Seguimiento

Taller 2 - Técnicas Robustas y No Paramétricas

Pablo Andrés Saldarriaga Aristizábal[†]
psaldar2@eafit.edu.co

[†]Ingeniería Matemática, Universidad EAFIT

30 de octubre de 2018

1. Elección de una provincia y estimar su densidad

Provincia elegida: Barcelona

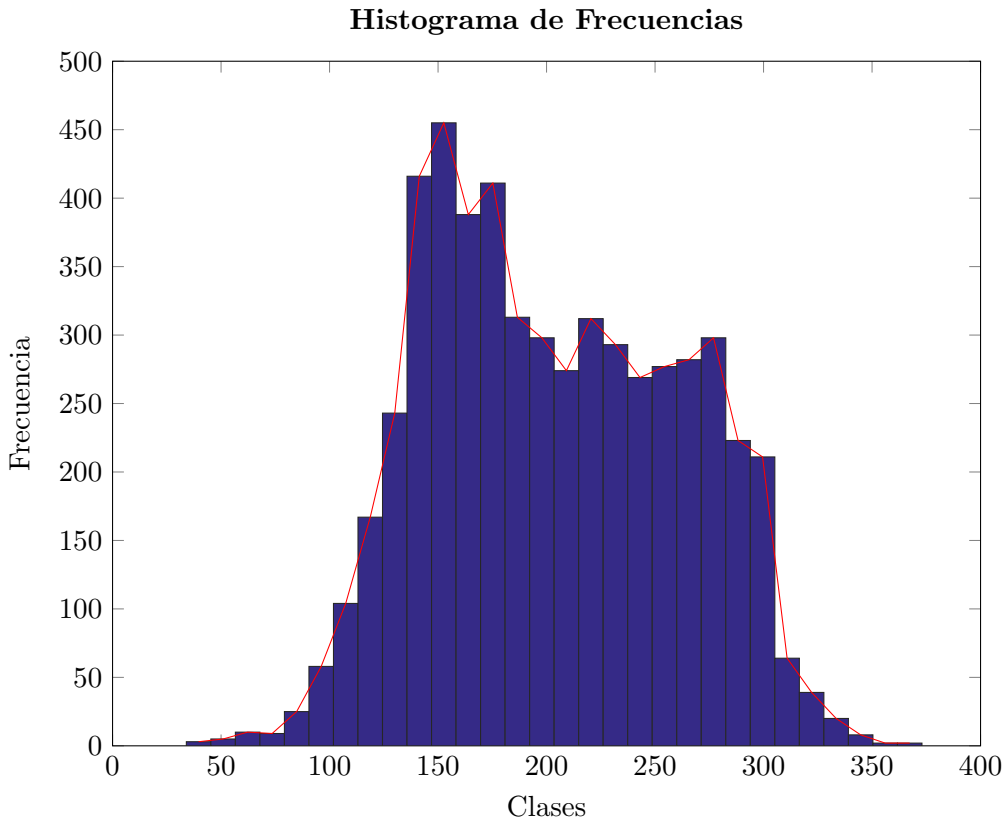

Figura 1: Densidad de temperaturas en la provincia de Barcelona

Cuadro 1: Densidad de la provincia de Barcelona

Límite Inferior	Límite Superior	Punto Medio	Frecuencia	Frecuencia Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
34	45.3	39.65	3	3	0.000547545	0.000547545
45.3	56.6	50.95	5	8	0.000912575	0.00146012
56.6	67.9	62.25	10	18	0.001825151	0.003285271
67.9	79.2	73.55	9	27	0.001642636	0.004927907
79.2	90.5	84.85	25	52	0.004562876	0.009490783
90.5	101.8	96.15	58	110	0.010585873	0.020076656
101.8	113.1	107.45	104	214	0.018981566	0.039058222
113.1	124.4	118.75	167	381	0.030480015	0.069538237
124.4	135.7	130.05	243	624	0.044351159	0.113889396
135.7	147	141.35	416	1040	0.075926264	0.18981566
147	158.3	152.65	455	1495	0.083044351	0.272860011
158.3	169.6	163.95	388	1883	0.070815842	0.343675853
169.6	180.9	175.25	411	2294	0.075013689	0.418689542
