Universidad de San Andrés

Departamento de Matemática y Ciencias

Maestría en Ciencia de Datos

Regresión Avanzada

Informe sobre el trabajo final

-

Alumnos:

Dominutti, Nicolás

Suárez Gurruchaga, Carlos Roque

Telechea, Hernán

Introducción

En este informe, estudiaremos una base de datos sobre la seguridad vial en ciudades europeas durante el 2018. Analizaremos las variables, modelos de regresión y transformaciones a partir del estudio de residuos, interacciones y colinealidad. El fin es proponer políticas públicas para disminuir los accidentes viales. Las tablas, cálculos y gráficos utilizados aquí están disponibles en el R Markdown.

Análisis de variables

Son 20 las variables analizadas para 24 observaciones. 11 de esas variables aparecen como categóricas, el resto como cuantitativas. Sin embargo, al observar la tabla de datos, solo 3 deberían ser categóricas. Las demás aparecen como tal por tener decimales. Por eso, decidimos transformarlas. Una vez hecho esto, estudiamos las variables y sus relaciones, de lo que destacamos lo siguiente:

- Las ciudades con mayor población observada son Londres, Madrid, Roma y París, cuatro capitales europeas. Salvo Roma, las restantes son también las que más accidentes presentan. Londres incluso lidera la incidencia de accidentes entre autos y peatones y vehículos de 2 ruedas.

- La densidad poblacional tiende a estar sujeta a la extensión territorial de las ciudades. Por fuera de capitales como Londres o París, los territorios menos extensos como Barcelona, León y Lille poseen una densidad elevada. Esta variable igualmente no parece influir en los accidentes viales.

- El transporte en vehículos motorizados es el medio más utilizado. En Barcelona, aunque predomina el transporte público, más del 75% de sus accidentes involucran a los medios motorizados de 2 ruedas. Le siguen Madrid, Niza, Marsella, París y Toulouse con más del 50%

- Entre las ciudades de mayor incidencia de ciclistas, encontramos Bristol y Estrasburgo. Esta última es, incluso, la de mayor superficie para el uso de bicicletas y, no sorpresivamente, la de mayor porcentaje de accidentes con este medio (junto con Londres, Manchester y Oslo).

- Sheefield lidera la lista con mayor incidencia de choques entre autos. Glasgow (48%), Liverpool (46%) y Bradford (42%) son las ciudades con mayor incidencia de accidentes con peatones.

- En las ciudades con temperaturas más altas, la gente camina más, mientras que en temperaturas más bajas (y/o precipitaciones más fuertes) el transporte a motor es más elegido. Para esta última condición, en realidad, el efecto más fuerte es que la gente camina menos.

- En promedio, 7 de 10 km2 de las ciudades tienen baja restricción de velocidad para conducir. Ninguna ciudad se aleja tanto de este promedio, por lo que no detectamos valores atípicos.

Análisis de linealidad

Luego del estudio exploratorio, analizamos la linealidad en las relaciones entre variables regresoras y de respuesta. Para ello, además de las 6 variables dependientes, creamos una séptima correspondiente al total de todos los accidentes por ciudad, "accidentes_viales".

En primer lugar, descartamos las variables categóricas "Pais", "Codigo Pais" y "Ciudad" porque no esperamos obtener información relevante para este análisis. Luego, comparamos las variables con "accidentes_viales" y detectamos que la relación más lineal ocurre con "Poblacion" ($\rho = 0.90$), seguida más débilmente por "PMTPublico" ($\rho = 0.42$), "ArCiclista" ($\rho = 0.32$), "Precipitacion" ($\rho = 0.29$) y "PMCiclistas" ($\rho = 0.23$).

En segundo lugar, estandarizamos todas las variables para hacerlas más comparables, para luego encarar el estudio de las relaciones lineales y los modelos correspondientes, con sus residuos. En cada uno de estos modelos, nuestro análisis consistió en: correr el ajuste con todas las variables (salvo las 3 que son categóricas), aplicar un estudio de la inflación de varianza (VIF), analizar los valores *outliers* a partir de un modelo robusto y el gráfico de *r-weights* vs residuos, y finalmente aplicar métodos de selección de variables a partir del criterio *stepwise* y la regularización Lasso. El detalle de este análisis para cada modelo está disponible en el Anexo 1.

Para cada instancia del estudio, analizamos tanto la significancia de los coeficientes, del test general y el R2 ajustado, para decidir qué modelo de cada target elegir. En algunos casos, también decidimos eliminar observaciones atípicas que impactaban fuertemente en los resultados del estudio. Finalmente, consideramos si las variables elegidas por los modelos de selección tenían sentido dentro del contexto de cada tipo de accidente.

