarlos en las observaciones con ingresos inválidos de g .
- **Por grupos con aleatoriedad.** Este método es una combinación de los anteriores. En este método, en g se calcula la mediana (no se utiliza la media para evitar problemas de sensibilidad a valores exageradamente grandes). Para evitar el problema de subestimar la verdadera varianza del ingreso, se le suma a esa mediana la desviación estándar observada de ese grupo, multiplicada por una variable que se distribuye normal estándar.
- **Pareamiento por promedios predictivos.** Si se tienen variables continuas se tendrían que categorizar en grupos. Si al hacer esto, se pierde información valiosa, se podrían

tener consecuencias en la medición del ingreso imputado. Para evitar este problema, este método realiza lo siguiente: (1) estimamos una regresión de ingresos observados contra explicativas mediante MCO y obtenemos los coeficientes ($\hat{\mathbf{b}}$) y la varianza residual ($\hat{\sigma}_u^2$); (2) obtenemos un parámetro de varianza s_1^2 de forma aleatoria tal que $s_1^2 = (n_0 - k) \frac{\hat{\sigma}_u^2}{z}$ donde $z \sim \chi_{n_0-k}^2$; (3) obtenemos nuevos coeficientes \mathbf{b}_1 de su distribución normal $\mathcal{N}(\hat{\mathbf{b}}, s_1^2(\mathbf{X}'\mathbf{X})^{-1})$; (4) con \mathbf{b}_1 hacemos predicciones del ingreso tanto de los que reportan como de los que no; y (5) realizamos un pareamiento de aquéllos sin ingreso con su vecino más cercano que sí reporta ingreso, utilizando la métrica estimada en 4 (ingreso predicho), y a los primeros les imputamos el ingreso observado de sus vecinos.

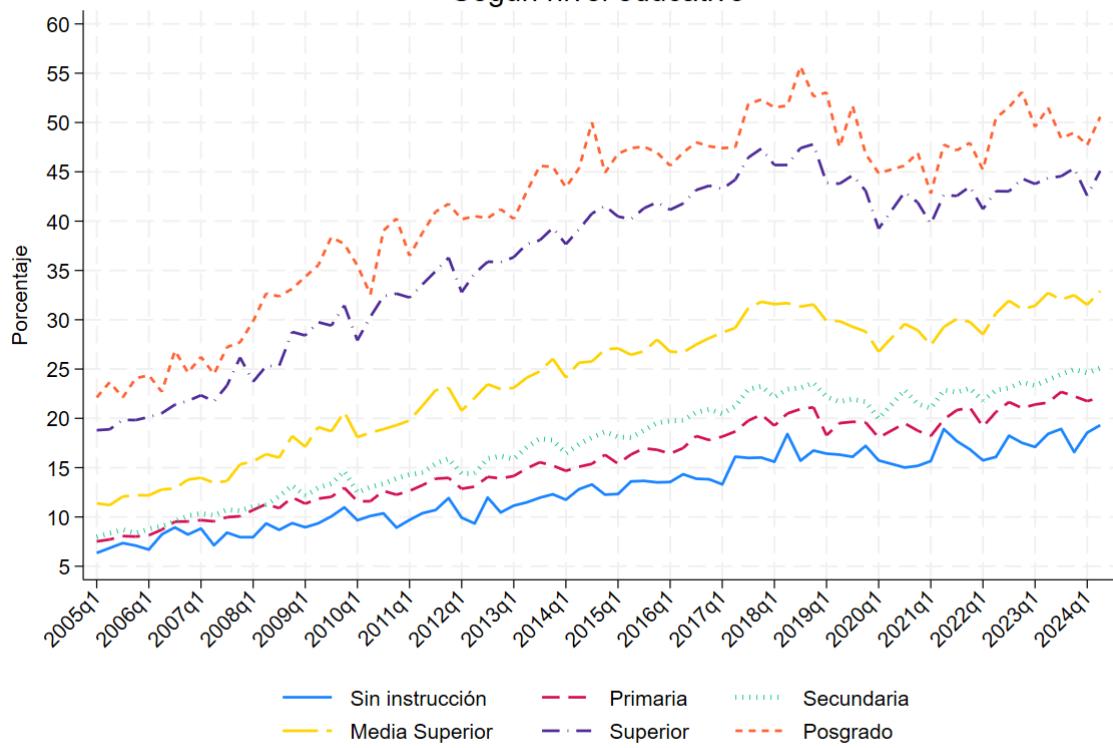
2. Para la ENOE 2005-2024, ¿cómo cambia el % de trabajadores que no reportan ingresos (pero sí son trabajadores)? Realiza análisis por grupo: formales vs informales; grupo educación; sexo; etc. ¿El cambio en el no reporte podrías considerarlo aleatorio, explica? ¿Cuántos o % no contestan la pregunta de ingreso pero sí contestan la pregunta de rangos de SM? ¿Cómo esa % ha cambiado en el tiempo del total que no contesta la pregunta?



