UNIVERSIDAD NACIONAL DE COLOMBIA

PREGRADO EN ESTADISTICA

Departamento de Estadistica
Facultad de Ciencias

Regresion Lineal

Proyecto - Entrega 3
Análisis RLM - Indicadoras
Encuesta Anual de Comercio (EAC)
Año 2012

Integrantes:

Santiago Vera Quiceno C.C. 1.000.205.497
Alejandro Velásquez Rendón C.C. 1.000.913.570
Iván Sebastián Palomino Umaña C.C. 1.105.787.354

Medellin, Colombia

31 de mayo de 2024

Índice

Índice de figuras

Índice de cuadros

1 Preambulo

Nos basaremos en la Encuesta Anual de Comercio (EAC) del año 2012, que fue hecha por el DANE (Departamento Administrativo Nacional de Estadística) y, pese a que hayamos tomado la encuesta realizada en el 2012, esta se hace periódicamente, dada la relevancia del sector comercial en nuestro país.

La primera anotación importante es que cada unidad de observación serán empresas, que se seleccionaron por muestreo probabilístico estratificado de las empresas dedicadas al comercio interior identificadas por el NIT.

Nuestro interés recae en las ventas de las empresas, que la trataremos como nuestra variable respuesta, y queremos saber qué tanta variabilidad de estas se puede explicar, en una parte, por los gastos en los que incurren las empresas en el desarrollo de su actividad comercial en diversos ámbitos (publicidad, transporte, etc.). Además también tomaremos en consideración al personal contratado por la empresa y cómo esta cantidad puede afectar las ventas.

No haremos el estudio sobre el total de empresas en el país, sino que estaremos interesados en trabajar con los tipos de sociedades mercantiles más comunes en el territorio. Entiéndase como sociedad mercantil a una persona jurídica conformada por dos o más personas (pueden ser personas naturales o jurídicas, o sea, una sociedad puede ser parte de otra sociedad). Los tipos de sociedades comerciales en las que estaremos interesados son:

Sociedad Limitada (S.L.): la responsabilidad de cada socio está limitada por el valor de sus aportes

Sociedad Anónima (S.A.): El capital de la sociedad se divide en acciones, y los socios son los dueños de estas acciones.

Sociedad por Acciones Simplificada (S.A.S.): Combina características de las Sociedades Anónimas y las Sociedades Limitadas.

En la tabla 1 presentaremos una ficha técnica con las variables involucradas en nuestro estudio:

Tabla 1: Ficha Técnica

VARIABLE	SIGNIFICADO
VENTA	Valor de las ventas causadas en un año, en unidades de miles de pesos corrientes, sin incluir impuestos indirectos.
BTA	Producción bruta, consiste en las ventas menos el costo de la mercancía vendida.
FLETE	Gasto en transporte, fletes y acarreos, en unidades de miles de pesos corrientes.
ENER	Gasto en energía eléctrica comprada, en unidades de miles de pesos corrientes.
PUBL	Gasto en propaganda y publicidad, en unidades de miles de pesos corrientes.
PERSO	Personal ocupado promedio durante un año por la empresa en su actividad comercial, se considera a propietarios, socios, familiares no remunerados, personal permanente y personal temporal. Y no considera a trabajadores con licencia limitada y no remunerada, en servicio militar, pensionados, miembros de la junta directiva a quienes se les paga únicamente por asistir a reuniones y socios o familiares que no desempeñan funciones dentro de la empresa.
IDOJ	Variable categórica que identifica al tipo de sociedad mercantil que es la empresa en cuestión siendo: 1 para Sociedad Limitada; 2 para Sociedad Anónima; 3 para Sociedad por Acciones Simplificada.

2 Análisis Descriptivo

2.1 Encabezado de los Datos

A través de la visualización del encabezado de los datos podemos darnos una primera impresión de la escala de estos y los cambios que hay entre ellos, en algunas observaciones se ven 0 como respuestas, y esto es lógico por la forma en la que se midieron los datos, en este caso, solamente vemos una empresa que no destino inversión a la publicidad, puede tratarse de alguna empresa pequeña o simplemente el dueño no lo reportó en la encuesta.

En todo caso, las escalas parecen concordantes con la premisa de la base de datos, pues, es de esperarse que las variables tengan una relación positiva, pues a mayor cantidad de ventas, es lógico pensar que hubo más gasto, producción bruta y personal, y en la tabla 2 se ve un poco esta tendencia.

Tabla 2: Encabezado de los datos

VENTA	BTA	FLETE	ENER	PUBL	PERSO	IDOJ
6124653	514168	169189.000	9865.00	0.00	11.0000	1
17610243	1288853	1920.000	194653.00	111077.00	48.0000	2
25206358	5077732	2765.764	59266.38	40301.14	237.0655	1
20182845	13326644	203083.000	35202.00	32370.00	50.0000	2
17610243	1288853	1920.000	194653.00	111077.00	48.0000	2
5454335	1939017	434.000	69537.00	83737.00	52.0000	3

2.2 Estadísticos descriptivos

Tabla 3: Vectores de medias

	medias
VENTA	20096865.86
BTA	6902300.00
FLETE	234618.59
ENER	104997.59
PUBL	320949.17
PERSO	95.81

A través del vector de medias visto en la tabla 3 podemos observar el valor de la media de cada una de las variables las cuales estudiaremos a lo largo de este proyecto. Se ve que el promedio de las ventas supera al de cualquier otra que sea medible en miles de pesos

corrientes, lo cual es lógico, y aún más supera la suma de los gastos, que es algo también esperable (puede pasar que hayan casos donde una empresa gaste más de lo que vende, pese a no ser una situación favorable, sí es factible).

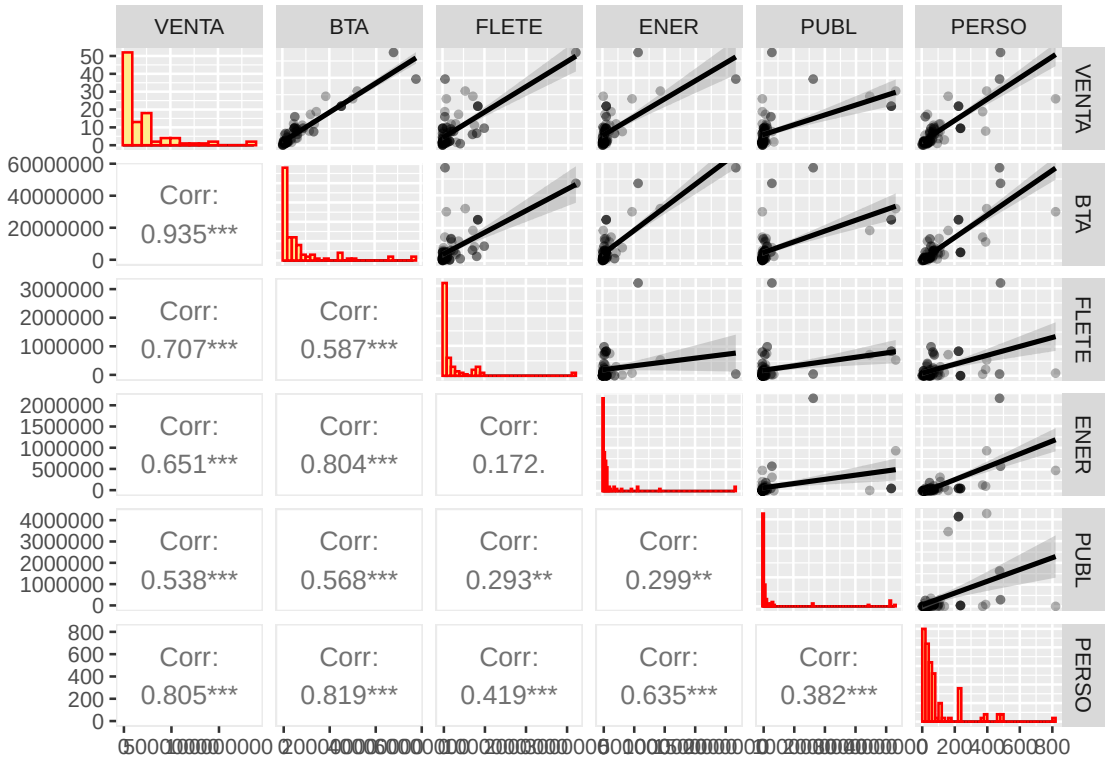


Figura 1: Matriz de correlación

De la figura 1 podemos sacar varias conclusiones, en primera instancia podemos analizar las correlaciones, absolutamente todas dieron positivas, lo que refuerza la hipótesis que planteamos recién vimos el encabezado de los datos, y es que las variables tienen una relación lineal de tendencia positiva, pero más allá del signo, hay correlaciones de más de 0.8 entre más de una pareja de variables que son:

- *VENTA* y *BTA*
- *VENTA* y *PERSO*
- *BTA* y *ENER*
- *BTA* y *PERSO*

En el caso de la variable "VENTA", el hecho de que tenga correlaciones altas lo podemos tomar como algo positivo, pues indica que hay relación lineal entre nuestra variable dependiente y las variables regresoras, y es más, pese a que no tiene una correlación alta con todas las variables, en todos los casos es mayor a 0.5, lo que indica que hay correlación al menos moderada entre la variable de las ventas y las demás.