Consideramos que los modelos que mejor explican cada variable de respuesta son los siguientes:

- "PeatAuto": modelo OLS con las regresoras "ArCiclista" y "PMPeatones".
- "CicAuto": modelo OLS con las regresoras "Temp" y "PMPeatones". Eliminamos Londres.
- "V2RMSM": modelo OLS con las regresoras "ArBajaVel", "Población" y "Temp" significativas, y "PMCiclistas" y "Precipitación" como variables no significativas. Eliminamos Marseille
- "V2RMAuto": modelo OLS con las regresoras "Población", "PMVMotor" y "Temp" significativas, y "PBI" como variable no significativa. Eliminamos Marsella y Niza de los datos.
- "AutoSM": modelo con sqrt(AutoSM) como respuesta. Regresoras son "ArCiclista", "ArBajaVel" y "PMPeatones", siendo esta última, no significativa estadísticamente.
- "AutoAuto": modelo OLS con las regresoras "ArCiclista", "PMPeatones" y "ArBajaVel", siendo esta ultima, no significativa estadisticamente. Eliminamos Birmingham y Sheffield de los datos.

Conclusiones y recomendaciones

A partir de los modelos, planteamos políticas públicas a implementar para disminuir estos accidentes:

- Modelo “PeatAuto”: observamos que la covariable “ArCicilista” es inversamente proporcional a los accidentes producidos entre peatones y autos. Por lo tanto, creemos que fomentar el uso de las bicicletas es primordial, a través del alquiler público, el lanzamiento de publicidad para concientizar, y la creación de espacios para circular y estacionar este vehículo. Asimismo, podríamos construir autopistas para agilizar la circulación de autos y agilizar el tránsito general.

- Modelo “CicAuto”: las altas temperaturas durante el verano europeo pueden estar impactando en que haya menos circulación de autos y más de bicicletas. La gente quizás prefiere pasear o ir de turismo a otras ciudades en sus vacaciones, incluso debido a esto, no hay clases y por tanto el tránsito vehicular de padres se ve reducido. Por lo tanto, la circulación de peatones y/o ciclistas podría aumentar, generando un aumento en los accidentes que los involucran.

- Modelo “V2RMSM”: a mayor tamaño poblacional, mayor cantidad de vehículos que interactúa entre sí. A menores restricciones a la velocidad, mayor probabilidad de accidentes. Por lo tanto, proponemos identificar las zonas calientes de accidentes para luego impulsar nuevos controles de velocidad (ayudados por cámaras de vigilancia para hacer cumplir la normativa). Finalmente, seguimos apostando por el fomento del uso de bicicletas. Sobre esto, podemos ofrecerlas gratis a lo largo de las ciudades durante la época veraniega, a manera de incentivo.

- Modelo “V2RMAuto”: a mayor cantidad de vehículos con motor, mayores chances de accidentes. Esta conclusión no es muy informativa y puede deberse a que contamos con tan pocas observaciones que algunas variables no puedan arrojar información coherente.

- Modelo “AutoSM”: la conclusión más importante es que la raiz cuadrada de los accidentes parecen aumentar a medida que hay menores controles de velocidad. Ya propusimos una respuesta a este problema.

- Modelo “AutoAuto”: en este caso, lo más coherente parece ser aumentar la participación modal de los peatones para reducir la presencia de autos. Fomentar el uso de bicicletas podría ayudar.

Ninguna de las propuestas planteadas funcionaría de manera aislada. Deberíamos encontrar aquellas que logren solucionar la mayor cantidad de problemáticas. Por ejemplo: si queremos fomentar el uso de bicicletas pero también reducir los accidentes con éste vehículo, entonces la mejor propuesta antes de fomentar el uso es buscar mayores espacios para que estos puedan circular por las ciudades. De todas formas, consideramos que el total de observaciones (y su diversidad) no es suficiente para proponer soluciones, dado que las normativas por país/ciudad pueden variar significativamente. Para contrarrestar esa variabilidad, deberíamos acotar la muestra a una sola región o de ser posible, sumar muchas más observaciones para poder trabajar.

Anexo 1: ajustes lineales para cada variable de respuesta

Modelo para "PeatAuto"

Partimos de un modelo que utiliza todas las variables. Se observa un R2 ajustado inicial de ~52% y un p-valor global significativo (0.02693), aunque no tan bueno. Varios estimadores no resultan significativamente distintos de 0, que finalmente logramos explicar con problemas de alta colinealidad entre las variables de participación modal: "PMPeatones", "PMCiclistas", "PMTPublic" y "PMVmotor" (colineales por construcción), a partir de un estudio VIF. Intentamos aplicar el criterio de selección de variables *stepwise* pero no solucionamos la colinealidad, aunque mejoramos el R2 ajustado. Luego, aplicamos regularización por Lasso (para solucionar el problema de colinealidad entre variables) y posteriormente un OLS con estas variables que nos seleccionó el modelo de regularización ("ArCiclista", "PMPeatones"). Logramos subir el R2 ajustado a 55%, sin potenciales outliers ni problemas de colinealidad. Finalmente, modelamos una regresión robusta que arrojó un mejor R2 ajustado aunque p-valores para los coeficientes más grandes que el modelo OLS anterior. Como en este trabajo buscamos realizar tareas de inferencia y no predicción, optamos por quedarnos con el modelo OLS.

