



Universidad Simón Bolívar  
Dpto. de Cómputo Científico y Estadística  
CO-3321 Estadística para Ingeniería  
Intensivo Julio-Agosto 2016

# Laboratorio 6: Regresión Lineal

Estudiantes:

Alessandra Marrero, 12-11091

Verónica Mazutiel, 13-10853

Profesor: Pedro Ovalles.

Sartenejas, 19 de agosto de 2016

## Laboratorio 6: Regresión Lineal

1. Realice un primer modelo con las variables cuyo  $|\rho| > 0,5$  con respecto a la variable respuesta. Para esto calcule la matriz de correlación y estudie las gráficas de las variables.

> (datos = read.delim("Datos.txt"))

	Country	HDI	ExpVida	Escol	PromEscol	GNI
1	Argentina	0.836	76.3	17.9	9.8	22050
2	Barbados	0.785	75.6	15.4	10.5	12488
3	Belize	0.715	70.0	13.6	10.5	7614
4	Bolivia (Plurinational State of)	0.662	68.3	13.2	8.2	5760
5	Brazil	0.755	74.5	15.2	7.7	15175
6	Canada	0.913	82.0	15.9	13.0	42155
7	Chile	0.832	81.7	15.2	9.8	21290
8	Colombia	0.720	74.0	13.5	7.3	12040
9	Costa Rica	0.766	79.4	13.9	8.4	13413
10	Cuba	0.769	79.4	13.8	11.5	7301
11	Dominican Republic	0.715	73.5	13.1	7.6	11883
12	Ecuador	0.732	75.9	14.2	7.6	10605
13	El Salvador	0.666	73.0	12.3	6.5	7349
14	Guatemala	0.627	71.8	10.7	5.6	6929
15	Guyana	0.636	66.4	10.3	8.5	6522
16	Haiti	0.483	62.8	8.7	4.9	1669
17	Honduras	0.606	73.1	11.1	5.5	3938
18	Jamaica	0.719	75.7	12.4	9.7	7415
19	Mexico	0.756	76.8	13.1	8.5	16056
20	Nicaragua	0.631	74.9	11.5	6.0	4457
21	Panama	0.780	77.6	13.3	9.3	18192
22	Paraguay	0.679	72.9	11.9	7.7	7643
23	Peru	0.734	74.6	13.1	9.0	11015
24	Trinidad and Tobago	0.772	70.4	12.3	10.9	26090
25	United States	0.915	79.1	16.5	12.9	52947
26	Uruguay	0.793	77.2	15.5	8.5	19283
27	Venezuela (Bolivarian Republic of)	0.762	74.2	14.2	8.9	16159

	GNI	HDI	PCA
1	11	0.71	
2	27	0.39	
3	9	0.43	
4	4	0.88	
5	-1	0.91	
6	11	0.31	
7	11	0.73	
8	-9	0.79	
9	10	0.67	
10	47	0.54	
11	-12	0.76	
12	7	0.53	
13	-3	1.02	

```

14 -11 1.10
15 -4 0.66
16 4 0.61
17 7 0.75
18 13 0.28
19 -4 0.65
20 12 1.02
21 1 0.72
22 -3 0.67
23 8 0.75
24 -25 0.57
25 3 0.26
26 7 0.57
27 -2 0.76

