2. Modelos lineales generalizados para datos continuos

Consideraremos la base de datos `Preg1B.csv` con información sobre 295 pacientes seleccionados de forma aleatoria. Se desea analizar si existe una asociación entre la presión arterial sistólica (bpsystol) y el índice de masa corporal (bmi), considerando el sexo (sex: 1-hombre, 2-mujer, con hombre como referencia) y la edad (age) de los pacientes.

i) Explorando modelos con variable dependiente contínua.

Para presentar un modelo que parezca adecuado para modelar E(bpsystol; bmi, sex, age), exploramos una malla de los diferentes modelos lineales generalizados comúnmente usados: para el componente aleatorio cuando la variable dependiente es continua exploramos las distribuciones normal, gamma, e inversa gaussiana; empleamos distintas funciones ligas tales como la inversa, identidad, logarítmica, y 1/mu^2(solo para IG); y consideramos el componente lineal tanto de potencias (-3, -2.5, ..., 2.5, 3) como de polinomios (grado 1 al 5). Consideramos por simplicidad que no hay interacción entre las covariables del modelo. En el siguiente Cuadro se muestraa el mejor modelo, con el menor AIC de 2484.009 (que coincide con el mejor modelo por su BIC de 2502.443), con la siguiente estructura:

glm(formula = bpsystol ~ age+sex+I(bmi^(1.5)), family = inverse.gaussian(link = identity), data = datos).

Sin embargo, se elige el modelo más simple o parsimonioso sin el exponente de 1.5 para la variable bmi, pues al considerar bmi sin modificación se obtiene un AIC de 2484.1, el cual no parece ser muy diferente a 2484.009. En el siguiente Cuadro se muestra el modelo final elegido.

glm(formula = bpsystol ~ age+sex+bmi, family = inverse.gaussian(link = identity), data = datos).

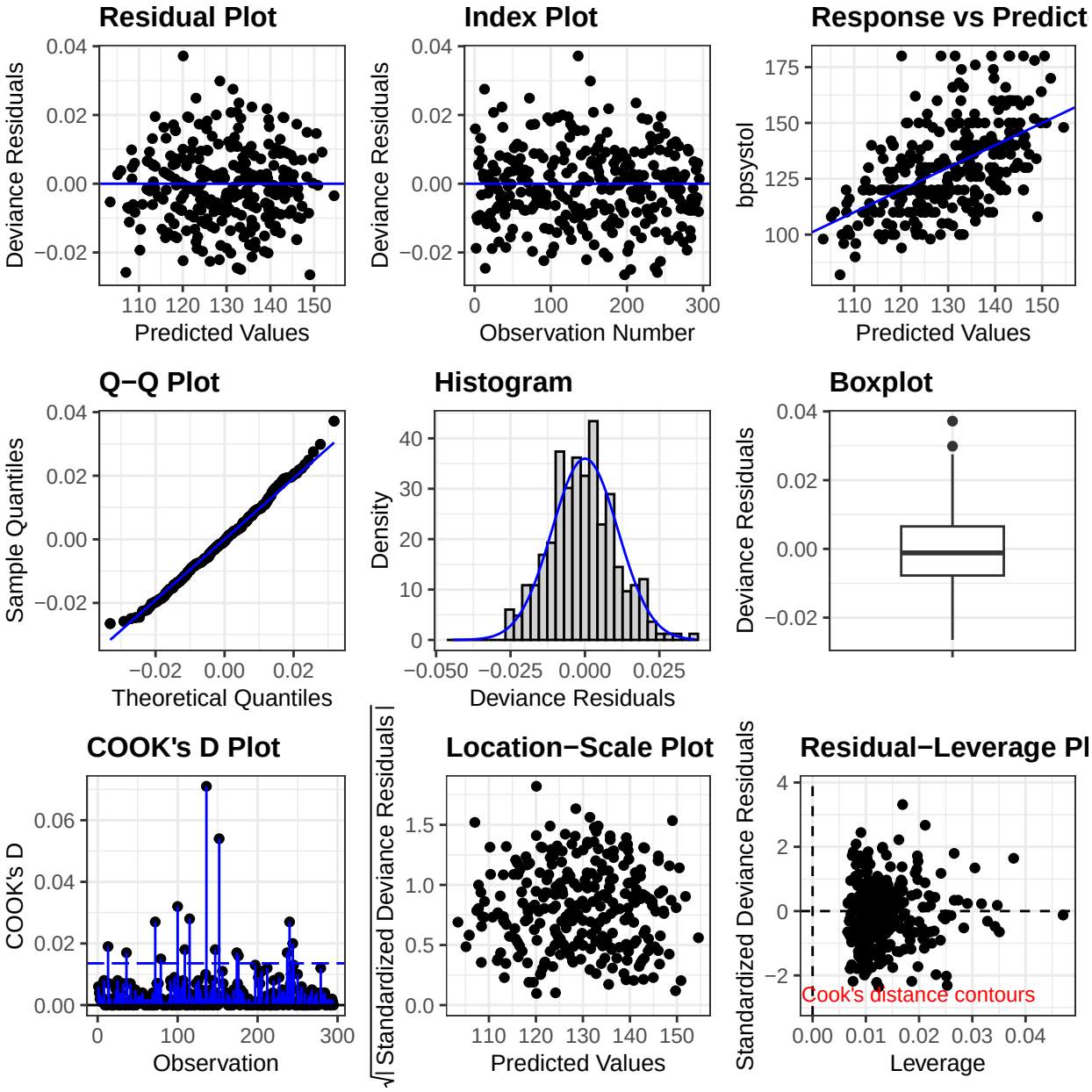
Table 1:

	Dependent variable:	
	bpsystol	
	(1)	(2)
age	0.48671***	0.48269***
	s.e. (0.057)	s.e. (0.057)
	p-value: 1.06e-15	p-value: 1.95e-15
sex2	−7.05833***	−6.88649***
	s.e. (1.908)	s.e. (1.906)
	p-value: 0.000258	p-value: 0.000356
I(bmi^(1.5))	0.15131***	
	s.e. (0.026)	
	p-value: 1.95e-08	
bmi		1.17620***
		s.e. (0.203)
		p-value: 1.68e-08
Constant	90.16891***	80.02163***
	s.e. (3.902)	s.e. (5.254)
	p-value: < 2e-16	p-value: < 2e-16
Observations	295	295
Log Likelihood	−1,238.004	−1,238.068
Akaike Inf. Crit.	2,484.009	2,484.136
Note:		*p<0.1; **p<0.05; ***p<0.01

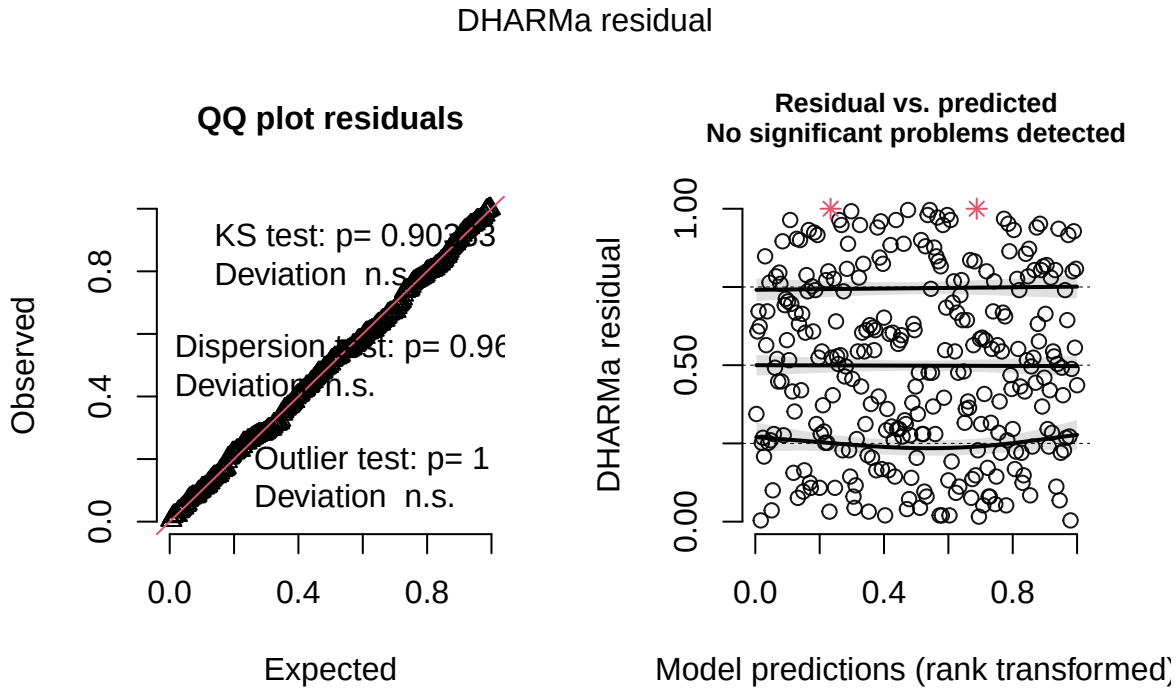
La prueba de hipótesis global con la chi-cuadrada del modelo lineal general **inversa gaussiana con liga identidad** muestra un valor Chisq de 142.2139 y un p-value muy pequeño (Pr(>Chisq): 1.259176e-30), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de los supuestos de este modelo reducido más sencillo.

