



UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSCO MATEMÁTICAS

Materia: Métodos Estadísticos.

Facilitador: MET. Alejandra Guadalupe Cerda Ruiz.

Proyecto: Washington D. C.

Integrantes del equipo:	Matrícula:
José Eduardo López Hernández	1871089
Valeria Guadalupe García Salazar	1850355
Jennifer Priscila de León Flores	1860533
Nancy Nohemí Dávila Juárez	1937883

14 Julio 2021

INDICE

Introducción	3
Desarrollo	4
1 Análisis exploratorio de los datos	4
2 Pruebas de bondad de ajuste	5
3 Multicolinealidad	10
4 Ajuste modelos lineales	11
5 Ajuste modelos NO lineales	14
6 Datos atípicos, palanca y de influencia	18
Conclusiones	20
Bibliografías	21

INTRODUCCIÓN:

Washington D. C., es uno de los cincuenta estados, que forman los Estados Unidos de América. Su capital es Olympia y su ciudad más poblada, Seattle. Está ubicado en la región Oeste del país, división Pacífico, limitando al norte con Canadá, al este con Idaho, al sur con Oregón y al oeste con el océano Pacífico. Fue admitido en la Unión el 11 de noviembre de 1889, como el estado número 42. El distrito posee un área total de 68.3 millas cuadradas.

Washington cuenta con enormes bosques de coníferas, que le han valido el apodo de Evergreen State (estado siempre verde, o estado de la hoja perenne). Estos bosques hacen de Washington un líder de la industria maderera estadounidense. Se encuentra cortado por varios ríos y salpicado por varios lagos, lo que crea un terreno propicio para la instalación de presas. Aquí se localiza la mayor del país, la presa Grand Coulee, en el río Columbia. Su economía, sin embargo, se centra en el turismo y en la industria aeroespacial. El segundo mayor fabricante de aviones del mundo, Boeing, tiene su sede en este estado, así como varias de sus fábricas.

La economía de Washington se concentra principalmente en el sector terciario. El producto interior bruto de Washington era, en 2004, de 262.000 millones de dólares, 10 lo que lo colocaba en el 14º puesto de la nación. La renta per cápita en 2004 era de 33.332 dólares. 11 La tasa de desempleo se sitúa en el 6,2%. El centro económico, financiero e industrial del estado es Seattle.

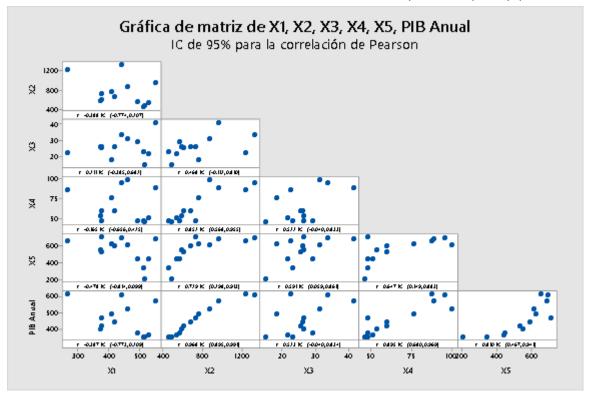
Los datos que se utilizarán son los correspondientes al porcentaje de participación en el PIB por actividad económica más importante de Washington desde el año 2008 al 2020. Cabe aclarar que todas las variables son cuantitativas y los datos están representados en millones de dólares. Las actividades económicas son: Minería, explotación de canteras y extracción de petróleo y gas, transporte terrestre y de tránsito de pasajeros, actividades de apoyo a la minería, transporte por tubería, fabricación de vehículos de motor, carrocerías, remolques y piezas.

Los datos son los siguientes:

	WASHINGTON (Millones de dólares)						
AÑO	Minería, explotación de canteras y extracción de petróleo y gas	Transporte terrestre y de tránsito de pasajeros	Actividades de apoyo a la minería	Transporte por tubería	Fabricación de vehículos de motor, carrocerías y remolques y piezas	PIB Anual	
	X1	X2	Х3	X4	X5		
2008	513.2	457	23.1	47.8	344.1	355.597	
2009	518.4	481.9	15	45.9	205.6	353.706	
2010	532.3	533.5	21.6	51.1	441.5	365.8	
2011	495.2	569.9	29.1	47.7	449.8	379.796	
2012	373.9	583.2	26.3	53.4	551.4	400.623	
2013	376.9	605.6	25.5	59.5	535.4	419.671	
2014	418.8	675.3	26.1	59.5	595.8	442.93	
2015	376.4	731	26.4	47.9	714.2	471.703	
2016	411.2	764.1	18.3	76.4	627.6	493.635	
2017	462.6	873.6	31.1	100.2	618	527.708	
2018	552.8	958	41.4	89	690.8	575.417	
2019	443.9	1327.4	33.5	95.7	700.5	612.997	
2020	267.2	1233.5	22.4	87.5	657.7	618.705	

