**ENTREGA 6 - MODELO DE REGRESIÓN - GRUPO 2**

**Información general:**

**Tema**: Desigualdad de género

**Objetivo**: Determinar los factores que explican e influyen en la desigualdad de género a nivel mundial.

**Link del repositorio:** <https://github.com/DanaeNorah/Estadistica-para-el-Analisis-Politico-2-/tree/master/TRABAJO/Hito%204>

**Objetivo del estudio:** Determinar los factores que explican e influyen en la desigualdad de género a nivel mundial.

**Hipótesis planteadas:**

**H1**: La desigualdad de género tiene relación con la falta de marcos institucionales que protegen a la mujer frente a abusos. Estos marcos se pueden desagregar en las variables observables autonomía reproductiva y violencia contra la mujer. Entonces a menor desarrollo de marcos institucionales, mayor desigualdad de género.

**H2:** La desigualdad de género tiene una relación indirecta con las libertades civiles. Es decir, a mayor desigualdad de género menor libertades civiles. Por ello, las mujeres no tienen confianza en el sistema de justicia, sufren de acoso callejero y necesitan mayor participación política para visibilizar sus libertades civiles.

**H3:** La desigualdad de género tiene una relación indirecta con la educación. Debido a que, a menor educación de las mujeres mayor desigualdad de género. Por ello, la diferencia en los grados y tipos de educación entre hombres y mujeres limita sus oportunidades en el futuro.

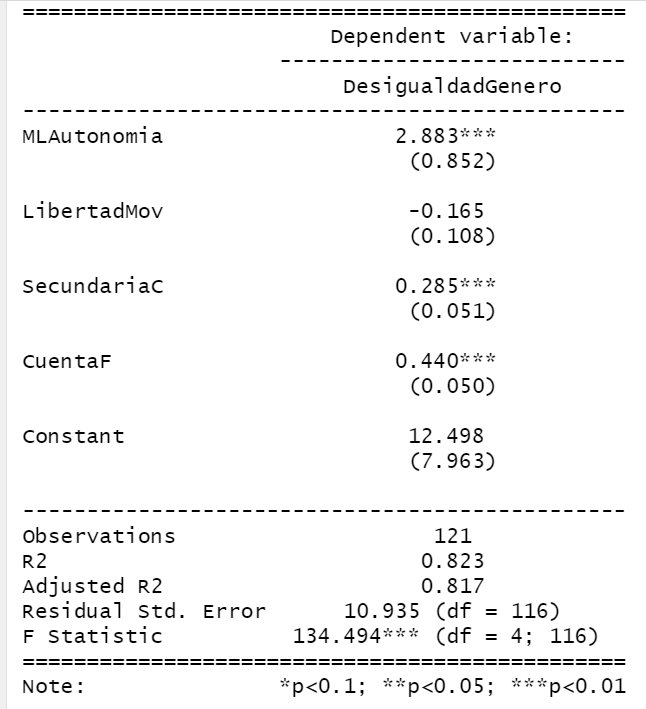
**H4:** La desigualdad de género está relacionada de manera indirecta con la independencia económica; en otras palabras, a mayor desigualdad de género, la mujer cuenta con menores oportunidades laborales y con menos posibilidades de disponer de servicios financieros. Esto responde a estructuras económicas que excluyen a las mujeres e impiden su empoderamiento y desarrollo económico.

**Modelos de regresión:**

Al no ser viable el Análisis Factorial Exploratorio se emplea la Regresión Lineal Múltiple con las variables observables (Anexo 1). Primero, analizamos si la desigualdad de género está correlacionada con las variables observables. Se encuentra que la desigualdad de género no tiene una correlación significativa con la desconfianza en el sistema de justicia y con el marco legal de autonomía reproductiva. Por lo tanto, con las variables que están correlacionadas con la desigual de género se crean los siguientes modelos:

