Modelos Estadísticos I (Fecha: 14/04/2021)

Tarea #5

Estudiante: Roberto Vásquez Martínez *NUA:* 424662

Problema 1

(a) Decimos hacer la regresión sin intercepto pues presentaba mayor R^2 y la diferencia en AIC es pequeña.

Los resultados de la regresión son los siguientes.

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
x1	1.7079	0.1775	9.6207	0.0000	1.3407	2.0751
x2	0.0161	0.0038	4.2593	0.0003	0.0083	0.0239

Con un $R^2 = 0.985$.

(b) La correlación entre las variables X_1 y X_2 fue de

$$\rho(X_1, X_2) = 0.8242,$$

(c)

El factor de inflación de la varianza lo obtenemos de la diagonal de $(X^T X)^{-1}$ cuando los datos estan escalados y centrados, luego

$$VIF(\hat{\beta}_1) = 3.1185 \text{ y } VIF(\hat{\beta}_2) = 3.1185,$$

que no es alto, por lo que esto da evidencia de que no tenemos problemas de colinealidad.

(d)

Ahora, calculamos el número de condición de la matriz $X^T X$, este es

$$\kappa(X^T X) = 3.2214,$$

y al ser menor que 100 podemos concluir que no hay evidencia de que halla problemas de multicolinealidad.

(e)

A pesar de que la correlación sea alta, vemos que tanto el factor de inflación de la varianza como el número de condición son aceptables, por lo que podemos decir que no hay evidencia contundente de que tenemos problemas de multicolinealidad.

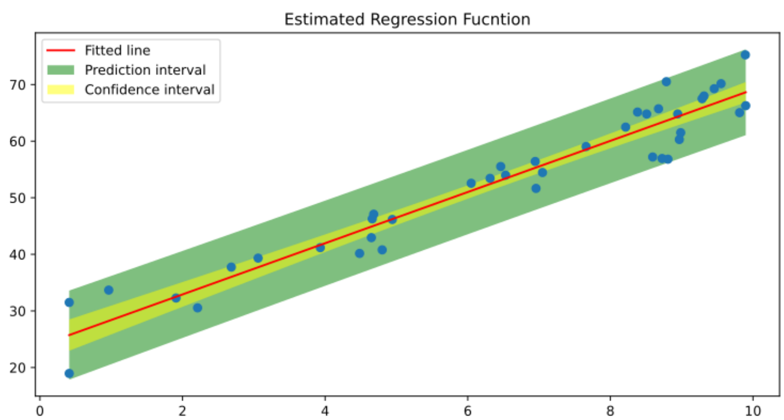
□

Problema 2

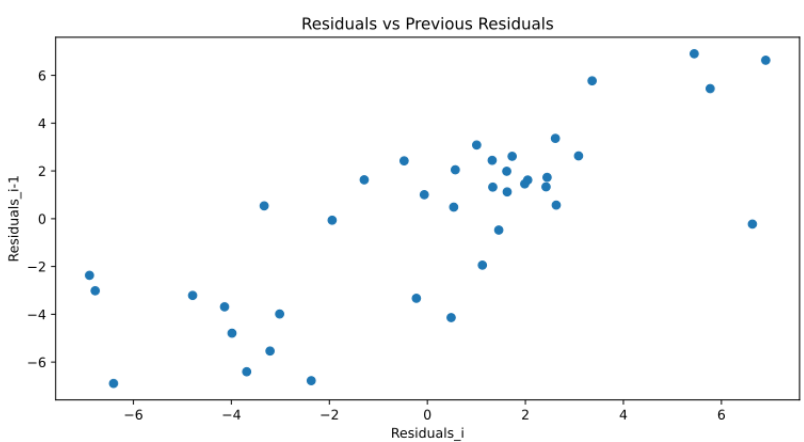
(a) Ajustamos el modelo de regresión simple a los datos de la table 2.
El resumen de resultados es

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	23.8701	1.4437	16.5342	0.0000	20.9475	26.7927
x1	4.5231	0.2033	22.2504	0.0000	4.1116	4.9347

Y la gráfica del ajuste es



Graficando los residuos contra los anterior tenemos lo siguiente



Donde vemos un patrón claramente creciente, por lo que es una evidencia de correlación positiva entre las observaciones.

Haciendo la prueba de Durbin-Watson llegamos a que

```
Durbin-Watson:          0.444,
```

que es señal de alarma pues es menor que 1 y que también nos dice que hay una correlación positiva al ser también menor que 2, luego debemos resolver un problema de autocorrelación.

(c) Se sugiere hacerlo un por un proceso autoregresivo de orden 1 de la forma

$$y_t - \rho y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + e_t, \tag{1.1}$$

donde para obtener ρ hacemos una regresión con los residuos ϵ del primer ajuste de la forma $\epsilon_t = \rho\epsilon_{t-1} + e_t$.