Por otra parte, las correlaciones altas entre covariables puede significar un problema, la variable que más parece tener esta tendencia es la de "BTA", pues tiene correlaciones altas con dos variables, lo que puede ser indicio de problemas de multicolinealidad. Pero esto puede encontrar su explicación en la forma en que definimos esta variable, pues su cálculo se hace directamente con las ventas, de ahí que, además, tenga una correlación tan alta con nuestra variable respuesta.

Además de la figura 1 también observamos, en la diagonal principal, los histogramas de las variables, todos muestran una misma tendencia con colas pesadas a la derecha, lo que puede indicar observaciones muy alejadas de la tendencia central.

2.3 Gráficos

2.3.1 Box-plot

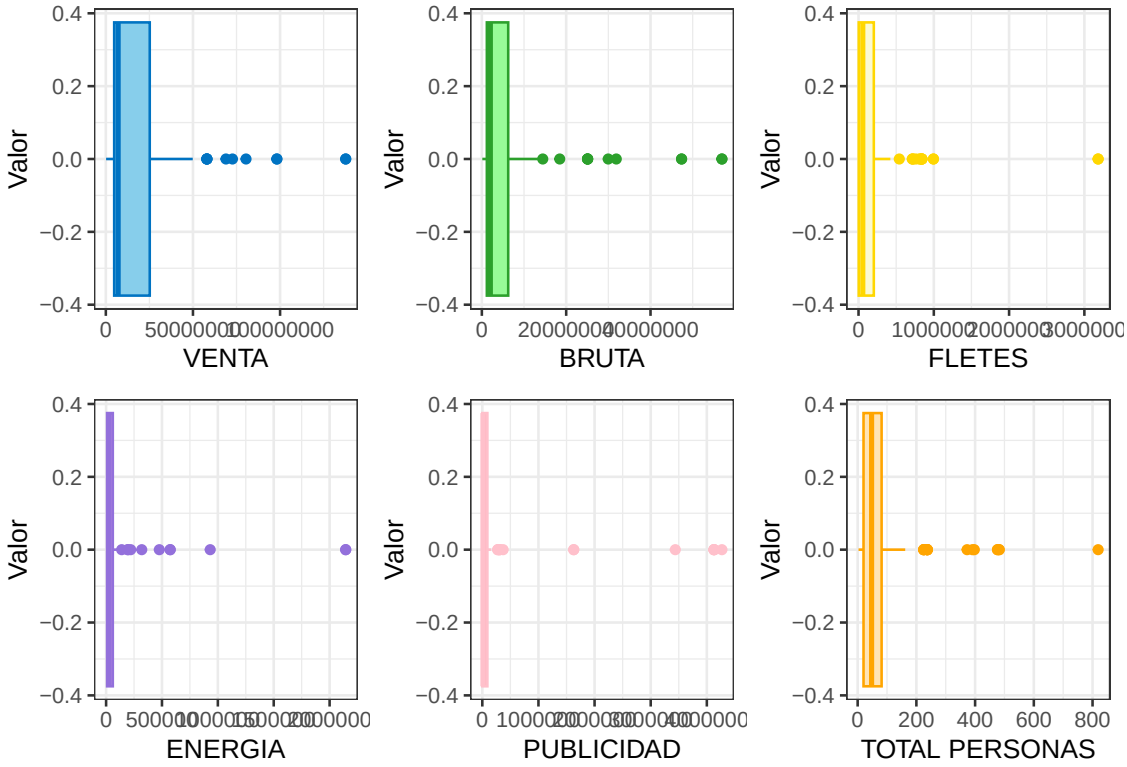


Figura 2: Boxplots

Gracias a los box plot vistos en la figura 2 en los cuales podemos ver los cuantiles 25, 50 y 75 de cada variables notamos la existencia de valores atípicos, que habíamos previsto con anterioridad, hay presencia de observaciones atípicas en todas las variables, incluyendo "VENTA", por lo que es posible que haya empresas en esta muestra que tengan gastos y ventas muy superiores a la norma, que podría derivar en problemas con observaciones atípicas y/o generar puntos influenciables o de balanceo.

2.3.2 Gráfico de barras

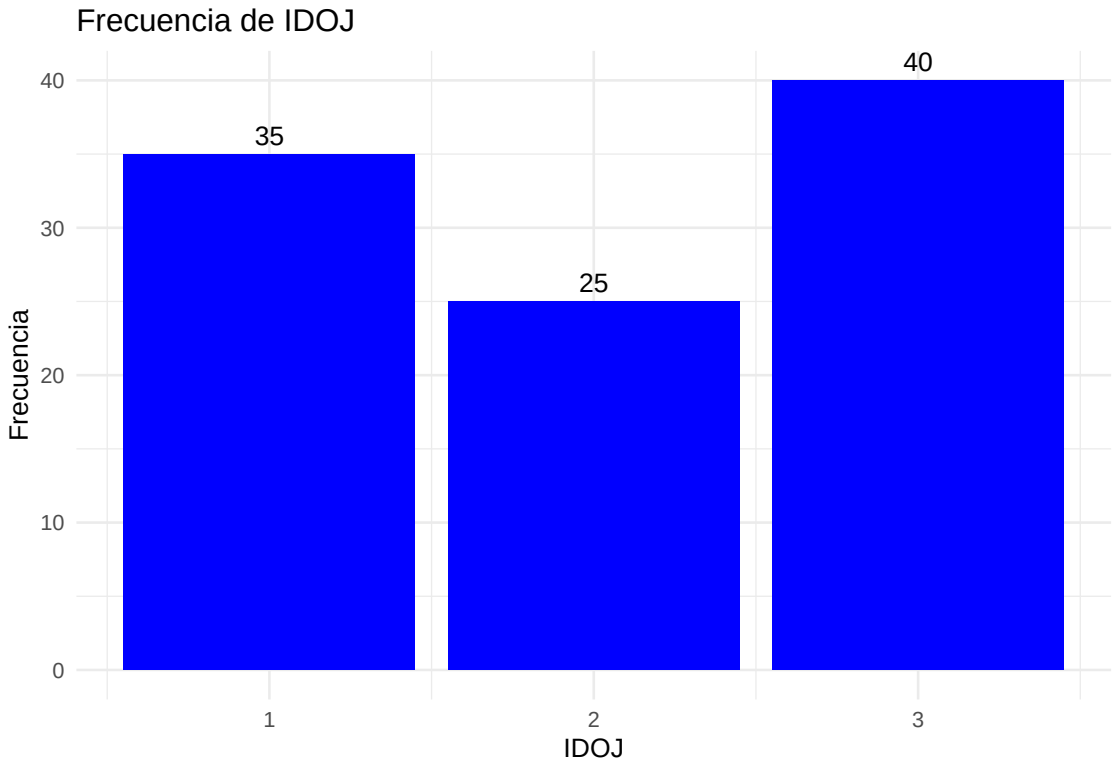


Figura 3: Gráfico de Barras

Finalmente, en el gráfico de barras 3 se observan las frecuencias en las que aparecen cada una de los distintos tipos de empresas en la base de datos, siendo más prominente la Sociedad por Acciones Simplificada, de la que contamos con 40 datos y la menos común es la Sociedad Anónima de la cual tenemos 25 observaciones.

3 Entrega 2

3.1 Punto 1

Como ya hemos dicho, queremos ver si las ventas de las empresas se pueden explicar por ciertas variables continuas, como los gastos o el personal contratado. Siendo coherente con la notación que definimos previamente, queremos establecer un modelo de regresión lineal de la siguiente forma:

$VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 FLETE_i + \beta_3 ENER_i + \beta_4 PUBL_i + \beta_5 PERSO_i + \xi_i$

$\xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ e $i = 1, 2, ..., 100$

Lo primero que haremos es determinar una ecuación ajustada o estimada de nuestro modelo, para esto es necesario determinar los estimadores de nuestros coeficientes, esto es, el vector

$\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_5 \end{bmatrix}$, que serán los estimadores de nuestros betas. Estos estimadores los calculamos haciendo uso del software R, y obtenemos los siguientes valores:

Coeficiente	Estimación
β_0	$\hat{\beta}_0$ =3.096e+06
β_1	$\hat{\beta}_1$ =1.6553
β_2	$\hat{\beta}_2$ =12.25
β_3	$\hat{\beta}_3$ =-5.685
β_4	$\hat{\beta}_4$ =0.7789
β_5	$\hat{\beta}_5$ =3.203e+04

Tabla 4: Coeficientes Estimados

Basados en la tabla 4, el modelo ajustado es, entonces:

$\hat{VENTA}_i = 3.096 \times 10^6 + 1.6553 BTA_i + 12.25 FLETE_i - 5.685 ENER_i + 0.7789 PUBL_i + 3.203 \times 10^4 PERSO_i$

Ahora queremos ver, si el modelo es significativo, esto es, si la variabilidad de las ventas de las empresas es explicada por nuestro modelo de regresión lineal.

