EYP2425 ANOVA y Diseño de Experimentos
Proyecto: Choques

Diego Aravena Alonso Campos Josefa Silva

July 10, 2023

1 Resumen Ejecutivo

Gracias a la información proporcionada por el Ministerio de Transportes, el cual mide las características de algunos modelos de autos y nivel de daño provocado en una persona al momento de un choque, es que se ha permitido evaluar los factores más decisivos a la hora de salir perjudicado en un accidente de tránsito.

Para esto se adquieren diferentes medidas del daño infligido, que luego son calificados según su importancia en términos de su riesgo mortal para posteriormente crear una sola variable respuesta que esté asociado al riesgo mortal en un accidente.

Luego, mediante un estudio de Análisis de Varianza mediante modelos de regresión, ya que es un modelo desbalanceado, se declara que los factores más determinantes corresponden al tamaño del auto, su tipo de protección, número de puertas y el asiento que la persona utiliza. De los cuales todos resultan presentar un impacto significativo sobre la respuesta con al menos un 95% de confianza. Y es dentro de esta misma línea que se logra encontrar el perfil de auto con mayor inseguridad al momento de afectar la variable respuesta durante un choque automovilístico. Finalmente, se realiza un validación del modelo, para estudiar si cumple con los supuestos que se proponen.

2 Introducción

En Chile entre el año 1987 y 1991 ocurrieron entre 32.790 y 40.368 accidentes automovilísticos, alcanzando una tasa de al menos 3.65 muertes involucradas por cada 100 accidentes [1]. De aquí surge la motivación sobre querer identificar las características del vehículo que poseen un impacto significativo a la hora de resguardar la vida de una persona durante un choque. Para ello se recolectó información sobre el efecto de los choques en muñecos colocados en los dos asientos delanteros de automóviles que colisionaron a 35 millas por hora. Se consideraron diversos modelos y marcas reconocidas a nivel mundial que fueron fabricados entre los años 1987 y 1991. Por lo tanto, dada esta información, el objetivo sería realizar un estudio para distinguir los factores con mayor influencia tal que la vida de una persona, situada en los asientos delanteros de un automóvil, logre salir perjudicada en lo más mínimo al momento de ocurrir un accidente de tránsito.

3 Análisis Exploratorio

3.1 Un primer acercamiento

El archivo de datos contiene información de 352 registros sobre el efecto de los choques en los muñecos colocados en los dos asientos delanteros. Por ende, para cada tipo de automóvil se obtienen dos observaciones, uno proveniente del piloto y otro del copiloto, lo cual significa que se posee información sobre

176 tipos de autos distintos. A cada uno de estos se le midieron las variables presentadas en la Tabla 1.

Una primera observación trata acerca de que el set de datos no es ajeno a los datos faltantes. Tal como se presenta en la Tabla 9, todas las variables asociadas al daño no están completas, incluyendo la variable asociada al número de puertas, la cual posee el mayor porcentaje de datos faltantes. Dada la ocurrencia de este fenómeno, en la sección 3.2 se inspecciona el comportamiento de las observaciones con datos faltantes con el fin de identificar patrones e imputar valores dentro de lo posible.

Variable	**Descripción**
Make	Marca del auto
Model	Modelo del auto
carID	Identifcación del auto, usualmente combinación de Marca y Modelo
carID&Year	Identifcación completa del auto
HeadIC	Medida de daño en la cabeza
Chestdecel	Desaceleración cardíaca
LLeg	Daño en la pierna izquierda
RLeg	Daño en la pierna derecha
D/P	Indica si el muñeco va en el asiento del conductor (D) o en del acompañante (P)
Protection	Tipo de protección (cinturones, airbag, etc)
Doors	Número de puertas del auto
Year	Año del auto
Wt	Peso del auto en libras
Size	Auto liviano o minivan

Tabla 1: Descripción de las variables contenidas en la base de datos

Variable	**Datos faltantes**	**Porcentaje**
HeadIC	12	3.4%
Chestdecel	11	3.1%
LLeg	9	2.5%
RLeg	11	3.1%
Doors	66	18.75%

Tabla 2: Porcentaje de datos faltantes en variables incompletas

3.2 Análisis de datos faltantes

3.2.1 Imputación Número de puertas

El conjunto de datos contiene 66 registros con valores faltantes para el número puertas, que equivalen al 18.75 % del total. Estos datos corresponden a vehículos de tamaño pick up o van. Ahora, dado que corresponden a un porcentaje considerable del total de registros se decide no eliminar aquellos datos que no contengan la información, pues es posible recuperar la información faltante complementándola con fuentes externas dado que se conoce la marca, el modelo y año del vehículo.
Dicho lo anterior, cada dato faltante se reemplaza con información recopilada de los sitios web oficiales donde venden este tipo de automóviles y de sitios especializados en esta materia.

3.2.2 Imputación variables relacionadas con daño

Como el conjunto de datos contiene una cantidad reducida de datos, se propone utilizar técnicas de imputación con el objetivo de reducir la incertidumbre. En particular, para este informe se propone la

técnica Miss Forest que, como su nombre lo indica, utiliza el método de Random Forest para imputar valores faltantes. Más precisamente, considera una variable e imputa los datos faltantes en ella utilizando la información de las demás variables, repitiendo el proceso sobre el resto de variables iterando en cada paso hasta converger.

