***ENTREGA 2 DISEÑO DE LA INVESTIGACIÓN SOCIAL***

Integrantes: Benjamín Carreño Canales y Ramón Jara López

**Parte I: Ejercicios manuales (No STATA)**

**1) Explique que representa el R2 en una regresión bivariada. De ejemplos**

Cuando hablamos de R2 se habla del coeficiente de determinación. Esta medida estadística se utiliza en el análisis de regresión para evaluar la bondad del ajuste de un modelo determinado. En este sentido, el R2 nos entrega información sobre cuánta de la variabilidad en la variable dependiente (Y) puede explicarse por el modelo de regresión en relación con la variabilidad total observada en los datos.

El valor que puede tomar esta medida varía entre 0 y 1, y en este sentido, si es 0 indica que el modelo no explica ninguna variabilidad en la variable dependiente. Si es 1, indica que el modelo explica toda la variabilidad en la variable dependiente. Ahora, si se encuentra entre 0 y 1 significa que explica una parte de la variabilidad en la variable dependiente (porción porcentual). La situación más común es la descrita anteriormente. Por ejemplo, si R2 es igual a 0.45 significa que el 45% de la variabilidad en la variable dependiente es explicada por el modelo de regresión, el resto se debe a otros factores no incluidos en el modelo.

**2) La definición matemática de R2 corresponde a**

**R2= 1 – SSR/TSS**

1. **Describa cada uno de los términos (por separado) que componen la definición matemática de R2.**

El término SSR (la suma de residuos al cuadrado) mide la variabilidad de los residuos. Entendido como la diferencia entre los valores observados (Y) y los valores predichos de Y. Por otro lado, el TSS (la suma total de cuadrados) mide la variabilidad total en los datos de la variable dependiente Y. El cálculo se realiza a partir de la suma de los cuadrados de las diferencias entre los valores observados (Y) y la media de Y.

Ahora bien, siguiendo la definición matemática de R2, ambos elementos se dividen entre sí porque miden la proporción de la variabilidad total que no es explicada por el modelo. En otras palabras, esta fracción representa la variabilidad que queda sin explicar luego del ajuste del modelo. Es, por tanto, que se resta por 1 para obtener la proporción de la variabilidad total que sí se explica por el modelo (y así pudiendo interpretar R2 como resultado de la medición de la bondad de ajuste).

1. **Explique a que refiere el componente: SSR/TSS**

Como se mencionó anteriormente, la división SSR/TSS sirve para obtener la proporción de la variabilidad total que no se explica por el modelo. Ahora, para explicarlo de manera más profunda habría que partir indicando que la esencia de la regresión lineal es descomponer la variabilidad total de la variable dependiente en dos partes, la variabilidad que se explica por el modelo y la que no (esta última denominada como residuos).

Por ejemplo, imaginemos el caso hipotético de que TSS (Suma total de cuadrados) es todo el “espacio” de la variabilidad que se busca entender. La SSR (suma residual de cuadrados) es la parte de dicho “espacio” que nuestro modelo no es capaz de explicar. Al calcular la proporción de este “espacio” no explicado y restarlo por 1 se responde básicamente a la pregunta “¿Cuánto del total hemos logrado captar con nuestro modelo?”.

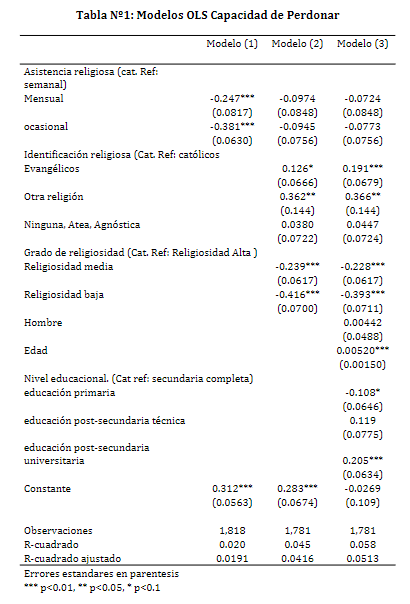
**3) Explique a que refieren los términos: resultados o valores observados (Y), resultados/valores esperados (‘Y) y errores de predicción (e).**

