

Modelización lineal del salario

Emanuelle Marsella, Maximiliano Saldaña, Lucia Tafernaberry

Noviembre 2020

Índice

1. Introducción	3
2. Descripción de los datos	3
2.1. Variables Indicadoras	3
3. Exploracion inicial	3
3.1. Variables categóricas	3
3.1.1. Región	3
3.1.2. Rama de actividad	4
3.1.3. Ocupación	5
3.1.4. Otras variables categóricas	5
4. Métodos	5
4.1. Presentación del modelo	5
4.2. Estimación del modelo	6
5. Resultados	7
5.1. Análisis de convergencia	7
5.2. Réplicas de los datos simulados	8
6. Conclusión	9
7. Bibliografía	10

1. Introducción

En este proyecto aplicaremos los conceptos y métodos bayesianos aprendidos en el curso Inferencia II para analizar los datos seleccionados, una base de datos sobre salarios del año 1976 en Estados Unidos extraída del libro *Introducción a la Econometría, Un enfoque moderno (2009)* de J.M. Wooldridge. Nuestra intención es tratar de explicar la variable principal, el salario por hora medido en dólares, a partir de las otras variables de la base. Para lograr esto haremos uso de modelos de regresión lineal y metodologías bayesianas asociadas, que implementaremos mediante el software R.

2. Descripción de los datos

La base cuenta con observaciones de 526 personas y con las siguientes 22 variables: Salario (promedio por hora, medido en dólares), Años de Educación, Años de Experiencia, Antigüedad, Raza (variable indicadora, 1 si la persona no es de raza blanca), Sexo, Estado Civil (vale 1 si la persona está casada), Número de Dependientes, Región Metropolitana (vale 1 si la persona vive en dicha región), Región (dividida en tres variables indicadoras: Norte, Sur y Oeste), Rama de Actividad (dividida en tres variables indicadoras: Construcción, Comercio y Servicios), Ocupación (tres variables indicadoras: Profesional, Administrativos y Servicios), el logaritmo de la variable Salario y los cuadrados de las variables Experiencia y Antigüedad.

2.1. Variables Indicadoras

Inicialmente modificamos la base para llevarla a una forma que nos parecía más práctica de trabajar, agrupando variables que refieren a una sola característica de los datos y estaban en una forma binaria. Estas variables son la Región, separada en la base original en Norte, Sur y Oeste; la Rama de Actividad, separada en Construcción, Servicios y Comercio; y la Ocupación, separada en Profesional, Administrativos y Servicios. Luego de consultar a las docentes de Modelos Lineales optamos por no descartar las variables indicadoras ya que nos permiten su utilización a la hora de aplicar un modelo lineal multivariado y a su vez mantenemos la variable agrupada para poder usarla a la hora de manipular y graficar datos.

3. Exploración inicial

Para familiarizarnos con la base utilizamos medidas de resumen.

3.1. Variables categóricas

Lo primero que hacemos es visualizar cómo se distribuye la población de acuerdo a las variables categóricas de las que disponemos (Región, Rama de actividad, Ocupación, Raza, Sexo y Estado civil). En cuanto a las primeras cuatro variables buscamos saber cuántos individuos pertenecen a cada rama, así como aplicar algunas medidas de resumen de la variable salario para cada categoría de las tres variables, como se puede ver en los cuadros a continuación.

3.1.1. Región

La variable región nos indica en qué región se ubican los individuos. Las categorías de esta variable son Norte, Sur, Este y Oeste.

Tenemos 3 variables indicadoras para esta región, Norte-Centro, Sur y Oeste. Este es la categoría de referencia, la cual no cuenta con variable indicadora ya que en este caso la matriz de datos \mathbb{X} no sería de rango completo y por lo tanto no sería invertible, lo que más adelante impediría la estimación única de los coeficientes. Esto ocurre análogamente para las otras variables cualitativas.

	Región	Cantidad	Mínimo	Media	Máximo
1	Norte	132	1.50	5.71	21.86
2	Sur	187	1.50	5.39	20.00
3	Este	118	1.43	6.37	24.98
4	Oeste	89	0.53	6.61	22.20
5	Todas	526	0.53	5.90	24.98

Cuadro 1: Algunas medidas de resumen para las distintas regiones.