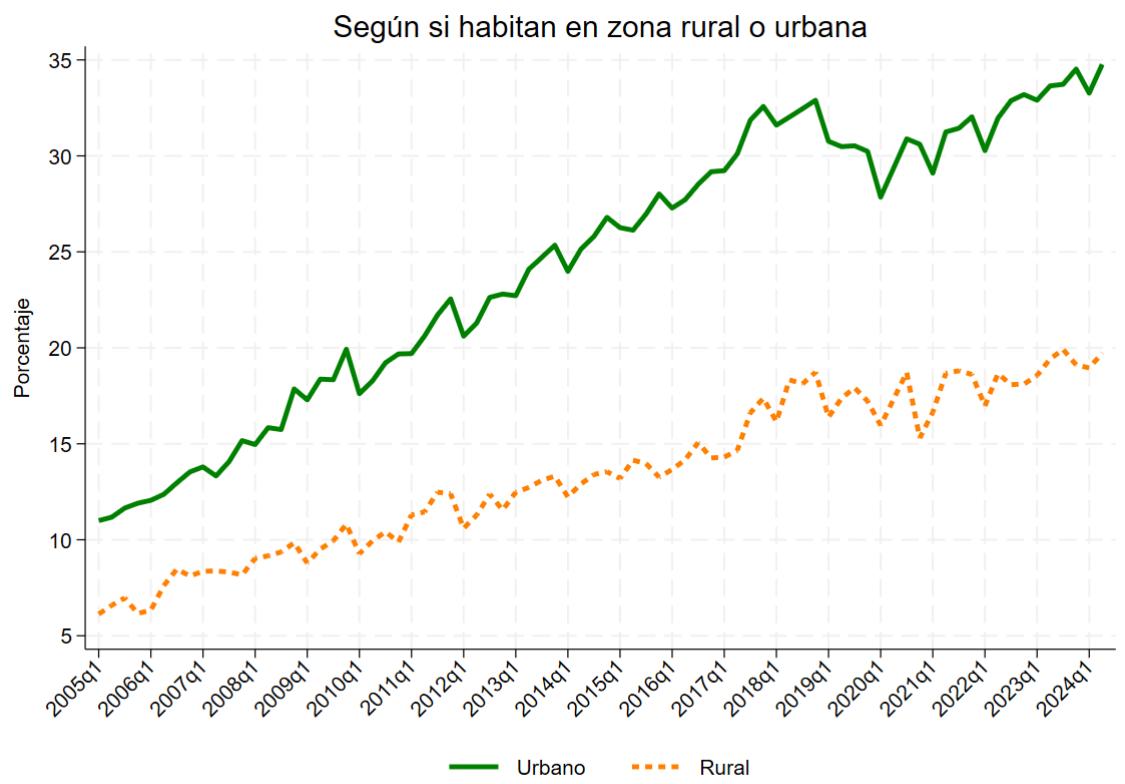
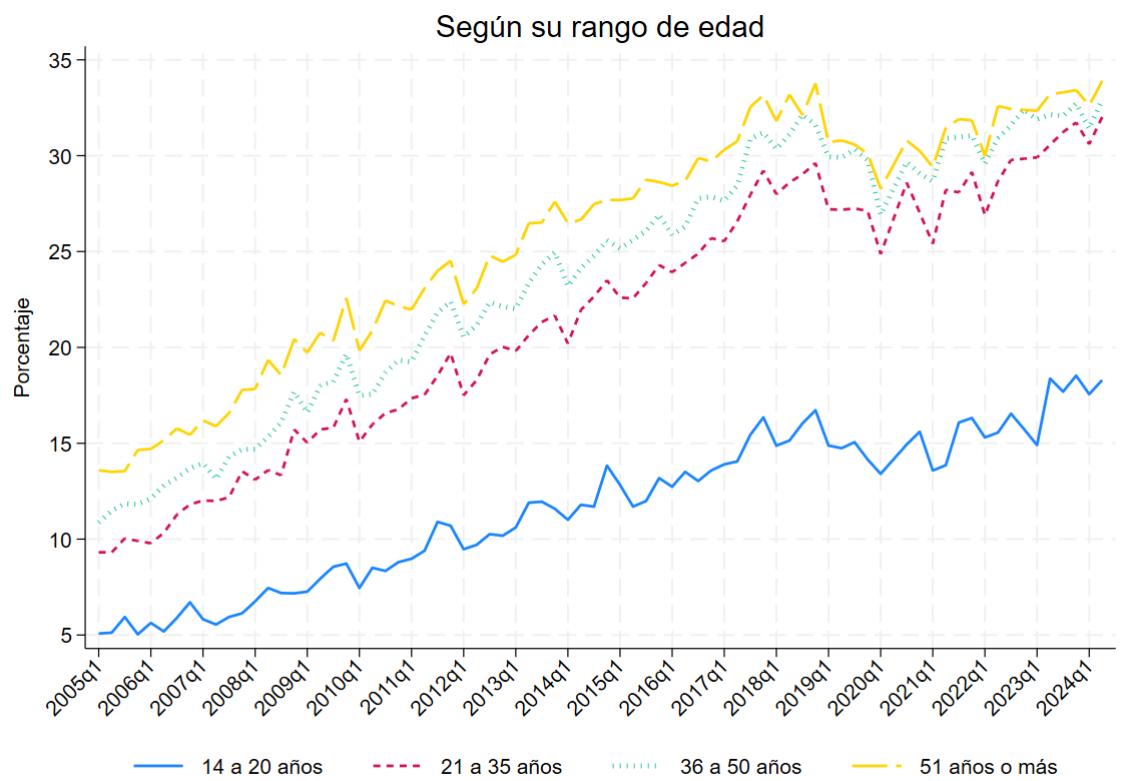
Según condición de formalidad en el empleo