DESARROLLO:

- 1-. Análisis exploratorio de los datos (incluyendo tablas, gráficas de dispersión y estadísticas) del cual se deberá plantear qué variables se sospecha deberán ser involucradas en "el mejor" modelo y de ser posible plantear un modelo.
- X1: Minería, explotación de canteras y extracción de petróleo y gas
- X2: Transporte terrestre y de tránsito de pasajeros
- X3: Actividades de apoyo a la minería
- X4: Transporte por tubería
- X5: Fabricación de vehículos de motor, carrocerías y remolques y piezas



Correlaciones

	Χi	X 2	X 3	X4	X5
X2	-0.388				
X3	0.211	0.468			
X4	-0.165	0.852	0.522		
X5	-0.478	0.729	0.591	0.647	
PIR Anual	-0.387	0.968	0.523	0.895	0.810

Al ver la matriz de correlación y sus respectivas correlaciones, nos damos cuenta de que la X2 y el PIB Anual tienen una mayor correlación, con un 0.968.

2-. Realice una prueba de bondad de ajuste por variable según la información obtenida en (1).

Haremos pruebas de bondad de ajuste para cada una de las variables, además de nuestra respuesta, que en este caso es el PIB.

1. X1

Minería, explotación de canteras y extracción de petróleo y gas		
X1		
513.2		
518.4		
532.3		
495.2		
373.9		
376.9		
418.8		
376.4		
411.2		
462.6		
552.8		
443.9		
267.2		

H0: La variable tiene distribución normal Ha: La variable tiene alguna otra distribución

media	441.753846
desviación	81.5554271

n	13	
# de clases	3.60555128	4
max	552.8	
min	267.2	
rango	285.6	
ancho	71.40	72

#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes
1	267.2	339.2	1	0.104290986	1.355782822	0.09336408
2	339.2	411.2	4	0.249673484	3.245755297	0.17527047
3	411.2	483.2	3	0.340377633	4.424909229	0.45884926
4	483.2	555.2	5	0.305657896	3.973552652	0.26515168
	·		13	1	13	0.99263549

Región de rechazo

Alfa	0.05
X2 de tabla	3.84145882

Rechazo H0 si EP=0.992>X2 de tabla=3.841

No rechazamos H0, la variable provienen de distribución normal

2. X2

Transporte terrestre y de			
tránsito de pasajeros X2			
, AZ			
457			
481.9			
533.5			
569.9			
583.2			
605.6			
675.3			
731			
764.1			
873.6			
958			
1327.4			
1233.5			

H0: La variable tiene distribución normal Ha: La variable tiene alguna otra distribución

media	753.384615
desviación	276.496446

n	13	
# de clases	3.60555128	4
max	1327.4	
min	457	
rango	870.4	
ancho	217.6	218

#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes	
1	457	675.0	6	0.388399739	5.049196607	0.17904375	Ī
2	675.0	893.0	4	0.304801384	3.962417988	0.00035645	
3	893.0	1111.0	1	0.208859302	2.71517093	1.08347187	
4	1111.0	1329.0	2	0.097939575	1.273214475	0.41486899	
			13	1	13	1.67774106	

Región de rechazo

Alfa	0.05	
X2 de tabla	3.84145882	

Rechazo H0 si EP=1.677>X2 de tabla=3.841

No rechazamos H0, la variable provienen de distribución normal

3. X3

Actividades de apoyo a
la minería
Х3
23.1
15
21.6
29.1
26.3
25.5
26.1
26.4
18.3
31.1
41.4
33.5
22.4

media	26.1384615	
desviación	6.77194439	

n	13	
# de clases	3.60555128	4
max	41.4	
min	15	
rango	26.4	
ancho	6.60	7

#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes
1	15	22.0	3	0.270560516	3.517286703	0.07607726
2	22.0	29.0	6	0.393130553	5.110697187	0.15474591
3	29.0	36.0	3	0.263645017	3.42738522	0.05329373
4	36.0	43.0	1	0.072663915	0.94463089	0.00324544
			13	1	13	0.28736233