```

```
> attach(datos)
```

```
> names(datos)
```

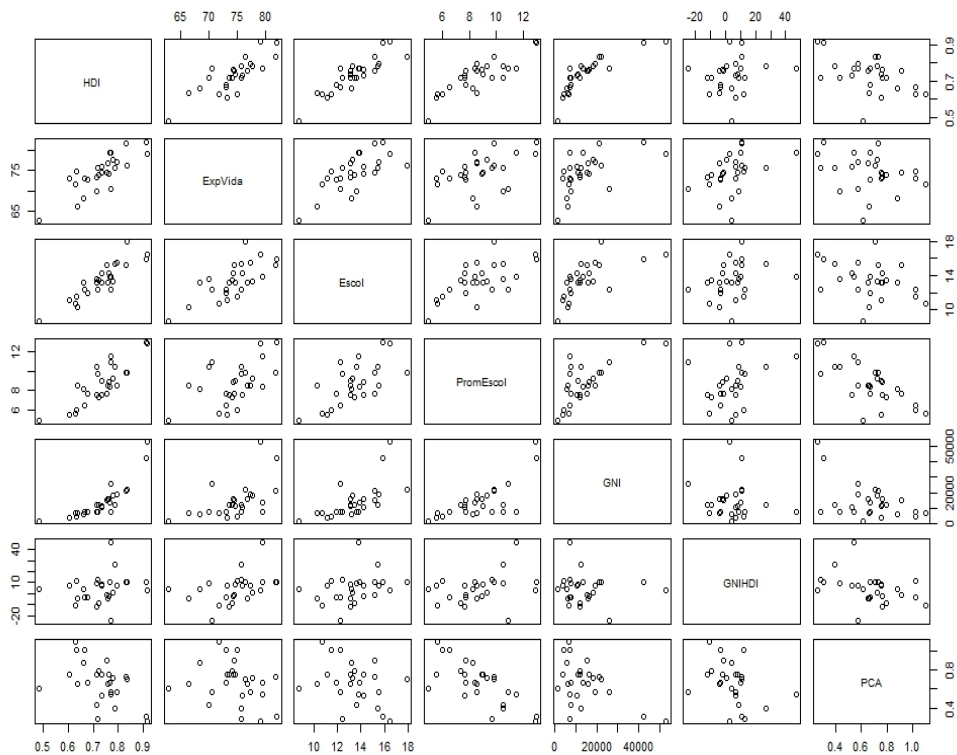
```
[1] "Country" "HDI" "ExpVida" "Escol" "PromEscol" "GNI"
```

```
[7] "GNIHDI" "PCA"
```

```
> #Diagrama de dispersión
```

```
> pairs(datos[2:8])
```

```
>
```



```
> #Correlación
```

```
> c=cor(datos[2:8])
```

> c

	HDI	ExpVida	Escol	PromEscol	GNI
HDI	1.0000000	0.8056660	0.8925582	0.8451962	0.83435919
ExpVida	0.8056660	1.0000000	0.7096762	0.5339212	0.55732821
Escol	0.8925582	0.7096762	1.0000000	0.6642338	0.66373973
PromEscol	0.8451962	0.5339212	0.6642338	1.0000000	0.72973858
GNI	0.8343592	0.5573282	0.6637397	0.7297386	1.00000000
GNIHDI	0.1925694	0.4096271	0.3010869	0.3236630	-0.08917148
PCA	-0.4882619	-0.2703955	-0.3514418	-0.7236550	-0.49027647
	GNIHDI	PCA			
HDI	0.19256939	-0.4882619			
ExpVida	0.40962713	-0.2703955			
Escol	0.30108687	-0.3514418			
PromEscol	0.32366305	-0.7236550			
GNI	-0.08917148	-0.4902765			
GNIHDI	1.00000000	-0.3492623			
PCA	-0.34926227	1.0000000			

Si estudiamos las gráficas de las variables, podemos decir que aunque para la gráfica (HDI,ExpVida) no es muy clara la línea que sigue se puede notar una recta con pendiente positiva. De igual forma para la gráfica (HDI,Escol) los datos se agrupan formando lo que parece una recta ascendente. Lo mismo se repite para las gráficas (HDI,PromEscol) y (HDI,GNI) que aunque existen datos dispersos se puede observar una tendencia a formar una recta con pendiente positiva. Por otro lado, para las gráfica (HDI,GNIHDI) no se observa una forma conocida sino más “ruido” y para la de (HDI,PCA) aunque también lo que se ve es “ruido” se pude ver una tendencia más cercana a una recta con pendiente negativa.

Todo esto tiene sentido en relación a las correlaciones obtenidas, pues las de (HDI,ExpVida), (HDI,Escol), (HDI,PromEscol) y (HDI,GNI) son positivas y bastantes cercanas a 1. Mientras que la correlación entre HDI y GNIHDI es menor a 0.5 y la de (HDI, PCA) es negativa y mayor a -0.5, lo cual tiene concordancia con las observaciones antes mencionadas.

