# Proyecto Econometría Aplicada y Ciencia de Datos

Diego Monroy y Sergio Sarmiento

December 2025

## Introducción

La estimación de los retornos de la educación es un tema de interés para practicamente todas las ciencias sociales; no obstante, ninguna disciplina ha sido tan atormentada y maravillada por esta problemática como la econometría. Debido que los niveles de educación y los salarios representan un problema de simultaneidad, realizar cualquier estimación sobre el impacto que un año más de educación tiene sobre el salario de una persona resulta una tarea que pocos econometristas se han atrevido a resolver. A pesar de que este escrito no pretende adentrarse en esta labor, el presente reporte puede ser de utilidad para cualquier econometrista interesado en la educación.

Como punto de partida, retomamos el enfoque clásico de Card (1993), quien propone explotar variación geográfica en el acceso a educación superior: en particular, la cercanía a una universidad como fuente de variación exógena en los años de escolaridad. Siguiendo esa intuición, en este reporte primero estimamos una ecuación tipo Mincer para el logaritmo del salario mediante MCO con errores estándar robustos, y posteriormente instrumentamos la educación con la variable `nearc4` (proximidad a una universidad) para mitigar el sesgo por endogeneidad.

A partir de esa base econométrica, el ejercicio se amplía con una lógica complementaria: utilizar métodos predictivos para organizar la selección de controles y, con ello, construir especificaciones más informadas sin perder de vista que la identificación descansa en el instrumento. En concreto, entrenamos un Random Forest para el log-salario, obtenemos métricas de importancia de covariables y re-estimamos el modelo IV incorporando controles seleccionados por dicho algoritmo (evitando colinealidades directas con la experiencia). Además, reforzamos la inferencia con un bootstrap del coeficiente de educación, reportamos explícitamente las primeras etapas y el poder del instrumento, y finalmente exploramos heterogeneidad en los efectos con un *Instrumental Causal Forest*. Por ende, el motivo de este

reporte es presentar cómo las técnicas de Machine Learning y ciencia de datos pueden complementar el repertorio de herramientas del economista al estudiar retornos a la educación: no como sustitutos de la identificación causal, sino como apoyos prácticos para seleccionar controles, evaluar estabilidad inferencial y explorar patrones de heterogeneidad que suelen quedar ocultos cuando solo se reporta un efecto promedio.

# Datos

Los datos utilizados en este proyecto provienen del archivo original de David Card (también disponible en el paquete wooldridge bajo el nombre `card`) y contienen información individual sobre salarios, educación, características familiares y condiciones demográficas. La base consta de 3010 observaciones y 34 variables, entre ellas el salario, años de educación, edad, raza, estado civil, escolaridad de los padres, ubicación geográfica y un indicador clave para el instrumento: si el individuo creció cerca de una universidad (`nearc4`). Estas variables permiten estimar modelos de MCO e IV para el retorno a la educación bajo distintas estrategias de control.

En estos datos, la variable de experiencia laboral se construyó como $\text{exper} = \text{age} - 6$, asumiendo que los individuos ingresan al mercado laboral inmediatamente después de la adolescencia temprana. Esta aproximación se utiliza en lugar de la fórmula original $\text{exper} = \text{age} - \text{educ} - 6$, utilizada por Card (1993) porque evita una fuerte colinealidad entre educación y experiencia, la cual puede distorsionar las estimaciones e inflar los errores estándar.

# La instrumentación de Card

Siguiendo el enfoque de Card, el objetivo de esta sección es fijar una especificación estándar de retornos a la educación con una ecuación tipo Mincer y, a la vez, dejar explícito el problema de identificación que motiva el uso de variables instrumentales. En particular, se reconoce que $educ_i$ es endógena en la ecuación salarial, de modo que la estimación por MCO podría capturar tanto el efecto causal de la educación como la influencia de factores no observados correlacionados con ella. Para mantener comparabilidad, primero se construyen las variables centrales y se depura la muestra: se trabaja con el logaritmo del salario y con experiencia potencial (y su cuadrado), eliminando observaciones con valores faltantes en salario, educación, experiencia, el instrumento y un conjunto básico de controles.

