**Conclusiones TP Final Estadística**

Hipótesis de estudio: A mayor nivel de castigo, menor cantidad de crime rates. Esto supone lo siguiente: a mayores valores de *Po1* y *Po2*, menor cantidad de *Crime*; a mayor *Prob*, menor *Crime*; a mayor *Time*, menor *Crime*.

A su vez, las variables *M*, *Ed*, *LF*, *M.F.*, *NW*, *U1* y *U2*, *Wealth* e *Ineq* también pueden influir, aunque no cumplan precisamente con el interés que tienen los criminólogos. De todas formas, se los va a incluir en el análisis.

Por su parte, las variables *So* y *Pop* las consideraremos a priori no importantes. No obstante, las vamos a incluir y veremos si son significantes.

Medimos So y lo sacamos:

Ya que es una variable binaria y a priori la consideramos como potencialmente no significativa, realizamos un boxplot comparando al nivel de crimen entre los Estados del sur y del norte. Pudimos observar que las medianas entre ambos son prácticamente iguales. Sin embargo, pareciera que para los estados del norte hay mucha más variabilidad, a diferencia del sur que tiene cuartiles bastante cercanos.

De momento removeremos a esta variable. Luego la utilizaremos.

Primer modelo:

A raíz de la quita de la variable *So*, realizamos el primer modelo. Tuvimos como resultado a un estadístico F significativo con p-valor de 1.118e-07, un RSE de 205.8 y un R2 de 0.8031.

El porcentaje de error es aproximadamente de un 23%. Esto significa que el modelo no es muy preciso. En cuanto a las regresoras, gran parte de ellas no son significativas al 5%.

Observando a los residuos contra los fitted values, podemos concluir que los datos tienen cierta forma, aunque se juntan bastante en el centro y se abren en el centro. No pareciera haber homocedasticidad. Por último, los residuos no parecieran seguir normalidad.

Transformaciones:

Debido a que el primer modelo presentó varias imprecisiones, intentaremos realizar algunas transformaciones para intentar mejorarlo.

1. Transformación sqrt(y):

Los residuos contra los fitted values parecieran ser mejores respecto del modelo inicial. El R2 se mantuvo igual, el RSE mejoró significativamente bajando a 3.264 y el estadístico F mantuvo un p-valor similar.

Sin embargo, la normalidad de los residuos empeoró considerablemente. Éstos están lejos de seguir una línea normal. Por consiguiente, decidimos no seguir con este modelo.

1. Transformación log(y)

Al igual que en la anterior transformación, los residuos contra los fitted values parecieran tener una buena forma. El R2 empeoró, el RSE bajó a 0.2271 y el estadístico F sigue siendo significativo.

La normalidad de los errores sigue siendo peor que la del modelo inicial. Tampoco consideraremos este modelo.

1. Transformación y2

El RSE y el R2 empeoraron. Los residuos tienen una forma parabólica y no siguen normalidad. Este modelo no es bueno.

Outliers:

Utilizando el primer modelo, podemos observar un outlier. Decidiremos sacarlo de la muestra.

Ya con la nueva muestra, podemos observar otro outlier. Para experimentar, lo quitaremos.

Ya habiendo sacado estos outliers, no se encuentran nuevos. No obstante, no hay normalidad en los residuos. Tampoco consideraremos este modelo.

Estudio de interacciones y luego selección de variables:

Como era de esperar, luego de hacer interacciones entre todas las regresoras del modelo inicial, R nos dejó de valor N/A para casi todas las variables a estudiar. Esto se debe a que muchas de ellas están relacionadas linealmente entre sí.

Selección de variables:

Utilizando los comandos de cor, observamos que hay ciertas variables que se correlacionan entre sí más que otras. Debido a esto, mediante la estrategia *Stepwise* seleccionaremos las variables más confiables con la ayuda del algoritmo.

Como resultado, en el modelo me quedaron las siguientes variables: *M.F*, *U1*, *Prob*, *U2*, *M*, *Ed*, *Ineq* y *Po1*.

El nuevo RSE bajó a 195.5, el R2 bajó a 0.7888 y el pvalor mejoró a 1.159e-10. Por su parte, *M.F* figura como la variable menos significativa con un pvalor ligeramente superior al 10%. De momento, este modelo pareciera haber mejorado al incial.

Observando a los residuos contra los fitted values, se puede observar una estructura medianamente aceptable, aunque cerca del medio se abre. Los residuos no parecieran ser normales, aunque sí parecieran serlo más que los residuos del primer modelo.

Tampoco se observan outliers en este nuevo modelo. Seguiremos con este modelo.

Interacciones luego de la selección:

Luego de realizar interacciones entre las variables luego de la selección, podemos observar como del modelo es notablemente menos significativo que antes de sus interacciones. Casi ninguna regresora es significativa a nivel 5%. No consideraremos este modelo y continuaremos con el anterior.

Nuevas transformaciones:

Como se explicó con anterioridad, el nuevo modelo pareciera ser el mejor hasta ahora. Sin embargo, sigue teniendo cierto problema de heterocedasticidad de acuerdo con el gráfico de los residuos contra los valores predichos. Por eso, haremos nuevas transformaciones sobre la variable respuesta:

1. Transformación log(y):

El RSE bajó considerablemente a 0.2241, el R2 también bajó a 0.7544, y el estadístico de F es significativo con un p-valor de 1.806e-09. En cuanto a las regresoras, la Intercept, *M.F* y *U1* son las únicas no significativas a nivel 5%.

El gráfico de los residuos contra los valores predichos pareciera ser relativamente bueno, pero la normalidad de los residuos no es muy buena. No obstante, puede considerársele mejor que la del anterior modelo. No parecen haber outliers.

1. Transformación sqrt(y)

El RSE está en 3.159, el R2 en 0.7813 y el estadístico F es significativo con un p-valor de 2.199e-10.

El gráfico de los residuos contra los valores predichos pareciera no tener muchos problemas de heterocedasticidad. Sin embargo, la normalidad de los residuos no es buena.

1. Transformación y2

Esta es la peor transformación de las tres. El gráfico de los residuos contra los valores predichos muestra una forma parabólica y los residuos claramente no siguen normalidad.

**HAY QUE VER SI NOS QUEDAMOS CON EL PRIMER MODELO, CON EL MODELO AJUSTADO POST-STEP O EL MODELO AJUSTADO POST-STEP CON TRANSFORMACIÓN LOG(Y)**