**MODELO 1:**

**Cuadro 1:**

****

**Cuadro 2:**

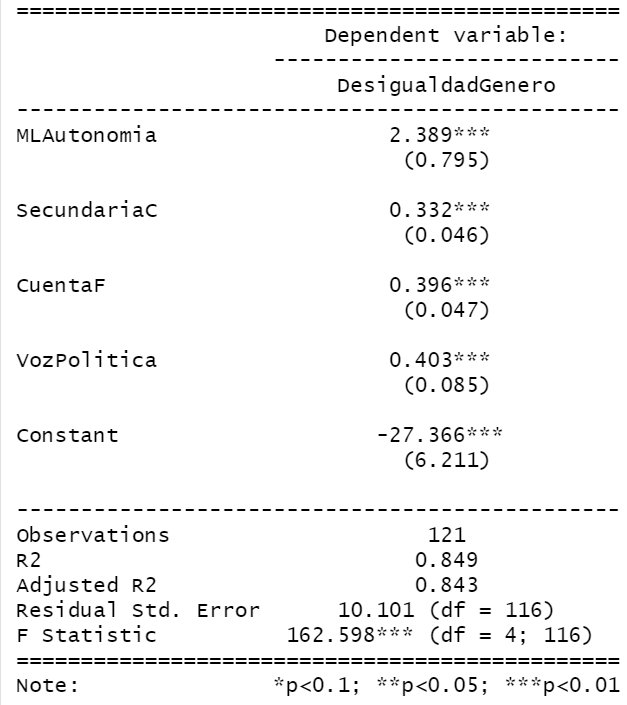
|  |  |
| --- | --- |
| Fórmula seguida | DesigualdadGenero ~ MLAutonomia + LibertadMov + SecundariaC + CuentaF |
| Interpretación de la tabla ANOVA | Se observa que los p-values de las cuatro variables (<2.2e-16; 1.026e-12, <2.2e-16; 1.878e-14) son significativas pues tienen un valor menor a 0.05. |
| Interpretación de Coeficientes | Constatamos que solo tres de las cuatro variables tienen un p value menor a 0.05. Incluso es menor a 0.01 lo que se toma como un indicador positivo y podemos concluir que estas sí son significativas. |
| Interpretación de poder explicativo | Este grupo de variables tienen un alto porcentaje de explicación (81.7%). Por ello, podemos decir que el modelo sí funciona. |
| Ecuación final | Modelo 1 = 12.49 + (2.883\*Autonomía reproductiva) - (0.165\*LibertadMov) + (0.285\*Secundaria Completa) + (0.440\*CuentaF) |

**Cuadro 3:**

|  |  |
| --- | --- |
| Linealidad | En el gráfico, la línea roja se aleja bastante a los extremos y al centro de la línea punteada. Por lo tanto, podemos concluir que el modelo no tiene linealidad. |
| Homocedasticidad | En el gráfico, gran parte de la línea roja está pegada a la línea punteada, pero se aleja en el extremo izquierdo dando indicio de que no habría homocedasticidad. Por esa razón, recurrimos al test de Breusch-Pagan para corroborar. Obtenemos un p value mayor a 0.05 (0.6923) por lo que no se rechaza la H0 y podemos decir que sí hay homocedasticidad. |
| Normalidad de los residuos | El criterio es que los puntos finales se encuentren sobre la línea punteada. En este caso, solo algunos de los puntos finales se alejan lo que deja lugar a dudas, por lo que recurrimos a la prueba Shapiro para corroborar. Obtenemos un p value menor a 0.05 por ello se concluye que los residuos no cuentan con una distribución normal. |
| No multicoli- nealidad | Los predictores no tienen una correlación muy alta entre sí y todos los valores obtenidos fueron menores a 5, por lo que podemos concluir que no hay multicolinealidad. |
| Valores  influyentes | En el gráfico, la línea roja se encuentra bastante cerca de la línea punteada lo que representa un buen indicio. Luego de revisar las distancias cook y hat podemos corroborar que no hay valores que distan mucho del resto del grupo y no están alterando o apalancando el modelo. |