Región de rechazo

Alfa	0.05
X2 de tabla	3.84145882

Rechazo H0 si EP=0.287>X2 de tabla=3.841

No rechazamos H0, la variable provienen de distribución normal

4. X4

Transporte por tubería
X4
47.8
45.9
51.1
47.7
53.4
59.5
59.5
47.9
76.4
100.2
89
95.7
87.5

media	66.2769231
desviación	20.4392659

n	13	
# de clases	3.60555128	4
max	100.2	
min	45.9	
rango	54.3	
ancho	13.58	13.6

#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes
1	45.9	59.5	8	0.370109284	4.811420698	2.11310517
2	59.5	73.1	0	0.260633781	3.388239148	3.38823915
3	73.1	86.7	1	0.210409952	2.735329372	1.10091606
4	86.7	100.3	4	0.158846983	2.065010781	1.81315435
			13	1	13	8.41541473

Región de rechazo

Alfa	0.05
X2 de tabla	3.84145882

Rechazo H0 si EP=8.415>X2 de tabla=3.841

Rechazamos H0, la variable proviene de alguna otra distribución

5. X5

Fabricación de vehículos de						
motor, carrocerías y						
remolques y piezas						
X5						
344.1						
205.6						
441.5						
449.8						
551.4						
535.4						
595.8						
714.2						
627.6						
618						
690.8						
700.5						
657.7						

media	548.646154
desviación	151.874118

n	13	
# de clases	3.60555128	4
max	714.2	
min	205.6	
rango	508.6	
ancho	127.15	127.2

							•
#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes	
1	205.6	332.8	1	0.077626770	1.009148006	8.2927E-05	
2	332.8	460.0	3	0.202090448	2.627175822	0.05290771	
3	460.0	587.2	2	0.320478590	4.166221671	1.12632421	
4	587.2	714.4	7	0.399804192	5.197454502	0.62514646	
			13	1	13	1.80446131	

Región de rechazo

Alfa	0.05
X2 de tabla	3.84145882

Rechazo H0 si EP=1.804>X2 de tabla=3.841

No rechazamos H0, la variable provienen de distribución normal

6. Y

PIB Anual
355.597
353.706
365.8
379.796
400.623
419.671
442.93
471.703
493.635
527.708
575.417
612.997
618.705

media	462.945231
desviación	96.0054193

n	13	
# de clases	3.60555128	4
max	618.705	
min	353.706	
rango	264.999	
ancho	66.25	66.3

#	LI	LS	Frecuencia	Probabilidad	Frecuencia esperada	Cocientes
1	353.706	420.0	6	0.327344248	4.255475219	0.71516495
2	420.0	486.3	2	0.268779944	3.494139268	0.63891333
3	486.3	552.6	2	0.228701584	2.973120591	0.31850833
4	552.6	618.9	3	0.175174225	2.277264922	0.22937428
			13	1	13	1.90196089

Región de rechazo

	Alfa	0.05
I	X2 de tabla	3.84145882

Rechazo H0 si EP=1.901>X2 de tabla=3.841

No rechazamos H0, la variable provienen de distribución normal

Solo una de nuestras variables se muestra que no tiene distribución normal, la X4, que sería la de Transporte por tubería, las demás, junto con el PIB, que es nuestra respuesta, muestran que tiene distribución normal con sus respectivos parámetros, que han sido calculados para la aplicación de esta prueba de bondad de ajuste.

3-. Revisar multicolinealidad

Coeficientes

		EE del		Valor	Valor
Término	Coef	coef.	IC de 95%	T	p FIV
Constante	160.8	58.5	(22.5, 299.0)	2.75	0.029
X1	-0.004	0.115	(-0.277, 0.268)	-0.04	0.971 2.82
X2	0.2115	0.0461	(0.1024, 0.3206)	4.58	0.003 5.20
Х3	-0.23	1.44	(-3.65, 3.18)	-0.16	0.876 3.05
X4	1.152	0.569	(-0.193, 2.497)	2.03	0.082 4.32
X5	0.1358	0.0775	(-0.0475, 0.3192)	1.75	0.123 4.43

Se puede observar en la tabla que los FIV´s son menores a 10, por lo que no existen problemas de multicolinealidad.

4-. Ajuste modelos lineales mostrando los resultados a modo de tabla y pruebe el cumplimiento de los supuestos para "el mejor" modelo.