180.9	192.2	186.55	313	2607	0.057127213	0.475816755
192.2	203.5	197.85	298	2905	0.054389487	0.530206242
203.5	214.8	209.15	274	3179	0.050009126	0.580215368
214.8	226.1	220.45	312	3491	0.056944698	0.637160066
226.1	237.4	231.75	293	3784	0.053476912	0.690636978
237.4	248.7	243.05	269	4053	0.04909655	0.739733528
248.7	260	254.35	277	4330	0.050556671	0.790290199
260	271.3	265.65	282	4612	0.051469246	0.841759445
271.3	282.6	276.95	298	4910	0.054389487	0.896148932
282.6	293.9	288.25	223	5133	0.040700858	0.93684979
293.9	305.2	299.55	211	5344	0.038510677	0.975360467
305.2	316.5	310.85	64	5408	0.011680964	0.987041431
316.5	327.8	322.15	39	5447	0.007118087	0.994159518
327.8	339.1	333.45	20	5467	0.003650301	0.997809819
339.1	350.4	344.75	8	5475	0.00146012	0.99926994
350.4	361.7	356.05	2	5477	0.00036503	0.99963497
361.7	373	367.35	2	5479	0.00036503	1

1.1. Simulación de Datos

A partir de la densidad estimada en la tabla de frecuencias, simularemos 1000 datos que provengan de la misma distribución a la de la provincia de Barcelona. Para esto se utiliza dos aproximaciones en la generación de datos (1) Uso de un Kernel Uniforme, (2) Uso de un Kernel Gausiano.