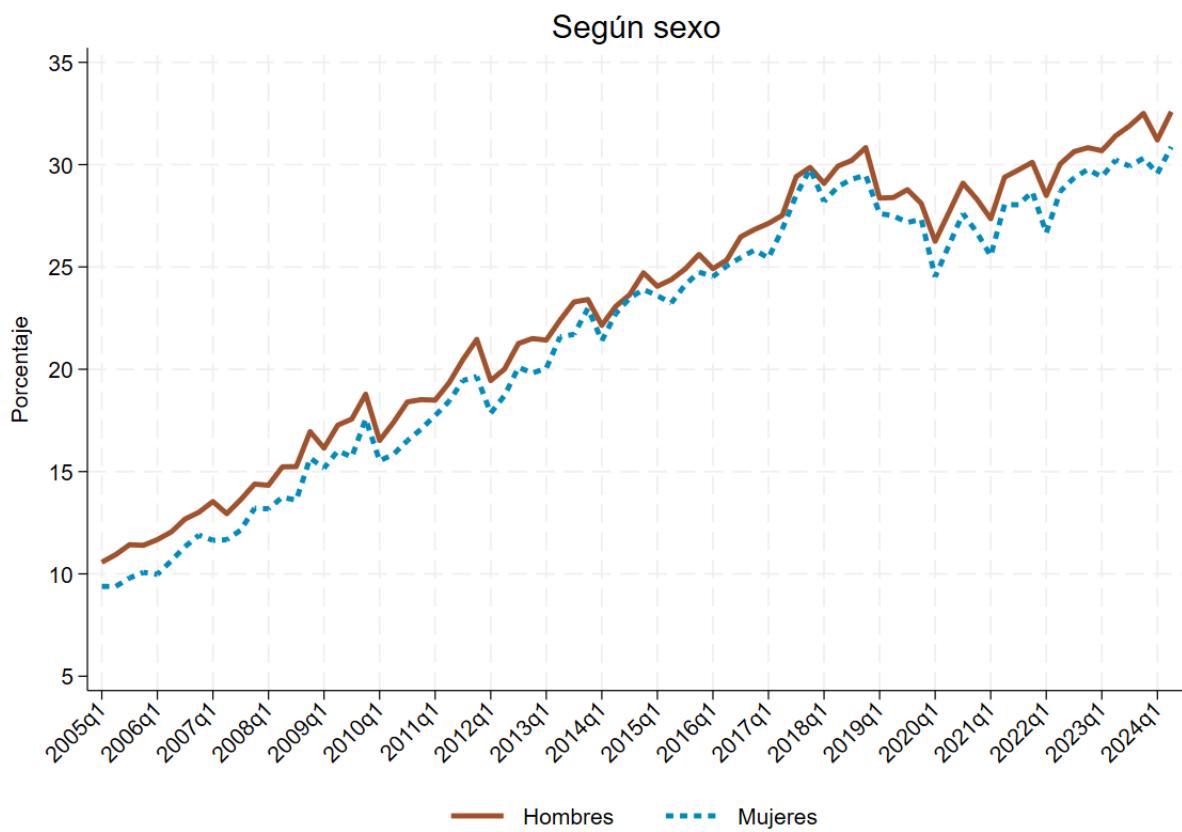
Según nivel educativo



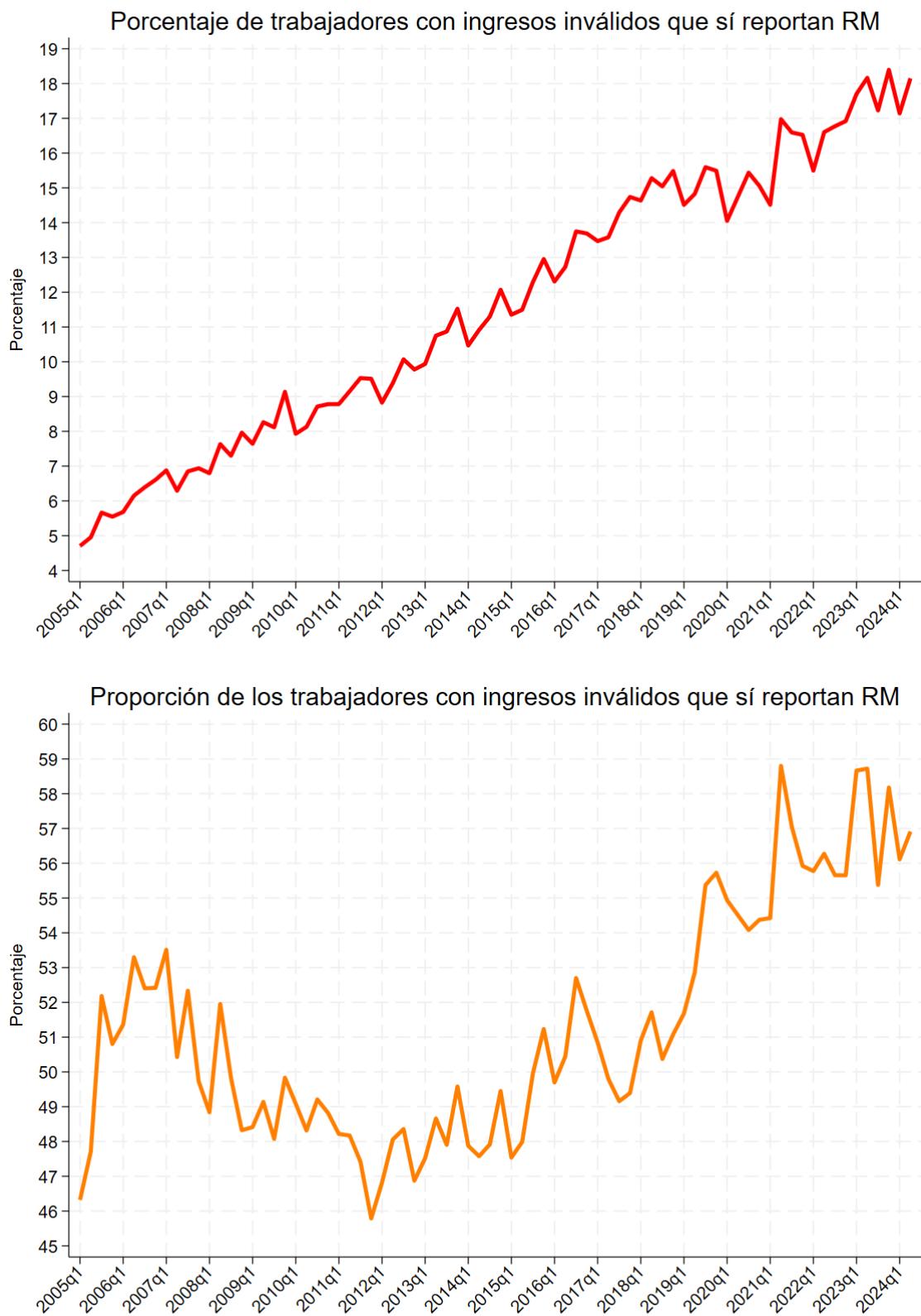
Fuente: Elaboración propia con datos la ENOE



Fuente: Elaboración propia con datos la ENOE



Espacio en blanco. Gráficas continúan en la siguiente página.



Fuente: Elaboración propia con datos la ENOE

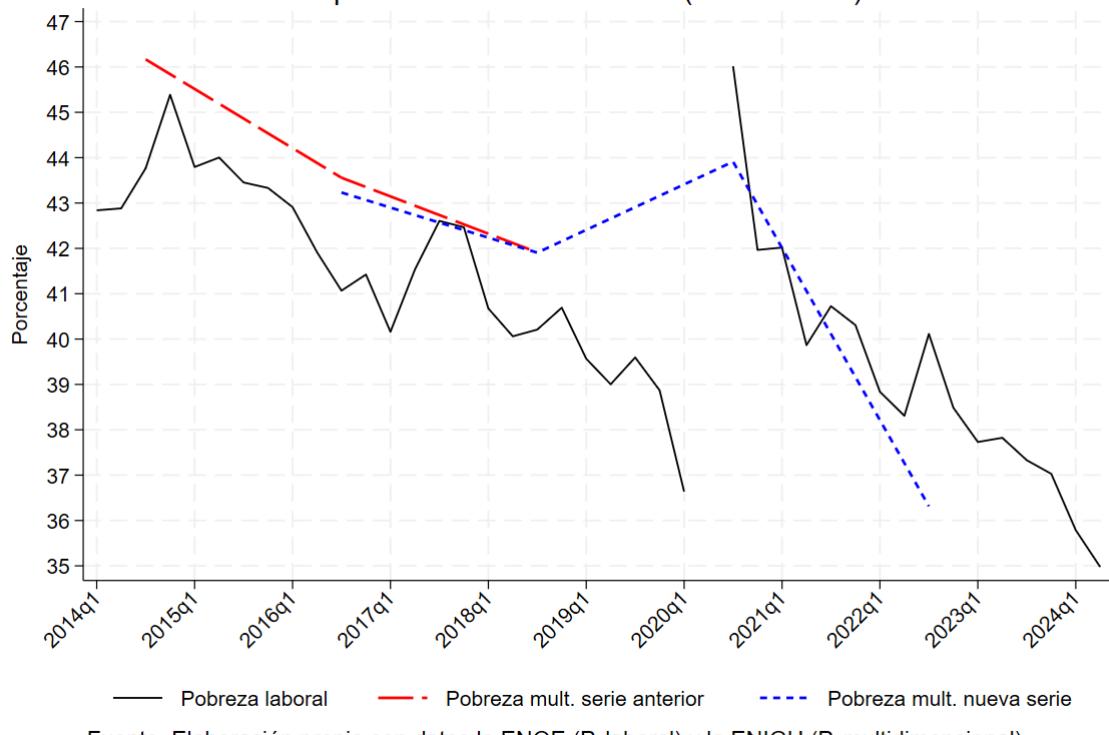
Las gráficas muestran que los trabajadores formales tienden a no reportar sus ingresos en una medida creciente, si comparamos con los informales; lo mismo ocurre con las personas que tienen educación superior y que cuentan con un posgrado. En cuanto a la edad, la creciente brecha entre grupos de edad con respecto al porcentaje de ingresos inválidos, indica que las personas mayores a 20 años cada vez más evitan reportar sus ingresos, en comparación con las más jóvenes. Adicionalmente, los trabajadores de zonas urbanas son más propensos a no reportar en comparación con los que laboran en zonas rurales. Por último, no parece existir diferencias entre géneros; no obstante, todo lo demás sugiere que la evolución de la composición de trabajadores con ingresos inválidos no es aleatoria, dado que tienden a poseer ciertas características: son trabajadores formales, con educación superior, mayores de 20 años y viven en zonas urbanas.