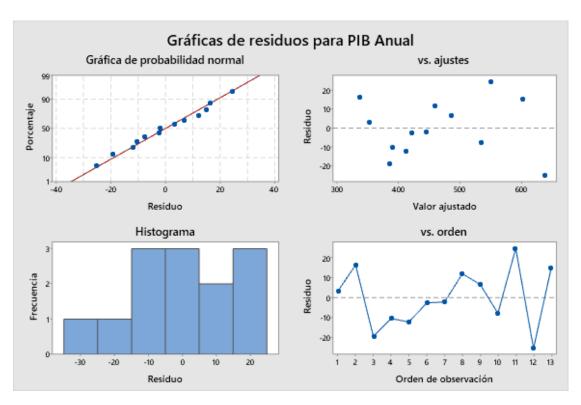
La respuesta es PIB Anual

	_	R-	R-						
				Cp de		X			
Vars	cuad.	(ajust)	(pred.)	Mallows	S 1	2	3	4	5
1	93.8	93.2	89.0	9.3	24.978	Χ			
1	80.2	78.4	68.5	49.4	44.660			Χ	
2	96.1	95.3	91.6	4.5	20.817	Χ			Χ
2	95.6	94.7	90.4	6.0	22.110	Χ		Χ	
3	97.6	96.8	92.8	2.1	17.174	Χ		Χ	Χ
3	96.2	95.0	87.2	6.1	21.562 X	Χ			Χ
4	97.6	96.4	90.2	4.0	18.132	Χ	Χ	Χ	Χ
4	97.6	96.4	88.9	4.0	18.164 X	Χ		Χ	Χ
5	97.6	95.9	87.1	6.0	19.382 X	Χ	Χ	Χ	Χ

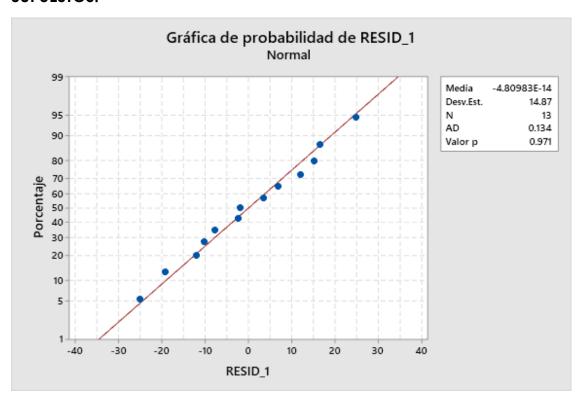
El modelo elegido fue el de la variable 3 ya que es el más bajo en la columna S.

Ecuación de regresión

PIB = 156.2 + 0.2137 X2 + 1.109 X4 Anual + 0.1316 X5



SUPUESTOS:



Prueba de normalidad

H₀: Los residuales siguen una distribución normal

H₁: Los residuales siguen otra distribución

Región de rechazo:

Rechazo Hosi p valor < alfa

0.971< 0.10

NO RECHAZO H₀

Los residuales siguen una distribución normal

<u>Media cero</u>

Media: -4.809E-14

Podemos decir que la media es relativamente pequeña.

Varianza constante

Para la gráfica de vs ajustes por lo que consideramos una varianza constante.

<u>Incorrelación</u>

Ho: Residuales incorrelacionados

H₁: Residuales correlacionados

RESID_1	RESID ²	RESTAS	RESTAS ²
3.39818326	11.5476495		
16.517196	272.817765	-13.1190128	172.108496
-19.2227427	369.513838	35.7399388	1277.34322
-10.3254514	106.614947	-8.89729132	79.1617928
-12.0320054	144.769154	1.70655399	2.91232651
-2.43264005	5.91773762	-9.59936535	92.1478152
-2.01526454	4.06129117	-0.41737551	0.17420232
12.1450291	147.501731	-14.1602936	200.513915

6.78142369	45.9877073	5.36360536	28.7682625
-7.68471398	59.054829	14.4661377	209.269139
24.834663	616.760488	-32.519377	1057.50988
-25.2328607	636.697259	50.0675237	2506.75693
15.2691838	233.147973	-40.5020445	1640.41561
	2654.39237		7267.08159

EP=7267.08159/2654.39237=2.73775711

Datos: n=13 alfa=0.05

n	dL	dU
13		
15	0.95	1.57

Región de rechazo:

Rechazo H_0 si 2.73 < dL

No Rechazo H_0 si 2.73 > dU

No se rechaza Ho

Con un nivel de significancia del 5% existe evidencia estadística que afirma que los residuales están incorrelacionados.