De manera que realizaremos nuestro modelo con las variables ExpVida,Escol,PromEscol y GNI.

> #Modelo con variables con  $|\rho| > 0,5$

> M=lm(HDI~ExpVida+Escol+PromEscol+GNI)

> M

Call:

lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNI)

Coefficients:

(Intercept)	ExpVida	Escol	PromEscol	GNI
-4.570e-02	5.663e-03	1.585e-02	1.334e-02	1.920e-06

> summary(M)

Call:

lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNI)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.033399	-0.007973	0.002521	0.010324	0.028578

Coefficients:

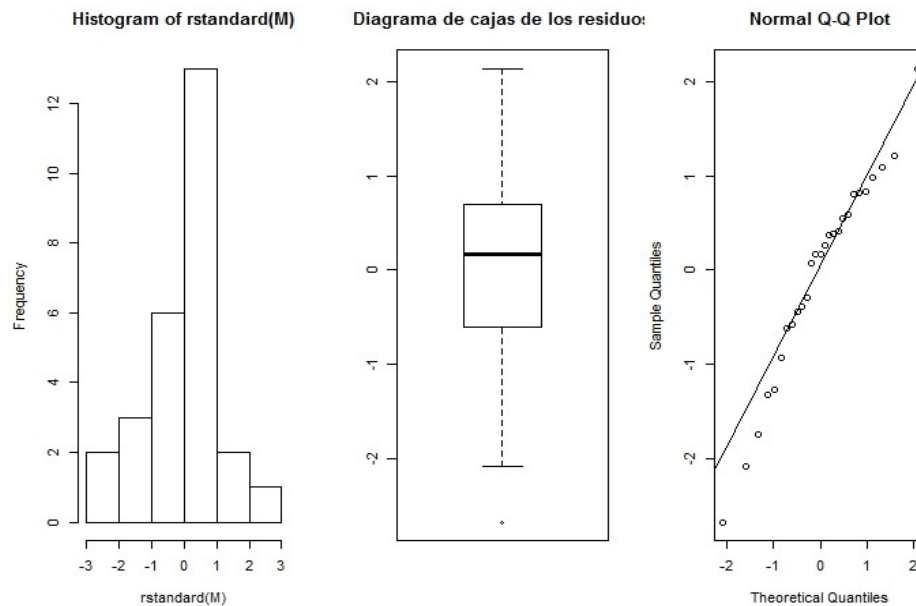
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-4.570e-02	6.184e-02	-0.739	0.467721
ExpVida	5.663e-03	1.013e-03	5.592	1.27e-05 ***
Escol	1.585e-02	2.537e-03	6.250	2.73e-06 ***
PromEscol	1.334e-02	2.321e-03	5.746	8.84e-06 ***
GNI	1.920e-06	4.252e-07	4.516	0.000171 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01574 on 22 degrees of freedom  
Multiple R-squared: 0.9756, Adjusted R-squared: 0.9712  
F-statistic: 219.9 on 4 and 22 DF, p-value: < 2.2e-16

```
> ##### Análisis de los residuos #####
> #Modelo con variables con |p| > 0,5
> #Para chequear la normalidad
> par(mfrow=c(1,3))
> hist(rstandard(M))
> boxplot(rstandard(M),main="Diagrama de cajas de los residuos")
> qqnorm(rstandard(M))
> qqline(rstandard(M))
```



