# Estadística Descriptiva Bivariante con STATGRAPHICS -Dependencia lineal y Regresión-

Fichero de datos empleado: VelVientos730.sf3

# 1. Introducción

En este documento se analizarán, utilizando Statgraphics, dos variables observadas simultáneamente. Se analizará su dependencia lineal y se construirá la recta de regresión que ayude a predecir una variable a partir de la segunda.

Los datos que se tienen son los registros de velocidades de viento de dos anemómetros colocados en dos parques eólicos cercanos. El fichero VelVientos.sf3 contiene el registro de 730 horas, donde en cada hora se tiene la velocidad del viento registrada en cada parque. Las velocidades, en metros por segundo (m/s), de cada parque se encuentran en las variables Parque1 y Parque2 respectivamente.

Se quiere disponer de un sistema informático que registre las velocidades del viento en esos parques en tiempo real. Esa información es muy importante para poder gestionar la producción energética del parque y también para detectar errores de funcionamiento de los aerogeneradores. El sistema informático que se instalará es muy costoso, pues requiere una red donde algunas etapas usan transmisión por microondas, calibraciones periódicas, y personal y procedimientos que monitoricen las transmisiones de los ficheros. Por esta razón se decide realizar esa instalación sólo para el Parque1. El objetivo final que se pretende con el análisis de los datos es utilizar las mediciones de viento del Parque1 para predecir las del Parque2 mediante una recta de regresión, y ahorrarnos así duplicar el coste del sistema.

Lo primero que hacemos es leer ese fichero de datos.

🌮 SI	IAIG	КАРН	ILS Plus -	Untitled 9	statFolic	)				
File	Edit	Plot	Describe	Compare	Relate	Special	View	Window	Help	
Op	pen					O	pen Sta	tFolio		Ctrl+F11
C	ose					► 0	pen Dal	ta File		Ctrl+F12
Sa	ave					► 0	pen Sta	tGallery		
Sa	ave As	5				► 0	oen Sta	tReporter	·	
		Open D Busca en: Vel	Data File ar 🔄 Proy Wientos Wientos730	grams Matlab		1	- +	È 🔺 🖩	<u>?</u> ] :: ::	-
		Nombre archivo	de Ve	Vientos730					Abrir	
		Tipo de	s sc	i PLUS Files (	*.sf3;*,sfx;*.	.sf)	•	- Ca	ancelar	
		archivu	15.					A	Avuda 🛛	

VelVientos730.sf3								
	Pargue1	Parque2		Co1_3				
1	3,50	4,03						
2	2,19	2,75						
3	1,93	1,99	Velocidad de viento					
4	2,08	1,92 en m/s						

# 2. Análisis Gráfico

En primer lugar vamos a hacer un histograma de cada variable. Para poder comparar mejor ambas distribuciones lo haremos colocando ambos histogramas en un mismo gráfico, lo que se consigue en Compare/Two samples/Two-Sample Comparison.

🌽 ST	🧭 STATGRAPHICS Plus - Untitled StatFolio									
File	Edit	Plot	De:	scribe	Compare	Relate	Special	View	Window	Help
2 🖬 🗃 🐮 🚿		Two Sa	Two Samples 🔹 🕨		Two-Sample Comparison					
		Multiple	Multiple Samples		Paired-Sample Comparison					
- Y V Blot Devoue1 uv			Analysis of Variance		ince 🕨	Нурс	othesis Tes	ts		

Como nuestros datos están dispuestos en dos columnas, una con cada variable, la introducción de las variables es de la siguiente forma:

Two-Sample Com	nparison
Parque1 Parque2	Sample 1:
	Sample 2:
	(Select.)
Sort	Two Data Columns  Data and Code Columns

En las opciones gráficas seleccionamos el histograma. El resultado es el siguiente



Los histogramas nos muestran que las propiedades estadísticas de la velocidad del viento en ambas localizaciones son muy parecidas. Son unimodales, con asimetría positiva, de rango similar, y con un intervalo modal en torno a 4-5 m/s. El que las distribuciones univariantes sean parecidas no implica en absoluto que ambas variables estén correladas. Simplemente nos dice que ambas variables tienen una naturaleza muy parecida, y por tanto, si estuviesen correladas, la recta de regresión va a ser una herramienta de predicción muy adecuada.