En este sentido vamos a plantear la siguiente prueba de hipótesis:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

$H_1 : \exists j$ tal que $\beta_j \neq 0$ para $j = 1, 2, 3, 4, 5$

Para esto, usaremos el estadístico $F_0 = MSR/MSE \sim f_{5,94}$, luego usamos el valor p ($p(f_{5,94} > F_0)$) para tomar una decisión, en este caso que nos da un valor de 2.2e-16, es decir, será más pequeño que cualquier α, por lo que hay evidencia suficiente para rechazar la hipótesis nula, o, lo que es lo mismo, el modelo es globalmente significativo para explicar la variabilidad de la variable respuesta

Una vez determinado que el modelo es significativo, es bueno mencionar el coeficiente de determinación (R^2), que tiene un valor de 0.9234, o sea, que el 92.34% de la variabilidad total de la variable "VENTA" es explicada por el modelo de regresión, el valor es bastante alto, lo que puede ser indicio de que el modelo es adecuado.

3.2 Punto 2

Si queremos saber cuál variable es la que tiene un mayor efecto parcial sobre la respuesta media, debemos considerar las escalas de medida, pues, por ejemplo, la variable PERSO, no mide ninguna cantidad monetaria, por lo que estará en otra escala de medida, para determinar entonces cuál es la variable que aporta más a la respuesta media, debemos estandarizar todas las variables (restarle su media y dividir sobre su desviación estándar), incluida la

variable respuesta, una vez estandarizadas las variables, determinaremos nuevos coeficientes estandarizados que denotaremos por β_i^*, y, igualmente, calcularemos sus estimadores, los resultados obtenidos se ven en la 5.

Variable asociada	Beta Estimado	Beta Estandarizado Estimado
Intercepto	$\hat{\beta}_0$ =3.096e+06	$\hat{\beta}_0^*$ =0
BTA	$\hat{\beta}_1$ =1.6553	$\hat{\beta}_1^*$ =0.7063
FLETE	$\hat{\beta}_2$ =12.25	$\hat{\beta}_2^*$ =0.2279
ENER	$\hat{\beta}_3$ =-5.685	$\hat{\beta}_3^*$ =-0.068
PUBL	$\hat{\beta}_4$ =0.7789	$\hat{\beta}_4^*$ =0.0283
PERSO	$\hat{\beta}_5$ =3.203e+04	$\hat{\beta}_1^*$ =0.1629

Tabla 5: Coeficientes Estimados Estandarizados

Ahora, del cuadro 5, el valor mayor de un coeficiente estimado estandarizado en valor absoluto es el de $\hat{\beta}_1^*$, con un valor de 0.7063, lo que hace lógica con lo visto en el análisis descriptivo, desde la otra perspectiva, el de menor valor es el que acompaña a la variable "PUBL" con un valor de 0.0283, lo que significa que la variable referente a la publicidad es la que menos aporta en la respuesta promedio.

3.3 Punto 3

Ya llegamos a la conclusión de que el modelo es globalmente significativo, ahora, veremos si, todas las variables son significativas de forma individual, para esto, para $j = 1, 2, ..., 5$ plantearemos el siguiente juego de hipótesis:

$H_0 : \beta_j = \beta_{j,0}$ vs $H_1 : \beta_j \neq \beta_{j,0}$

Y usaremos como estadístico de prueba a:

$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{s.e(\hat{\beta}_j)} \sim t_{94}$

Donde $s.e(\hat{\beta}_j)$ es un estimador de la varianza de $\hat{\beta}_j$.

Puesto que queremos ver la significancia de cada una de las variables, haremos que $\beta_{j,0} = 0$ para $j = 1, 2, ..., 5$. Usaremos como criterio de rechazo al Valor P ($p(|t_{94}| > |T_0|)$), rechazaremos si este es pequeño.

Los resultados de las pruebas de hipótesis aplicadas a cada coeficiente se ven en la tabla 6.

Variable Asociada	Estadístico T_0	Valor P
BTA	5.733	1.19e-07
FLETE	4.492	2.01e-05
ENER	-0.936	0.3514
PUBL	0.687	0.494
PERSO	3.023	0.0032

Tabla 6: Tests de Significancia individual

De la tabla 11 vemos que los valores p asociados a las pruebas de hipótesis de los coeficientes de "BTA", "FLETE" y "PERSO" son menores que un α de referencia (supongamos 0.05). Por tanto serán estadísticamente significativos para el modelo de manera individual. Mientras que, las variables "ENER" y "PUBL" no lo son, pues sus valores P serán mayores que cualquier α, por tanto "ENER" y "PUBL" no ayudan a explicar la cantidad promedio de las ventas en las empresas.

3.4 Punto 4

Hemos visto que tano β_3 y β_4 no son significativas de forma individual, o sea, que las variables que acompañan a estos coeficientes (ENER y PUBLI) no ayudan a explicar la respuesta media de la variable "VENTA", ahora, siguiendo estos resultados, queremos plantear una nueva prueba de hipóteis, cómo sigue:

$$\begin{cases} \mathbf{H_0} : \beta_3 = \beta_4 = 0 \\ \mathbf{H_1} : \beta_3 \neq 0 \ o \ \beta_4 \neq 0 \end{cases}$$

Para esto, usaremos lo siguiente:

$$\begin{cases} \text{Modelo Full será: } VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 FLETE_i + \beta_3 ENER_i + \beta_4 PUBL_i + \beta_5 PERSO_i + \xi_i, \\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ \text{Modelo Reducido será: } VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 FLETE_i + \beta_5 PERSO_i + \xi_i, \ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ \text{Estadístico de prueba: } F_0 = \dfrac{[SSE(MR) - SSE(MF)]/\nu}{MSE(MF)} = 1.3597 \\ \text{Distribución del estadístico: El estadístico } F_0 \sim f_{2,94} \\ \text{Valor p: al calcular: } P(f_{2,94} > 1.3597) = 0.2618 \end{cases}$$

Al obtener un valor-p de 0.2618 no tenemos evidencia suficiente para rechazar la hipótesis nula y por lo tanto verificamos que de manera conjunta β_3 como β_4 son iguales a 0, y por tanto, las variables ENER y PUBL no son conjuntamente significativas para el modelo.

3.5 Punto 5

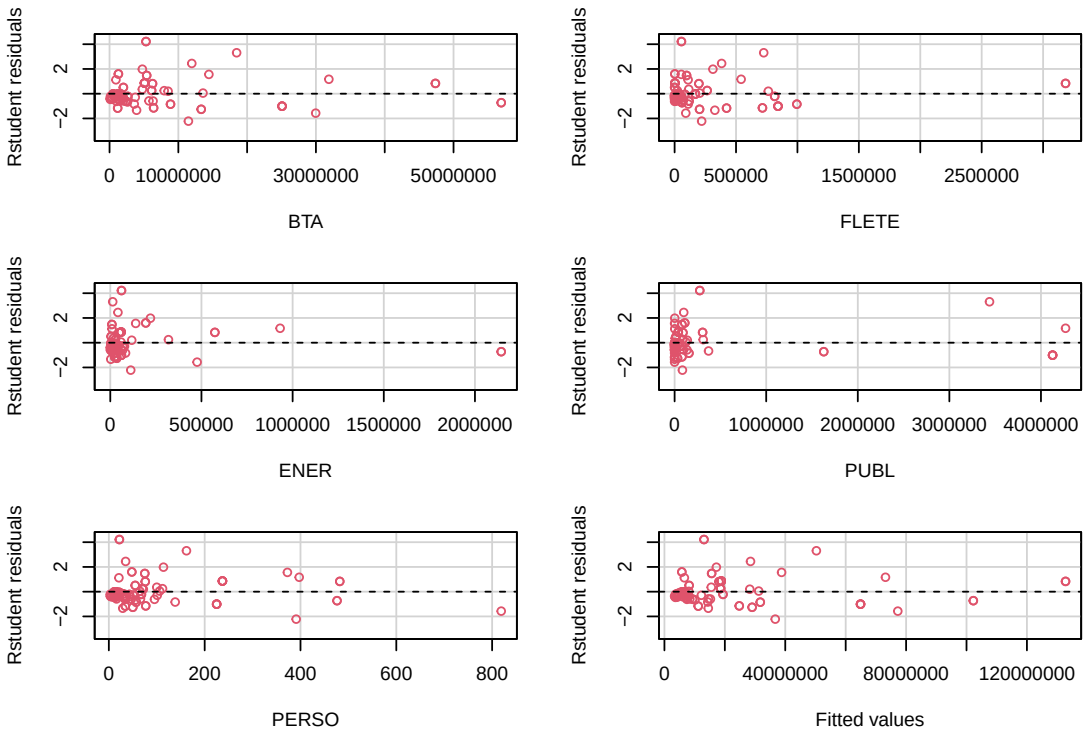


Figura 4: Residuales Estudentizados vs Valores Ajustados

De la gráfica 4 observamos:

-Que no hay patrones parabólicos o en forma de "U", por lo tanto es una evidencia de que no hay problemas de carencia de ajuste en el modelo.