La Tabla 3 muestra las medidas de resumen del conjunto de datos original, omitiendo registros nulos por variable, en contraste con el de los datos imputados mediante Miss Forest. Se aprecia que los cambios no son significativos, por lo que la imputación no llega a modificar la distribución de cada variable. Es importante destacar que los estadísticos máximo y mínimo son iguales para ambos conjuntos

	Sin Imputar				Imputación Miss Forest			
	Head_IC	Chest_decel	L_leg	R_leg	Head_IC	Chest_decel	L_leg	R_leg
Media	903.07	48.37	1054.01	740.92	901.99	48.33	1052.83	741.91
Primer cuartil	583.00	42.00	688.50	450.00	585.00	42.00	693.25	450.75
Mediana	792.50	47.00	996.00	644.00	796.50	47.00	996.00	652.50
Tercer cuartil	1074.00	54.00	1371.00	959.00	1074.00	54.00	1355.00	957.50
Desviación Estándar	464.97	9.59	544.92	427.01	458.58	9.45	539.18	421.40

Tabla 3: Resumen comparación de los datos imputados

3.3 Elección de la variable respuesta

El objetivo principal está orientado a determinar los factores que tengan un impacto significativo sobre la vida de una persona en un accidente de tránsito, por ende la variable respuesta quedará asociada al daño recibido, lo cual sirve como una medida sobre cuanto se pudo ver afectada la vida de la persona. Por lo tanto, se estarían involucrando todas las variables asociadas al daño, considerando la desaceleración cardíaca. Por este motivo se llegó a un consenso sobre considerar un promedio ponderado de las 4 variables, quedando así una sola variable respuesta que estaría representando un riesgo hacia la vida de la persona más que a la integridad física.

Como es posible apreciar en la Figura 1, el daño en la cabeza y la desaceleración cardíaca poseen una asociación lineal positiva con un aumento en la variabilidad, esto es posible corroborarlo con la correlación que es 0.569. Sin embargo, el resto de variables no presenta mayor correlación, asimismo, estudiando lo gráficos de dispersión, tampoco se aprecia asociaciones o patrones.

Dada la información anterior, para las ponderaciones se consideró que el daño producido en la cabeza junto con la desaceleración cardíaca pueden llegar a ser lo más determinante cuando se trata de la vida de una persona, en cambio un daño en las piernas tiende a tener mayor repercusión sobre la integridad física. Por lo tanto, la respuesta que se propone queda representada por la ecuación (1). No obstante, tal como se presenta en la tabla 5, la variable asociada a la desaceleración cardíaca no está en la misma escala, por ende en (1) se estarían considerando las variables estandarizadas.

$$Daño = HeadIC \cdot 0.4 + Chestdecel \cdot 0.4 + LLeg \cdot 0.1 + RLeg \cdot 0.1 \tag{1}$$

3.4 Análisis gráfico

3.4.1 Gráficos de dispersión

En la Figura 2 no es posible observar ningún tipo de asociación entre la variable respuesta con el peso del vehículo, y cuyo coeficiente de asociación lineal es de 0.23, lo cual es útil para determinar a primera

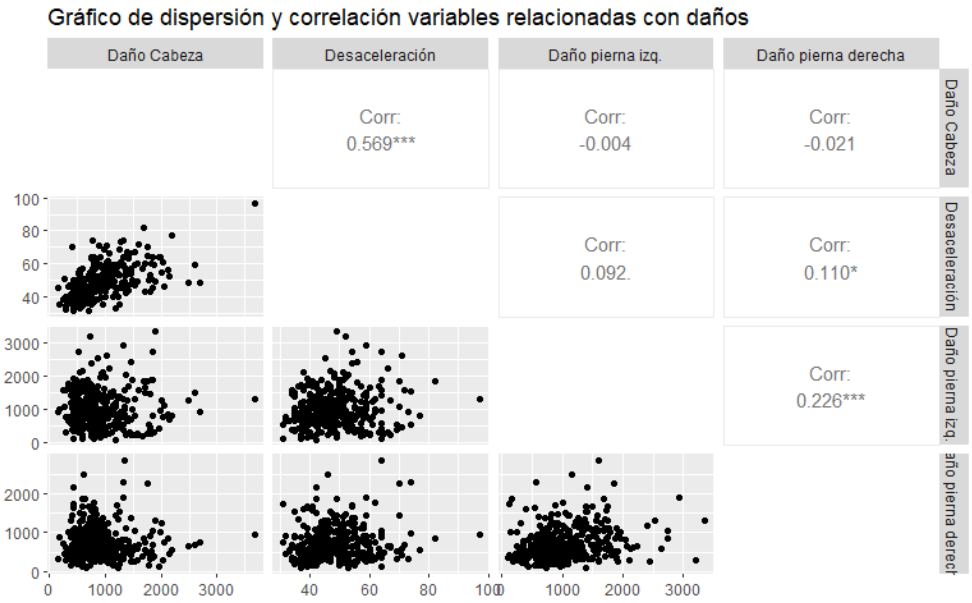

Figura 1: Gráfico de dispersión y correlación entre las variables Daño Cabeza, Desaceleración del corazón, Daño pierna izquierda y derecha

vista que el peso no viene siendo un factor determinante.