El parámetro estimado es

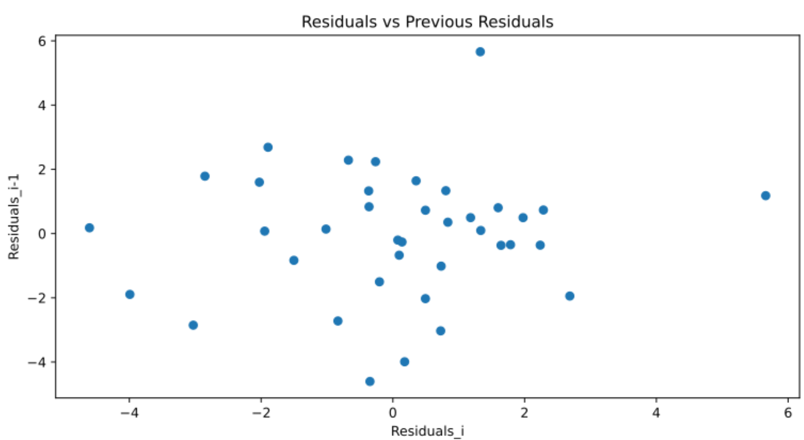
$$\hat{\rho} = 0.7942,$$

después hacemos la regresión en (1.1).

Con esto obtenemos el siguiente resumen

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
x1	21.3169	1.6629	12.8195	0.0000	17.9476	24.6861
x2	4.8487	0.0879	55.1737	0.0000	4.6706	5.0268

(d) Volvemos a gráficar residuos contra los anteriores con esta regresión



,

donde ya no vemos un patrón aparante y la prueba de Durbin-Watson nos dice

```
Durbin-Watson:          1.744,
```

que es cercano a 2 que es cuando la prueba nos dice que no hay autocorrelación, con este procedimiento hemos mejorado notablemente el problema de autocorrelación.

□

Problema 3

(a)

Para determinar que observaciones son potencialmente influyentes recurrimos a la diagonal de la matriz de proyección, usando el criterio de que observaciones influyentes tienen un valor en la diagonal mayor a $2p/n$ tenemos que los puntos potencialmente influyentes son

```
Puntos␣palanca:
[␣9.␣10.␣16.␣22.]
```

(b) Para decir que observaciones se tiene un mayor desplazamiento de la respuesta $\hat{Y}_i$ recurrimos a la D de Cook con un corte en $4/n$, en este caso los puntos con una valor D de Cook considerable son

```
Puntos influyentes en la respuesta (D-Cook):
[ 9. 20. 22.]
```

(c) Vemos que las observaciones 9 y 22 llaman la atencion analizmos sus valores $DFBETAS$ y $DFFITS$

```
OBSERVACION 9:
DFBETA_1: 0.0917, DFBETA_2: 0.3040, DFFITS: 1.0956

OBSERVACION 22:
DFBETA_1: -1.1902, DFBETA_2: 0.8466, DFFITS: -1.4029

Corte DFBETA: 0.4000

Corte DFFITS: 0.5657
```

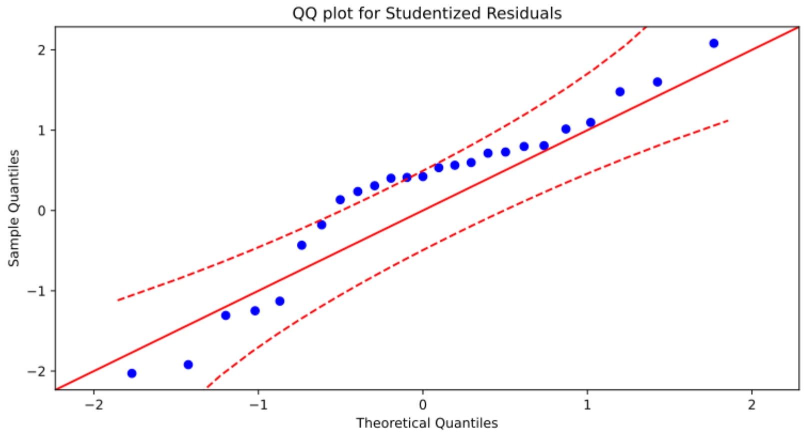
Recordamos que observaciones $|DFBETA| > 2/\sqrt{n}$ y $|DFFITS| > 2\sqrt{p/n}$ se deben considerar observaciones influyentes y vemos al menos la observación 22 es influyente con todos los criterios y la nueve 9 debería considerar descartarse por su valor DFFITS.

(d) Además vemos que observaciones demás de la 9 y 22 tienen DFBETA y DFFITS considerables, estas son

```
Observaciones descartables DFBETA1:
[10. 22.]

Observaciones descartables DFBETA2:
[10. 22. 24.]

Observaciones descartables DFFITS:
[ 9. 20. 22.]
```



,

Con todos los criterios anteriores debemos descartar la observación $22y9$, y dado que tenemos una muestra pequeña deberíamos no hacer mas descartes.

(e)

Hacemos un análisis de los residuos estudentizados, primero nos fijamos en la gráfica QQ, que no muestra un buen ajuste

además dado que los residuos estudentizados siguen una distribución t_{23}, los valores los podemos determinar aquellos que en valor absoluto son mayores que el cuantil de probabilidad 0.95, y aquí solo tenemos a la cuarta observación, por lo que este análisis hubiera sido insuficiente para determinar otras observaciones influyentes.