Cuando hablamos de valor observado (Y), se refiere al valor real de la variable dependiente que se ha medido en un conjunto especifico de valores de la variable dependiente. El valor observado es importante ya que sirve como referencia para evaluar el modelo predictivo comparando los valores predichos por el modelo. Además, sirve para calcular los residuos (se explicará más adelante) entre el valor observado y el predicho.

Por otro lado, cuando hablamos de valor esperado (‘Y), se hace referencia a los valores de la variable dependiente que se estiman o calculan utilizando un modelo de regresión lineal a partir de los valores de la variable independiente en ese modelo. Es decir, son las estimaciones de Y basadas en la relación matemática establecida por el modelo de regresión lineal.

Finalmente, los errores de predicción o los residuos (e) son las diferencias entre los primeros dos conceptos mencionados, es decir, los valores observados y los esperados. Los residuos también permiten evaluar la precisión del modelo, para ver qué tan bien predicen los valores observados, por otro lado, también ayudan a medir y analizar los errores en el modelo.

**4) La Encuesta Bicentenario del 2016 incorporó una serie de preguntas acerca de la concepción del perdón. La variable de interés se denomina “Escala de perdón”, la cual es una variable continua que va desde 0 a 7 puntos. Los investigadores desean conocer si la asistencia religiosa influye en la capacidad de perdonar para ello estimaron tres modelos de regresión.**



1. **Interprete para los tres modelos el coeficiente de regresión de la variable asistencia religiosa. ¿Qué ocurre con este coeficiente en el modelo 2 con la incorporación de las variables Identificación religiosa y Grado de religiosidad?**

Para el primer modelo, las personas que asisten mensualmente al servicio religioso ven disminuidos en 0,247 puntos su escala del perdón (la capacidad de perdonar) respecto a aquellos que asisten semanalmente. La capacidad de perdonar para las personas que asisten ocasionalmente disminuye aún más, en este caso en 0,381 puntos de la escala del perdón, a diferencia de las personas que asisten al menos una vez a la semana. Ambos coeficientes de regresión reportados son significativos a un 99% de confianza.

Con respecto al segundo modelo, se observa que la capacidad de perdonar de las personas que asisten al menos una vez al mes a servicios religiosos disminuye en 0,097 puntos la escala de perdón respecto a las personas que van al menos una vez a la semana, dejando el resto de variables constantes. En el caso de la capacidad de perdonar de las personas que van ocasionalmente disminuye en 0,095 unidades menos que las personas que van al menos una vez a la semana, ceteris paribus. Cabe destacar que ninguno de estos dos coeficientes de regresión es significativo.

Finalmente, en el último modelo, la capacidad de las personas para perdonar que van mensualmente al servicio religioso disminuye en 0,072 puntos de la escala del perdón, respecto a quienes van semanalmente, ceteris paribus. En el caso de las personas que van ocasionalmente, disminuye en unas 0,077 unidades, dejando el resto de variables constantes. Al igual que en el modelo anterior, ninguno de los dos coeficientes es significativo.

Respondiendo a la última pregunta, se puede ver que al agregar más variables al modelo los coeficientes disminuyen notablemente y dejan de ser estadísticamente significativos. Esa situación puede tener muchas explicaciones. Una de ellas es que es probable que las nuevas variables integradas capten mejor la variabilidad en la capacidad de perdonar respecto a lo que explicaba el modelo 1. Es decir, las variables de identificación religiosa, el grado de religiosidad, el sexo, la edad, el nivel educacional, están explicando mejor parte del efecto que antes se atribuía a la asistencia a servicios religiosos.