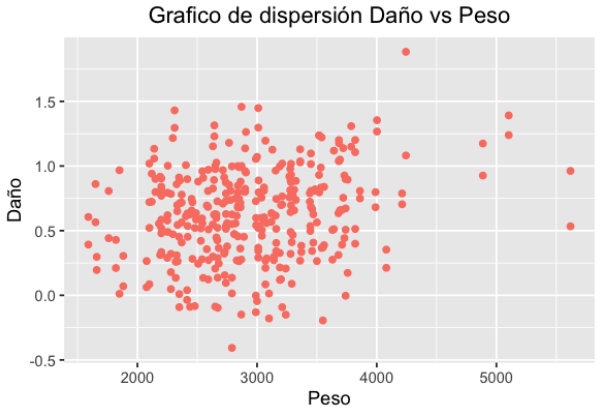

Figura 2: Gráfico de dispersión

3.4.2 Gráficos de Boxplot

De la Figura 3, se observa que algunas categorías exhiben grupos que muestran un comportamiento similar. Por ejemplo, para la variable `Size`, los niveles *comp*, *hev*, *it*, *med* y *mini* parecieran tener comportamientos bastante similares, mientras que el resto tiende a tener valores más altos de daño. Por otro lado, se observa que el comportamiento de 2 y 4 puertas es similar, pudiendo notar además un leve aumento en el daño para autos de 3 puertas. Luego, es posible notar medianas diferencias en la ubicación del muñeco, siendo el que se ubica en el asiento del piloto quien recibe mayor daño. Finalmente, el uso de cinturón manual, motorizado y pasivo llegan a estar igualmente centrados pero con una variabilidad diferente, mientras que el uso de airbags disminuye levemente el daño.

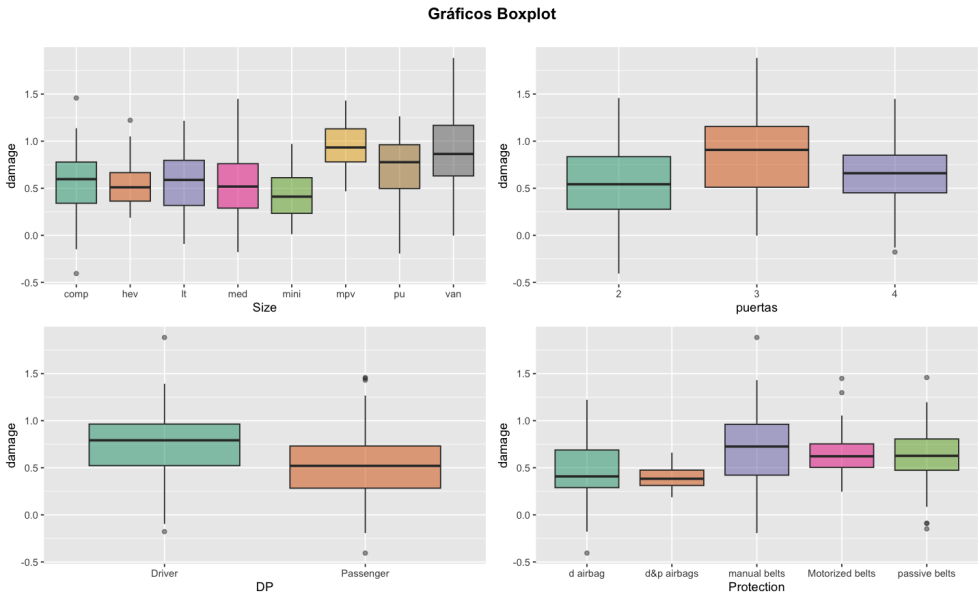

Figura 3: Gráficos de Boxplot

3.4.3 Interacciones

Notar que si bien, a simple vista en la Figura 4 se puede observar interacción entre casi todas las variables, se debe tener en cuenta que esto se puede deber a un desbalance existente por cada categoría, es decir, en categorías con un mayor número de observaciones, es más probable que las estimaciones sean más precisas y, por lo tanto, las líneas de interacción puedan mostrarse más cercanas o incluso superpuestas. Por otro lado, en categorías con un menor número de observaciones, las estimaciones pueden ser menos precisas y las líneas de interacción pueden mostrar una mayor variabilidad o no juntarse.