Modelo para "CicAuto"

Realizamos un modelo que utiliza todas las variables y observamos un R2 ajustado muy bajo (~0.29) y un p-valor no significativo (0.1535), con un alfa 5%. Aplicamos el criterio de selección de variables *stepwise*, que mejora estas métricas pero que, al realizar un análisis VIF, observamos que no considera colinealidades (por lo tanto, no es el más adecuado para hacer inferencia). Finalmente, solucionamos el problema de colinealidad al aplicar una regularización Lasso para la selección de variables y, al aplicar un estudio de outliers a partir de una regresión robusta, decidimos eliminar la observación "Londres" por tener un residuo muy elevado y r-weight 0. Nos quedamos con las variables "Temp" y "PMPeatones". El modelo OLS final logra un R2 ajustado de ~52%, un p-valor chico (0.0002587) , así como también variables significativas.

Modelo para "V2RMSM"

Realizamos un modelo que utiliza todas las variables y observamos un p-valor = 0.08913 (no significativo) y R2 ajustado = 40%. Al aplicar un análisis VIF, *stepwise*, regularización Lasso y un estudio de residuos a partir de un modelo robusto (*r-weigths* vs residuals), concluimos eliminar la observación "Marsella", por tener un residuo muy alto y *r-weight* 0. Finalmente, a partir de un análisis Box-Tidwell, aplicamos una transformación a las variables "Precipitación", "Temp" y "PMCiclistas". El modelo termina con un R2 ajustado de 63.58%, aunque no todas sus variables lo son a nivel

individual. Dado que el modelo pierde interpretación al eliminar una observación y convertir variables, podríamos optar por quedarnos con todas las ciudades y un R2 ajustado menor (57.74%).

Modelo para "V2RMAuto"

Realizamos un modelo que utiliza todas las variables y observamos un p-valor = 0.008055 y un R2 ajustado de ~ 62%. Tenemos varios coeficientes no significativos y, al realizar un análisis VIF y un estudio robusto de outliers, encontramos problemas de colinealidad y outliers que rompen nuestros supuestos para realizar inferencia. Quitando a Marsella y Niza, logramos un modelo con R2 ajustado a 76% y el p-valor disminuye a 6.9e-06. Los residuos dejan de tener estructura alguna y no se observan valores atípicos.

Modelo para "AUTOSM"

Realizamos un modelo que utiliza todas las variables y resulta no significativo (p-valor = 0.1026 y un R2 ajustado de 35.41%). Al aplicar un estudio VIF, una regularización Lasso y la transformación del target aplicando raíz cuadrada, a partir de un análisis Box y Cox, logramos mejorar el R2 ajustado (51.7%) y el p-valor (0.0004977). No consideramos accionar sobre residuos porque no encontramos valores atípicos ni estructura alguna en los gráficos de los residuos del modelo, ni al aplicar un análisis robusto de outliers.

Modelo para "AutoAuto"

Los coeficientes del modelo base no son significativos, por lo que luego de aplicar los análisis de colinealidad y utilizando Lasso para seleccionar variables, nos sugiere conservar 2 variables que sabemos tienen altos problemas de colinealidad ("PMCiclistas" y "PMPeatones"). Decidimos descartar "PMCiclistas" para quedarnos con "PMPeatones". También eliminamos 2 observaciones consideradas outliers a partir del gráfico de *r-weights* vs residuos: Birmingham y Sheffield. Con estos cambios, el modelo pierde colinealidad y gana significancia (p-valor = 0.001 y R2 ajustado = 51%).

Modelo para "Accidentes viales"

Por último, analizamos la variable que combina todos los tipos de accidentes y, a partir de la regularización Lasso, nos quedamos con "Poblacion", "ArCiclista", "ArBajaVel" y "PMPeatones" como regresoras. Observamos un valor con alto leverage y un residuo 2 pero no lo suficientemente grande como para considerarlo outlier. Los residuos parecen homocedásticos y aproximadamente normales, por lo que no aplicamos más transformaciones. Logramos mejorar el R2 ajustado con una regresión robusta pero preferimos continuar con OLS para tareas de inferencia por la significancia de sus coeficientes.