Se cumplen todos los supuestos.

5-. Ajuste modelos NO lineales mostrando los resultados a modo de tabla y pruebe el cumplimiento de los supuestos para "el mejor" modelo.

Usando x2=transporte terrestre y de tránsito de pasajeros ya que es el mejor modelo para una variable predictora.

	_	Modelo lineal	Prueba de significancia			
MODELO	Ecuación estimada	asociado	hipótesis	p valor	conclusión	R ²
Lineal	y =209.59+0.336x2	****	H ₀ : Regresión significativa H _a : Regresión no significativa	1.11E- 07	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	93.80%
Potencia	y=9.393x2 ^{0.59}	y=2.24+0.59x2*	H₀: Regresión significativa Hₐ: Regresión no significativa	4.31E- 09	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	96.55%
Exponencial	y=267,73e ^{0.00070x2}	Y*=5.59+0.00070x2*	H ₀ : Regresión significativa H _a : Regresión no significativa	9.89E- 07	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	90.78%
Logaritmo	y=- 1376.51+280.02LN(x2)	Y*=- 1376.51+280.02x2*	H ₀ : Regresión significativa H _a : Regresión no significativa	6.33E- 10	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	97.57%
Recíproco	y=x/(0.00083x+0.95)	Y*=0.00083+0.95x2*	H ₀ : Regresión significativa H _a : Regresión no significativa	8.94E- 10	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	97.41%

		Prueba de significancia			
MODELO	Ecuación estimada	hipótesis	p valor	conclusión	R ²
Cuadrático	y=-6.40+0.887x2-0.00031x2 ²	H ₀ : Regresión significativa H _a : Regresión no significativa	1.03E- 09	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	98.61%
Cúbico	y=335.72-0.457x2+0.00133x2 ² - 6.30E-07X2 ³	H ₀ : Regresión significativa H _a : Regresión no significativa	6.03E- 10	La regresión es significativa, las variables se relacionan de forma lineal al 95% de confianza	99.38%

Podemos notar que el R^2 más grande sería el del modelo cubico, entonces tomaremos ese como el mejor modelo que se ajusta a nuestros datos.

Verificando el cumplimiento de supuestos para el modelo cúbico MEDIA CERO Y DIST NORMAL

H₀: Los residuales provienen de población normal con media cero H_a: Los residuales provienen de alguna otra distribución

# de clases	3.60555128	4
	-	
Min	13.0529015	
Max	10.3576468	
Rango	23.4105483	
Ancho	5.85263708	5.8527

Media	0
Desv Est	7.56287644

Clases	Li	Ls	Frecuencia	Probabilidad	Esperado	Cociente
	-	-				
1	13.0529015	7.20020152	3	0.17053699	2.21698081	0.27655586
	-	-				
2	7.20020152	1.34750152	2	0.25875651	3.36383465	0.55295374
	-					
3	1.34750152	4.50519848	4	0.29501836	3.83523872	0.00707812
4	4.50519848	10.3578985	4	0.27568814	3.58394581	0.04829903
			13	1	13	0.88488675

EP	0.88488675
X ² de tabla	5.99146455

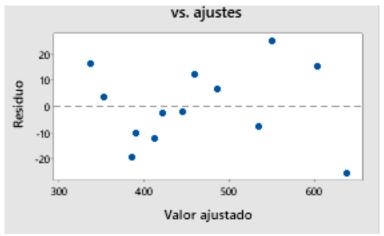
Región de Rechazo

Rechazo H₀ si EP=0.884>X² de tabla=5.99

No rechazo H₀

Los residuales provienen de una distribución normal con media cero y con una confianza del 95%

VARIANZA CONSTANTE



Dada la gráfica, no se aprecian patrones visibles, por lo que se cumple el supuesto de varianza constante

INCORRELACION

H₀: Los residuales están incorrelacionados

Ha: Los residuales están correlacionados

Residuos	Residuos^2	Restas	Restas^2
9.65329445	93.1860937		
-1.73729521	3.01819466	-11.3905897	129.745533
-10.9426382	119.741331	-9.20534299	84.7383396
-13.0529015	170.378238	-2.11026332	4.45321127
1.7082511	2.9181218	14.7611526	217.891627
10.3576468	107.280847	8.64939569	74.8120457
0.22254246	0.04952515	-10.1351043	102.72034
1.83106579	3.35280194	1.60852333	2.5873473
7.7131878	59.493266	5.882122	34.5993593
-8.86345071	78.5607584	-16.5766385	274.784944
4.71275348	22.2100454	13.5762042	184.31332
2.72666399	7.43469651	-1.98608949	3.94455147
-4.32912019	18.7412816	-7.05578418	49.7840904
	686.365201	<-den	1164.37471

<-num

alfa=0.05		
n dL dU		
13		
15	0.81	1.07

EP>dL EP>dU

EP>du>dL

No rechazo H_0 , con 95% de confianza

Los residuales están incorrelacionados, es decir son independientes.