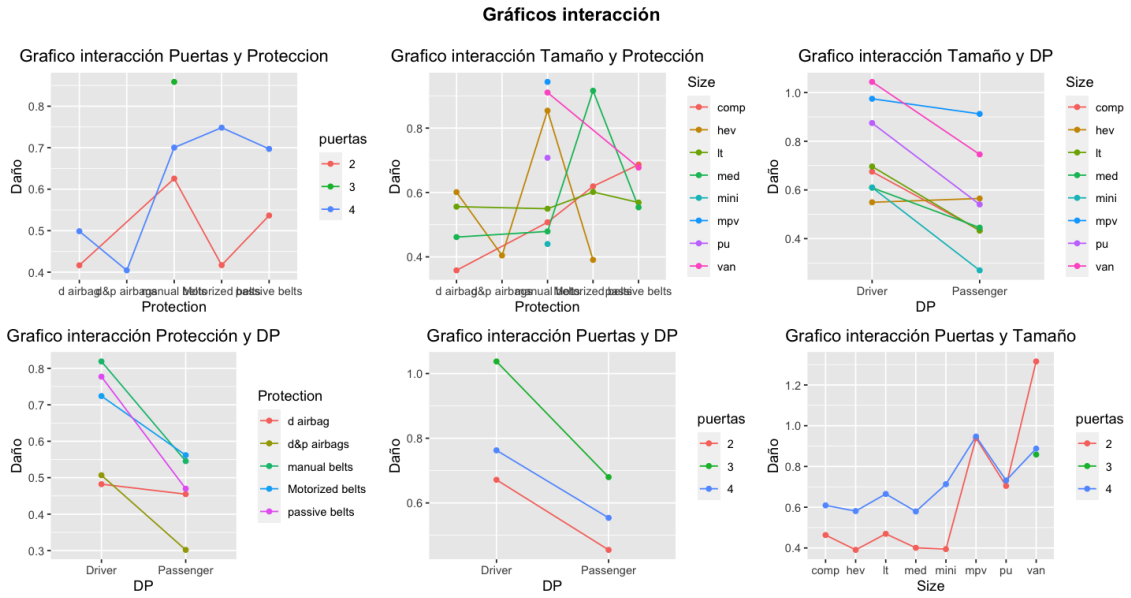

Figura 4: Gráficos de interacción

3.5 Transformación variable respuesta

Como se puede apreciar en la Figura 5, en el gráfico izquierdo la variable respuesta *Daño* presenta una leve asimetría a la derecha lo que la aleja de una distribución normal, por lo que se opta por llevar a cabo alguna transformación. Mediante el método de Box-Cox, se exploraron diferentes transformaciones y se determinó que la transformación óptima consiste en la función logaritmo. No obstante, la variable posee valores negativos producto de la estandarización por lo que se considera sumar 2 unidades. De este modo, se solucionaría la falta de normalidad en la respuesta, aunque se vuelve más complejo interpretar este valor.

$$\text{Daño} = log(\text{Daño} + 2) \tag{2}$$

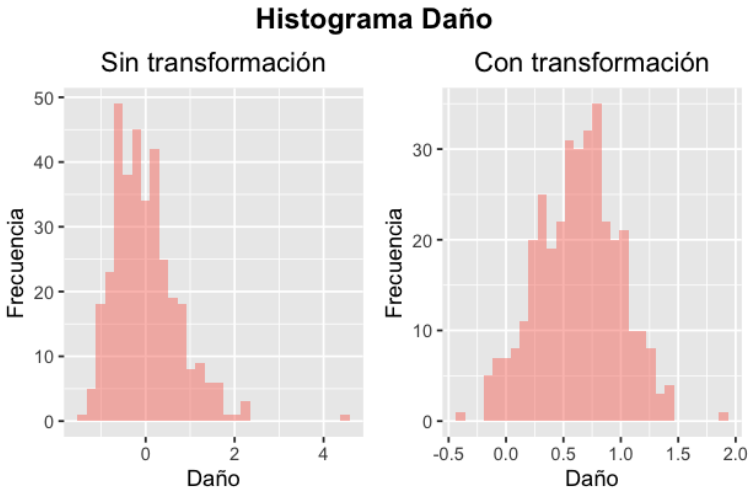

Figura 5: Histograma de la variable Daño

4 Metodología

Para la elección de un modelo adecuado a los objetivos del estudio, en un principio se pensó en considerar un modelo con efectos mixtos, debido a que podemos suponer que los autos fueron escogidos al azar. Sin embargo, considerar dicha suposición equivaldría a utilizar metodologías más complicadas, ya que corresponde a un modelo de efectos aleatorios con desbalance en las celdas lo conllevaría (...)

Por este motivo, se plantea un modelo de efectos fijos. Cabe destacar que las clases son desbalanceadas, por ende, para estudiar la incorporación de variables, así como las interacciones entre ellas se debe utilizar modelos de regresión para su ajuste.

Primero se estudian variables que podrían incorporarse al modelo. En consecuencia, se utiliza la metodología backward, con el fin de discriminar, mediante test F parciales, variables que no son significativas una vez incluidas las demás. Así, es posible notar que las variables año y peso del vehículo no son significativas, utilizando una confianza de 95%.

Dado lo anterior, se procede a utilizar un modelo que contemple sólo interacción entre las variables y para evaluar su presencia se llevaron a cabo pruebas de significancia mediante el método de modelos anidados, probando interacciones cuádruples, triples y dobles, respectivamente. Sin embargo, al intentar incluir al menos una interacción triple empiezan a surgir problemas en las estimaciones de los coeficientes debido

a la cantidad de parámetros, lo que además generaba que la matriz de diseño fuese no invertible.

Finalmente el modelo propuesto considera las variables `D/P`, `Protection`, `Doors` y `Size`, de este modo quedaría representado como

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijklm}, \qquad \epsilon_{ijklm} \overset{iid}{\sim} N(0, \sigma^2) \tag{3}$$

Para $i = 1, ..., 7$, $j = 1, 2$, $k = 1, 2, 3, 4, 5$, $l = 1, 2, 3$, donde

$$\sum_{i=1}^{7} \alpha_i = 0 \qquad \sum_{j=1}^{2} \beta_j = 0 \qquad \sum_{k=1}^{5} \gamma_k = 0 \qquad \sum_{l=1}^{3} \delta_l = 0$$

Posteriormente se realiza la implementación en `R` del modelo planteado en la ecuación 3, cuyos resultados se presentan en la tabla 4. Esta información sirve como una referencia acerca de los factores que tienen un impacto tanto positivo como negativo en el daño provocado por el choque. Se observa que los autos de tamaño *MPV*, *Pick up* o *Van*, cuyo pasajero está ubicado en el asiento del conductor, que sólo posee cinturón y que además cuenta con más de 2 puertas corresponderían al tipo de auto con mayor peligro para las personas. Posteriormente se estudiará la significancia del factor y si existe evidencia para concluir que las medias son distintas en cada factor.

