Ejercicio 6

Con datos de `Ex6.csv` se considera un modelo de regresión lineal con las covariables X_1 a X_6 sin interacción. Se muestra el resultado en la primera columna del Cuadro de `MODELOS`, en donde se observa la prueba global F asociado a la tabla ANOVA cuyo p-value es menor a 0,05, por lo que se rechaza la hipótesis nula $H_0 : \hat{\beta}_i = 0, \forall i = 1, 2, ..., p$, a favor de la alternativa de que al menos un coeficiente estimado $\hat{\beta}_i$ es distinto de cero. Se observa además que dado que están las otras variables, la variable X3 no agrega información adicional al modelado, lo mismo para el caso de X5. (Chunk modelo1, línea de código 34).

Si hacemos una prueba visual de los supuestos del modelo, tales como la linealidad (Residuals vs Fitted), homocedasticidad (Scale-Location), normalidad (Q-Q Residuals) y presencia de outliers influyentes (Residuals vs Leverage), se observa que no se cumple la linealidad y normalidad, aunque parece no haber problemas con la homocedasticidad y la presencia de outliers influyentes (que se salgan de la distancia de Cook). (Chunk plotsmodelo1, línea de código 47)

De acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.1576004 por lo que no se rechaza la hipótesis nula de homocesaticidad, mientras que las pruebas de normalidad Jarque-Bera, Shapiro-Wilk y Kolmogorov-Smirnov rechazan la hipótesis nula de normalidad, con p-value de 0, $5{,}3610845 \times 10^{-7}$ y 0.0284986, respectivamente. (Chunk pruebasmodelo1, línea de código 56).

Para el caso de la linealidad, la prueba de Tukey rechaza la hipótesis nula de linealidad, particularmente con X4 y X6, el p-value asociado es menor a 0.05. (Chunk residualplotsmodelo1, línea de código 82).

```
##           X3           X4           X5           X6           X1           X2
## 7.775409e-01 1.227254e-03 9.239395e-02 6.168694e-06           NA           NA
##   Tukey test
## 1.185248e-29
```

Esto se refleja en las gráficas individuales, para X4 y X6 es evidente la no linealidad. (Chunk residualPlots-modelo1, línea de código 96).

En las gráficas crPlots, se muestra que es posible ajustar tal vez un polinomio para X4 y X6. (Chunk crPlotsmodelo1, linea de código 109)

Haciendo la prueba `powerTransform` se rechaza la hipótesis nula de $\lambda = 1$ por lo que hay que hacer una transformación a la variable dependiente Y, en la misma prueba se rechaza la hipótesis nula de la transformación logarítimica $\lambda = 0$, y se presenta un valor sugerido de 2 como exponente. (Chunk powerTransform1, linea de código 119).

Luego, al hacer la prueba de `boxTidwell` para ver si se requiere transformar X4 y X6, tenemos que para el caso de X6 no se rechaza la hipótesis nula de una $\lambda = 1{,}032$, por lo que no requiere transformación, sin embargo, para el caso de X4, se muestra una $\lambda = 0{,}38193$ que redondearemos a $1/2$, por lo que tomaremos la raíz cuadrada de la variable X4 en el modelado. (Chunk boxTidwell1, linea de codigo 125).

Entonces, planteamos un segundo modelo que considera Y^2 en función de las covariables $X1$ a $X6$ considerando $\sqrt{X4}$ y los demás covariables sin cambios. Los resultados se muestran en la segunda columna del Cuadro de `MODELOS`, en donde se observa la prueba global F asociado a la tabla ANOVA cuyo p-value es menor a 0,05, por lo que se rechaza la hipótesis nula $H_0 : \hat{\beta}_i = 0, \forall i = 1, 2, ..., p$, a favor de la alternativa de que al menos un coeficiente estimado $\hat{\beta}_i$ es distinto de cero. Se observa además que dado que están las otras variables, la variable X3 no agrega información adicional al modelado, lo mismo para el caso de X5. (Chunk modelo2, linea de código 137).

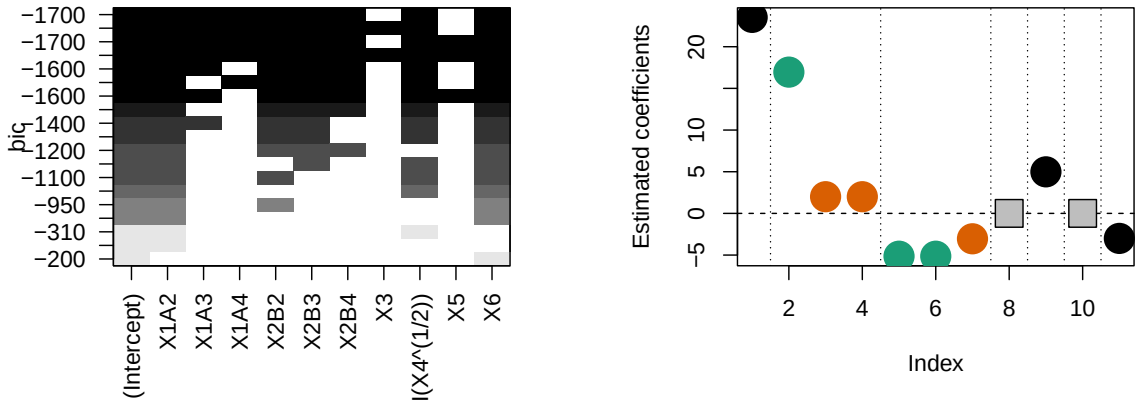
Con este segundo modelo se cumple linealidad con la prueba Tukey, al igual que para todas las variables individuales. (Chunk residualplotsmodelo2, linea de código 144).

```
##          X1          X2          X3 I(X4^(1/2))          X5          X6
##          NA          NA   0.4045252   0.6665737   0.1516197   0.2226187
##  Tukey test
##   0.6844529
```