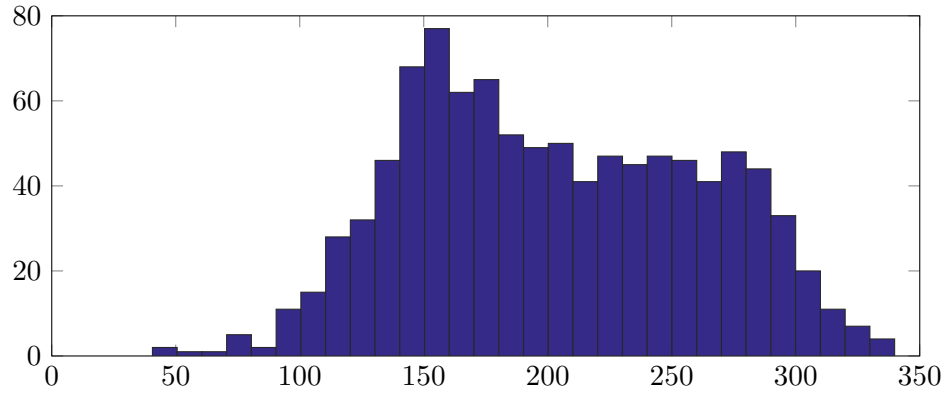

Figura 2: Simulación con Kernel Uniforme

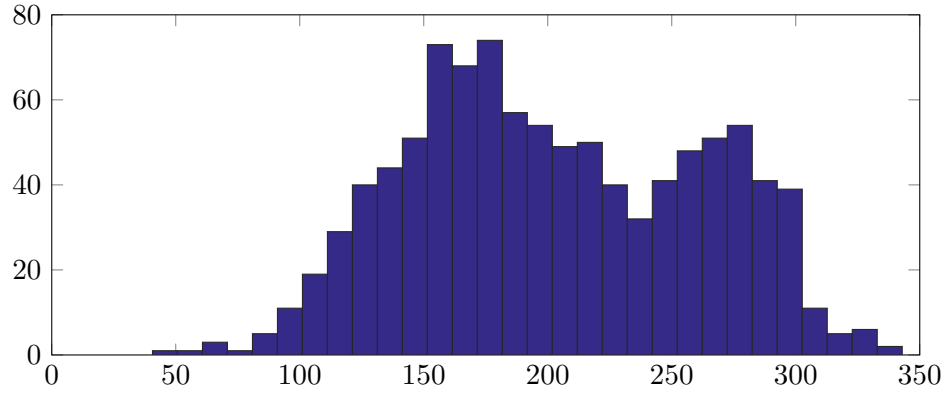

Figura 3: Simulación con Kernel Gaussiano

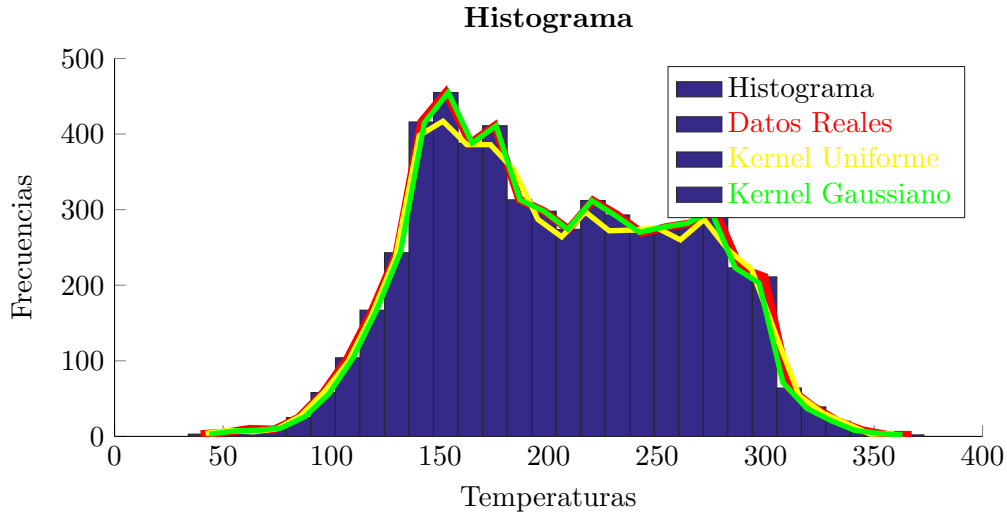

Figura 4: Datos VS Densidades Estimadas

Podemos ver que utilizando la densidad estimada a partir de la tabla de frecuencias, encontramos una distribución similar al generar datos simulados, esto evidenciado en los histogramas anteriores, comparados con el histograma de la distribución de temperaturas en Barcelona. Adicional a esto, existen métodos más sofisticados al momento de estimar la densidad de un conjunto de datos, otra aproximación está dada por la estimación de un *Smooth Kernel Function*. Para hacer uso de esta aproximación, se utiliza el toolbox *ksdensity* de Matlab, ajustando los datos con diferentes kernels, vemos entonces en el siguiente histograma que los resultados obtenidos por todos los kernels son similares, además de que se ajustan a la distribución de los datos de temperatura de la provincia de Barcelona.