Finalmente, como observaremos en las gráficas de la siguiente página, con respecto al total de trabajadores, el porcentaje con ingresos inválidos que sí reportan rango de salario mínimo ha tenido una tendencia al alza, con un estancamiento entre 2019 y 2021 para después retomar el crecimiento (lo cual es bueno). Por otro lado, la proporción de los trabajadores que reportan RM con respecto a los que no reportan ingresos válidos, ha tenido un comportamiento distinto: el primer ciclo corresponde al sexenio de 2006-2012, donde hay un crecimiento, un pico y una caída, mientras que el segundo a los dos sexenios posteriores, el cual aún continúa al alza. **Idealmente**, la segunda gráfica debería acercarse al 100%, de tal modo que, aunque existan trabajadores con ingresos inválidos, sea posible ubicarlos dentro de algún grupo de rangos de salarios mínimos, lo cual podría facilitar su imputación.

3. Replica el cálculo de pobreza laboral de CONEVAL con ENOE 2005-2024. Los programas están disponibles en la red. Asegúrate de entender los supuestos que realizan para los que no declaran ingresos. Descarga la serie de tiempo de pobreza multidimensional usando ENIGH para los últimos 10 años. Realiza dos figuras y dos mapas: una en serie de tiempo con dos líneas de pobreza, ENOE y ENIGH, y otra donde sea **scatter** con ENOE en eje x y ENIGH en eje y y contrastes la correlación entre ambas. Los mapas usa R para calcular la pobreza a nivel entidad federativa para el último año.

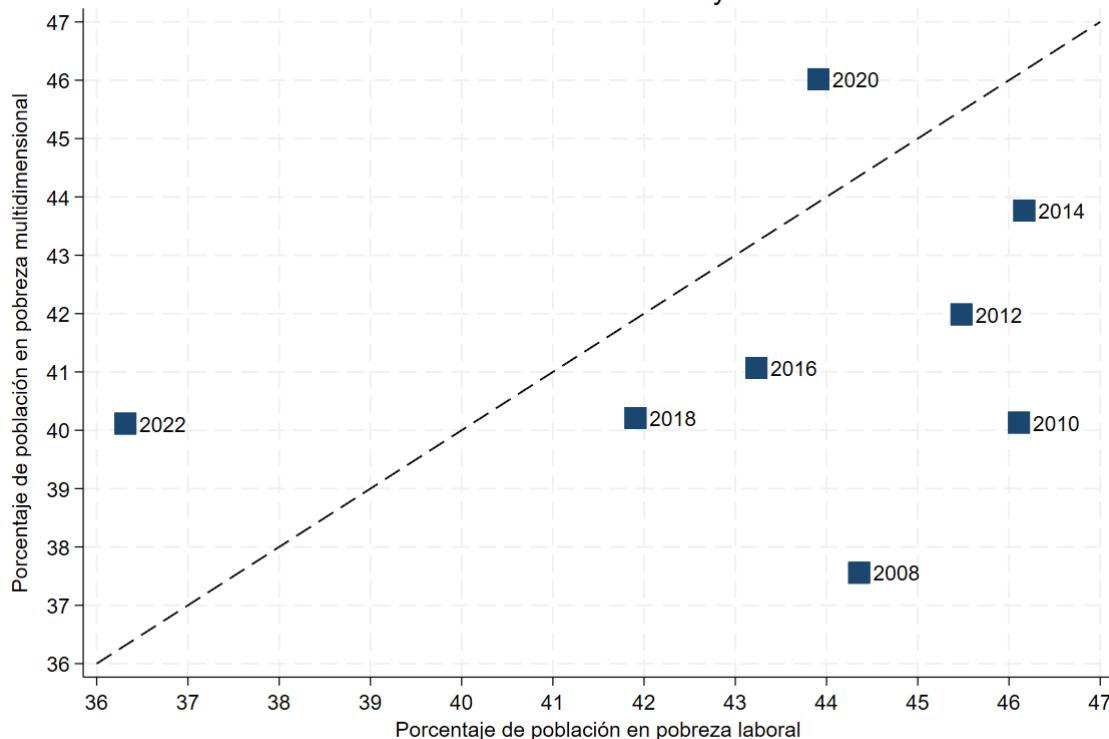
Figuras en la siguiente página.