Además, de acuerdo con la prueba `studentized Breusch-Pagan` se tiene un p-value de 0.9656083 por lo que no se rechaza la hipótesis nula de homocesaticidad, mientras que las pruebas de normalidad Jarque-Bera, Shapiro-Wilk y Kolmogorov-Smirnov no rechazan la hipótesis nula de normalidad, con p-value de 0.5337811, 0.3336509 y 0.109361, respectivamente. (Chunk pruebasmodelo2, línea de código 158).

A partir de este segundo modelo, haremos una selección de variables.

Con la función `regsubsets` de la biblioteca `leaps` se analizaron los mejores subconjuntos, se presenta la siguiente gráfica izquierda mostrando los resultados, con esto decidimos tomar las covariables X1, X2, X4 y X6, omitiendo las variables X3 y X5, con lo que se podría tener un tercer modelo que muestra un BIC de 1236.7351928 y que se muestra en la columna 3 del Cuadro de MODELOS. (Chunk subconjuntosleaps, linea de código 187). Este resultado se refuerza con el análisis hecho con los métodos por pasos forward y backward, con la función `step` y el criterio BIC. (Chunk stepBIC, linea de código 207). Adicionalmente se realizó con la biblioteca `smurf` un modelo lasso de seleccion via penalizacion en la logverosimilitud, considerando la familia gausiana, pesos glm.stand, y lambda is.bic, cuyo BIC es de 1226.1559243 y resultado se muestra se presenta en la gráfica de la derecha (Chunk glm_lambdabic, linea de código 221). Como se puede observar en la gráfica derecha, las variables X3 y X5 tienen un valor de cero, además podemos notar que X1A3 y X1A4 pueden combinarse en una sola variable, al igual que X2B2 y X2B3 en otra, así se realizó el ajuste con el modelo lasso final cuyos resultados se presentan en la cuarta y última columna del Cuadro de MODELOS y cuyo BIC es de 1226.0633129. (Chunk modsellasso, linea de código 238).



Finalmente, se presentan los resultados de los modelos referidos anteriormente en el Cuadro de MODELOS. El modelo final elegido es el de la última columna donde la variable dependiente es $V1 = I((Y)\hat{}2)$, podemos observar que la prueba F tiene un p-value menor a 0,05, los tres asteriscos indican que incluso menor que 0,01, por lo que al menos un coeficiente estimado es distinto de cero. Además, cada uno de los coeficientes, en una análisis individual, dado que están las otras variables, agregan información al modelo.

Cuadro 1: MODELOS

	Dependent variable:			
	Y	I((Y)^2)		V1
	(1)	(2)	(3)	(4)
X3	0.002	0.036		
	(0.003)	(0.028)		
X4	0.066***			
	(0.003)			
I(X4^(1/2))		5.014***	5.018***	5.006***
		(0.167)	(0.167)	(0.167)
X5	−0.002	−0.016		
	(0.002)	(0.018)		
X6	−0.238***	−2.999***	−3.005***	−3.002***
	(0.003)	(0.028)	(0.028)	(0.028)
X1A2	1.295***	16.979***	16.967***	16.969***
	(0.014)	(0.152)	(0.151)	(0.151)
X1A3	0.177***	2.032***	2.030***	
	(0.014)	(0.151)	(0.151)	
X1A4	0.177***	2.019***	2.014***	
	(0.014)	(0.152)	(0.151)	
X2B2	−0.410***	−5.078***	−5.073***	
	(0.014)	(0.151)	(0.151)	
X2B3	−0.410***	−5.248***	−5.244***	
	(0.014)	(0.151)	(0.151)	
X1A3.4				2.023***
				(0.131)
X2B2.3				−5.159***
				(0.131)
X2B4	−0.235***	−3.035***	−3.047***	−3.048***
	(0.014)	(0.152)	(0.152)	(0.151)
Constant	5.550***	23.262***	23.410***	23.448***
	(0.030)	(0.548)	(0.541)	(0.540)
Observations	400	400	400	400
R^2	0.982	0.988	0.988	0.999
Adjusted R^2	0.981	0.988	0.988	0.999
Residual Std. Error	0.102 (df = 389)	1.065 (df = 389)	1.066 (df = 391)	1.065 (df = 393)
F Statistic	2,101.510*** (df = 10; 389)	3,204.967*** (df = 10; 389)	3,999.207*** (df = 8; 391)	89,411.300*** (df = 7; 393)

Note: *p<0.1; **p<0.05; ***p<0.01