Labor Income Prediction

Predicting Income

Sany León, Andrés Suárez, and Juan Rueda

2026-02-21

Research Question

Can prediction-based income models reliably identify individuals with anomalous earnings patterns, and do large prediction errors reflect misreporting or structural model limitations?

Data

Results

	LOOCV RMSE		Test RMSE	
	With outliers	Without outliers	With outliers	Without outliers
Modelo 1	0.662	0.662	0.846	0.846
Modelo 2	0.676	0.676	0.855	0.855
Modelo 3	0.399	0.399	0.666	0.666
Modelo 4	0.391	0.391	0.655	0.655
Modelo 5	0.390	0.390	0.655	0.655
Modelo 6	0.389	0.389	0.652	0.652
Modelo 7	0.390	0.390	0.653	0.653
Modelo 8	0.388	0.388	0.650	0.650
Modelo 9	0.377	0.377	0.648	0.648

Train Obs. (Outliers-No outliers): 10334-10334

Dos observaciones fueron eliminadas en LOOCV para evitar leverage de 1 en el test set

Model 9 achieves the lowest validation RMSE (0.648). Notably, its LOOCV RMSE (0.377) is considerably smaller, indicating that LOOCV may understate the true out-of-sample prediction error relative to the validation-sample benchmark.

Results

Modelos

$$(1) \quad \log(ingresos_i) = \beta_0 + \beta_1 Edad_i + \beta_2 Edad_i^2 + u_i$$

$$(2) \quad \ln(salario)_{i,f} = \beta_0 + \beta_1 Mujer_{i,f} + u_i$$

$$\begin{aligned}(3) \quad \ln(salario)_{i,f} = {} & \beta_0 + \beta_1 Mujer_{i,f} + \beta_2 Edad_{i,f} + \beta_3 Edad_{i,f}^2 \\ & + \beta_4 NivelEduc_{i,f} + \beta_5 Oficio_{i,f} + \beta_6 Relab_{i,f} \\ & + \beta_7 TamFirma_{i,f} + u_i\end{aligned}$$

$$\begin{aligned}(4) \quad \ln(salario)_{i,f} = {} & \beta_0 + \beta_1 Mujer_{i,f} + \beta_2 Edad_{i,f} + \beta_3 Edad_{i,f}^2 \\ & + \beta_4 NivelEduc_{i,f} + \beta_5 Oficio_{i,f} + \beta_6 Relab_{i,f} \\ & + \beta_7 TamFirma_{i,f} + \beta_8 NumMen_f + \beta_9 NumMay_f + u_i\end{aligned}$$

$$\begin{aligned}(5) \quad \log(y_i) = {} & \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad_i^2 + \beta_4 Edad_i^3 \\ & + \beta_5 NivelEduc_i + \beta_6 Oficio_i + \beta_7 Relab_i \\ & + \beta_8 TamFirma_i + \beta_9 Formalidad_i + u_i\end{aligned}$$

Results

Modelos

$$(6)\quad \log(y_i) = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad_i^2 + \beta_4 (NivelEduc_i \times Formalidad_i) + \beta_5 Oficio_i + \beta_6 Relab_i + \beta_7 TamFirma_i + u_i$$

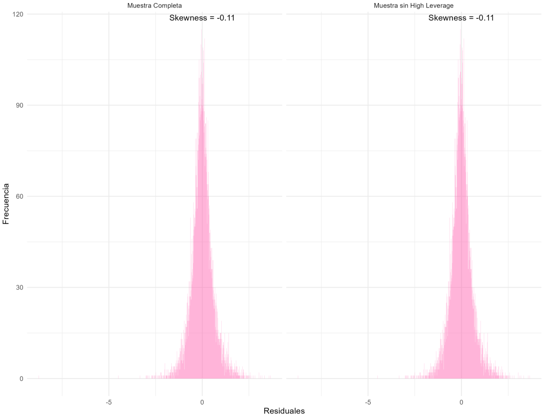
$$(7)\quad \log(y_i) = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad_i^2 + \beta_4 NivelEduc_i + \beta_5 Oficio_i + \beta_6 (Relab_i \times TamFirma_i) + \beta_7 Formalidad_i + u_i$$

$$(8)\quad \log(y_i) = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad_i^2 + \beta_4 Oficio_i + \beta_5 Formalidad_i + \beta_6 (NivelEduc_i \times Formalidad_i) + \beta_7 (Relab_i \times TamFirma_i) + u_i$$

$$(9)\quad \log(y_i) = \beta_0 + \beta_1 Sexo_i + \beta_2 Edad_i + \beta_3 Edad_i^2 + \beta_4 Oficio_i + \beta_5 Formalidad_i + \beta_6 (NivelEduc_i \times Formalidad_i \times Edad_i) + \beta_7 (Relab_i \times TamFirma_i \times Edad_i) + u_i$$

Results

Figure 1: Distribution of the Prediction Residuals of Model from Equation (9)



Discussion