En la prueba de normalidad `Lilliefors (Kolmogorov-Smirnov) normality test` tenemos que el p-value es de 0.560576574569643, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, pa la prueba de normalidad de `Shapiro-Wilk normality test` el p-value es de 0.44368350946798, lo que también no rechaza la hipótesis nula de normalidad.

En las siguientes gráficas podemos observar en **Residual Plot** que se conserva la linealidad y varianza constante. En **Q-Q Plot** y **Histogram** se observa un buen comportamiento de la normalidad de los errores. En **Index Plot** no hay patrones relacionados con la forma en que se ordenaron los datos, lo que puede proporcionar información sobre tendencias adicionales en los datos que no se han tenido en cuenta en el modelo, no hay una tendencia obvia en el gráfico. En **Location-Scale Plot** se observa que hay homoscedasticidad. En el **Boxplot** se puden observar algunos aoutliers, sin embargo en **COOK'S D Plot** y en **Residuals-Leverage Plot** parece no haber outliers influyentes.



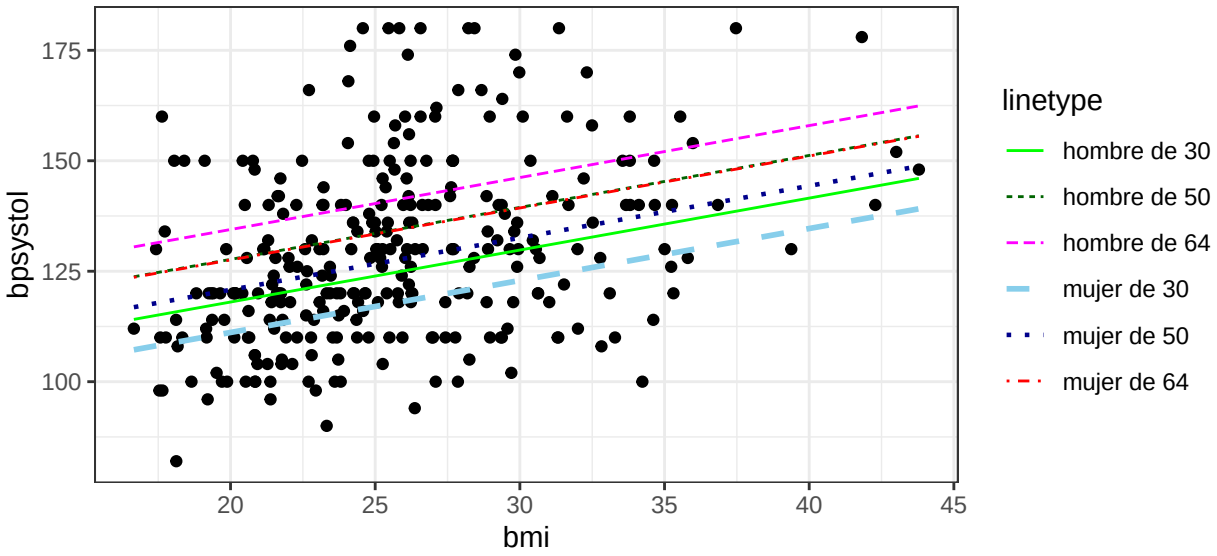
Además en las siguientes gráficas se comprueba las observaciones de las gráficas anteriores.