Factor	Nivel	Estimación
Intercepto	Intercepto	0.631
Size	Compacto	-0.145
Size	Pesado	-0.053
Size	Light	-0.111
Size	Mediano	-0.14
Size	Mini	-0.164
Size	MPV	0.29
Size	Pick up	0.108
Size	Van	0.215
DP	Conductor	0.111
DP	Pasajero	-0.111
Protection	Airbag Conductor	-0.045
Protection	Airbag Pasajero y conductor	-0.221
Protection	Cinturón manual	0.027
Protection	Cinturón motorizado	0.123
Protection	Cinturón pasivo	0.117
Puertas	2	-0.073
Puertas	3	0.016
Puertas	4	0.057

Tabla 4: Estimaciones coeficientes modelo Anova

Como se mencionó anteriormente los coeficientes se estiman mediante la restricción de celda de referencia. Por este motivo se puede interpretar como la diferencia con respecto a la media global al pertenecer a dicha variable. Si un individuo tiene un auto de tamaño MPV, entonces aumenta la media del daño logaritmo del daño más dos en un 0.29 unidades, por el contrario, si un individuo tiene un auto compacto, éste disminuye -0.145 unidades del logaritmo del daño más dos con respecto a la media global..

Test para variable Size

Se desea evaluar si el tamaño del auto, registrado mediante categorías, incide en el daño provocado por los accidentes de auto. Las hipótesis a plantear son:

$$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_7$$
$$H_1 : \text{Al menos un } \alpha_i \text{ es distinto de cero}$$

Los resultados de la Tabla 5 arrojan los valores de interés. Se aprecia que valor-p asociado al tamaño es menor que un 5%, lo que implica que existe evidencia para concluir que el tamaño del vehículo influye en el daño de un choque en al menos un 95% de confianza.

Test para variable Pasajero o Conductor

Otra pregunta que se podría plantear es si influye el asiento en el auto en un accidente de tráfico. Para ello se plantean las siguiente hipótesis,

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{Al menos un } \beta_j \text{ es distinto de cero}$$

La hipótesis nula plantea que no existe diferencia según el asiento que utiliza una persona al momento de ocurrir un choque. De esta forma, en la Tabla 5 se aprecia que existen diferencias significativas, pues el valor-p es menor que un 5%. En otras palabras, existe evidencia suficiente para rechazar la hipótesis nula con al menos un 95% de confianza.

Test para el tipo de protección

Para evaluar si el tipo de protección posee un efecto significativo se plantea la siguiente hipótesis

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = 0$$
$$H_1 : \text{Al menos un } \gamma_k \text{ es distinto de cero}$$

Con la información contenida en la Tabla 5 es posible concluir con un nivel de significancia del 5% la existencia de diferencias significativas, es decir, se rechaza H_0.

Test para el número de puertas

Finalmente para evaluar si el número de puertas del vehículo posee un efecto significativo sobre la respuesta se plantea la siguiente hipótesis

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0$$
$$H_1 : \text{Al menos un } \delta_l \text{ es distinto de cero}$$

Al igual que en los casos anteriores, se rechaza la hipótesis nula, por lo cual existe evidencia suficiente para determinar que el número de puertas influye en la respuesta con al menos un 95% de confianza.

Fuente variación	g.l	Sumas Cuadráticas	Media Cuadrática	Estadístico F	Valor-p
Size	7	7.68	1.1	12.111460	0.00000
DP	1	4.303842	4.3038424	47.493107	0.00000
Protección	4	1.119535	0.2798838	3.088531	0.01614

Puertas	2	1.139537	0.5697684	6.287421	0.00209
Residuo	337	30.539061	0.0906204		
Total	351	44.785			

Tabla 5: Tabla ANOVA modelo ajustado

4.1 Test de comparación de medias

Anteriormente se realizaron las estimaciones de los parámetros. En esta sección se estudiará si existe diferencias entre las medias de cada nivel del factor, pues es importante destacar que para rechazar la hipótesis nula en un test F, al menos un coeficiente de debe ser significativo. Por lo que en este apartado se desarrolla un análisis más exhaustivo acerca de las medias en cada grupo. Para ello se utilizarán los test de comparaciones de medias de Tukey, ya que, su implementación en `R` realiza aproximaciones para casos desbalanceados.

Es importante destacar el sentido practico de este análisis, pues de esta manera se pueden determinar recomendaciones para los conductores y pasajeros para un viaje más seguro.

Comparación según Pasajero o Conductor

En este caso se desea evaluar si existe diferencias entre las medias de quienes son Pasajeros y Conductores. La significancia de este test está estrechamente relacionada con la significancia de la variable, pues al considerar la restricción suma se estaría estimando sólo un coeficiente. Los límites del intervalo considerando un 95% confianza son (-0.28, -0.16), como no contiene el valor 0, entonces se concluye que existen diferencias significativas entre los promedios asociados al daño de las personas según el asiento que utilizan.

Comparación según número de puertas

Ahora se desea evaluar si existen diferencias significativas en el promedio de los daños entre aquellos autos con 2, 3 o 4 puertas.