En particular, definimos:

$$lwage_i \equiv \log(wage_i), \tag{1}$$

$$exper_i \equiv age_i - 6, \tag{2}$$

$$expersq_i \equiv exper_i^2. \tag{3}$$

Con estas variables, el primer punto de referencia es la estimación por MCO de una ecuación de Mincer para el log-salario, incorporando controles demográficos y de origen familiar, y reportando errores estándar robustos (HC1). La especificación se escribe como

$$lwage_i = \alpha + \beta\, educ_i + \delta_1\, exper_i + \delta_2\, expersq_i + \gamma' X_i + \varepsilon_i \tag{4}$$

Esta especificación incluye $X_i = \big(black_i,\ south_i,\ smsa_i,\ fatheduc_i,\ motheduc_i\big)$. Es decir, además de experiencia, se controla por raza, región, urbanidad y educación de los padres, con el objetivo de absorber diferencias sistemáticas de entorno que podrían sesgar la relación entre educación y salario.

Posteriormente, se instrumenta $educ_i$ utilizando `nearc4`, interpretada como un indicador de cercanía a una universidad de cuatro años. La primera etapa describe el efecto de la cercanía sobre los años de escolaridad, manteniendo constantes los mismos controles:

$$educ_i = \pi_0 + \pi_1\, nearc4_i + \pi_2\, exper_i + \pi_3\, expersq_i + \lambda' X_i + v_i. \tag{6}$$

La segunda etapa sustituye $educ_i$ por su componente predicho $\widehat{educ_i}$ para estimar el efecto causal sobre salarios:

$$lwage_i = \alpha + \beta\, \widehat{educ_i} + \delta_1\, exper_i + \delta_2\, expersq_i + \gamma' X_i + u_i. \tag{7}$$

La justificación del instrumento descansa, primero, en su relevancia: crecer cerca de una universidad reduce costos de acceso a educación superior y, por tanto, debe afectar sistemáticamente la escolaridad. En segundo lugar, bajo la restricción de exclusión condicional a $X_i$, se asume que la cercanía impacta el salario principalmente a través de la educación y no mediante canales directos. Aun así, la credibilidad de esta estrategia requiere cautela: al tratarse de una variable ligada a la geografía, es natural considerar posibles fuentes de sesgo asociadas a selección residencial y a diferencias persistentes en mercados laborales locales. En consecuencia, el estimador IV debe interpretarse como un efecto causal local (para los individuos cuya escolaridad se ve afectada por la cercanía), y su validez depende de la plausibilidad de la restricción de exclusión una vez incorporados los controles.