ii) Asociación entre masa corporal y presión arterial sistólica, y estimación puntual.

En esta sección describiremos la asociación entre masa corporal y presión arterial sistólica y la prueba de hipótesis de esta relación. Dado que lo que buscamos responder es si tener un indice de masa corporal alto se relaciona con tener una presion sistolica alta agregaremos una prueba de hipótesis con dirección, donde la hipótesis nula es $H_0 : \beta_4 < 0$ contra la alternativa $H_0 : \beta_4 > 0$. El p-value asociado a la prueba es de Pr(>z) = 3.22e-09, por lo tanto, rechazamos la hipótesis nula y por lo tanto hay relación asociación entre la masa corporal alta y la presión arterial sistólica alta para una persona de cierta edad y sexo.

Por otra presentaremos una gráfica resumen con la estimación puntual de la relación bpsystol y bmi, considerando edades de 30, 50 y 64, así como la diferenciación entre mujeres y hombres.



iii) Comparativo modelo de regresión lineal múltiple contra modelo lineal generalizado.

En esta sección compararemos el modelo de regresión lineal múltiple del ejercicio anterior (ejercicio 1) contra el modelo lineal generalizado con base en sus AIC (ejercicio 2). Además, compararemos las conclusiones e interpretaciones de ambos modelos, para indicar cuál nos parece más adecuado y fácil de interpretar. El AIC del primer modelo de regresión lineal es de -376.2099, el cual no es directamente comparable al AIC del modelo lineal general de $2,484.136$, pero haciendo el cambio de escala podemos notar que el modelo de menor AIC es el del modelo OLS. Sin embargo, se observa que el modelo más sencillo de interpretar sus coeficientes de manera directa es el mlg, sin tener que hacer transformaciones adicionales, por lo que nos parece adecuado elegir este como el mejor modelo.

Tomando en cuenta el modelo elegido y habiendo mostrado el cumplimiento de los supuestos del modelo, además de una prueba de hipótesis global satisfactoria, podemos concluir que la relación entre bmi es directa (positiva) con bpsystol, el incremento en una unidad en mbi, incrementa el bpsystol en 1.76 unidades. Por otra parte, el ser mujer, con respecto a ser hombre, tiene una relación inversa (negativa) con bpsystol, es decir, ser mujer disminuye en 6.886 unidades la bpsystol. Con respecto a la edad, un incremento en una unidad de edad, incrementa bpsystol en 0.483 unidades.

Table 2:

	Dependent variable:	
	I(log(bpsystol))	bpsystol
	OLS	*glm: inverse.gaussian link = identity*
	(1)	(2)
bmi	0.009***	1.176***
	(0.002)	(0.203)
sex2	−0.049***	−6.886***
	(0.015)	(1.906)
age	0.004***	0.483***
	(0.0004)	(0.057)
Constant	4.461***	80.022***
	(0.042)	(5.254)
Observations	295	295
R^2	0.321	
Adjusted R^2	0.314	
Log Likelihood		−1,238.068
Akaike Inf. Crit.		2,484.136
Residual Std. Error	0.127 (df = 291)	
F Statistic	45.922*** (df = 3; 291)	
Note:		*p<0.1; **p<0.05; ***p<0.01