Comparación	Diferencia	Límite inferior	Límite superior
3 - 2 puertas	0.0594	-0.0961	0.2150
4 - 2 puertas	0.1216	0.0432	0.2000
4 - 3 puertas	0.0622	-0.0919	0.2163

Tabla 6: Diferencia de medias Método de Tukey según Número de puertas

Como se aprecia en la Tabla 6 existen diferencias significativas en la media del daño entre vehículo de 4 y 2 puertas, pues con un 95% de confianza no contiene el valor cero. Por el contrario, las comparaciones entre aquellos vehículos de 4 y 3 puertas, y con los de 3 y 2 puertas no resultan ser significativas, pues ambos intervalos contienen el valor cero, lo cual significa que pueden llegar a tener los mismos promedios.

Comparación según Protección

En este segmento se estudia los tipos de protecciones que difieren en el daño promedio utilizando el test de Tukey.

Como se aprecia en la Figura 6 no existen diferencias significativas en cuanto a la protección, salvo en un caso. Utilizando la información de la Figura 3 notamos que existe un patrón entre los gráficos boxplots, en particular aquellos cuyo tipo de protección es el airbag resultan tener un daño levemente menor que aquellos que poseen sólo cinturón, por lo que se decide separar dichas categorías en dos grupos.

Lo anterior tiene como fin dar mayor interpretación a la variable, lo que resulta ser significativo en el modelo de regresión y, como se estudiará próximamente para estimar las varianzas por grupo, pues el modelo no es homocedástico. Con los cambios realizados se obtienen que el intervalo de un 95% de confianza es (0.0151, 0.1786) que no contiene el valor cero, por ende, existen diferencias significativas en las diferencias de las medias de cada grupo.

Figura 6: Intervalo de Tukey para variable Protección

Comparación según tamaño

Finalmente se estudia si existen diferencias en las medias de los daños en los diferentes tipos de vehículo. Empleando nuevamente el test de Tukey para comparaciones múltiples se obtiene los siguientes resultados.

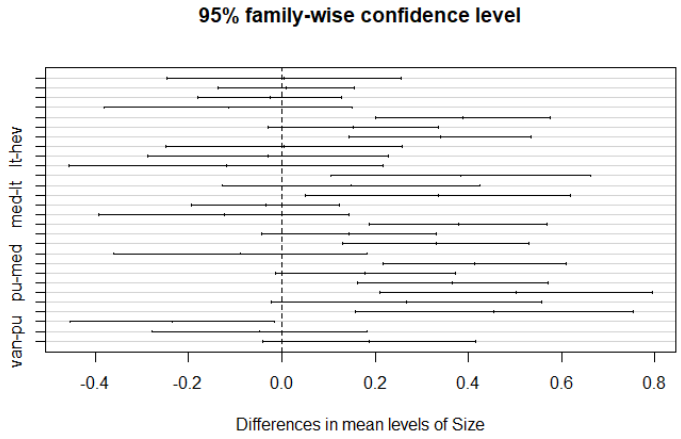

Figura 7: Intervalo de Tukey para variable Tamaño del auto

Como se aprecia en la Figura 7 en muchos casos no existe diferencia entre las medias de dos tamaños de vehículo al considerar una confianza de de 95%. Esto se debe principalmente a que el tamaño del auto, muchas veces determinado por el volumen, no cambia significativamente entre categorías de tamaño. Lo anterior es posible evidenciarlo en la Figura 3, donde se aprecia que los vehículos más pequeños tienden a tener comportamientos similares, mientras que por su parte los autos de mayor volumen, en este

caso las categorías van y pick up, también muestran comportamientos similares. Utilizando los mismos argumentos mencionados anteriormente, se agrupan las categorías de los tamaños `comp`, `hev`, `lt`, `medy` `mini` como liviano y las categorías `van` y `pu` como pesado. De esta forma en la Tabla 7 notamos que existe diferencias significativas entre cada categoría de tamaño, pues el intervalo no tiene el valor cero el 95% de las veces.

Comparación	Diferencia	Intervalo Inferior	Intervalo Superior
mpv-liviano	0.3981	0.2687	0.5275
pesado-liviano	0.2476	0.1497	0.3455
pesado-mpv	-0.1505	-0.3000	-0.0010

Tabla 7: Diferencia de medias Método de Tukey según tamaño

5 Validación de supuestos

Recordando que la variable respuesta fue estandarizada y posteriormente transformada, ahora se busca comprobar los supuestos del modelo.

5.1 Normalidad

Considerando los residuos estandarizados del modelo, en la Figura 8 se presenta un gráfico de residuos que involucran los de una distribución normal. Es posible notar un buen ajuste del modelo para la gran mayoría de observaciones a excepción de las colas que, como suele suceder, tienden a distanciarse del centro. No obstante, vemos que esto no corresponde a un peligro importante, por lo cual no hay indicios de una falta de ajuste. Por su parte, la Tabla 5.1 muestra los resultados obtenidos luego de aplicar un test de Shapiro-Wilk a los residuos del modelo, obteniendo un valor p no significativo, por lo que no existe evidencia suficiente para rechazar la normalidad al considerar un 95% de confianza.