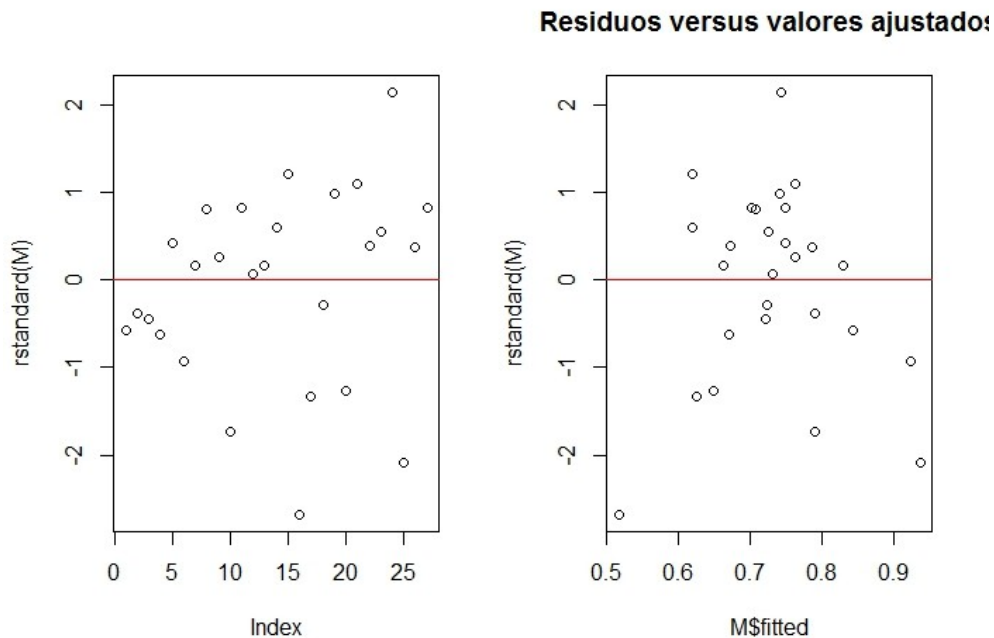
-Vemos que los residuos se distribuyen Normal.

```
> # Para chequear la homocedasticidad
> par(mfrow=c(1,2))
```

```

> plot(rstandard(M))
> abline(h=0,col=2)
> plot(M$fitted,rstandard(M))
> title("Residuos versus valores ajustados")
> abline(h=0,col=2)

```

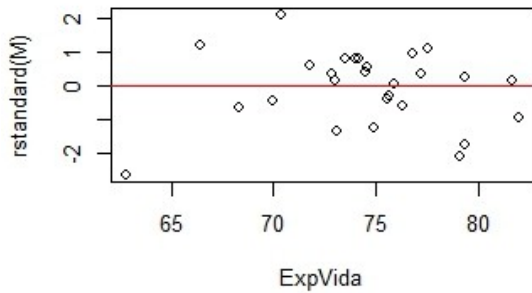


```

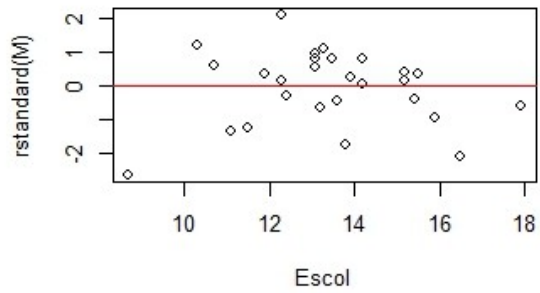
> #Para chequear la independencia
> par(mfrow=c(2,2))
> plot(ExpVida,rstandard(M),main="Residuos estandarizados versus var ind.")
> abline(h=0,col=2)
> plot(Escol,rstandard(M),main="Residuos estandarizados versus var ind.")
> abline(h=0,col=2)
> plot(PromEscol,rstandard(M),main="Residuos estandarizados versus var ind.")
> abline(h=0,col=2)
> plot(GNIHDI,rstandard(M),main="Residuos estandarizados versus var ind.")
> abline(h=0,col=2)

```

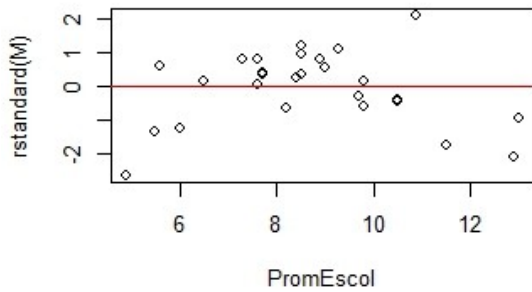
**Residuos estandarizados versus var ind.**