$\square$

Problema 4

(a)

Para cada grupo ajustamos una regresión lineal simple. Obteniendo los siguientes resultados

Para Colonia

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	7.9004	15.4509	0.5113	0.6444	-41.2711	57.0719
x1	0.9207	0.1119	8.2243	0.0038	0.5644	1.2770

con un $R^2 = 0.958$.

Para CentroCom.

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	50.6302	17.3528	2.9177	0.0616	-4.5942	105.8545
x1	0.8290	0.0918	9.0262	0.0029	0.5367	1.1213

con un $R^2 = 0.964$.

Y para el Centro

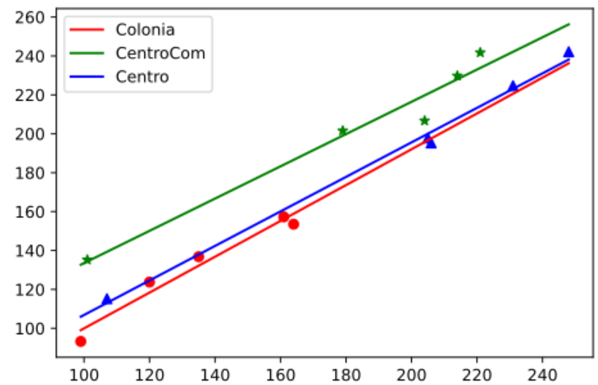
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	18.1555	8.3570	2.1725	0.1182	-8.4403	44.7512
x1	0.8871	0.0407	21.7938	0.0002	0.7575	1.0166

con un $R^2 = 0.99$.

En todos los casos resulta que el modelo lineal parece sensato.

(b)

Hacemos la gráfica de dispersión que se pide sobreponiendo las rectas.



La gráfica da evidencia, de que las rectas tienen la misma pendiente y que entre Centro y Colonia se tiene el mismo intercepto, y que hay una diferencia significativa del intercepto de CentroCom con Centro y con Colonia.

(c) Ahora utilizaremos el modelo

$$Y = \beta_0 + \beta_1 X + \alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon,$$

Con Z_1, Z_2 las variables de asignación que vimos en clase.

El resumen es el siguiente

```
Coef. Std.Err. t P>|t| [0.025 0.975]
const 21.8415 8.5585 2.5520 0.0269 3.0044 40.6785
x1 0.8686 0.0405 21.4520 0.0000 0.7795 0.9577
x2 -6.8638 4.7705 -1.4388 0.1780 -17.3635 3.6360
x3 21.5100 4.0651 5.2914 0.0003 12.5628 30.4572
```

con un $R^2 = 0.98$.

Con estos datos podemos obtener las diferencias entre los interceptos, sus desviaciones estándar y los t-valores correspondiente.

```
(Col-CentroC): -28.3738, std(Col-CentroC):4.4613, t-valor: -6.3599

(Col-Centro): -6.8638, std(Col-Centro): 4.7705, t-valor: -1.4388, p-valor: 0.1780

(CentroC-Centro): 21.5100, std(CentroC-Centro): 4.0651, t-valor: 5.2914, p-valor: 0.0003.
```

Lo que indica que la diferencia entre Colonia-CentroCom, así como Centro-CentroCom, es significativa, no así la diferencia entre Colonia-Centro, que ya habíamos anticipado.

(d) Para investigar la diferencia entre pendientes consideramos el modelo

$$Y = \beta_0 + \beta_1 X + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X).$$

Aplicamos el modelo de regresión con estos datos notando $\hat{\gamma}_1$ es la estimación de la diferencia de pendientes entre Colonia y Centro, $\hat{\delta}_1$ es la estimación de la diferencia entre CentroCom y Centro, y $\hat{\gamma}_1 - \hat{\delta}_1$ es la diferencia entre Colonia y CentroCom.

El resumen es el siguiente

```
    Coef. Std.Err. t P>|t| [0.025 0.975]
const 18.1555 12.7585 1.4230 0.1885 -10.7062 47.0171
x1 0.8871 0.0621 14.2753 0.0000 0.7465 1.0276
x2 -10.2550 21.2832 -0.4818 0.6414 -58.4010 37.8909
x3 32.4747 18.3034 1.7742 0.1098 -8.9304 73.8798
x4 0.0336 0.1382 0.2434 0.8132 -0.2790 0.3462
x5 -0.0581 0.0932 -0.6233 0.5486 -0.2689 0.1527
```

con $R^2 = 0.988$.

De estos datos obtenemos algo análogo al inciso anterior para las pendientes

```
(Col-CentroC):0.0917  std(Col-CentroC):0.1416, t-valor: 0.6474

(Col-Centro): 0.0336, std(Col-Centro): 0.1382, t-valor: 0.2434, p-valor: 0.8132

(CentroC-Centro): -0.0581, std(CentroC-Centro): 0.0932, t-valor: -0.6233, p-valor: 0.5486.
```

Y en ningún caso podemos rechazar la hipótesis de nulidad, por lo que hay evidencia a favor de que todas las pendientes son iguales.

$\square$