**MODELO 2:**

**Cuadro 1:**

****

**Cuadro 2:**

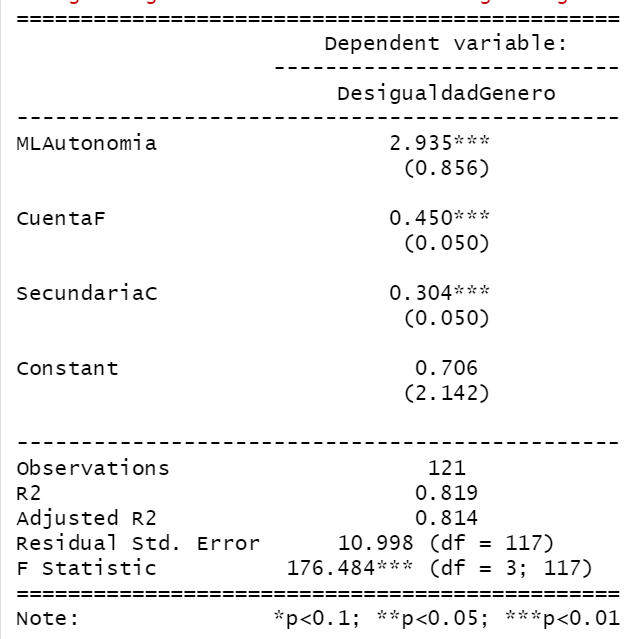
|  |  |
| --- | --- |
| Fórmula seguida | DesigualdadGenero ~ MLAutonomia + SecundariaC + CuentaF + VozPolitica |
| Interpretación de la tabla ANOVA | Se observa que los p-values de las cuatro variables (<2.2e-16; <2.2e-16, <2.2e-16; 5.512e-06) son significativas pues tienen un valor menor a 0.05. |
| Interpretación de Coeficientes | Constatamos que las cuatro variables tienen un p value menor a 0.05, lo que se toma como un indicador positivo y podemos concluir que estas sí son significativas y aportan al modelo. |
| Interpretación de poder explicativo | Este grupo de variables tienen un alto porcentaje de explicación (84.3%). Por ello, podemos decir que el modelo sí funciona. |
| Ecuación final | Modelo 2 = -27.36 + (2.389\*Autonomía reproductiva) + (0.332\*Secundaria Completa) + (0.396\*CuentaF) + (0.403\*VozPolitica) |

**Cuadro 3:**

|  |  |
| --- | --- |
| Linealidad | En el gráfico, la línea roja está casi totalmente pegada a la línea punteada y solo se aleja un poco a los extremos y al centro. Por lo tanto, podemos concluir que el modelo tiene linealidad. |
| Homocedasticidad | En el gráfico, gran parte del extremo izquierdo de la línea roja se aleja de la línea punteada dando indicio de que no habría homocedasticidad. Por esa razón, recurrimos al test de Breusch-Pagan para corroborar. Obtenemos un p value mayor a 0.05 (0.4805) por lo que no se rechaza la H0 y podemos decir que sí hay homocedasticidad. |
| Normalidad de los residuos | El criterio es que los puntos finales se encuentren sobre la línea punteada. En este caso, solo hay unos pocos puntos extremos se alejan, lo que deja lugar a dudas, por lo que recurrimos a la prueba Shapiro para corroborar. Obtenemos un p value mayor a 0.05 (0.5825), por ello se concluye que los residuos sí cuentan con una distribución normal. |
| No multicoli- nealidad | Los predictores no tienen una correlación muy alta entre sí y todos los valores obtenidos fueron menores a 5, por lo que podemos concluir que no hay multicolinealidad. |
| Valores  influyentes | En el gráfico, la línea roja se encuentra pegada casi en su totalidad a la línea punteada lo que representa un buen indicio. Luego de revisar las distancias cook y hat podemos corroborar que sí hay un valor que dista mucho del resto del grupo y está alterando o apalancando el modelo. |