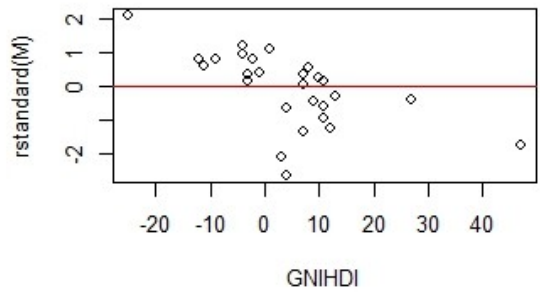
**Residuos estandarizados versus var ind.**



**Residuos estandarizados versus var ind.**



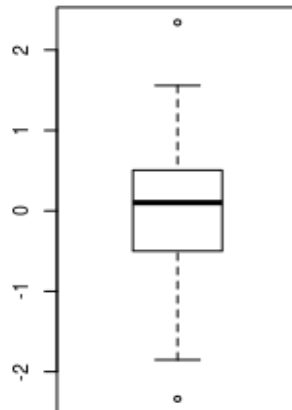
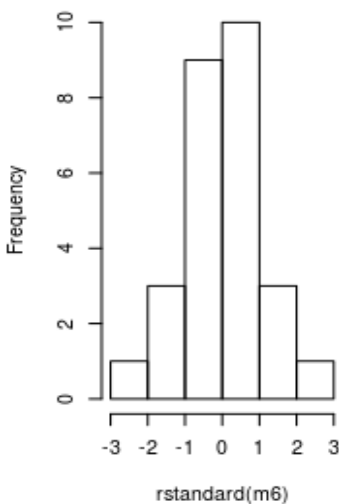
**Residuos estandarizados versus var ind.**



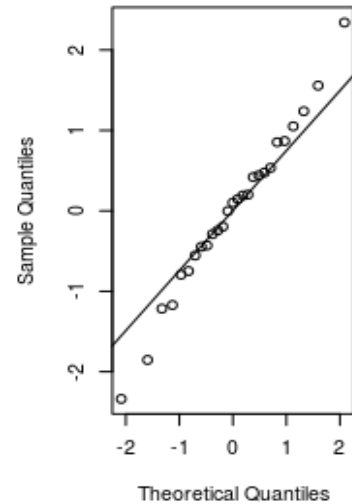
Se verifica que se cumplen todas las condiciones para poder usar este modelo.

- > #Modelo regresion paso a paso
- > #Para chequear la normalidad
- > par(mfrow=c(1,3))
- > hist(rstandard(m6))
- > boxplot(rstandard(m6),main="Diagrama de cajas de los residuos")
- > qqnorm(rstandard(m6))
- > qqline(rstandard(m6))