En primer lugar, podemos observar que la región Oeste cuenta con un número considerablemente pequeño de observaciones y la región sur cuenta con muchas observaciones, en comparación con las demás.

En cuanto al salario por hora, vemos que las medias para las regiones Norte y Sur son menores a la media del total de observaciones, mientras que para las regiones Este y Oeste son mayores. Por otro lado, los máximos son similares, si bien apreciamos que el máximo total se observa en la región este.

Los mínimos también son similares para todas las regiones excepto para la región Oeste, para la cual podemos apreciar que el mínimo es casi 3 veces menor que para las demás.

Como parte de la exploración inicial, también realizamos algunos gráficos de dispersión del salario en función de algunas variables cuantitativas. En ellos podemos apreciar que no hay demasiada diferencia entre las regiones a la hora de explicar el salario a partir de las distintas variables explicativas de forma individual, por lo cual optamos por no incluirlos en este documento.

3.1.2. Rama de actividad

	Rama de Actividad	Cantidad	Mínimo	Media	Máximo
1	Construcción	24	3.00	5.96	17.71
2	Servicios	53	0.53	4.34	12.50
3	Comercio	151	1.43	4.79	21.86
4	Otros	298	1.50	6.73	24.98
5	Todas	526	0.53	5.90	24.98

Cuadro 2: Algunas medidas de resumen para las distintas ramas de actividad.

Lo primero que observamos, es que la cantidad de observaciones para las distintas categorías varía mucho. En particular, para la categoría “Otros”, que es la de referencia, tenemos un gran número de individuos, más de la mitad. Consideramos que esto se debe a que la diferenciación de la que disponemos para las diferentes ramas de actividad no es exhaustiva (como sí sucedía para las regiones, que solo son 4). Por lo tanto, es lógico que suceda que muchos individuos no pertenezcan ni al sector de Construcción, ni al de Servicios, ni al de Comercio; sino que pertenecen a alguna otra categoría que no está especificada y queda incluida dentro de “Otros”.

Comparando las regiones en media, observamos que la media de la categoría de referencia es mayor que en las demás ramas de actividad. La categoría Servicios tiene la menor media, y en esta categoría se observa el mínimo total, y el menor de los máximos. La categoría Comercio es similar en media a Servicios, si bien tanto su mínimo como su máximo son mayores. Finalmente, la categoría Construcción tiene la mayor media de las 3 indicadoras de las que disponemos, y también el menor mínimo.

3.1.3. Ocupación

	Ocupación	Cantidad	Mínimo	Media	Máximo
1	Servicio	74	0.53	3.59	7.81
2	Administrativos	88	2.65	4.74	12.50
3	Profesional	193	2.23	8.04	24.98
4	Otros	171	1.43	5.07	15.00
5	Todas	526	0.53	5.90	24.98

Cuadro 3: Algunas medidas de resumen para las distintas ocupaciones.

Podemos observar que para la variable Ocupación, la media de la categoría Profesional es considerablemente mayor respecto de las demás, así como también su máximo. Es razonable esperar que así sea, que las personas que se desempeñan en un empleo profesional perciban un mayor salario, si bien no tenemos fundamentos para afirmar que esta diferencia sea estadísticamente significativa, dado que aún no hemos realizado inferencia a partir de nuestros datos.

Por otro lado, la categoría Administrativos tiene el mayor mínimo y el menor máximo, y una media menor que la de la categoría de referencia, lo que sugiere que los niveles de salario no varían demasiado, y en general son menores que los de la categoría de referencia.

Luego, la categoría Servicio tiene el menor mínimo, la menor media, y el menor máximo, lo que sugiere que los individuos que se encuentran en esta categoría perciben en promedio un menor salario.