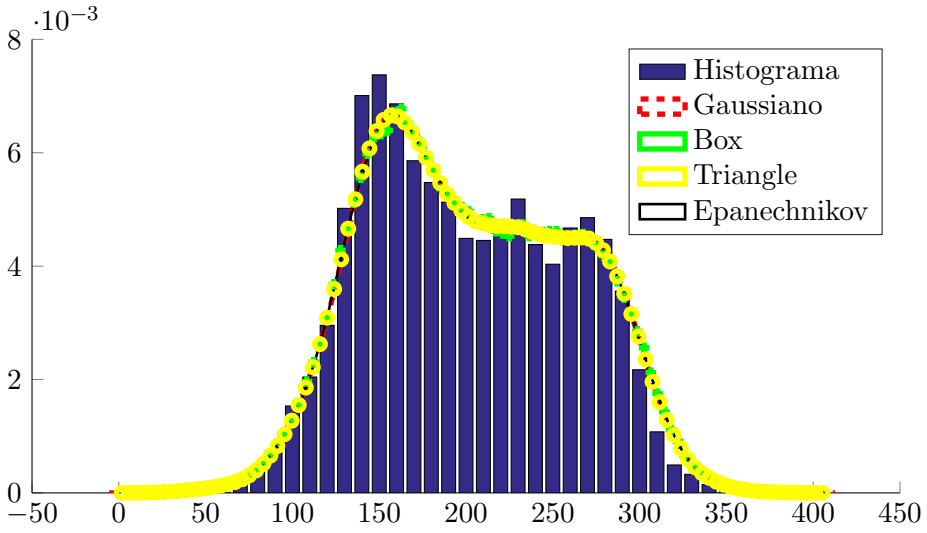

Figura 5: Estimación de Densidad - *ksdensity*

2. Método de los rangos para determinar que provincias son distintas a las elegidas

El método de rangos, considera que se posee n observaciones pareadas de la forma (X_i, Y_i), además se define $D = X_i - Y_i$. Se está interesados probar la hipótesis de que las variables X y Y poseen la misma distribución contra la hipótesis de que las distribuciones son diferentes. Para realizar este test se realiza lo siguiente:

- Se obtiene el vector $D = X - Y$
- Del vector D se eliminan las entradas cuyo valor sea cero
- Se clasifican los valores absolutos de las diferencias asignando 1 al valor más pequeño, 2 al valor que le sigue, y así sucesivamente. En caso de que se encuentre un empate, entonces la clasificación será el promedio de las clasificaciones (Ej, supongamos que las diferencias empatan en los puestos 2 y 3, en este caso, a cada diferencia se le asigna el valor de 2.5).
- T^+ se define como la suma de los rangos de las diferencias positivas
- T^- se define como la suma de los rangos de las diferencias negativas

Definimos la prueba de hipótesis cómo:

- H_0**:** Las distribuciones poblacionales de X y Y son iguales.
- H_1**:** Las distribuciones poblacionales de X y Y son diferentes.

Estadístico de prueba:

$$T = min(T^+, T^-)$$

Se rechaza la hipótesis nula si el valor de significancia es mayor al valor-p obtenido en la prueba. Teoría tomada del capítulo de Estadística no Paramétrica de Wackerly *et al.* (2010)

Cuadro 2: Resultados mediante el Test de Rangos Barcelona VS Otras provincias

Provincia	**Valor-p**	**Provincia**	**Valor-p**	**Provincia**	**Valor-p**	**Provincia**	**Valor-p**
alava5	0	lugo31	0	cordoba18	0	salamanca43	9.4229E-121
albacete6	1.45116E-12	lleida32	1.2007E-113	coruña19	0	tenerife44	0
alicante7	0	madrid33	9.72591E-09	creal20	1.0654E-101	cantabria45	9.4776E-264
almeria8	0	malaga34	0	cuenca21	1.28306E-49	segovia46	5.3117E-268
avila9	0	melilla35	0	gerona22	4.38654E-90	sevilla47	0
badajoz10	0	murcia36	0	granada23	5.1669E-218	soria48	0
baleares11	0	orense37	5.42907E-61	guadalajara24	2.7001E-202	tarragona49	0
burgos13	0	asturias38	0	guipuzcoa25	0	teruel50	8.8145E-39
caceres14	7.2171E-120	palencia39	1.0919E-237	huelva26	0	toledo51	7.5512E-147
cadiz15	2.7653E-202	laspalmas40	0	huesca27	1.20028E-95	valladolid52	3.022E-100
castellon16	0	navarra41	3.3171E-203	jaen28	6.56737E-67	valencia53	0
ceuta17	2.18726E-51	pontevedra42	1.5165E-99	leon29	0	vizcaya54	2.04687E-68
rioja30	1.79712E-27	zamora55	1.87617E-69				

De los resultados, podemos ver que no hay suficiente información para rechazar la hipótesis de que la distribución de temperatura en Barcelona es igual a la distribución el las demás provincias.