**Modelo 3**

**Cuadro 1:**

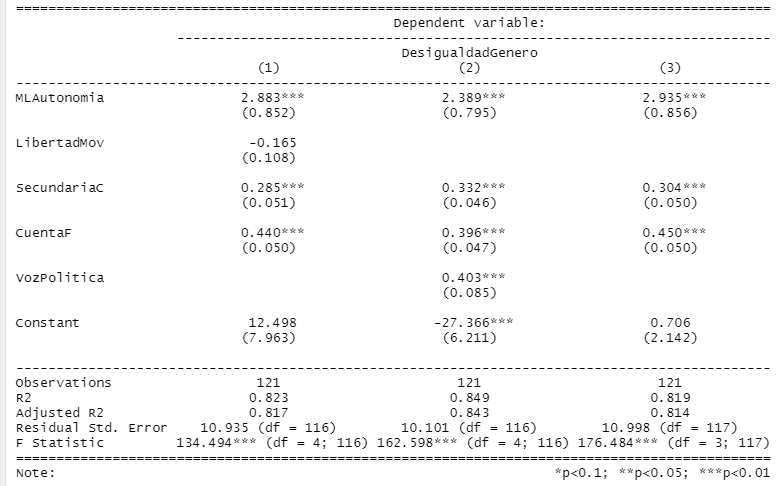


**Cuadro 2:**

|  |  |
| --- | --- |
| Fórmula seguida | DesigualdadGenero ~ MLAutonomia + CuentaF + SecundariaC |
| Interpretación de la tabla ANOVA | Se observa que los p-values de las 3 variables (<2.2e-16; <2.2e-16; 1.532e-8) son significativas pues tienen un valor menor a 0.05. |
| Interpretación de Coeficientes | Se observa que los coeficientes de las variables tienen un p-value menor a 0.01, por lo que son significativos y aportan al modelo. |
| Interpretación de poder explicativo | Este modelo explica el 81.4 % de la variabilidad de la desigualdad de género, lo cual es bastante alto. |
| Ecuación final | Modelo 3= 0.71 + (2.935\*Autonomía reproductiva)+(0.45\*Cuenta financiera)+(0.304\*Secundaria Completa) |

**Cuadro 3:**

|  |  |
| --- | --- |
| Linealidad | La línea que se obtiene es parecida a una línea horizontal, por lo tanto si cumple con el supuesto de linealidad |
| Homocedasticidad | De acuerdo al gráfico, la línea no es clara como para comprobar la homocedasticidad, pero si se aplica el test de Breusch-Pagan se obtiene que el p-value es mayor a 0.05 (0.372), por lo que aceptamos la H0 que nos dice que el modelo es homocedástico. |
| Normalidad de los residuos | Preliminarmente, con el gráfico se observa que una gran parte de los residuos están en la línea de normalidad. Aplicando el Shapiro test, se obtiene que el p-value es mayor a 0.05 (0.085), por lo que aceptamos la H0 que nos dice que los residuos tienen distribución normal. |
| No multicolinealidad | Los valores (1.457; 2.086; 2.238) muestran que no existe los predictores no tienen una correlación muy alta entre ellos, por lo que no hay multicolinealidad. |
| Valores influyentes | No se observan valores influyentes que apalanquen el cálculo del modelo. |

**Comparación de modelos**

Se puede observar que el modelo que mejor explica la variabilidad de la desigualdad de género, tomando en cuenta el principio de parsimonia, es el Modelo 3 debido a que explica un 81.9% con solo 3 variables (MLAutonomía, SecundariaC y CuentaF).

**Conclusión:**

Las hipótesis planteadas al inicio de la investigación se cumplen de forma parcial. Luego de realizar el análisis factorial y posteriormente los modelos de regresión, podemos concluir que algunas de las variables observables no terminan aportando mucha información. Las variables que tienen explican mejor la desigualdad de género son la autonomía reproductiva, secundaria completa y cuenta bancaria, que corresponden a las variables latentes marcos institucionales, educación e independencia económica respectivamente. Estas tres podrían ser tomadas como ejes principales en donde los estados deberían poner mayor énfasis. Se necesitan políticas públicas que promuevan el empoderamiento de la mujer y le permitan tener una mayor autonomía sobre sus decisiones tanto personales como profesionales. Cabe mencionar que sería interesante ampliar un poco más la investigación sobre la necesidad de implementar marcos legales que realmente protejan a las mujeres en situaciones de abuso y vulnerabilidad. Adicionalmente, el grupo sugeriría explorar más el impacto de la presencia de mujeres en el sistema político y en el mercado laboral.

**ANEXO**

**Anexo 1: Análisis Factorial Exploratorio:**

En trabajos anteriores hemos reportado la particularidad de la variable que mide la proporción de mujeres a hombres que están desempleadas (*DesempleoMuj*). Esta ha llamado la atención porque al momento en que realizamos el EFA, se puede observar que esta variable aporta a más de un factor, pero no significativamente. Asimismo, otro problema con “*DesempleoMuj*” es que en el "resultado mejorado" la variable no aporta a ninguno de nuestros tres factores planteados en esa ocasión. Además, esta misma es la segunda que menos aporta a los factores. A pesar de los resultados de esas tres situaciones ya expuestas, en la prueba *"complexity*", se obtuvo que la proporción desempleo de mujeres sobre hombres contribuye a más de un factor. Estos resultados, en primera instancia, nos resultaron interesantes pues altera el desarrollo y resultados obtenidos durante el EFA. Por ello, hemos decidido ejecutar nuestras operaciones para el análisis a través de factores excluyendo a esta variable. Según la recomendación del programa, estos factores se agrupan en tres.