1. **Interprete para el tercer modelo el coeficiente de edad.**

Un año más de vida aumenta en 0,0052 unidades de medida la capacidad de perdonar de las personas, manteniendo el resto de las variables constantes.

El coeficiente es significativo (no incluye el cero) y positivo, pero tiene una magnitud muy pequeña, por lo que el efecto en cuestión de la edad respecto a la capacidad de perdonar es mínimo, ceteris paribus.

1. **Determine cuál de los tres modelos tiene un mejor ajuste e interprete.**

El R2 del primer modelo es 0,020, lo que implica que sólo el 2% de la variabilidad total de la variable dependiente es explicada por el modelo. El R2 ajustado del modelo es de 0,019.

Para el segundo modelo, el R2 aumenta a 0,045, es decir el modelo explica un 4,5% de la variabilidad total de la variable dependiente. El R2 ajustado asociado es 0,041.

El R2 para el tercer modelo (el que presenta mejor ajuste) es de 0,058, esto nos indica que el 5,8% de la variabilidad total de la variable dependiente es explicada por el modelo. Esto nos podría llevar a inferir que el modelo no está capturando mucha información relevante, y que hay mucha variabilidad en la variable que no está siendo explicada por las variables dependientes en cuestión. El R2 ajustado es de 0,0513. La pequeña diferencia entre ambos números (R2 y R2 ajustado) indica que el modelo no se ve afectado de manera significativa con la inclusión de variables adicionales.

**Parte II: Ejercicios STATA**

**Para responder a las siguientes preguntas utilice STATA y la base de datos escogida para realizar el trabajo de investigación durante el semestre. Es relevante que den cuenta de manera sintética y entendible de todas las decisiones metodológicas realizadas en su trabajo. Presente los resultados de sus análisis en una tabla editada.**

1. **Indique la variable dependiente e independiente (y de control) que utilizará en su análisis. Defina cuál es la pregunta de investigación que quiere responder en sus análisis ¿Esta pregunta des descriptiva o predictiva?**

Para el análisis de la investigación en curso se toma como variable dependiente la Educación por años (variable continua a partir de una construcción categórica), como variable independiente la Asistencia religiosa (variable categórica) y como variables de control el Nivel socioeconómico (NSE), el Sexo (dicotómico) y la Posición política (categórica y distribuida en escala likert).

La pregunta de investigación se formula así: **¿cómo influye la religiosidad activa, es decir, la intensidad de asistencia a eventos religiosos, en la cantidad de años educativos en Chile?** En tal sentido, se busca describir la posible influencia que tiene la actividad religiosa persistente, controlando por sexo, nivel socioeconómico y posición política, en los años de escolaridad alcanzados por persona.

1. **Estime un primer modelo de regresión, el que debe ser bivariado. Describa los resultados del modelo. Llame a este modelo, Modelo 0 (este es el modelo base). Describa la o las hipótesis que está evaluando en este modelo. Escriba la ecuación que representa a este Modelo. Estime un segundo modelo de regresión, el que debe ser multivariado. Describa los resultados del modelo. Llame a este modelo, Modelo 1. Describa la o las hipótesis que está evaluando en este modelo. Escriba la ecuación que representa a este Modelo. Proponga un tercer modelo de regresión. Llame a este Modelo 2, el que debe ser multivariado. Evalúe dos hipótesis en este modelo. Describa sus resultados respecto a sus hipótesis.**