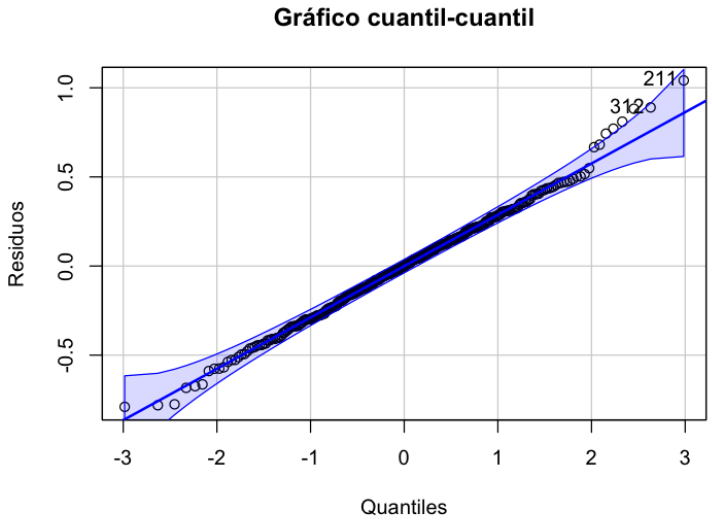

Figura 8: QQ-plot para los residuos del modelo

Estadistico	Valor - p
0.994	0.233%

Tabla 8: Test Shapiro-Wilk para los residuos del modelo

5.2 Homocedasticidad

Recordemos que el modelo supone que $\epsilon_{ijklm} \overset{iid}{\sim} N(0, \sigma^2)$, en otras palabras, se supone que el modelo tiene una varianza constante. Por lo que es necesario comprobar el supuesto. Para ello se utilizará el test de Barlett, cuyas hipótesis son

$$H_0 : \textbf{El modelo es homocedasticidad}$$
$$H_1 : \textbf{El modelo es heterocedasticidad}$$

En otras palabras en la hipótesis nula tenemos que $H_0 : \sigma^2_{ijkl} = \sigma^2_{i'j'k'l'}$, para $i, j, k, l \neq i', j', k', l'$. En la tabla 9 notamos que en algunos grupos se rechaza la hipótesis nula, es decir, existe indicios de heterocedasticidad. Para ello, se propone implementar un ajuste de Mínimo Cuadrados Ponderados, ya que este modelo no asume una varianza constante.

	Bartlett's K-squared	df	Valor - p
Variable Size	9.0036	7	0.2524%
Variable DP	1.325	1	0.2497%
Variable Protection	12.389	4	0.01468%
Variable Puertas	9.8454	2	0.007279%

Tabla 9: Test Barlett

Posteriormente se ajusta el modelo con las transformaciones en los predictores antes mencionados.

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_{ijklm}, \qquad \epsilon_{ijklm} \overset{iid}{\sim} N(0, \sigma^2_{ijkl}) \tag{4}$$

Para $i = 1, ..., 7$, $j = 1, 2$, $k = 1, 2, 3, 4, 5$, $l = 1, 2, 3$, donde

$$\sum_{i=1}^{7} \alpha_i = 0 \qquad \sum_{j=1}^{2} \beta_j = 0 \qquad \sum_{k=1}^{5} \gamma_k = 0 \qquad \sum_{l=1}^{3} \delta_l = 0$$

De la ecuación (6) notamos que se define una varianza que depende de la celda, es decir, el modelo que se ajusta considera la heterocedasticdad de los datos.

5.2.1 Gráficos de Homocedasticidad

En la Figura 9 los residuos estandarizados versus los valores ajustados. Es posible apreciar que, en primer lugar, la media está bien identificada, es decir, que los errores están entorno al valor cero en la recta. Por otra parte notamos que no todos las celdas tienen la misma varianza, sin embargo, al ajustar el modelo con mínimo cuadrados ponderados notamos que no es un problema, pues no se emplea el supuesto de varianza constante. Finalmente, es de interés identificar observaciones outliers, por lo que se muestran las bandas de confianza, se identifican varias observaciones que están por debajo o por alto de ellas, por lo que se debería realizar un análisis más detallado en este aspecto.

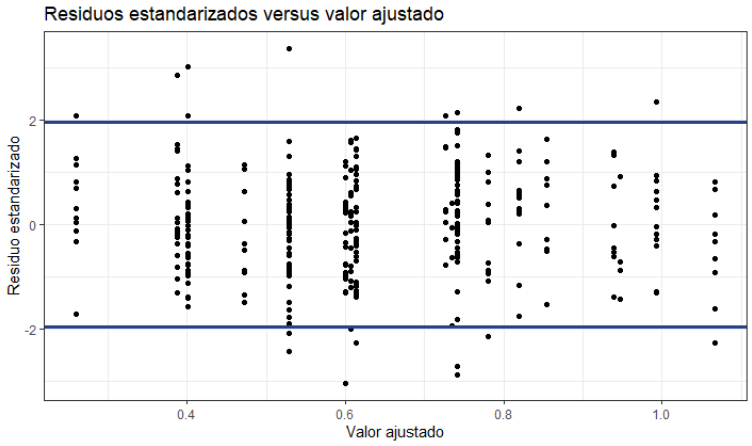

Figura 9: Gráfico Homocedasticidad variable Size

5.3 Puntos Palanca

Finalmente, para la validación del modelo, otro tópico de interés es analizar observaciones palanca. Para ello se dispone de los Figuras 10, 11, 12 y 13.