## *Random Forest* y controles seleccionados

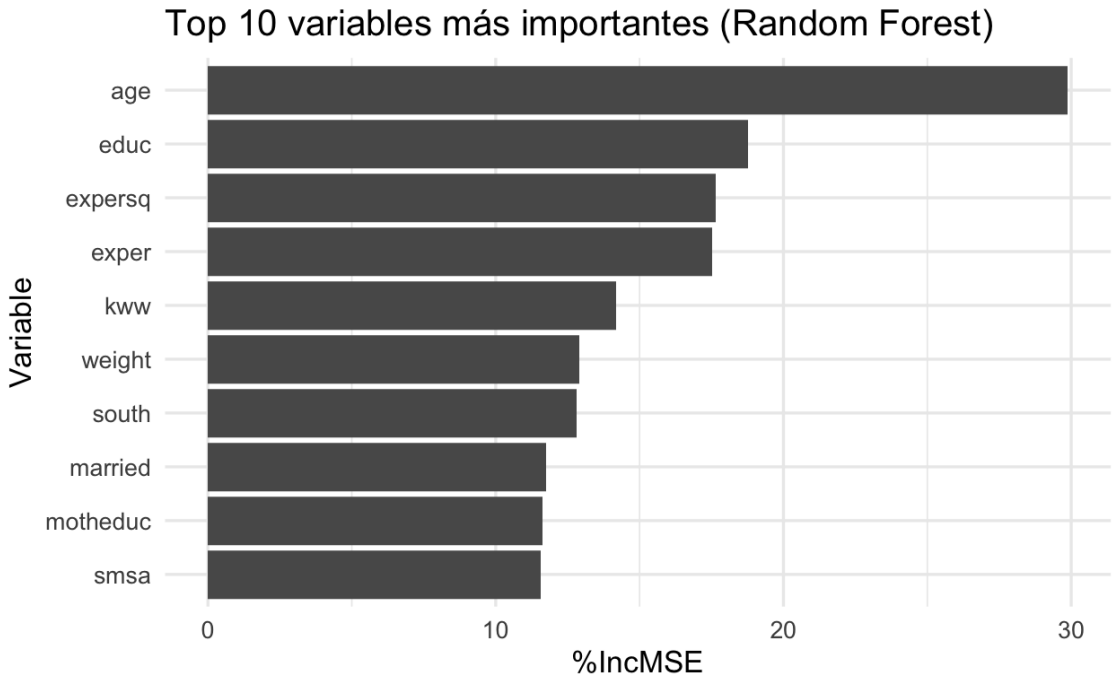


Figure 1: Top 10 variables más importantes (Random Forest) según %IncMSE.

Con el fin de identificar, de manera sistemática, qué covariables observables aportan mayor poder predictivo para explicar el log-salario (*lwage*), se estimó un *Random Forest*. La motivación de este paso es meramente práctica: cuando se cuenta con muchas variables candidatas, el *Random Forest* permite jerarquizarlas mediante una medida de importancia (en este caso, %IncMSE), lo que facilita proponer un conjunto parsimonioso de controles para especificaciones posteriores.

La Figura 1 muestra las 10 variables más importantes según %IncMSE, dichas variables serán utilizadas como controles para estimar la especificación de IV Card (controles RF):

- `age`: edad del individuo (en años).
- `educ`: años de educación.
- `expersq`: experiencia potencial al cuadrado (`exper`$^2$).
- `exper`: experiencia potencial (en años).
- `kww`: puntaje en el test KWW (medida de habilidad/capital humano).
- `weight`: ponderador muestral (peso de la observación en la encuesta).

- `south`: indicador de residir en el sur (dummy).
- `married`: indicador de estar casado (dummy).
- `motheduc`: años de educación de la madre.
- `smsa`: indicador de residir en un área metropolitana (dummy).

Por último, para ejemplificar el *Random Forest*, la figura 3 en el apéndice muestra un árbol representativo del *Random Forest* utilizado para evaluar la importancia de las covariables en la predicción del salario. Cada división refleja la variable que mejor separa a los individuos en términos de su resultado esperado, revelando patrones no lineales e interacciones entre características como experiencia, estado civil o nivel educativo. Aunque el modelo final promedia cientos de árboles, este ejemplo sirve para ilustrar cómo el Random Forest identifica de manera flexible cuáles variables contienen mayor poder predictivo dentro del conjunto de datos.

# Resultados

La siguiente tabla 1 resume los resultados principales de las distintas especificaciones utilizadas para estimar el retorno a la educación. Se presentan estimaciones por MCO, versiones ampliadas con controles al estilo de Card, así como estimadores IV con y sin controles seleccionados mediante Random Forest. La comparación entre modelos permite evaluar cómo cambian los coeficientes al incorporar más controles y al corregir por endogeneidad.

Table 1: Estimaciones MCO e IV de los retornos a la educación

| | MCO | MCO Card | IV básico | IV Card | IV Card (controles RF) |
|---|---|---|---|---|---|
| educ | 0.052*** | 0.028*** | 0.174*** | 0.072 | 0.032 |
| | (19.067) | (7.761) | (7.211) | (0.785) | (0.389) |
| exper | 0.105** | 0.127** | 0.006 | 0.092 | 0.036 |
| | (2.756) | (2.973) | (0.121) | (1.089) | (0.429) |
| expersq | -0.001+ | -0.002+ | 0.001 | -0.001 | -0.000 |
| | (-1.694) | (-1.947) | (0.667) | (-0.634) | (-0.009) |
| Num.Obs. | 3010 | 2220 | 3010 | 2220 | 1600 |
| Controles Adicionales | NO | SÍ | NO | SÍ | SÍ (RF) |
| Baseline mean | 6.262 | 6.285 | 6.262 | 6.285 | 6.343 |