-No hay patrones lineales, entonces no existe evidencia en contra del supuesto de varianza constante; así mismo, la nube de puntos está sobre su media esperada que es 0.

-Hay varios residuales que sobrepasan las bandas del valor absoluto de 3 en varias gráficas (están por encima de 3), o sea que es posible que haya presencia de observaciones atípicas.

-Posible existencia de valores de balanceo debido a sus valores muy alejados de la nube de puntos en el eje x. De igual manera notamos valores influenciables debido a sus altos valores tanto en el eje x como en el eje y.

3.6 Punto 6

El siguiente supuesto que queremos probar es el de la distribución de los errores:

$$\text{Nuestra hipótesis a probar será: } \begin{cases} \mathbf{H_0} : \xi_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}) \\ \mathbf{H_1} : \xi_\mathbf{i} \nsim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}) \end{cases}$$

El resultado de la prueba de hipótesis se ve en la figura 5.

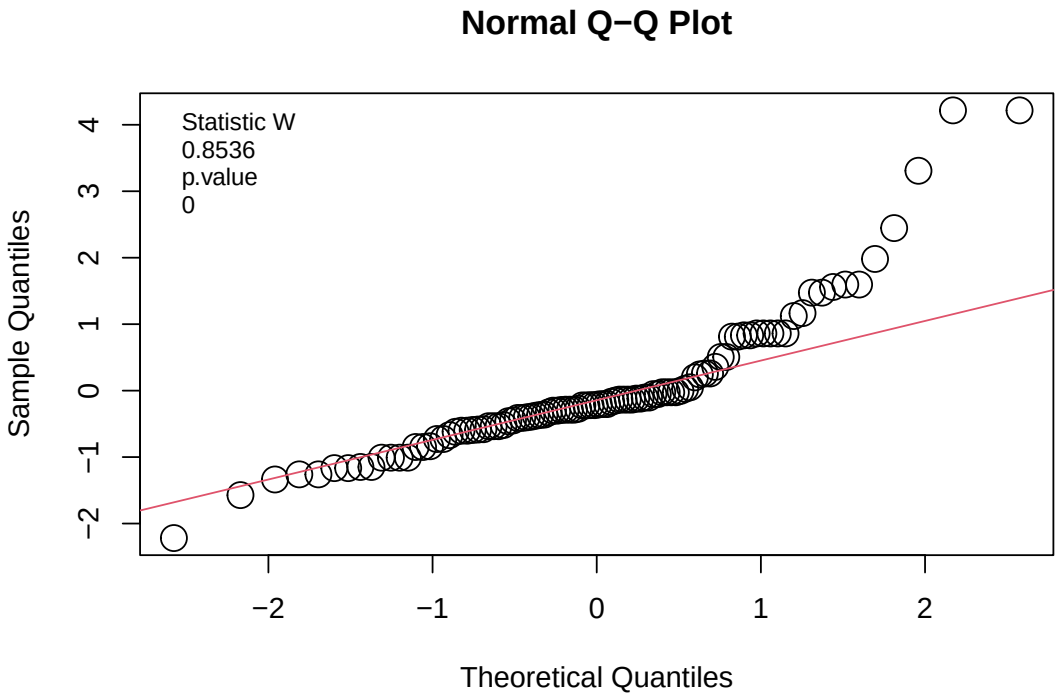


Figura 5: Q-Q Plot y Prueba de Normalidad

Al obtener un Valor-p ≈ 0, y siendo este menor que cualquier α, tenemos evidencia suficiente para rechazar H_0, por consiguiente nuestros errores no seguirán una distribución normal.

Al observar el gráfico de probabilidad normal (5) soportamos esta conclusión, dado que observamos bastantes residuales que se alejan de la línea de 45° que forman los cuantiles teóricos de la distribución normal, y con una forma que puede indicar que los residuales pueden seguir una distribución de colas pesadas.

3.7 Punto 7

Para este punto, realizaremos un análisis para determinar si existen, y cuales son, los datos outliers u observaciones atípicas, los datos con problemas de balanceo y las observaciones influenciables.

Tabla 7: Diagnóstico sobre los datos

	dfb.1_	dfb.BTA	dfb.FLET	dfb.ENER	dfb.PUBL	dfb.PERS	dffit	cov.r	cook.d	hat
11	0.47	0.38	-0.33	-0.27	-0.19	-0.41	0.63	0.38*	0.06	0.02
36	-0.02	-0.5	0.32	0.51	0.7	0.34	0.81*	1.45*	0.11	0.33*
48	-0.17	0.05	0.34	-0.02	-0.21	0.01	0.69	1.73*	0.08	0.41*
53	0.47	0.38	-0.33	-0.27	-0.19	-0.41	0.63	0.38*	0.06	0.02
54	0.07	0.02	0.03	-0.2	0.87	-0.04	1.28*	0.63*	0.25	0.13
55	-0.17	0.05	0.34	-0.02	-0.21	0.01	0.69	1.73*	0.08	0.41*
59	0.46	-0.09	0.47	0.43	0.37	-1.13*	-1.61*	1.87*	0.43	0.51*
60	0.02	-0.08	0.05	0.14	-0.28	0.03	-0.49	1.23*	0.04	0.19*
63	0	-0.08	0.12	-0.22	0.05	0.14	-0.67	1.9*	0.07	0.46*
64	0	-0.08	0.12	-0.22	0.05	0.14	-0.67	1.9*	0.07	0.46*
65	0.02	-0.08	0.05	0.14	-0.28	0.03	-0.49	1.23*	0.04	0.19*
66	0.09	0.17	0.04	0.08	0.04	-0.68	-0.8*	0.89	0.1	0.12
74	0.02	-0.08	0.05	0.14	-0.28	0.03	-0.49	1.23*	0.04	0.19*
86	0.21	0.69	-0.44	-0.56	-0.48	-0.56	0.76*	0.8*	0.09	0.09
92	0.02	-0.08	0.05	0.14	-0.28	0.03	-0.49	1.23*	0.04	0.19*

Tabla 8: Diagnóstico adicional sobre los datos

	StudRes	Hat	CookD
11	4.2166530	0.0220725	0.0567540
53	4.2166530	0.0220725	0.0567540
54	3.3082047	0.1309219	0.2484932
59	-1.5722578	0.5121058	0.4257769
63	-0.7281779	0.4573018	0.0748418

Para determinar si una observación tiene alguno de estos problemas, haremos uso de ciertos chequeos y medidas que se observan en las tablas 7 y 8 donde se ve la distancia de Cook, los $DFBetas_j$, etc. con esto concluimos que:

Las observaciones outliers son: 11, 53 y 54.

Puntos de balanceo son: 36, 48, 55, 59, 60, 63, 64, 65, 74 y 92.