3.1.4. Otras variables categóricas

Con el resto de las variables encontramos que las proporciones entre las categorías son:

- Para la variable Sexo hay un 47.91 % de mujeres y 52.09 % de hombres. La proporción de la mitad cada sexo.
- Para la variable Estado Civil hay un 60.84 % de casados y 39.16 % de no casados.
- Para la variable Área Metropolitana hay un 72.24 % de personas que habitan en la misma y 27.76 % que no. Predomina bastante la población metropolitana.
- Para la variable Raza hay un 10.27 % de personas no blancas y 89.73 % blancas. El resultado es de esperarse dado que las personas no blancas son una minoría en los Estados Unidos.

4. Métodos

4.1. Presentación del modelo

Vamos a plantear un modelo de la forma

$$\log(Y) = X\beta + \varepsilon$$

donde hacemos los supuestos clásicos sobre los errores: normalidad, esperanza nula, homocedasticidad e incorrelación entre los mismos.

Nuestra variable explicada es el logaritmo del salario, que se puede interpretar de la siguiente forma: dado un aumento unitario en una de las variables explicativas x_i , la variación porcentual del salario equivale aproximadamente a $100\beta_i$, a niveles constantes de todas las demás variables ¹, es decir:

$$\% \Delta y \simeq (100\beta_i) \Delta x$$

X es nuestra matriz de datos, que contiene los valores de las siguientes variables:

¹Introducción a la econometría: Un enfoque moderno. Wooldridge, Jeffrey. (2009). pág. 43.

- Experiencia
- Antigüedad
- Sexo
- Reg. metropolitana
- Norte/Sur
- Comercio/servicios
- Ocup. profesional
- Exper. al cuadrado
- Educ. al cuadrado
- Indic. Obs. 128
- Indic. Obs. 381
- Indic. Obs. 440

Hubo un trabajo previo en el curso de Modelos Lineales para llegar a esta forma del modelo. La selección de las variables se hizo en base al criterio de información de Akaike y contrastes de significación conjunta, que sirvieron para agrupar las variables Norte con Sur y Comercio con Servicios. Además, para que se cumplieran los supuestos se tuvo que realizar una transformación de Box-Cox, excluirse una observación influyente de la base e incluirse un conjunto de variables indicadoras para representar un posible cambio en media y/o varianza en varias observaciones. Estas últimas variables indicadoras son de la forma:

$$z_t = \begin{cases} 1 & \text{si } t = i \\ 0 & \text{si } t \neq i \end{cases}$$

Lo que nos dejó con un modelo de la forma.

$$\log(y_t) = x'_t\beta + \Delta_i z_t + \varepsilon_t$$

4.2. Estimación del modelo

La estimación de los parámetros del modelo se realizará desde un enfoque bayesiano, es decir, se les asignará una distribución previa para cuantificar la incertidumbre que se tiene sobre sus valores y utilizando los datos se obtendrá una distribución posterior. Se diferencia del caso de un modelo lineal clásico dado que en este último se trabaja con la distribución en el muestreo de los estimadores de los parámetros y no con la distribución de los parámetros en sí.

Realizamos la estimación mediante simulaciones Monte Carlo basadas en cadenas de Markov (MCMC), simulaciones en las cuales cada resultado depende únicamente del anterior. Para la estimación de cada uno de los parámetros se plantean 4 cadenas iniciadas en valores distintos, cada una con 4000 iteraciones. A partir de estos valores se avanza iterativamente hasta llegar a los 4000 valores simulados, de los cuales los primeros 2000 se usan como “calentamiento” para desligar las simulaciones de los valores iniciales.

Se desea que las cadenas converjan a la distribución posterior de los parámetros, y aunque nunca podemos asegurar que se da al ser un concepto teórico podemos evaluar si hay indicios de que no se da. Dos métodos para evaluar la convergencia son el \hat{R} y el \hat{n}_{eff} .

$$\hat{R} = \sqrt{\frac{Var^+(\theta|y)}{W}}$$

Donde:

$$Var^+(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

En la expresión anterior n indica el número de iteraciones en cada cadena, W la media de las distintas varianzas dentro de las cadenas y B la varianza entre las medias de las cadenas. Si el \hat{R} se acerca a uno, eso

significa que la variabilidad entre las cadenas es pequeña y por lo tanto la variabilidad de las simulaciones es explicada por la variabilidad dentro de las cadenas.