Los resultados muestran diferencias claras entre las estimaciones MCO y los modelos IV. En MCO, el retorno estimado de un año adicional de educación es positivo y significativo, pero se reduce notablemente al incluir los controles socioeconómicos utilizados por Card. Al pasar a las estimaciones por variables instrumentales, el coeficiente de educación aumenta sustancialmente en el modelo IV básico, lo que es consistente con la idea de que la educación está subestimada en MCO debido a sesgos por habilidad no observada. Sin embargo, cuando se incorporan los controles propuestos por Card o los seleccionados mediante Random Forest, el retorno por IV disminuye y pierde significancia estadística, lo que refleja menor poder del instrumento y mayor variabilidad en estas especificaciones. En general, la tabla sugiere que los retornos a la educación son sensibles al conjunto de controles y al método de estimación, y que la identificación basada en `nearc4` puede debilitarse al introducir controles adicionales.

Los errores estándar usados en la tabla son robustos a heterocedasticidad (HC1) en todos los modelos, salvo en la especificación *IV Card (controles RF)*. Para este último caso se aplicó un procedimiento de bootstrap: se generaron múltiples réplicas de la muestra mediante remuestreo y se volvió a estimar el modelo en cada una de ellas. Con los coeficientes obtenidos se construyó una distribución empírica que permite calcular un error estándar "promedio" y un intervalo de confianza más flexible. Esta estrategia sirve como un chequeo adicional de robustez, especialmente útil cuando el modelo incluye controles seleccionados de manera automática y cuando la estabilidad del estimador puede ser más sensible a la composición de la muestra.

La Tabla 2 presenta las primeras etapas de las tres especificaciones IV. En los tres modelos, el instrumento nearc4 muestra una asociación positiva con los años de educación, aunque con distinta magnitud al incluir controles adicionales. Los modelos con controles, tanto la especificación de Card como la basada en Random Forest, reducen el coeficiente del instrumento, lo que sugiere que parte de la variación inicial estaba correlacionada con características observables. Finalmente, los valores de la F parcial confirman que el instrumento conserva fuerza razonable, aunque es más débil en el modelo con más controles.

De igual forma, se realizó una prueba de igualdad entre los coeficientes de las distintas especificaciones por IV. Los resultados se encuentran en el apendice en la tabla 3. Dicha prueba muestra que ninguna de las diferencias entre especificaciones es estadísticamente significativa. Aunque los valores puntuales de los estimadores difieren entre el IV básico, el IV con controles tipo Card y el IV con controles seleccionados por RF, los *p-value* en todos los contrastes son mayores que los niveles convencionales de significancia. Por lo tanto, no se rechaza la hipótesis de que los coeficientes estimados son iguales. En conjunto, estos resultados indican que las distintas estrategias de control generan estimaciones consistentes del retorno a la educación.

# Discusión

La variable instrumental utilizada por Card, crecer cerca de una universidad (`nearc4`), opera bajo el supuesto de que este instrumento afecta la educación (exógenamente) pero no afecta directamente los salarios excepto a través de la educación. Sin embargo, tanto Kitagawa (2015) como Sloczynski (2021) y Sloczynski et.al. (2025) señalan que este supuesto puede ser delicado en la práctica: si la cercanía a una universidad también está correlacionada con características no observadas (como oportunidades locales de trabajo, recursos familiares o redes sociales), entonces la independencia del instrumento podría estar comprometida. En otras palabras, nearc4 podría estar capturando factores socioeconómicos más amplios que influyen en los salarios además de la educación.

Además, la condición de monotonía es decir, que todos los individuos reaccionen al instrumento en la misma dirección, -por ejemplo "la proximidad de la universidad no debería reducir los logros educativos de ninguna persona"-es difícil de verificar y puede no sostenerse completamente en estos datos. Słoczyński (2021, 2023) argumenta que, bajo heterogeneidad de efectos, los estimadores lineales con instrumentos pueden terminar promediando efectos muy distintos entre subgrupos, lo que complica la interpretación causal tradicional de Card. En esa misma línea, Card (2001) reconoce que distintos instrumentos y especificaciones pueden arrojar estimaciones muy distintas de los retornos a la educación y que la interpretación de los coeficientes debe ser cauta.