Vamos a introducir un par de modificaciones en el histograma. Como los datos no pueden ser negativos, vamos a forzar que el límite inferior sea cero. Además, como tenemos muchos datos, vamos a aumentar hasta 27 el número de clases ( $\sqrt{730} \approx 27$ ). El resultado es el siguiente:



Ahora puede verse que aunque los histogramas presentan ciertas diferencias, siguen mostrándonos dos variables muy parecidas. Parece que el Parque1 tiene una distribución multimodal, con una moda en torno a 3, otra en torno a 6 y otra en torno a 9. Sin embargo en el Parque2 el comportamiento es más unimodal. La multimodalidad del Parque1 no es tampoco muy acusada.

Veamos a continuación el gráfico de dispersión de estas dos variables. El Statgraphics permite hacer gráficos de dispersión en varios lugares. El lugar más sencillo es el siguiente: Plot/Scatterplots/X-Y Plot

🥖 s	TATG	RAPH	ICS Plus -	StatF	olia	)				
File	Edit	Plot	Describe	Compare	Rela	te	Special	View	Window	F
a		Sc	atterplots		•		Univariat	e Plot		
		Exploratory Plots			•		X-Y Plot			
	Two-9	Bu	isiness Cha	irts	•		X-Y-Z Plo	t		
		Pr	Probability Distributions				Multiple X	-Y Plot		
	Response Surfaces					Multiple X-Y-Z Plot				
Sur	marv	Custom Charts					Polar Coo	ordinate	s Plot	

Como nuestro objetivo es usar al Parque1 como variable explicativa, y al Parque2 como variable respuesta, llamaremos X=Parque1 e Y=Parque2, aunque a efectos de hacer el gráfico esa distinción sea arbitraria

-Y Plot		>
Parque1 Parque2	Y: Parque2 X: Parque1 (Select) Select	
Sort 🗸		
OK	Cancel Delete Transform	Help

#### El gráfico que resulta es





donde se aprecia que la relación entre ambas variables es lineal y muy fuerte. Parece entonces sensato utilizar una recta de regresión para predecir Y en función de X. El que las distribuciones de ambas variables sean parecidas sin duda ayuda a que la relación sea mayor.

### 3. Medidas características bivariantes

Para buscar las medidas características que resuman esta relación lineal vamos a

🌽 S	🛃 STATGRAPHICS Plus - Untitled StatFolio									
File	Edit	Plot	Describe	Compare	Relate	Speci	al View	Window	Help	
8	🚗 💼 🙀 Numeric Data						One-Va	ariable Ana	lysis	
_	🖉 🗖 💆		Catego	Categorical Data		•	Multiple	Multiple-Variable Analysis		
	K-Y PI	ot - P	Distribu	utions		•	Subset	Analysis		
	Life Data					•	Row-W	/ise Statisti	ics	
			Hypoth	Hypothesis Tests			Power Transformations			
line	lvei	5 S117	Sample		Statisti	cal Tolerar	nce Limits			

y allí seleccionamos nuestras dos variables

×
×

Las opciones gráficas de esta sección muestran también un diagrama de dispersión, pero no tan claro como el anterior. Las opciones numéricas (Tabular Options) que nos interesan son la matriz de covarianzas y, sobre todo, la matriz de correlaciones.

📲 Multiple-Variable Analysis				
📰 📃 🖬 🎬 🏤 🕅	<b>56?</b>	Row:		
Analysis ummary	Tabular Options	×		
Data variables: Parquel Parque2 There are 730 complete cases	Analysis Summary     Summary Statistics     Confidence Intervals		Parquel	
The StatAdvisor  This procedure is designed quantitative data. It will of correlations, covariances, ar the procedure are a number of	<ul> <li>Correlations</li> <li>Rank Correlations</li> <li>Covariances</li> <li>Partial Correlations</li> </ul>			Parque2
interesting views into the de Graphical Options buttons on	OK Cancel	All Help		
La matriz de covarianzas	S es: Covariances			

	Parquel	Parque2
Parquel	10,5057 ( 730)	9,84153 ( 730)
Parque2	9,84153 ( 730)	10,5948 ( 730)

De la información de esta matriz podríamos ya deducir las correlaciones e incluso el coeficiente de regresión. Por ejemplo, la correlación entre ambos parques será

$$corr(x, y) = \frac{cov(x, y)}{s_x s_y} = \frac{9.84153}{\sqrt{10.5057}\sqrt{10.5948}} = 0.9328$$

Esta correlación coincide con el resultado que suministra el Statgraphics

	Parquel	Parque2
Parquel		0,9328 ( 730) 0,0000
Parque2	0,9328 ( 730) 0,0000	

Un gráfico de dispersión tan lineal y una correlación tan alta harán que la recta de regresión vaya a ser muy precisa.