**Histogram of rstandard(m6) Diagrama de cajas de los residu**

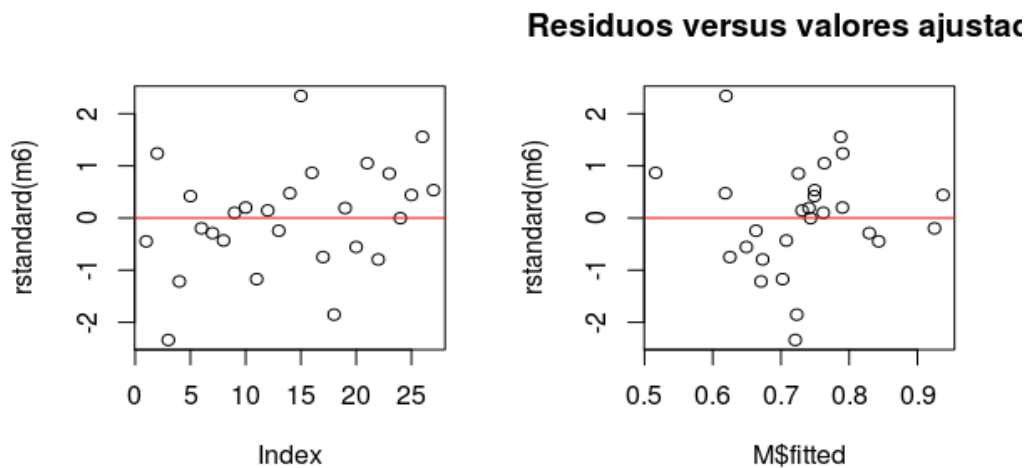


**Normal Q-Q Plot**



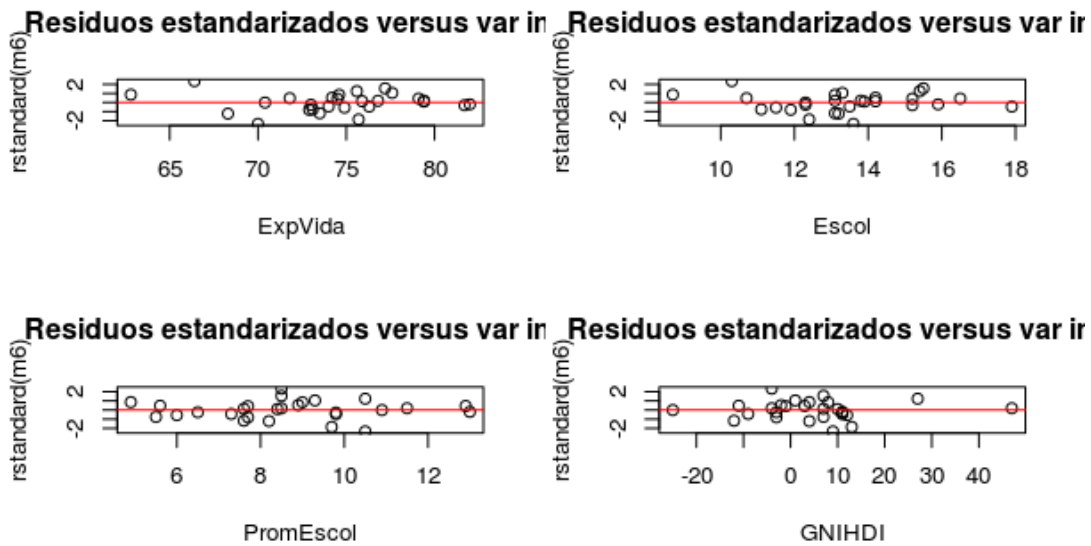
-Vemos que los residuos se distribuyen Normal.

```
> # Para chequear la homocedasticidad  
> par(mfrow=c(1,2))  
> plot(rstandard(m6))  
> abline(h=0,col=2)  
> plot(M$fitted,rstandard(m6))  
> title("Residuos versus valores ajustados")  
> abline(h=0,col=2)
```



```
> #Para chequear la independencia  
> par(mfrow=c(2,2))  
> plot(ExpVida,rstandard(m6),main="Residuos estandarizados versus var ind.")  
> abline(h=0,col=2)  
> plot(Escol,rstandard(m6),main="Residuos estandarizados versus var ind.")  
> abline(h=0,col=2)  
> plot(PromEscol,rstandard(m6),main="Residuos estandarizados versus var ind.")  
> abline(h=0,col=2)  
> plot(GNIHDI,rstandard(m6),main="Residuos estandarizados versus var ind.")  
> abline(h=0,col=2)
```





Se verifica que se cumplen todas las condiciones para poder usar este modelo.

- Realice un segundo modelo con regresión paso a paso. En cada paso elimine la variable menos significativa y continúe hasta que todas las restantes tengan un nivel de significancia menor a 0.05.