En resumen, aunque nearc4 es intuitivamente plausible y ha sido muy influyente, existe preocupación en la literatura de que sus supuestos de validez e independencia no siempre se cumplan estrictamente. Esto sugiere que los resultados obtenidos con este instrumento deben interpretarse como evidencia parcial del retorno causal a la educación y acompañarse de análisis de robustez o de métodos que permitan heterogeneidad en los efectos.

# *Instrumental Causal Forest*

Por último, para explorar si el retorno a la educación varía entre distintos tipos de personas, implementamos un Causal Forest con variables instrumentales (Instrumental Forest). Esta herramienta permite estimar cómo cambia el efecto causal de la educación sobre el salario según las características individuales, controlando al mismo tiempo la endogeneidad mediante el instrumento de proximidad a una universidad. A diferencia de un IV tradicional, que solo reporta un coeficiente promedio, el Causal Forest produce un estimador individualizado del efecto causal, denotado como $\hat{\tau}_i$ para cada persona de la muestra.

El modelo utilizó como variables explicativas las mismas que entraron en el IV con con-

troles seleccionados por Random Forest. Después de entrenar el bosque, se obtuvieron los efectos causales individuales y se calculó su promedio: el valor estimado de $\tau$ fue cercano a 0.033, lo que sugiere que, en promedio, un año adicional de educación incrementa el salario en aproximadamente 3.3% para los individuos que están en el "margen" inducido por el instrumento. Aunque este promedio es positivo, la desviación estándar fue relativamente alta, lo que indica bastante heterogeneidad en los retornos educativos dentro de la población estudiada.

El bosque también reportó medidas de importancia de variables, que ayudan a entender qué características explican mejor la variación en los retornos. Los resultados de esta especificación se encuentran en la figura2. Entre las variables más influyentes aparecen peso, IQ, conocimiento del mercado laboral (kww) y educación de los padres. Esto sugiere que los beneficios de estudiar un año extra no son iguales para todos: parecen ser mayores para personas con mayor habilidad cognitiva, mejor información laboral o una base socioeconómica más favorable.