Por otro lado definimos el indicador:

$$n_{\text{eff}} = \frac{n \cdot m}{1 + 2 \sum_{k=1}^{\infty} \rho_k}$$

Donde n es el número de iteraciones, m es el número de cadenas y ρ_k la autocorrelación k -ésima. Al estimar $\sum_{k=1}^{\infty} \rho_k$ conseguimos \hat{n}_{eff} . Este indicador es una estimación de la cantidad de muestras independientes equivalentes a las del método MCMC.

5. Resultados

	Parámetro	Parámetros clásicos	Parámetros bayesianos estimados	Cuantil 2.5 %	Cuantil 97.5 %
1	Intercepto	1.057426	1.057820	0.9197	1.192
2	Experiencia	0.030918	0.030822	0.0225	0.0388
3	Antigüedad	0.013780	0.013834	0.0087	0.0186
4	Sexo	-0.277087	-0.277403	-0.339	-0.2159
5	Región Metropolitana	0.157804	0.158204	0.0896	0.2278
6	Norte/Sur	-0.087530	-0.087912	-0.1506	-0.0247
7	Comercio/Servicios	-0.235756	-0.234436	-0.2978	-0.1712
8	Ocupación Profesional	0.195501	0.195010	0.1199	0.2689
9	Exp. al cuadrado	-0.000629	-0.000626	-8e-04	-4e-04
10	Educ. al cuadrado	0.002175	0.002175	0.0016	0.0027
11	Indic. 128	-1.235298	-1.241326	-1.9452	-0.5514
12	Indic. 381	1.139318	1.140625	0.4554	1.8112
13	Indic. 440	1.322450	1.330452	0.6362	2.0035

Cuadro 4: Cuadro comparativo entre las estimaciones de los parámetros realizados a través de inferencia clásica y bayesiana.

En el cuadro anterior podemos observar los resultados de la estimación de los coeficientes de las variables explicativas, tanto del modelo clásico por medio de mínimos cuadrados ordinarios, como del modelo bayesiano a través de estimación MCMC. Podemos ver que la diferencia es mínima entre ambas estimaciones, siendo los resultados iguales hasta al menos el segundo dígito después de la coma en todos los casos.

Además, planteamos para cada parámetro el intervalo de credibilidad al 95 %, y en todos los casos vemos que tanto el parámetro estimado mediante simulación bayesiana como el parámetro estimado mediante inferencia clásica están comprendidos en dicho intervalo.

5.1. Análisis de convergencia

En primer lugar, nos valdremos de distintos indicadores y visualizaciones que nos permitirán disgnósticar si existe o no falta de convergencia en nuestro modelo.

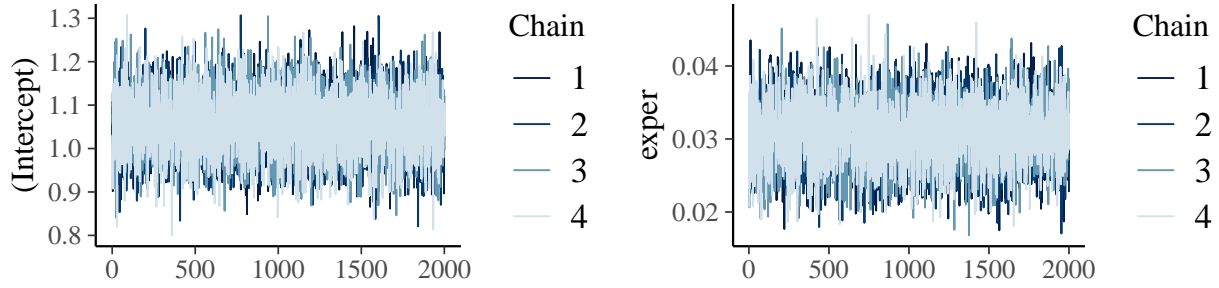


Figura 1: Gráficos que muestran la evolución de las cadenas para dos variables explicativas, el intercepto y la experiencia.