> `m5=lm(HDI~ ExpVida+ Escol+ PromEscol + GNI+GNIHDI+PCA) )`

Call:

`lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNI + GNIHDI + PCA)`

Coefficients:

	ExpVida	Escol	PromEscol	GNI
(Intercept)	7.656e-03	1.708e-02	1.979e-02	4.118e-07
GNIHDI	-1.412e-03			
PCA	7.129e-03			

> `summary(m5)`

Call:

`lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNI + GNIHDI + PCA)`

Residuals:

Min	1Q	Median	3Q	Max
-0.0109479	-0.0038418	0.0006767	0.0029871	0.0140140

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-2.435e-01	3.330e-02	-7.313	4.54e-07 ***
ExpVida	7.656e-03	4.966e-04	15.417	1.45e-12 ***
Escol	1.708e-02	1.131e-03	15.097	2.14e-12 ***
PromEscol	1.979e-02	1.350e-03	14.656	3.69e-12 ***
GNI	4.118e-07	2.486e-07	1.656	0.113
GNIHDI	-1.412e-03	1.558e-04	-9.062	1.61e-08 ***
PCA	7.129e-03	9.760e-03	0.730	0.474

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006786 on 20 degrees of freedom  
Multiple R-squared: 0.9959, Adjusted R-squared: 0.9946  
F-statistic: 805 on 6 and 20 DF, p-value: < 2.2e-16

-Dado que las variables GNI y PCA tienen un valor de significancia muy alto probamos con un nuevo modelo que no toma esas variables

> #Modelo 6:

> (m6=lm(HDI ~ ExpVida+ Escol+ PromEscol+GNIHDI))

Call:

lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNIHDI)

Coefficients:

(Intercept)	ExpVida	Escol	PromEscol	GNIHDI
-0.272420	0.008005	0.017706	0.020481	-0.001596

> summary(m6)

Call:

lm(formula = HDI ~ ExpVida + Escol + PromEscol + GNIHDI)

Residuals:

Min	1Q	Median	3Q	Max
-0.0143878	-0.0031472	0.0006626	0.0032920	0.0140772

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.2724201	0.0271659	-10.03	1.15e-09 ***
ExpVida	0.0080046	0.0004598	17.41	2.37e-14 ***
Escol	0.0177056	0.0010849	16.32	8.90e-14 ***
PromEscol	0.0204809	0.0008884	23.05	< 2e-16 ***
GNIHDI	-0.0015955	0.0001135	-14.05	1.81e-12 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006917 on 22 degrees of freedom  
Multiple R-squared: 0.9953, Adjusted R-squared: 0.9944  
F-statistic: 1161 on 4 and 22 DF, p-value: < 2.2e-16

-Escogemos el modelo m6 pues todas variables son significativas con un nivel de significancia menor a 0.05 y el  $R^2$  ajustado de m6 es mayor que el  $R^2$  ajustado de m5.

### 3. Compare los modelos

Observando los summary de ambos modelos podemos compararlos. Se puede notar que todas las variables del modelo m6 son significativas y las del modelo M no, además el  $R^2$  ajustado de m6 es mayor que el  $R^2$  ajustado de M. Luego m6 es un mejor modelo que M.

### 4. Use los datos que aparecen en la misma hoja de datos, subrayados en azul, para hacer una predicción con ambos modelos.

```
> #HDI ~ ExpVida + Escol + PromEscol + GNI
> new=data.frame(ExpVida=80,Escol=15,PromEscol=13,GNI=13000)
> new
  ExpVida Escol PromEscol  GNI
1    80    15      13 13000
> (Temp1=predict(M,new,interval="prediction"))# Intervalo de prediccion
      fit      lwr      upr
1 0.8434604 0.8039724 0.8829483

> #HDI ~ ExpVida + Escol + PromEscol + GNIHDI
> new2=data.frame(ExpVida=80,Escol=15,PromEscol=13,GNIHDI=3)# valores con los que se
van a predecir la concentraci??n de Ozono
> new2
  ExpVida Escol PromEscol GNIHDI
1    80    15      13      3
> (Temp2=predict(m6,new2,interval="prediction"))# Intervalo de prediccion
      fit      lwr      upr
1 0.894997 0.8787348 0.9112593
```

-El intervalo de predicción para el modelo M es [0.8039724, 0.8829483]

-El intervalo de predicción para el modelo m6 es [0.8787348, 0.9112593]