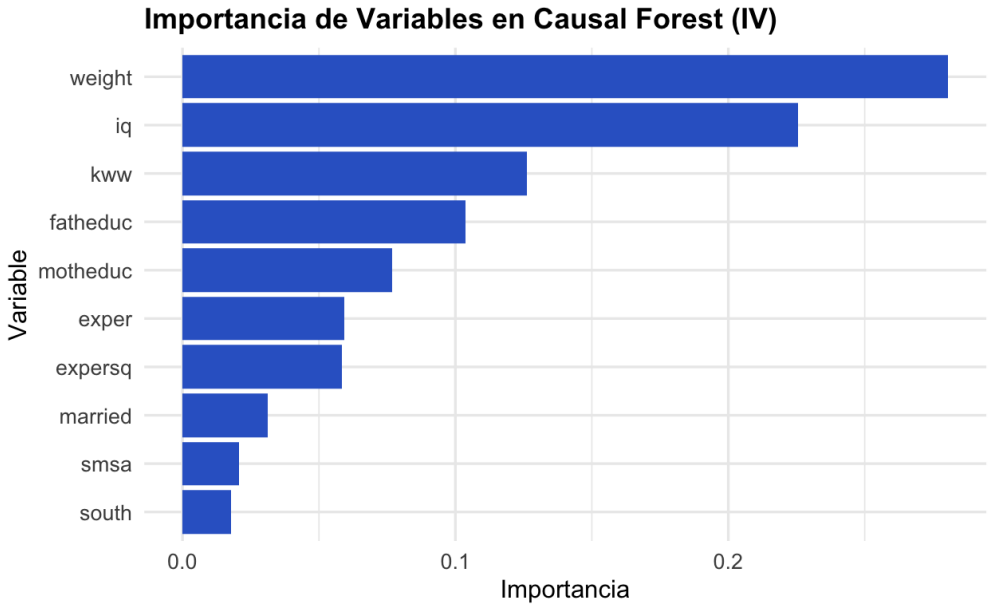

Figure 2: Importancia de Variables

Al dividir la muestra en terciles de kww (conocimiento económico), se observa que el primer tercil muestra un retorno casi nulo, mientras que los dos superiores presentan efectos cercanos al 5%. La Tabla 4 resume esta heterogeneidad y muestra además que la variabilidad del efecto disminuye en los grupos con mayor kww, lo que sugiere que quienes poseen más habilidades o información económica tienden a beneficiarse de manera más consistente de continuar estudiando.

# Conclusión

En este reporte se replicó la estrategia clásica de Card (1993) para estimar retornos a la educación, comparando un punto de partida por MCO con una estimación por variables instrumentales usando `nearc4` (cercanía a una universidad) como fuente de variación exógena en escolaridad. Los resultados confirman que el retorno estimado depende de manera importante tanto del método de estimación como del conjunto de controles: en MCO, el coeficiente de educación es positivo y estadísticamente significativo, pero cae al incorporar controles demográficos y de origen familiar en el espíritu de Card. Al pasar a IV, el estimador básico arroja un retorno sustancialmente mayor, mientras que al introducir controles adicionales (ya sea el set tipo Card o controles seleccionados con Random Forest) la estimación se vuelve mucho más imprecisa y pierde significancia estadística.

Una lectura central del ejercicio es que, en estos datos, la identificación basada en `nearc4` es sensible a la especificación. En particular, las primeras etapas muestran que el poder del instrumento puede deteriorarse al ampliar el vector de controles, lo que sugiere posibles problemas de instrumentos débiles y obliga a interpretar con cautela los coeficientes IV con mayor cantidad de controles. En este sentido, el reporte no solo ilustra la lógica de IV y su interpretación como efecto local (LATE), sino también la importancia de reportar explícitamente la primera etapa, estadísticos de relevancia y chequeos de robustez, pues la evidencia causal puede variar sustancialmente con cambios razonables en la especificación.

Finalmente, se incorporó una lógica de ciencia de datos para complementar (no sustituir) la identificación econométrica: el Random Forest se utilizó como herramienta práctica para ordenar covariables candidatas y proponer especificaciones más parsimoniosas, y el Instrumental Forest permitió explorar heterogeneidad potencial en los retornos a la educación. El mensaje general es doble: por un lado, los métodos de Machine Learning pueden ser útiles para organizar controles y detectar patrones que se pierden con un único efecto promedio; por otro, la validez causal sigue descansando en los supuestos del instrumento y en su fuerza empírica. Por ello, una extensión natural sería reforzar el análisis con pruebas y diagnósticos específicos para validez/relevancia del instrumento, así como con análisis de sensibilidad y, cuando sea posible, instrumentos alternativos o diseños complementarios. En otras palabras, estos metodos pueden ayudarnos a complementar las herramientas del econometrista para descartar instrumentos débiles, así como encontrar instrumentos fuertes, que, combinado con un buen *setting* causal, es esperable que puedan mejorar la investigación en las ciencias sociales en el futuro.

# Referencias

Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (NBER Working Paper No. 4483). National Bureau of Economic Research. https://doi.org/10.3386/w4483

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica, 69*(5), 1127–1160. http://www.jstor.org/stable/2692217

Kitagawa, T. (2015). A test for instrument validity. *Econometrica, 83*(5), 2043–2063. http://www.jstor.org/stable/43616996

Sloczynski, T. (2021). *When should we (not) interpret linear IV estimands as LATE?* IZA Discussion Paper No. 14349. Institute of Labor Economics (IZA). https://www.ifo.de/DocDL/cesifo1_wp9064.pdf

Słoczyński, T., Uysal, S. D., & Wooldridge, J. M. (2025). Abadie's kappa and weighting estimators of the local average treatment effect. *Journal of Business & Economic Statistics, 43*(1), 164–177. https://doi.org/10.1080/07350015.2024.2332763

# Apendice

## Primera Etapa de los modelos IV

Table 2: Primera etapa de los modelos IV: Especificación básica, Card y controles seleccionados por RF.