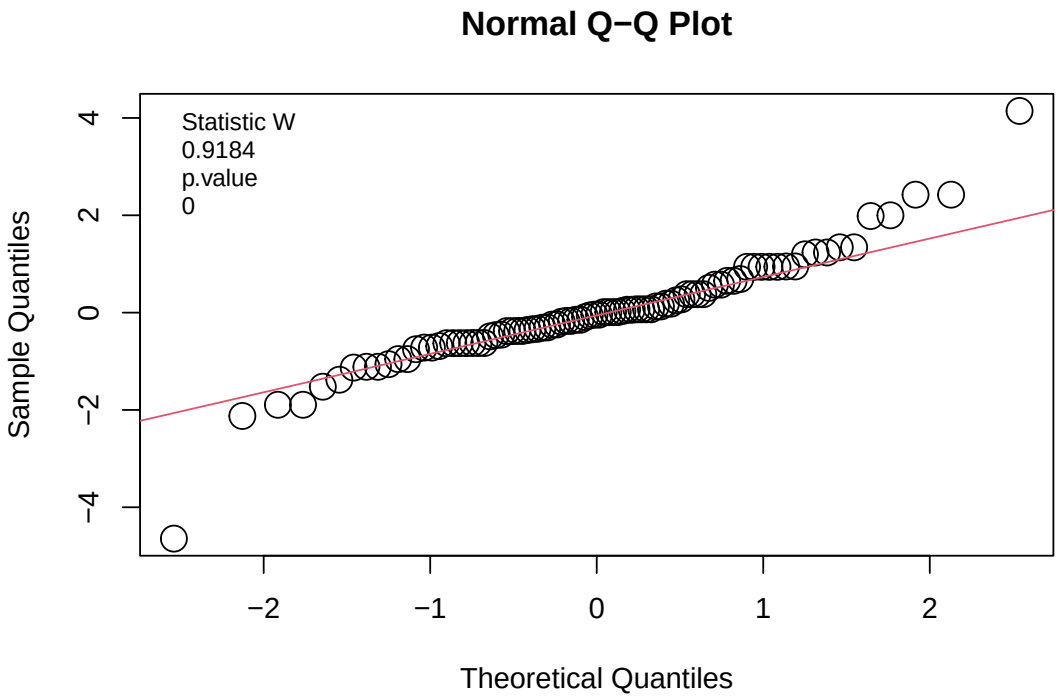
Para los influenciables tenemos que: 11, 36, 48, 53, 54, 55, 59, 60, 63, 65, 66, 74, 86 y 92.

3.8 Punto 8

Para la resolución de este ejercicio, asumimos que los datos mal digitados son los siguientes: 11, 36, 48, 53, 54, 55, 59, 60, 63, 64, esto seleccionado a raiz del apartado anterior en donde priorizamos los outliers y de balanceo, dado que los de balanceo también son influenciables.

Presentaremos el siguiente gráfico de normalidad, junto con el p-value resultante de la verificación de normalidad para nuestro modelo. Sea el siguiente juego de hipótesis

$$H_0 : \xi_i \sim N(0, \sigma^2) \quad vs \quad H_1 : \xi \nsim N(0, \sigma^2)$$



Nuevamente, se evidencian problemas de normalidad, esto siendo observable con el p-value resultante que es tan pequeño que tiende a 0, por lo tanto no hay evidencia suficiente para concluir que se cumple el supuesto de normalidad.

Ahora bien, para realizar una comparativa de los modelos, procederemos a mostrar un resumen de los parámetros del modelo de regresión sin los datos mal digitados

Coeficientes	**Estimate**	**Std. Error**	**t value**	**Pr(>t)**
(Intercept)	1.476205e+06	848008.2018	1.741	0.08538
BTA	1.7549	0.2316	7.577	0.0000000000424
FLETE	6.3966	2.7853	2.297	0.02413
ENER	29.8415	12.1907	2.448	0.01645
PUBL	-0.6389	1.2252	-0.521	0.60340
PERSO	3.5186e+04	10380.7602	3.390	0.00107

Tabla 9: Resumen salida Summary del modelo 2

Adicionalmente, obtenemos los parámetros estimados resultantes:

Coeficiente	Estimación Datos Filtrados	Estimación Datos
β_0	$\hat{\beta}_0$ =1.476205e+06	$\hat{\beta}_0$ =3.096e+06
β_1	$\hat{\beta}_1$ =1.7549	$\hat{\beta}_1$ =1.6553
β_2	$\hat{\beta}_2$ =6.3966	$\hat{\beta}_2$ =12.25
β_3	$\hat{\beta}_3$ =29.8415	$\hat{\beta}_3$ =-5.685
β_4	$\hat{\beta}_4$ =-0.6389	$\hat{\beta}_4$ =0.7789
β_5	$\hat{\beta}_5$ =3.5186e+04	$\hat{\beta}_5$ =3.203e+04

Tabla 10: Coeficientes Estimados

Y mostrarémos a continuación, el cambio porcentual de las estimaciones mostradas anteriormente.

Coeficiente	Estimación
β_0	$\hat{\beta}_0 = -52.31\%$
β_1	$\hat{\beta}_1 = 6.01\%$
β_2	$\hat{\beta}_2 = -47.78\%$
β_3	$\hat{\beta}_3 = 624.92\%$
β_4	$\hat{\beta}_4 = -182.03\%$
β_5	$\hat{\beta}_5 = 9.85\%$

Tabla 11: Cambios Porcentuales en las Estimaciones

En donde, se puede evidenciar que en las variables BTA, FLETE, ENER, PUBLI, PERSO, e incluso el intercepto, sus $\hat{\beta}_i$ tuvieron cambios notorios, esto es evidenciable mucho más drástico en determinados $\hat{\beta}_i$ como bien podría ser por ejemplo $\hat{\beta}_2$ donde su estimación pasó de 12.25 a 6.39 disminuyendo casi que a la mitad, o bien, con $\hat{\beta}_3$ con -5.685 aumentando hasta 29.84, siendo este el cambio más drástico con un aumento de 624% aproximadamente. Y si bien, tenemos en cuenta que la escala está en miles de pesos, cualquier cambio en cuanto al modelo será notorio.

Estás observaciones depuradas, que presentaban problemas de balanceo, influenciables o eran outliers, si bien presentaban ser negativas para el modelo, luego de quitarlas más allá del cambio notorio de los parámetros, sigue persistiendo el problema de normalidad.

3.9 Punto 9

Haremos uso de distintos métodos para diagnosticar multicolinealidad, esto lo haremos al modelo de regresión sin las observaciones "problemáticas" que eliminamos en el punto anterior.

3.9.1 Apartado 9.a

Por medio de la correlación.

Tabla 12: Matriz de correlación entre las covariables

	BTA	FLETE	ENER	PUBL	PERSO
BTA	1.0000000	0.5969451	0.2966055	0.7733136	0.6352110
FLETE	0.5969451	1.0000000	0.0857670	0.5030747	0.2095509
ENER	0.2966055	0.0857670	1.0000000	0.1170594	0.4520174
PUBL	0.7733136	0.5030747	0.1170594	1.0000000	0.3917956
PERSO	0.6352110	0.2095509	0.4520174	0.3917956	1.0000000

En la tabla 9 se observa la matriz de correlación entre las covariables, si consideramos que las correlaciones cuyo valor oscile entre 0.5 y 0.7 será considerada moderada, y si es mayor que 0.7 será considerada alta, las siguientes parejas de variables tienen correlaciones altas.

- *BTA* y *PUBL*
- *BTA* y *PERSO*
- *BTA* y *FLETE*
- *FLETE* y *PUBLI*

además hay algunas variables con una relación lineal moderada. Todo esto puede ser indicio de problemas de multicolinealidad.

3.9.2 Apartado 9.b

Tabla 13: VIF's

	x
BTA	4.667022
FLETE	1.681477
ENER	1.277023
PUBL	2.636395
PERSO	2.089336

Se puede observar en la tabla 10 que a través de los VIF no se pueden concluir problemas de multicolinealidad, dado que todos los valores son menores que 10.

3.9.3 Apartado 9.c

A pesar de evienciar alta correlación entre variables regresoras, los análisis de VIF y descomposición de varianza no permiten concluir que haya problemas de multicolinealidad en el modelo planteado. Pero al análizar los índices de condición, podemos observar que para $\sqrt{k_5} < 10$, no se evidencias problemas serios de multicolinealidad. Por otro lado, no se encuentra evidencia de multicolinealidad dado que para algún λ_i asociado, existan dos o más coeficientes $\pi_{ij} > 0.5$

Condition.Index	BTA	FLETE	ENER	PUBL	PERSO
1.000	0.008	0.018	0.018	0.011	0.014
2.663	0.002	0.353	0.108	0.085	0.059
3.296	0.001	0.211	0.667	0.001	0.121
3.878	0.007	0.176	0.084	0.303	0.382
6.328	0.972	0.209	0.000	0.448	0.411

Tabla 14: Proporciones de Varianza e Índice de Condición

3.10 Punto 10

Presentaremos la siguiente tabla con los posibles modelos a consideración:

covariables	"Modelo"	R^2_{adj}	C_p	$\lvert C_p - p \rvert$
1	(1) VENTA=BTA	0.8	27.1276	25.1276
2	(6) VENTA=BTA+PERSO	0.8328	11.629	8.629
3	(16) VENTA=BTA+FLETE+ENER	0.8436	7.4901	3.4901
4	(26) VENTA=BTA+FLETE+ENER+PERSO	0.8527	4.272	0.728
5	(31) VENTA=BTA+FLETE+ENER+PUBL+PERSO	0.8532	6	0

Tabla 15: Posibles modelos

3.10.1 Apartado 10.a. y 10.b.

Para elegir un modelo en base al R^2_{adj} y al C_p, nos basaremos en la tabla 11, donde, para cada número de covariables, elegimos el mejor modelo, y luego los comparamos.