Medición de la pobreza laboral y pobreza multidimensional (2014 - 2024)



Fuente: Elaboración propia con datos la ENOE (P. laboral) y la ENIGH (P. multidimensional)

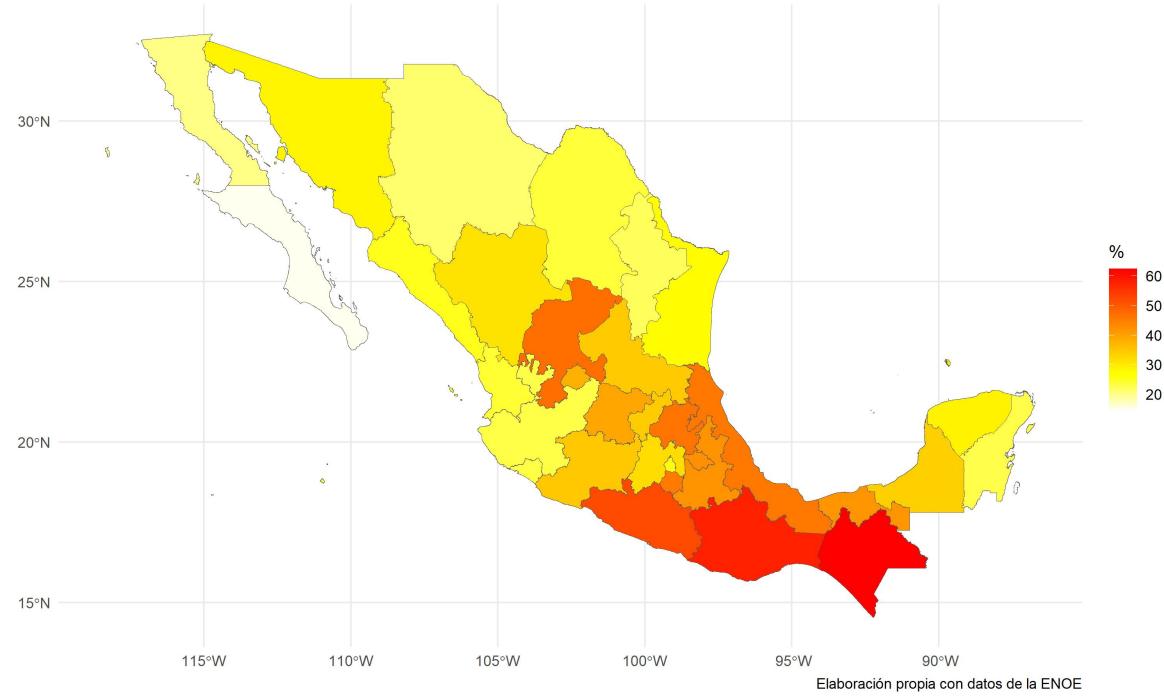
Pobreza Multidimensional y Laboral



Fuente: Elaboración propia con datos de la ENOE y la ENIGH
(De la ENOE, solo se considera el tercer trimestre de cada año que corresponde con un levantamiento de la ENIGH)