## 4. La recta de regresión

Para calcular la recta de regresión, también llamada de regresión simple (por tener una sola variable explicativa), nos vamos a Relate/Simple Regression

🥖 S	STATGRAPHICS Plus - Untitled StatFolio									
File	Edit	Plot	Describe	Compare	Relate	Special	View	Window	H	
3			234 😾		Simp	le Regres	sion			
_			ם መ	4.4	Polynomial Regression					
					Box-Cox Transformations					
					Multi	ple Regre	ssion			

y allí seleccionamos las variables implicadas. Variable respuesta=Y=Parque 2; Variable explicativa=X=Parque1.

La técnica de regresión tiene muchos más implicaciones teórico-prácticas que las que se exponen en este documento, por lo que la mayoría de las opciones que nos muestra el Statgraphics no nos son de ayuda. En lo que respecta a opciones numéricas (Tabular Options), seleccionamos sólo el resumen de los resultados

📑 Simple Regression - I	Parque2 vs. Parque1
	الله الله الله الله الله الله الله الله
Regression inalysis	Tabular Options
Dependent variable. Independent variabl	Analysis Summary
Parameter Est	Lack-of-Fit Test Forecasts
Intercept 0,1 Slope 0,9	Comparison of Alternative Models
	Unusual Residuals
Predicted Values	OK Cancel All Help

La ecuación que queremos estimar es la recta de mínimos cuadrados

$$\hat{y}_i = a + bx_i$$

donde

$$b = \frac{\operatorname{cov}(x, y)}{s_x^2}$$
$$a = \overline{y} - b\overline{x}$$

El cálculo de estos valores a,b que nos proporciona el programa es:

Regression Ar	nalysis - Linea	r model: Y = a	+ b*X 		
Dependent variable: Parque2 Independent variable: Parquel					
Parameter	Estimate	Standard Error	T Statistic	P-Value	
Intercept Slope	0,197868 0,936776	0,0891091 0,0134105	2,22051 69,8538	0,0264 0,0000	
	Regression Ar Dependent var Independent v Parameter Intercept Slope	Regression Analysis - Linea Dependent variable: Parque2 Independent variable: Parque Parameter Estimate Intercept 0,197868 Slope 0,936776	Regression Analysis - Linear model: Y = a Dependent variable: Parque2 Independent variable: Parque1 Standard Parameter Estimate Error Intercept 0,197868 0,0891091 Slope 0,936776 0,0134105	Regression Analysis - Linear model: Y = a + b*X         Dependent variable: Parque2         Independent variable: Parque1         Standard       T         Parameter       Estimate         Intercept       0,197868       0,0891091       2,22051         Slope       0,936776       0,0134105       69,8538	

donde sólo nos interesan los valores correspondientes a la columna *Estimate*. La pendiente de la recta, es decir, el parámetro **b**, es *Slope*, y el punto de corte cuando x=0, es decir, el parámetro **a**, es el *Intercep*. Nuestra recta de regresión es entonces

Velocidad Prevista en Parque2=0.198+0.937x(Velocidad del Parque1)

La recta de regresión aparece dibujada en las opciones gráficas



El gráfico que resulta tiene unas líneas auxiliares que al nivel en que estamos manejando la recta de regresión en este documento tampoco nos son de utilidad. Para quitarlas, nos colocamos en el gráfico y pulsamos el botón derecho del ratón. Accedemos así a Pane Options. Allí accedemos a las opciones de este gráfico, donde eliminamos las curvas que no nos interesan



Para mejorar la visualización de la recta de regresión nos colocamos de nuevo sobre el gráfico, pulsamos el botón derecho del ratón y seleccionamos Graphics Options

Pane Options Analysis Options			
Graphics Options			
Undo			
Select Locate Zoom In Undo Zoom			
Reset Scaling/Viewpoint			
Сору	Ctrl+C		
Print Print Preview	F4 Shift+F3		

y allí modificamos el grosor y el color de la línea



El gráfico que resulta es ahora más claro.



La recta junto con la nube de puntos muestra que, si bien el ajuste es bastante bueno en casi todo el rango de velocidades, a velocidades altas del Parque1, la recta subestima la velocidad del Parque2. Esa zona es, sin embargo, poco frecuente, como se pudo ver en los histogramas de ambas velocidades. Utilizando conocimientos más avanzados de regresión se podría mejorar esta relación, aumentando la precisión de las predicciones en esa zona. Las modificaciones que habría que introducir se escapan del alcance de este documento.