En primer lugar, si nos basamos en el R^2_{adj}, realmente todos tienen valores muy similares, por tanto, no necesariamente elegiremos al que tenga mayor valor en este estadístico, el mejor modelo es entonces el modelo (6), o sea:

$VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_5 PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Esto es, porque tiene un R^2_{adj} apenas menor al de sus contrapartes con mayor cantidad de variables, entonces por principio de parsimonia, lo elegimos a él, a continuación se muestra (en la tabla 16) el resumen del modelo.

Tabla 16: Tabla de coeficientes-Modelo 6

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2601658.756083	790051.1162971	3.293026	0.0014340
BTA	1.928088	0.1455439	13.247464	0.0000000
PERSO	38897.187791	9748.8289149	3.989934	0.0001374

Adicionalmente, el modelo tiene un valor del $\sqrt{MSE} = 552600$, un valor del R^2 y del R^2_{adj} de 0.8328 y de 0.829 respectivamente, un valor del estadístico F_0 de 216.7 asociado al test ANOVA del modelo, y, finalmente, un valor p de 2.2e-16.

Ahora, si nos basamos en el C_p, también necesitamos $|C_p - p|$, y ambas cantidades deben ser mínimas, con esto, el mejor modelo sería el (26), pues es el que cuenta con un C_p más bajo, y $|C_p - p|$ también es bastante baja (solo superado por el modelo (31), pero es más parsiomonioso que este). Por tanto, basados en el criterio C_p, el modelo elegido es de la forma:

$VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 FLETE_i + \beta_3 ENER_i + \beta_5 PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Y el resumen de este se ve en la tabla 17.

Tabla 17: Tabla de coeficientes-Modelo 26

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1608547.739321	805672.3519180	1.996528	0.0490768
BTA	1.676373	0.1752146	9.567538	0.0000000
FLETE	6.360595	2.7725137	2.294162	0.0242472
ENER	30.622819	12.0463218	2.542089	0.0128326
PERSO	35935.684703	10236.6742912	3.510484	0.0007183

Adicionalmente, se extraen los siguientes valores:

-$\sqrt{MSE} = 5249000$

-$R^2 = 0.8527$

-$R^2_{adj} = 0.8458$

-$F_0 = 123$

-Valor P=2.2e-16

3.10.2 Apartado 10.c. 10.d. y 10.e.

Para estos 3 numerales, se elegirán los modelos mediante métodos de selección automática, (stepwise, forward y backward), en los 3 casos el modelo resultante es el mismo y es el que hemos asociado al número (26) (o sea, el mismo que el modelo que elegimos basados en el C_p), su resumen ya fue presentado previamente.

3.11 Punto 11

Los modelos que seleccionamos para competir serán:

modelo 6: $VENTA_i = \beta_0 + \beta_1 \cdot BTA_i + \beta_5 \cdot PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ ya que es el modelo mas parsimonioso con el mayor R^2_{adj}

modelo 26: $VENTA_i = \beta_0 + \beta_1 \cdot BTA_i + \beta_2 \cdot FLETE_i + \beta_3 \cdot ENER_i + \beta_5 \cdot PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ ya que es el modelo escogido a través de los métodos de selección automática, además de tener el menor puntaje en el estadístico C_p

modelo 17: $VENTA_i = \beta_0 + \beta_1 \cdot BTA_i + \beta_2 \cdot FLETE_i + \beta_5 \cdot PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Ya que es el modelo que cuenta con las covariables individualmente significativas previo a la eliminación de los 10 datos.

Modelo	Normalidad	Varianza cte.	Media 0	$\|R^2_{adj}$	Multicol.
6	No	Sí	Sí	0.829	No
17	No	Sí	Sí	0.836	No
26	No	Sí	Sí	0.8458	No

Tabla 18: Comparativa de Modelos

Dado que en ninguno de los modelos cumplimos el supuesto de normalidad en los errores, y que las demás validaciones de los supuestos son similares (ver tabla 18), se sugerirá usar el modelo 6 ya que cuenta con R^2_{adj} muy similar al modelo 26. De igual manera, el modelo 26 cuenta con un menor puntaje en el estadístico C_p, sin embargo, el modelo 26 cuenta con un mayor número de variables regresoras, y ya que ninguno cumple con el supuesto de normalidad, optamos por el modelo más parsimonioso, que es el 6 ($VENTA_i = \beta_0 + \beta_1 \cdot BTA_i + \beta_5 \cdot PERSO_i + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$)

3.12 Punto 12

Dado que los errores no siguen una distribución normal, que presentamos observaciones atipicas, de balanceo e influenciables, es decir, no cumplimos todos los supuestos necesarios de la teoría, realizar intervalos de confianza y predicción no sería algo recomendado, dado que podemos presentar problemas como la falta de precisión y la falta de sentido entre estos. Sin embargo, por fines academicos e ilustrativos, evidenciaremos los siguientes intervalos para la observación 71, esta seleccionada aleatoriamente.

Tabla 19: Matriz de correlación entre las covariables

	x
fit	2989453
lwr	-6288618
upr	12267524

Como podemos observar, obtenemos un intervalo nada preciso y demasiado amplio, cosa que no tiene mucho sentido si observamos que el intervalo es: (-6288618, 12267524) y sabiendo que, nuestra escala está en miles de pesos, no tiene sentido esta predicción.

Ahora bien, mostraremos el intervalo de confianza para la respuesta media:

Tabla 20: Matriz de correlación entre las covariables

	x
fit	2989453
lwr	1700264
upr	4278642

Obtenemos un intervalo de (1700264, 4278642), donde nuevamente teniendo en cuenta que nuestra escala está en miles de pesos, no tiene sentido tener un intervalo tan amplio y poco preciso, pierde utilidad.

3.13 Extrapolación Oculta

En adición, debido a que nuestro intervalo no es muy concluyente de igual forma, corroboraremos, si la observación numero 71 presenta extrapolación oculta, para ello realizaremos la comparación entre x_0 y h_{max}, que para este caso $x_0 =$ observación 71.

$$x_0^T \cdot (X^T X)^{-1} \cdot x_0 \leq \max_{i=1,2,3} h_{ii}$$

En donde,

$$0.01410447 < 0.2344665$$

Por lo tanto, el vector x_0 no está fuera del rango de experimentación, es decir, no hay extrapolación oculta.

4 Entrega 3

Ahora, nuestro objetivo, tomando la misma base de datos, es estudiar la relación que existe entre la ventas causadas en un año por cada empresa (VENTA) con la producción bruta (BTA), ambas, como ya hemos dicho, medidas en miles de pesos corrientes, esto, según el tipo de empresa que sea (IDOJ), siendo 1 para Sociedad Limitada, 2 representa a una Sociedad Anónima y 3 hace referencia a una Sociedad por Acciones Simplificada.