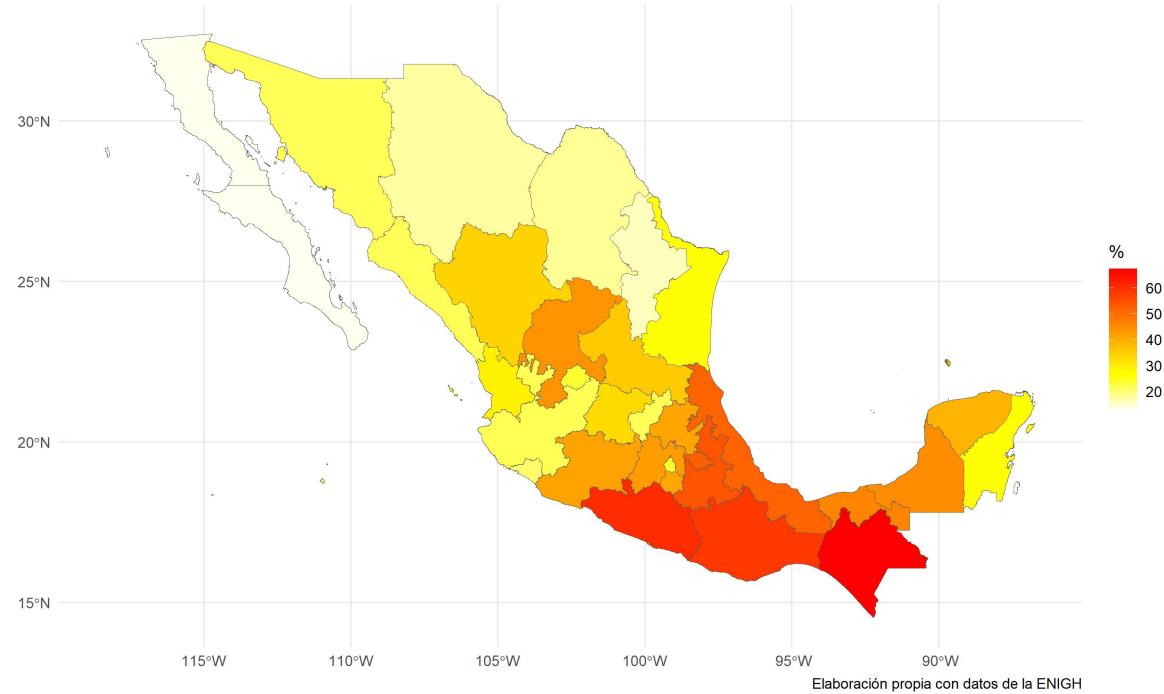
Población en pobreza laboral.

Segundo trimestre de 2024



Población en pobreza multidimensional.