De acuerdo a los datos obtenidos del KMO podemos decir que la relación entre las variables puede explicar el EFA porque su Overall MSA es mayor a 0.5. Asimismo, con las pruebas *Test de Bartlett* y *Test For Singular Square Matrix* podremos saber si la matriz de correlaciones es adecuada. En ambos casos, las hipótesis nulas han resultado falsas; es decir, la matriz de correlación no es una matriz identidad y tampoco una matriz singular. Luego, el comando *fa.parallel* nos recomienda 3 factores con los que podríamos redimensionar la subdata.

Visualizando el cuadro *loadings,* para notar el poder explicativo de nuestros factores recomendados, se obtiene que la variable que mide el marco legal con respecto a la autonomía reproductiva (*MLAutonomia*) contribuye a los factores 1 y 3, con 0.663 y 0286 respectivamente. La variable que mide el marco legal con respecto a la violencia (*MLViolencia*) contribuye solamente al factor 3, con un 0.427. Además, la que se refiere a la voz política (VozPolitica) contribuye a los factores 1 y 3, con 0.149 y 0.504 respectivamente. La variable que mide la desconfianza de las mujeres en el sistema de justicia (*ConfianzaSJ*) contribuye solo al factor 2 con 0.884. La variable que mide el porcentaje de mujeres de 25 años a más que ha llegado al nivel de educación secundaria (*SecundariaC*) contribuye a los factores 1 y 2 con 0.912 y -0.326 respectivamente. Finalmente, la variable que mide el acceso a los servicios financieros (*CuentaF*) contribuye a los factores 1, 2 y 3, con 0.708, -0.108 y 0.373 respectivamente. Respecto a esto, se observa que las variables no son muy significativas, pues pocas de ellas son mayores a 0.7.

Agregado a ello, el cuadro anterior nos brinda una varianza acumulada de 57.3%, lo cual es un problema para explicar la variabilidad de los factores, pues se pierde información. Por ello, hemos decidido que lo más viable es no usar el EFA. Por otro lado, cuando mejoramos el resultado con un corte en 0.5, se puede observar que las variables “*ConfianzaSJ”, “SecundariaC” y “CuentaF”* son las únicas que contribuyen significativamente a un factor, con un valor a mayor 0.7. Asimismo, gracias al diagrama de Análisis factorial, sabemos que en el primer factor encontramos a las variables *"MLAutonomia", "SecundariaC" y "CuentaF"*; mientras que el segundo está compuesto por "*LibertadMov*" y "*DesconfianzaSJ*"; y finalmente, el tercero contiene a *"MLViolencia"* y "*VozPolitica"*. Es interesante resaltar en este punto que los tres factores obtenidos en el análisis factorial no corresponden a las variables propuestas inicialmente según nuestra revisión bibliográfica.

Asimismo, es necesario evaluar la validez del EFA solicitado; para ello, esta debe de cumplir con ciertas condiciones. En primer lugar, nuestro análisis debe de cumplir con una raíz de error cuadrático corregido cercano a cero, pero, en este caso, el valor de la raíz es de 0.02, por ello, se debe de reportar. El segundo requisito a cumplir es que la raíz del error cuadrático medio de aproximación (RMSEA) sea menor a 0.5 para que sea considerado como óptimo. En este caso, el índice RMSEA obtenido es 0.00; el cual es menor a 0.5. El siguiente requisito a cumplir es que el índice de Tucker- Lewis debe de ser mayor a 0.9. En nuestro caso, es 1.06; es decir, cumple con la condición. Todas nuestras variables cuentan con presencia de decimales, es decir, cada una de ellas aportan a más de un factor; empero las que presentan un valor de "*complexity*" más elevado son *"MLAutonomia", "CuentaF"* y *"LibertadMov"*.