| | **IV básico** | **IV Card** | **IV Card (controles RF)** |
|---|---|---|---|
| (Intercept) | 3.310 | -0.428 | -5.125$^{+}$ |
| | (1.175) | (-0.154) | (-1.726) |
| nearc4 | 0.840*** | 0.213$^{+}$ | 0.262* |
| | (8.109) | (1.937) | (2.466) |
| exper | 0.853*** | 0.794** | 0.792** |
| | (3.399) | (3.207) | (3.055) |
| expersq | -0.019*** | -0.016** | -0.017** |
| | (-3.456) | (-2.983) | (-2.934) |
| black | | -0.167 | |
| | | (-1.181) | |
| south | | -0.039 | 0.339*** |
| | | (-0.372) | (3.382) |
| smsa | | 0.414*** | 0.099 |
| | | (3.632) | (0.890) |
| fatheduc | | 0.210*** | 0.121*** |
| | | (12.379) | (7.116) |
| motheduc | | 0.197*** | 0.090*** |
| | | (9.832) | (4.533) |
| kww | | | 0.050*** |
| | | | (6.416) |
| married | | | 0.085*** |
| | | | (3.683) |
| weight | | | -0.000** |
| | | | (-2.634) |
| iq | | | 0.054*** |
| | | | (15.042) |
| Num.Obs. | 3010 | 2220 | 1600 |
| F | 25.584 | 102.795 | 82.725 |
| F parcial nearc4 | 65.76 | 3.75 | 6.08 |

## Comparación de resultados

Table 3: Comparación de coeficientes IV para la estimación del retorno a la educación

| Comparación | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $t$ | *p-value* |
|---|---|---|---|---|
| IV básico vs IV Card | 0.1736 | 0.0718 | 1.111 | 0.2667 |
| IV básico vs IV Card (RF) | 0.1736 | 0.0321 | 1.705 | 0.0882 |
| IV Card vs IV Card (RF) | 0.0718 | 0.0321 | 0.334 | 0.7386 |

Nota: prueba t basada en la diferencia entre coeficientes utilizando errores estándar robustos. No se rechaza la igualdad de coeficientes en ningún caso.

## Heterogeneidad del efecto

Table 4: Heterogeneidad del efecto causal por terciles de KWW

| **Tercil de KWW** | **Media de $\hat{\tau}$** | **Desv. Est.** | **N** |
|---|---|---|---|
| 1 | 0.0013 | 1.0152 | 534 |
| 2 | 0.0496 | 0.9306 | 533 |
| 3 | 0.0492 | 0.3292 | 533 |

## Random Forest plot

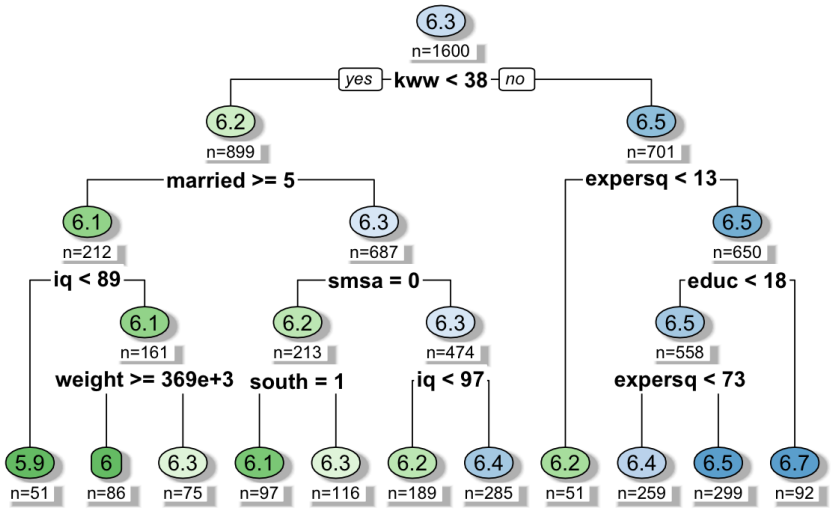


Figure 3: Random Forest