LOS RESIDUALES CUMPLEN LOS SUPUESTOS DEL MODELO

6-. Analizar la existencia de datos atípicos, palanca y de influencia.

Observación	X2	Х4	X5	PIB Anual	RESID	RESIDEST	ні	соок
1	457	47.8	344.1	355.597	3.39818326	0.22549979	0.23002131	0.00379771
2	482	45.9	205.6	353.706	16.517196	1.48844636	0.58247387	0.77267911
3	534	51.1	441.5	365.8	- 19.2227427	-1.2024334	0.13346515	0.05567291
4	570	47.7	449.8	379.796	- 10.3254514	- 0.65320986	0.15279204	0.01923783
5	583	53.4	551.4	400.623	- 12.0320054	-0.7610286	0.15247705	0.02604924
6	606	59.5	535.4	419.671	- 2.43264005	- 0.15104608	0.1205439	0.00078179
7	675	59.5	595.8	442.93	- 2.01526454	- 0.12619427	0.13530644	0.00062298
8	731	47.9	714.2	471.703	12.1450291	1.06137582	0.55604788	0.35273979
9	764	76.4	627.6	493.635	6.78142369	0.43557498	0.17814862	0.01028147
10	874	100.2	618	527.708	- 7.68471398	- 0.70667761	0.59905007	0.186533
11	958	89	690.8	575.417	24.834663	1.63235168	0.21518649	0.18264845
12	1327	95.7	700.5	612.997	- 25.2328607	2.10100003	0.51094438	1.1529421
13	1234	87.5	657.7	618.705	15.2691838	1.18132927	0.43354279	0.26702185

En la observación 12 tenemos un dato atípico, el residuo estandarizado - 2.10100003 ya que no se encuentra en el intervalo de (-2,2).

Para verificar si hay palancas y/o influencias, para el primero el promedio $\overline{h}=k/n$, donde k es el número de regresores y n el número de observaciones (En caso de que la regresión no sea al origen, entonces $\overline{h}=k+1$ n, por $\beta 0$), que en este caso es

$$\bar{h} = \frac{5}{13}$$

Entonces, existen observaciones palanca si $2\overline{h} \le hi$, donde hi es el HI de la observación i, siempre y cuando, $2\overline{h} \le 1$. Entonces, de la tabla, no hay observaciones que sean palancas

Para saber si una observación tiene influencia, Di > 1, donde Di es la distancia COOK de la observación i. Entonces, no hay observaciones que sean influencias

CONCLUSIÓN:

Después de realizar el análisis correspondiente concluimos que el modelo cubico con un 99.38% de ajuste es el mejor para estos datos, tiene como ecuación PIB=335.72-0.457x2+0.00133x2²-6.30E-07X2³ con x2=transporte terrestre y de tránsito de pasajeros. Se puede ver como la variable B_0 y B_1 afectan de forma positiva mientras que B_2 afecta de forma negativa la ecuación del modelo.

Dado el análisis de los residuales la prueba nos arrojó que no hay datos atípicos, esto nos indica que los datos que tomamos son los correctos, esto hace que nuestro modelo sea preciso a la hora de tomar una decisión para alguna estrategia.

Con este proyecto pudimos aplicar los conocimientos de análisis de datos para la implementación de estrategias, con base a la información que analizamos podemos tomar decisiones acertadas para mejorar en el área del proyecto.

Esto es muy productivo debido a que la demanda en el sector de análisis de datos últimamente está en aumento, siempre que podamos analizar los datos con un buen modelo que fue lo que aprendimos a lo largo del curso.

BIBLIOGRAFÍAS:

https://datosmacro.expansion.com/pib/usa-estados/washington

https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&acrd n=8

https://es.wikipedia.org/wiki/Washington (estado)#Econom%C3 %ADa

http://www.imarranz.com/indicadores valoracion modelo.html #apalancamiento o puntos palanca (leverage)