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLA N°1: MODELOS OLS AÑOS EDUCATIVOS** | | |  |
|  | Modelo 0 | Modelo 1 | Modelo 2 |
| Religiosidad activa (Cat. Ref.: Alta) |  |  |  |
| Media | 0.439 | 0.520\*\* | 0.510\*\* |
|  | (0.278) | (0.253) | (0.253) |
| Baja o nula | 0.860\*\*\* | 0.798\*\*\* | 0.784\*\*\* |
|  | (0.248) | (0.227) | (0.229) |
| NSE (Cat. Ref.: AB) |  |  |  |
| C1a-C1b |  | -0.381 | -0.167 |
|  |  | (2.180) | (2.181) |
| C2 |  | -2.561 | -2.405 |
|  |  | (2.145) | (2.145) |
| C3 |  | -4.510\*\* | -4.349\*\* |
|  |  | (2.139) | (2.139) |
| D |  | -6.100\*\*\* | -5.920\*\*\* |
|  |  | (2.143) | (2.143) |
| E |  | -7.117\*\*\* | -6.877\*\*\* |
|  |  | (2.231) | (2.233) |
| Mujer |  | -0.0557 | -0.0522 |
|  |  | (0.166) | (0.165) |
| Posición política (Cat. Ref.: Izquierda) |  |  |  |
| Centro |  |  | 0.294 |
|  |  |  | (0.192) |
| Derecha |  |  | -0.0718 |
|  |  |  | (0.230) |
| Constante | 11.80\*\*\* | 16.27\*\*\* | 15.98\*\*\* |
|  | (0.219) | (2.139) | (2.146) |
|  |  |  |  |
| N | 1,420 | 1,420 | 1,420 |
| R-cuadrado | 0.009 | 0.183 | 0.185 |
| R-cuadrado ajustado | 0.008 | 0.178 | 0.179 |
| Errores estándares en paréntesis | |  |  |
| \*\*\* p<0.01, \*\* p<0.05, \* p<0.1 | |  |  |

a.) Respecto al modelo 0 (regresión bivariada) se observa en la Tabla 1 que la relación entre las variables religiosidad activa, que mide la intensidad de la asistencia religiosa, y la educación, medida en años, es significativa al 99% de confianza solo si se considera la asistencia religiosa baja o nula (religiosidad baja) en referencia a la religiosidad alta. En este sentido, pertenecer al grupo con baja o nula asistencia religiosa incrementa en 0,86 los años de escolaridad respecto a quienes mantienen una alta asistencia religiosa. El modelo es particularmente simple y con un ajuste poco relevante dado que al observar el R2 se tiene una cifra de 0,009 (0,008 si se toma el R ajustado), lo que implica que el modelo solo explica cerca del 1% de la variabilidad de la variable dependiente Años de educación.

Esto permite sostener solo parcialmente la hipótesis de fondo de que la intensidad de la asistencia religiosa describa parte de los años asociados a la educación. El modelo no ajusta bien, pero sí revela la tendencia de que quienes no asisten a eventos religiosos (religiosidad baja o nula) tienen, en promedio, más años de educación comparado con quienes asisten regularmente a eventos religiosos (religiosidad alta), lo que no se sigue de quienes presentan una religiosidad media. Es decir, existe una tendencia solo entre quienes tienen más años educativos y quienes menos asisten a eventos religiosos.

La ecuación del modelo está dada por:

**Educación en años = (11,8) + (0,43)\*Religiosidad media + (0,86)\*Religiosidad baja o nula**

b.) Ahora bien, respecto al modelo 1 (regresión multivariada) se consideran, además de la variable independiente Religiosidad activa, las variables sociodemográficas Nivel socioeconómico y Sexo. Este modelo revela un incremento de la bondad de ajuste significativo respecto al modelo bivariado, dado que al observar el R2 se explica la variabilidad de los años de educación en un 18.3% por el modelo y la consideración de dichas variables sociodemográficas.

Además, se observa que, en este caso, todos los indicadores de la religiosidad activa son significativos al 99% de confianza (tanto media como baja o nula) y ciertos indicadores de las variables sociodemográficos son igualmente relevantes (confiables al 99%). Por un lado, se tiene la magnitud del coeficiente de religiosidad baja o nula atenuada pero igualmente relevante. Esta se interpreta, de nuevo, en que quienes no asisten nunca o casi nunca a eventos religiosos tienen, en promedio, 0,78 años educativos más en comparación con los que asisten con regularidad, controlando por el resto de variables. Además, se considera que aquellos que asisten ocasionalmente (religiosidad media) también presentan, en promedio, más años educativos, en específico en torno a 0,52 años de escolaridad más que quienes presentan una religiosidad alta, ceteris paribus.