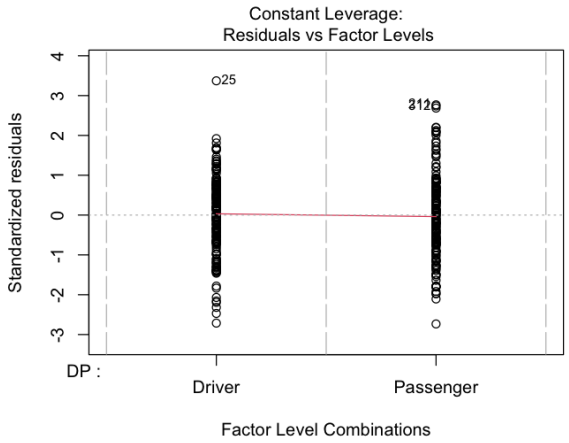

Figura 10: Gráfico Puntos Influyentes variable DP

Se puede observar, que para todas las categorías existen puntos palanca, para ello se debe destacar que los valores de dichos valores se encuentran en las abscisas, por lo que se encuentran observaciones que están más alejas del resto de las observaciones.

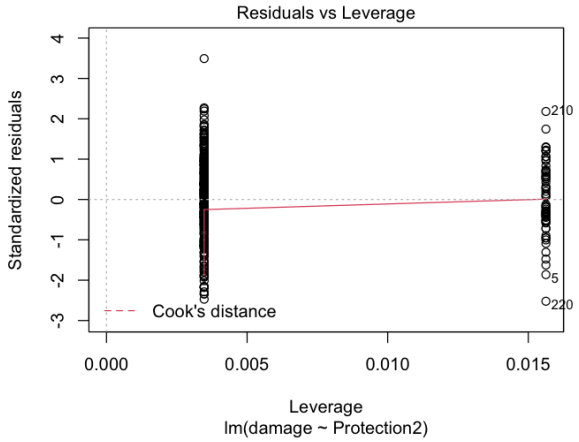

Figura 11: Gráfico Puntos Influyentes variable Protection

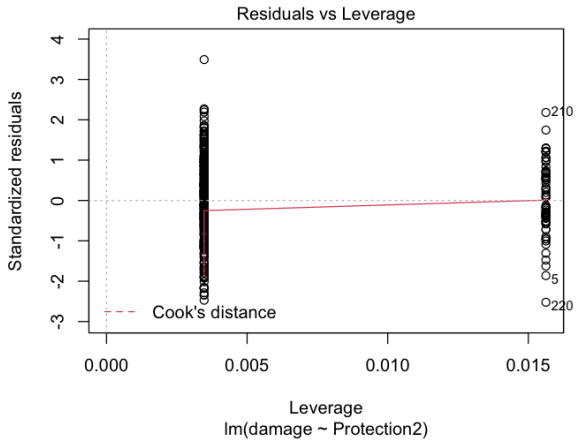

Figura 12: Gráfico Puntos Influyentes variable Size

6 Conclusión

En síntesis en este informe, en primer lugar se realiza un análisis exploratorio de los datos, con el fin encontrar patrones y estudiar las variables en el conjunto de datos. Además, en dicha sección se determina una variable respuesta que corresponde a una combinación lineal de las variables relacionadas con el daño, otorgándole más pesos a aquellas variables que estuvieran mayor relacionadas con daños mortales.

Luego de pasar por todos los modelos posibles con la información que se posee, se logró aterrizar en el modelo propuesto (3), el cual terminó considerando tan solo 4 variables, éstas son el asiento del pasajero o conductor, la cantidad de puertas, el tamaño del vehículo y el tipo de protección que posee, cuya significancia logró satisfacer nuestras necesidades. Se pudo conseguir un perfil del tipo de automóvil que más nos importaba conocer, el cual se puede catalogar como el tipo de auto con mayor índice de peligro para quienes hagan uso de él.

Posteriormente se ajusta el modelo y se interpretan los coeficientes asociados a ellos. Posteriormente se realiza inferencia sobre la significancia de cada factor y se estudia la diferencia entre los niveles en cada uno de ellos.

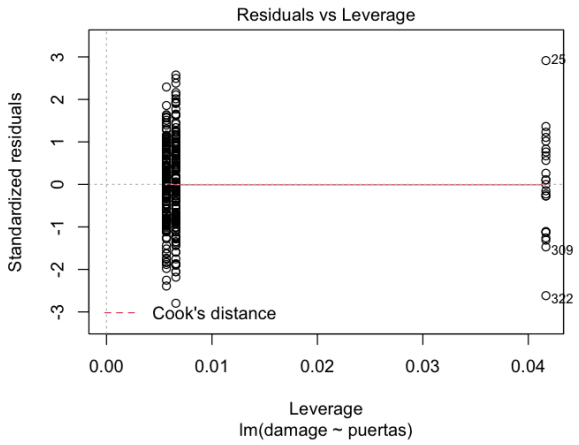

Figura 13: Gráfico Puntos Influyentes variable Puertas

Finalmente, para asegurar que el modelo propuesto es válido y por consiguiente las conclusiones que se puedan extraer de él, se procede a validar el modelo, en particular se estudia la homocedasticidad, la identificación de la media y observaciones palanca y outliers.

References

[1] Estadísticas Generales. (s. f.). Conaset. Recuperado 10 de julio de 2023, de https://www.conaset.cl/programa/observatorio-datos-estadistica/biblioteca-observatorio/estadisticas-generales/