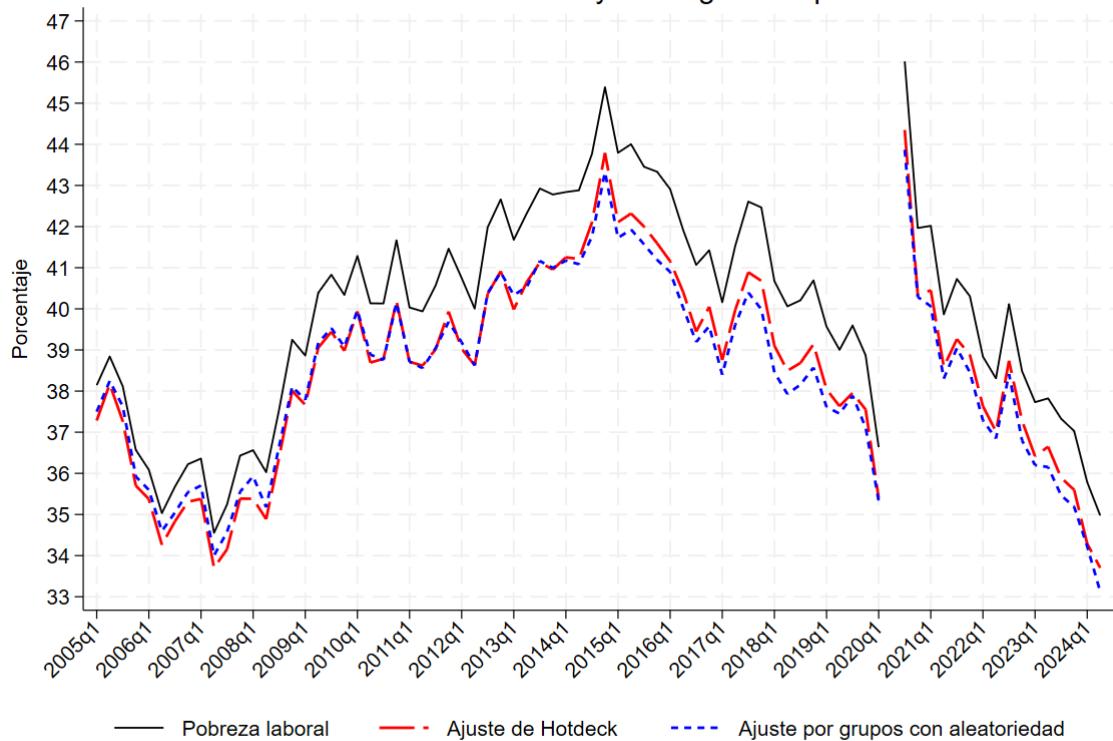
Segundo trimestre de 2024



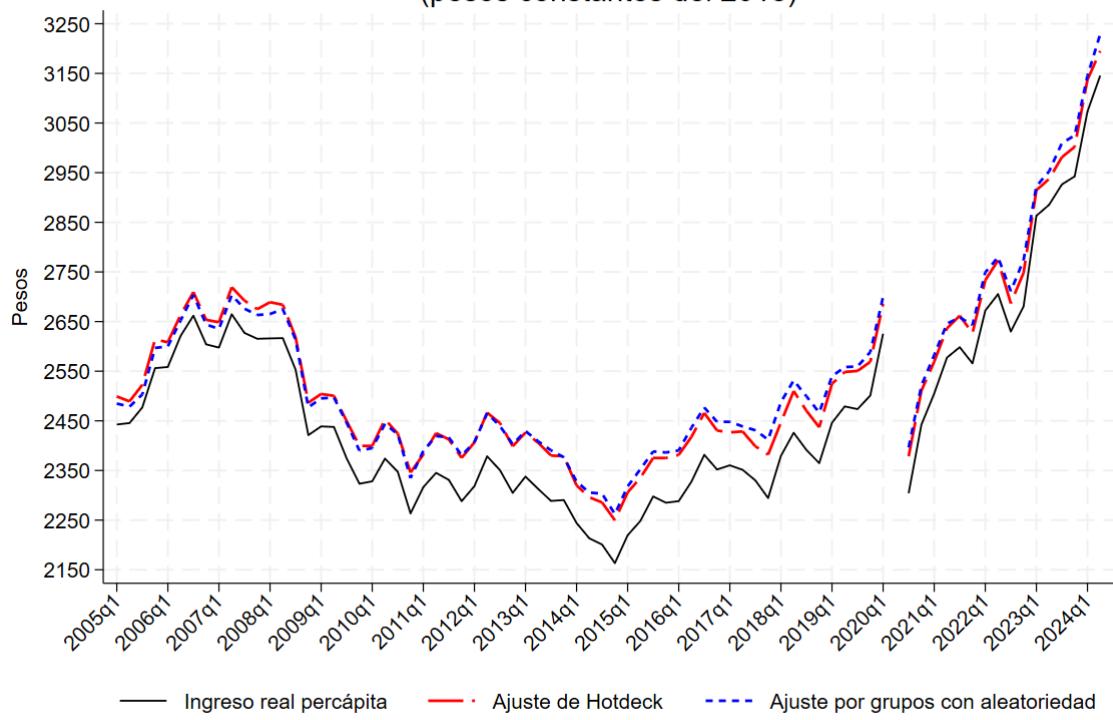
4. Replica el artículo de Campos Vázquez (2013) para el periodo 2005-2024, utilizando todos los trimestres. Utilizarás 3 métodos de imputación: 1) *Hotdeck* y 2) Medias de grupos específicos, si 2) no te gusta puedes sustituirlo por uno de tu preferencia. En dos gráficas por separado incluirás lo siguiente: a) % Pobreza (laboral) de acuerdo a cada método (3 líneas en la figura, la oficial y los 2 métodos adicionales), b) Ingreso laboral per cápita (3 líneas en la figura, oficial y los 2 métodos). Gráficas que no se entiendan o no se vean bonitas serán castigadas.

Figuras en la siguiente página

Pobreza laboral oficial y con ingreso imputado



Ingreso real per cápita oficial e imputado (pesos constantes del 2018)



Fuente: Elaboración propia con datos la ENOE

5. ¿Qué puedes concluir del problema de los ingresos faltantes y de la comparación entre métodos? También analiza la comparabilidad entre ENIGH y ENOE.

Respuesta. El problema de los ingresos faltantes debe ser abordado en todos las investigaciones que involucren trabajar con datos de México; esto como consecuencia de la no aleatoriedad que vimos caracteriza a la composición de trabajadores que deciden no reportar un ingreso válido. En cuanto a la comparabilidad entre la ENIGH y la ENOE, sería necesario un mayor análisis del que hemos hecho en este inciso para revisar si los mismos resultados cualitativos en términos de ingresos inválidos se mantienen en la ENIGH, aunque es probable que así sea. Finalmente, como observamos en las figuras de las páginas anteriores, el método utilizado genera hallazgos muy similares, por lo que la elección del método, aunque quizás no sea excesivamente importante, debería estar siempre guiada por una preferencia a utilizar más de uno, de tal forma que se contrasten distintas estimaciones y se robustezca la imputación.

6. ¿Crees que los supuestos del método de imputación se cumplan? Explica y argumenta tu respuesta.

Respuesta. El supuesto *Missing at Random*, "MAR", de que las razones de la ausencia de Y_i no dependen de Y_i sino de otras características observables \mathbf{X}_i , es decir, que

$$\mathbb{P}[Y_i \text{ sea faltante} | Y_i, X_i] = \mathbb{P}[Y_i \text{ sea faltante} | X_i];$$

puede que no sea totalmente cierto. Es posible que, a medida que el ingreso sea mayor, la probabilidad de que éste no sea reportado crezca, por lo que MAR sería violado. Este es el caso de México, donde los ingresos de la distribución más alta no son capturados por las encuestas del INEGI.

7. La pobreza laboral aumentó con COVID y luego disminuye en 2022-4. ¿A qué se debe? A nivel entidad realiza un **scatter** entre disminución de pobreza laboral y cambios en empleo y salario con IMSS y ENOE, con ENOE cambios en participación laboral y con cambios en proporción que no declaran ingresos. ¿Qué podemos decir al respecto?