4.1 Punto 1

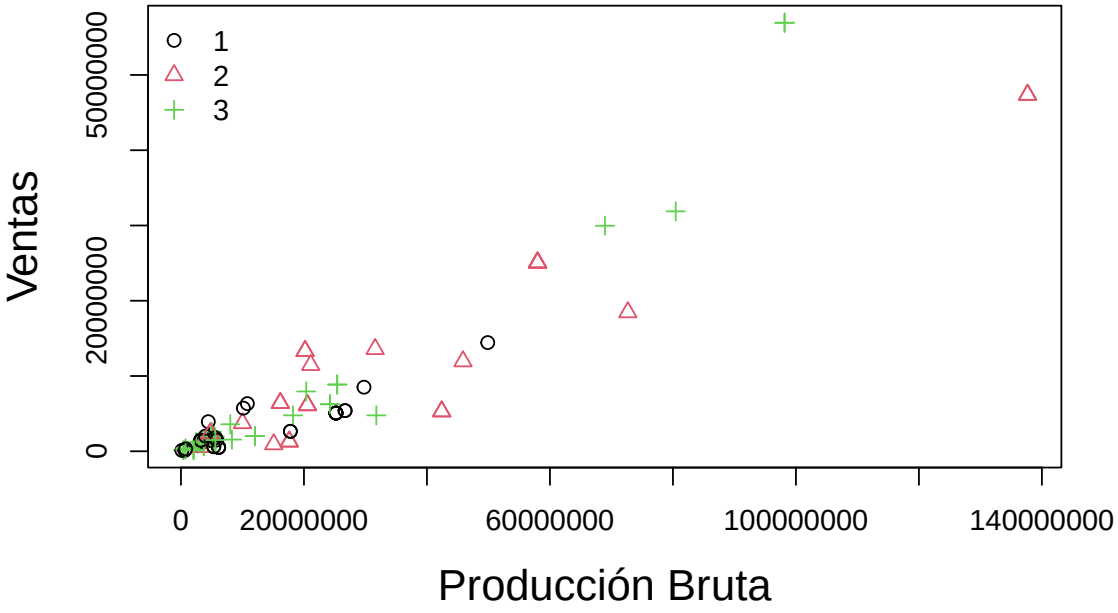


Figura 6: VENTA vs BTA discriminado por IDOJ

Nuestro análisis inicia en la figura 6, de esta, queremos ver si hay alguna diferencia de tendencia entre las distintas categorías de IDOJ, lo primero, y lo que ya hemos visto antes, es que en líneas generales, la relación entre VENTA y BTA es positiva, y, por cada categoría también se cumple esto, sin embargo, sí se observan diferencias, principalmente en los valores últimos, pues, por ejemplo, las categorías 2 y 3 tienen valores mucho más alejados en ambos ejes (más aún si consideramos la escala), también parece que la variabilidad es distinta, y la categoría 2 parece ser más "volatil" en este sentido. En cuanto a la diferencia de pendientes, nos parece que la gráfica no es suficientemente ilustrativa para tomar una postura, pero no parece que existan diferencias muy significativas en este aspecto.

4.2 Punto 2

En principio, nuestro modelo de regresión va a tomar en cuenta tanto a la variable BTA como a la variable cualitativa IDOJ, y, inicialmente supondremos que existe diferencia entre las tendencias o rectas de VENTA vs BTA según la categoría, para este modelo definiremos como nivel de referencia a 3 (S.A.S.), y, sean I_1, I_2 e I_3 tres variables aleatorias indicadores correspondientes a las categorías 1, 2 y 3 respectivamente, con todo esto, el modelo de regresión a plantear es tal que:

$$VENTA_i = \beta_0 + \beta_1 BTA + \beta_2 I_1 + \beta_3 I_2 + \beta_{1,1} BTA \cdot I_1 + \beta_{1,2} BTA \cdot I_2 + \xi,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Y sus correspondientes rectas por nivel

Recta 1: $I_1 = 1,\ I_2 = 0$

$$VENTA_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1}) \cdot BTA_i + \xi_i$$

Recta 2: $I_1 = 0,\ I_2 = 1$

$$VENTA_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2}) \cdot BTA_i + \xi_i$$

Recta 3: $I_1 = 1,\ I_2 = 0$

$$VENTA_i = \beta_0 + \beta_1 BTA_i + \xi_i$$

Y finalmente, la ecuación general ajustada:

$$\hat{VENTA_i} = (4.32e{+}06) + (1.82)BTA + (-2.87e{+}06)I_1 + (4.93e{+}0.5)I_2 + (1.64)BTA \cdot I_1 + (0.73)BTA \cdot I_2$$

4.3 Punto 3

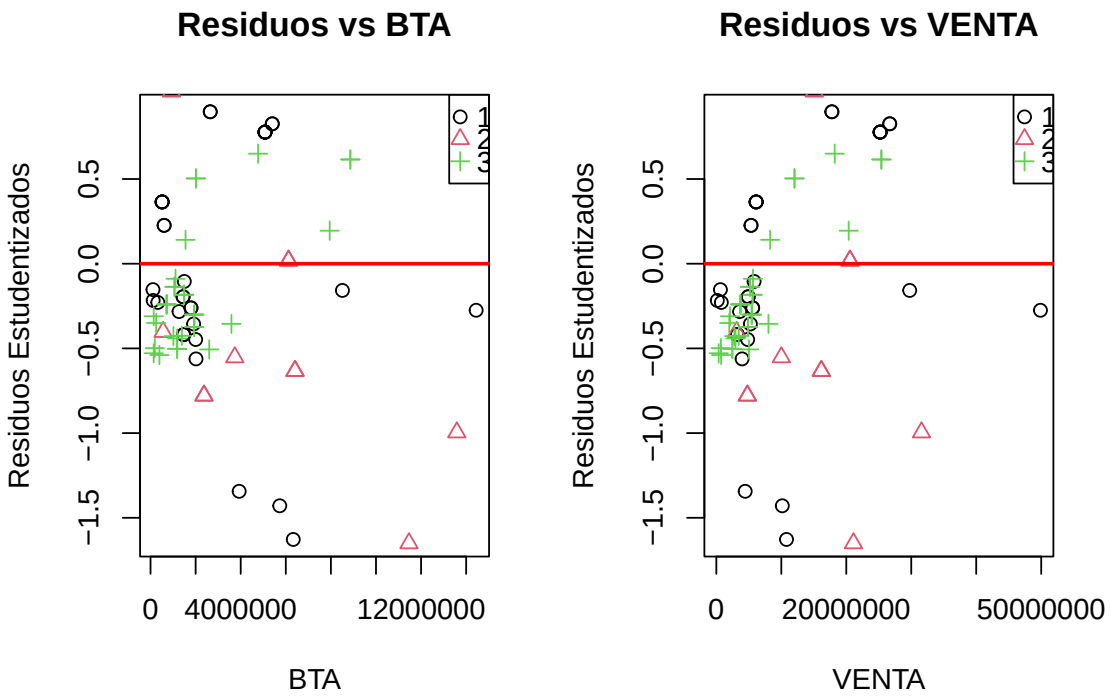


Figura 7: Gráfico de residuos

El gráfico 7 muestra los gráficos de residuales estudentizados, de este sacamos ciertas conclusiones:

-La media global de los puntos parece no ser 0.

-Así mismo, cada categoría parece tener una media inferior a 0.

-No parece que haya presencia de valores atípicos.

-Puede haber problemas de heterocedasticidad, pues las observaciones que corresponden a la categoría 2 parecen mucho más variables que las de la categoría 3, que, al contrario, no parecen tener mucha varianza.

-Hay observaciones que se alejan mucho en ambos ejes, puede ser síntoma de puntos de balanceo y/o influenciables.

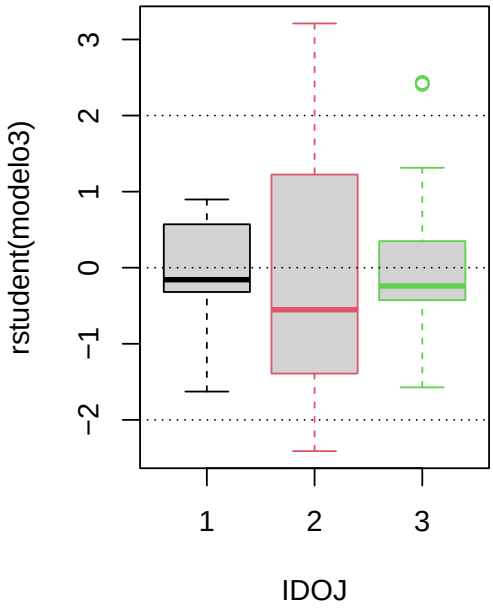


Figura 8: Boxplot

La figura 8 ayuda complementar la información y supuestos que teníamos sobre las medias y las varianzas, pues, como habíamos dicho, las medias no están alineadas en el 0 (están más abajo), mostrando un sesgo, y, la variabilidad en cada categoría es muy distinta, sobre todo, la categoría 2 (S.A.), sus errores difieren bastante.

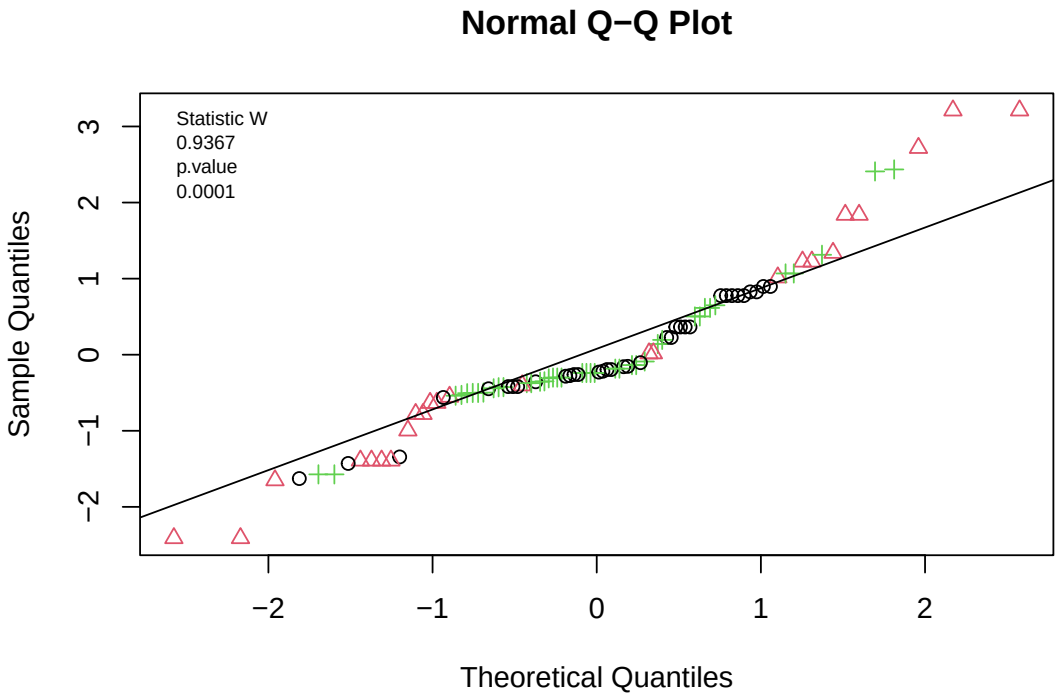


Figura 9: Q-Q plot

De la gráfica 9, simplemente concluímos que nuestros errores no siguen una distribución normal, el estadístico que se muestran en la gráfica y el valor p están asociados a la prueba de hipótesis que sigue:

$$\begin{cases} \mathbf{H_0} : \xi_\mathbf{i} \sim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}) \\ \mathbf{H_1} : \xi_\mathbf{i} \nsim \mathbf{N}(\mathbf{0}, \sigma^\mathbf{2}) \end{cases}$$

De igual manera, la gráfica también apoya esta conclusión, en los extremos de la misma, los valores se alejan bastante de la recta de 45º, y, hasta en el medio de esta, básicamente, nuestro modelo no cumple el supuesto distribucional de normalidad.

4.4 Punto 4

Al momento de analizar si existe diferencia entre las ordenadas en el origen de las rectas correspondientes a los diferentes niveles de la variable *IDOJ* debemos reescribir las ecuaciones de tal forma que podamos analizar su comportamiento individual, esto se evidenció en el numeral 2

Obteniendo de esta manera ecuaciones de regresion simplen notamos que si queremos evidenciar la existencia de diferencia entre las ordenadas en el origen de las rectas basta con descubrir si existe diferencia estadistica entre β_2 y β_3, por lo tanto nos podemos valer del test lineal general, con las siguientes hipotesis:

$$\begin{cases} \mathbf{H_0} : \beta_3 = \beta_4 = 0 \\ \mathbf{H_1} : \beta_3 \neq 0 \ o \ \beta_4 \neq 0 \end{cases}$$

Y usaremos:

$$\begin{cases} \text{Modelo Full será: } VENTA_i = \beta_0 + \beta_1 BTA + \beta_2 I_1 + \beta_3 I_2 + \beta_{1,1} BTA \cdot I_1 + \beta_{1,2} BTA \cdot I_2 + \xi_i, \\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2 \\ \text{Modelo Reducido será: } VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_{1,1} BTA \cdot I_1 + \beta_{1,2} BTA \cdot I_2 + \xi_i, \ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ \text{Estadístico de prueba: } F_0 = \dfrac{[SSE(MR) - SSE(MF)]/\nu}{MSE(MF)} = 0.8771 \\ \text{Distribución del estadístico: El estadístico } F_0 \sim f_{2,94} \\ \text{Valor p: al calcular: } P(f_{2,94} > 0.8771) = 0.4193639 \end{cases}$$

Con un valor p de 0.4193639 no tenemos evidencia suficiente con la cual rechazar nuestra hipotesis nula, por lo tanto no hay diferencia entre las ordenadas en el origen de las rectas de nuestra variable *IDOJ*. En el contexto de nuestro caso, no existe una interpretación para el intercepto, dado que no contamos con valores "cero" en la covariable *BTA*

4.5 Punto 5

De igual manera para evidenciar la existencia de diferencia en las pendientes de las rectas correspondientes a los diferentes niveles de la variable IDOJ y observando las ecuaciones previamente escritas en el punto 2 notamos que si $\beta_{1,1} = \beta_{1,2} = 0$ obtendremos las mismas pendientes por lo tanto usaremos el test lineal general con el siguiente juego de hipótesis:

$$\begin{cases} \mathbf{H_0} : \beta_{1,1} = \beta_{1,2} = 0 \\ \mathbf{H_1} : \beta_{1,1} \neq 0 \ o \ \beta_{1,2} \neq 0 \end{cases}$$

Usando:

$$\begin{cases} \text{Modelo Full será: } VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} BTA_i \cdot I_{i1} + \beta_{1,2} BTA_i \cdot I_{i2} + \xi_i, \\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2 \\ \text{Modelo Reducido será: } VENTA_i = \beta_0 + \beta_1 BTA_i + \beta_2 I_{i1} + \beta_3 I_{i2} + \xi_i,\ \xi_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ \text{Estadístico de prueba: } F_0 = \dfrac{[SSE(MR) - SSE(MF)]/\nu}{MSE(MF)} = 14.835 \\ \text{Distribución del estadístico: El estadístico } F_0 \sim f_{2,94} \\ \text{Valor p: al calcular: } P(f_{2,94} > 14.835) = 0.000002516 \end{cases}$$

Con un valor p de 0.000002516 tenemos evidencia suficiente con la cual rechazar nuestra hipótesis nula, por lo tanto existe diferencia en las pendientes de las rectas correspondientes a los diferentes niveles nuestra variable *IDOJ*. En otras palabras, el cambio promedio en las *VENTAS* por unidad de cambio en BTA es diferente para al menos alguna de los tipos de sociedades mercantiles.

4.6 Punto 6

Teniendo en cuenta que, basadas en las pruebas realizadas anteriormente, tenemos diferentes pendientes Véase punto 5 pero nuestras ordenadas al origen son las mismas Véase punto 4, es decir, no son paralelas pero son coincidentes en su intercepto a nivel general.

Dicho lo anterior, encontramos pertinente comprobar si la recta 1: $VENTA_I = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1}) \cdot BTA_i + \xi_i$ es igual a la recta 2: $VENTA_I = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2}) \cdot BTA_i + \xi_i$

Por lo tanto, el juego de hipótesis a evaluar es el siguiente:

$$\begin{cases} \mathbf{H_0} : \beta_{1,1} - \beta_{1,2} = 0, \beta_1 - \beta_2 = 0 \\ \mathbf{H_1} : \beta_{1,1} - \beta_{1,2} \neq 0 \text{ ó } \beta_1 - \beta_2 = 0 \end{cases}$$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
96	6372130542481805				
94	6139323867879188	2	232797674602617	1.7822	0.1739

Como podemos apreciar, obtenemos un p-value de 0.1739, que al compararlo con un $\alpha = 0.05$, concluimos que hay evidencia suficiente para no rechazar nuestra hipótesis nula, en otras palabras, la recta 1 y la recta 2 son iguales.

4.7 Punto 7

Para probar que la recta *VENTA* vs *BTA* es diferente para cada uno de los niveles de nuestra variable categorica *IDOJ*, planteamos el siguiente juego de hipótesis a probar tomando como referencia *IDOJ3* que en este contexto sería *Sociedad por Simplificada*:

$$\begin{cases} \mathbf{H_0} : \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0 \\ \mathbf{H_1} : \text{Algún } B_j \neq 0 \ ; \ j = 2, 3 \ \text{ ó } \ \text{Algún } \beta_{1,k} \neq 0 \ con \ k = 1, 2 \end{cases}$$

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
98	9010552302930450				
94	6139332867879188	4	2871219435051262	10.99	0.0000002354***

$$F_0 = \frac{\frac{SSE(MR) - SSE(MF)}{(n-2) - (n-k-1)}}{MSE(MF)} \overset{H_0}{\sim} f_{4,94}$$

$$F_0 = \frac{\frac{2871219435051262}{4}}{\frac{6139332867879188}{94}} \overset{H_0}{\sim} f_{4,94}$$

$$F_0 = 10.99 \sim f_{4,94}$$

Entonces, $Pr(> F) = 0.0000002354$

En donde nuestro p-value es 0.0000002354 que, al compararlo con un $\alpha = 0.05$, observamos que $0.0000002354 < 0.05$, es decir, hay evidencia suficiente para rechazar nuestra hipótesis nula, por lo tanto al menos alguna de nuestras 3 rectas es diferente recpecto a cada nivel de *IDOJ.*

DANE. (2012). *Encuesta Anual de Comercio - EAC - 2012*. DANE, Bogotá, CO.

Luque-Calvo, P.L. (2017). *Escribir un Trabajo Fin de Estudios con R Markdown*. Disponible en http://destio.us.es/calvo.

RStudio Team. (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.