Para observar el comportamiento de las cadenas usadas en la estimación, comenzamos con gráficos que nos permiten observar la evolución de las mismas a lo largo de las iteraciones. En los mismos podemos ver que las cadenas para los parámetros del intercepto y de la experiencia se mantienen en torno a un cierto valor en cada caso, lo cual sugiere que no hay falta de convergencia. El comportamiento para los demás parámetros es similar, presentándose estos dos casos en el documento de forma de que sirvan de visualización.

Además de las visualizaciones, nos valemos de dos medidas que nos permiten identificar la falta de convergencia o indicios de lo contrario: el \hat{R} y el \hat{n}_{eff} .

Parámetro	\hat{R}	n_{eff}
Intercepto	0.999962	8743
Experiencia	0.999883	6048
Antigüedad	0.999631	9652
Sexo	0.999988	10735
Región Metropolitana	1.000015	10467
Norte/Sur	0.999736	10541
Comercio/Servicios	0.999939	10101
Ocupación Profesional	0.999729	8976
Exp. al cuadrado	0.999974	6128
Educ. al cuadrado	1.000216	8528
Indic. 128	0.999891	10472
Indic. 381	0.999971	11287
Indic. 440	0.999742	10776

Cuadro 5: Indicadores sobre convergencia para los distintos parámetros del modelo bayesiano.

En este cuadro, podemos ver que para todos los parámetros del modelo bayesiano (del intercepto y las variables explicativas) el valor del \hat{R} está en torno a 1 pero sin superar 1.1, lo cual sirve como regla empírica para desestimar indicios de falta de convergencia. En cuanto al n_{eff} , los valores son altos en comparación con las 4000 iteraciones dependientes realizadas, del orden de 6000, 8000 y hasta 10000 simulaciones independientes.

5.2. Réplicas de los datos simulados

A continuación, veremos la capacidad predictiva de nuestro modelo, a través de la comparación de estadísticos aplicados tanto a nuestros datos originales como a una muestra de datos simulados.

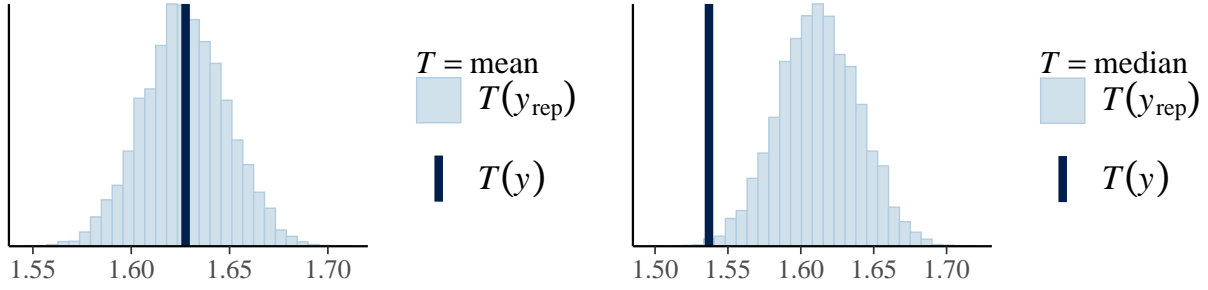


Figura 2: Comparación de estadísticos entre los datos originales y los simulados.

En la visualización anterior, se presentan dos visualizaciones del histograma del logaritmo del salario, en base a las 100 simulaciones realizadas con nuestro modelo lineal bayesiano.

En primer lugar, vemos que la media de los datos originales, representada con una línea azul oscura, se ubica sobre los valores centrales del histograma, donde se dan los intervalos con mayor frecuencia relativa. Esto quiere decir que nuestros datos simulados son similares en media a nuestros datos originales. No sucede así con la mediana, que se distribuye de forma similar a la media en nuestros datos simulados contrariamente a lo que sucede en los originales, lo cual podemos ver en que la mediana original se ubica sobre los valores menores de la distribución, en lugar de la parte central como queríamos.

6. Conclusión

7. Bibliografía

- Introducción a la econometría: Un enfoque moderno. Wooldrige, Jeffrey. (2009).
- Materiales del curso 2020 de Modelos Lineales.