Por otro lado, al considerar el nivel sociodemográfico se observan magnitudes robustas y significativas si se consideran los grupos C3, D y E en comparación al grupo de referencia AB (clase alta). En este caso, controlando por el resto de las variables se tiene que: 1) el factor de pertenecer a un nivel C3 disminuye, en promedio, un 4,5 los años educativos respecto al grupo AB; 2) el factor de pertenecer al nivel D, disminuye, en promedio, un 6,1 los años educativos respecto a AB; y, 3) el factor de pertenecer al nivel E disminuye, en promedio, 7,1 los años educativos respecto al grupo de clase alta. Es interesante destacar que ni el grupo C1a-C1b ni el C2 presentan coeficientes significativos, lo que podría explicarse, en parte, en que existe un grupo socioeconómico medio y medio-alto que logra los mismos años educativos que la clase alta y, por tanto, no existe diferencia estadística alguna entre estos grupos.

Sólo como excepción queda la variable Sexo/Mujer, que no presenta un coeficiente estadístico significativo (el intervalo considera el valor cero) y, por tanto, aun cuando la variable refiere una dirección (el coeficiente es negativo, lo que induce a considerar que el factor de ser mujer, respecto a ser hombre, disminuye la cantidad de años de educación, ceteris paribus), la magnitud de dicho coeficiente es muy pequeño, por lo que no hay significancia alguna.

Finalmente, se observa que tanto la religiosidad activa como el nivel socioeconómico son variables relevantes para describir los años educacionales alcanzados. Las diferencias observadas dan cuenta de tendencias a considerar respecto de ciertos grupos. La hipótesis central de que grupos menos participativos en eventos religiosos presentan una mayor cantidad de años educativos se afirma, según el modelo, tanto para quienes presentan una religiosidad media como entre quienes afirman una religiosidad nula o casi nula. Por otra parte, se confirma fuertemente la hipótesis de la estratificación social y la educación: mientras el nivel socioeconómico de las personas se distancie más de la clase alta, los años educativos serán, en promedio, menores. Observación que no presenta diferencias significativas en los grupos cercanos al nivel socioeconómico alto (C1a-C1b y C2), y por tanto se específica para los grupos más extremos (C3, D y E).

Con todo lo anterior, se establece que la ecuación del modelo está dada por:

**Educación en años = (16,2) + (0,52)\*Religiosidad media + (0,79)\*Religiosidad baja o nula + (-0,05)\*Mujer + (-0,38)\*NSE C1a-C1b + (-2,56)\*NSE C2 + (-4,51)\*NSE C3 + (-6,10)\*NSE D + (-7,11)\*NSE E**

c.) El modelo 2, que expone la misma regresión multivariada, solo que ahora añadiendo la variable de control Posición política, sirve para hacer una comparación más detallada entre modelos, además de que permite evaluar la relevancia de considerar patrones político-ideológicos en la variabilidad de los años educativos alcanzados.

En principio, los coeficientes de regresión asociados al modelo son bastante similares que los del modelo 1. Solo se perciben ciertas atenuaciones que no perturban ni la significación de los coeficientes al 99% de confianza ni siguen una direción diferente a lo desarrollado en b.). Por ejemplo, el coeficiente de religiosidad baja o nula solo cambia de 0,79 a 0,78. Lo mismo ocurre con el coeficiente de religiosidad media (de 0,52 a 0,51) y con el coeficiente asociado a la variable género (que se mantiene igual). Y, en general, los coeficientes asociados a cada NSE atenúan su magnitud en torno a dos decimales (de –6,1 pasa a –5,9 el coeficiente del NSE D, por ejemplo), por lo que las interpretaciones son prácticamente iguales respecto a lo comentado en el modelo 1.

La diferencia aquí reside en la incorporación de la variable Posición política y, por tanto, el modelo presenta más coeficientes a interpretar. En este sentido, el factor de denominarse en el centro político aumenta en 0,29 los años educativos en comparación con quienes se denominan de izquierda, mientras que el factor de denominarse de izquierda disminuye en 0,07 los años educativos en comparación al mismo grupo de referencia, ambas interpretaciones dejando el resto de variables constantes. Sin embargo, dichas interpretaciones de la posición política no resultan significativas en el modelo propuesto y, por tanto, no aportan sustantivamente a la explicación del modelo respecto a los años educacionales. Esto se observa, además, en que el ajuste del modelo 2 no cambia casi nada respecto al ajuste del modelo 1 (como sí había ocurrido entre el modelo 0 y el 1). Así es que puede concluirse que, respecto a los años educativos alcanzados, no existe influencia alguna a partir del clivaje político, en otras palabras, la posición política no muestra relación alguna con el alcance de años educativos.

Esto repercute en reafirmar la línea de la hipótesis central, dado que, al incluir una nueva variable de control, intuitivamente interesante para testear, se reafirma la relación descriptiva específica entre los años educacionales, la religiosidad activa y el nivel socioeconómico (no se revierte la relación multivariada con la variable de control). Y, por otro lado, como se advirtió anteriormente, el testeo del modelo 2 permite rechazar la hipótesis secundaria de que exista alguna relación entre los años educacionales y el clivaje político-ideológico.

Con todo, se establece que la ecuación del modelo está dada por:

**Educación en años = (15,9) + (0,51)\*Religiosidad media + (0,78)\*Religiosidad baja o nula + (-0,05)\*Mujer + (-0,167)\*NSE C1a-C1b + (-2,40)\*NSE C2 + (-4,39)\*NSE C3 + (-5,92)\*NSE D + (-6,87)\*NSE E + (0,29)\*Centro + (-0,07)\*Derecha**

1. **Considerando los resultados de sus análisis, indique cual es el mejor modelo de los modelos presentados. Justifique su decisión. Desarrolle análisis de comparación de modelos.**

Con todo lo dicho anteriormente, queda indicar el modelo más conveniente. En este sentido, los resultados de los análisis en los tres modelos permiten confirmar que el modelo 1 es la mejor decisión. Esto porque, además de presentar una bondad de ajuste considerable respecto al modelo bivariado y casi idéntico al modelo 2, refleja una correcta disposición entre la cantidad de variables significativas con aquello que se intenta describir. Con esto se quiere decir que la incorporación de la variable de control Posición política puede servir justamente para corroborar que los años educativos no tienen clivaje alguno respecto al factor político-ideológico y, por tanto, la inclusión en un modelo final podría estar aportando ruido a los efectos descriptivos de la regresión al plantear mayor número de variables e indicadores. Esto se observa al comparar los R2 del modelo 1 y 2, donde la consideración de la variable Posición política no significa un mejor ajuste (sabiendo que en los modelos OLS cualquier incorporación de una variable de por sí aumenta el R2, pero, viendo los resultados de ambos modelos, el ajuste es casi el mismo) y, por tanto, por parsimonia sería preferible escoger el modelo 1 sin la posición política como variable de control. Claramente el modelo 1 es un mejor modelo que el 0, dado que existe mayor robustez en las variables consideradas y, por tanto, se entiende de mejor manera la explicación de la variable dependiente a partir de la variable independiente y las variables sociodemográficas en conjunto. Sin embargo, la incorporación de una variable de control (posición política) no significa una mayor robustez en el modelo y, por tanto, puede llevar a perjudicar la limpieza y especificación que se quiere del modelo descriptivo sobre los años educacionales. En concreto, no queda más que elegir el modelo 1, a la vez robusto y parsimonioso.