3. Método de Kolmogorov-Smirnoff para determinar que provincias son distintas a la elegida

El método de Kolmogorov-Smirnov, toma dos muestras, y prueba si ambas provienen de la misma distribución. Cabe aclarar que en ningún momento se especifica la distribución que siguen.

Para esto, se toma la distribución empírica de cada conjunto de datos. Sea E_1 y E_2 las distribuciones empíricas de la muestra 1 y 2 respectivamente. Estamos interesados en comparar las distribuciones:

$$D = |E_1 - E_2| \tag{1}$$

De manera más formal, podemos ver el test de Kolmogorov-Smirnov como:

H_0 : *Ambas muestras provienen de la misma distribución*
H_1 : *Ambas muestras no provienen de la misma distribución*

Estadístico:

$$D = |E_1 - E_2|$$

Se rechaza la hipótesis nula, siempre y cuando el nivel de significancia (α) sea mayor al valor-p de la prueba[1].

Cuadro 3: Kolmogorov-Smirnov test: Temperatura de Barcelona VS otras provincias

Provincia	**Valor-p**	**Provincia**	**Valor-p**	**Provincia**	**Valor-p**	**Provincia**	**Valor-p**
alava5	0	rioja30	0	cadiz15	0	palencia39	0
albacete6	0	lugo31	0	castellon16	0	laspalmas40	0
alicante7	0	lleida32	0	ceuta17	0	navarra41	0
almeria8	0	madrid33	0	cordoba18	0	pontevedra42	0
avila9	0	malaga34	0	coruña19	0	salamanca43	0
badajoz10	0	melilla35	0	creal20	0	tenerife44	0
baleares11	0	murcia36	0	cuenca21	0	cantabria45	0
burgos13	0	orense37	0	gerona22	0	segovia46	0
caceres14	0	asturias38	0	granada23	0	sevilla47	0
guadalajara24	0	soria48	0	vizcaya54	2.68E-13	toledo51	0
guipuzcoa25	0	tarragona49	0	zamora55	0	valladolid52	0
huelva26	0	teruel50	0	zaragoza56	0	valencia53	0
huesca27	0	jaen28	0	leon29	0		

Del cuadro anterior, podemos concluir que no hay evidencia suficiente para decir que la distribución de la temperatura de alguna otra provincia en España es diferente a la distribución de temperatura de Barcelona.

4. Identificar mediante regresión no paramétrica, qué temperaturas explican la elegida

Una regresión no paramétrica es una forma de análisis de la regresión en el que el predictor no tiene una forma predeterminada, sino que se construye de acuerdo a la información derivada de los datos. La regresión no paramétrica requiere tamaños de muestra más grandes que los de una regresión sobre la base de modelos paramétricos porque los datos deben suministrar la estructura del modelo, así como las estimaciones del modelo.[2]

En la regresión no paramétrica, al igual que en su versión paramétrica, estamos interesados en explicar una variable Y en términos de otras variables $X_1, X_2, ..., X_n$ de la forma:

$$Y = r(X_i) + \epsilon_i \tag{2}$$

[1] https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/ks2samp.htm
[2] $https://es.wikipedia.org/wiki/Regresi\%C3\%B3n_no_param\%C3\%A9trica$

En el caso no paramétrico, consideramos como función r suavizadores lineales, algunos métodos no paramétricos son los métodos de regresión local y los métodos de suavización. Por lo tanto, definimos que un estimador $\hat{r}_n$ de r es un suavizador lineal si existe un vector $l(x) = (l_1(x), l_2(x), ..., l_n(x))^T$ tal que:

$$\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i \tag{3}$$

Tenemos el vector de valores ajustados r como: $r = (r_n(x_1), r_n(x_2), ..., r_n(x_n))^T$. Así, podemos decir que $r = LY$.

Donde L es una matriz de dimensión nxn, donde cada entrada es $l_{ij} = l_j(x_i)$, llamada matriz de suavización. Es importante aclarar que los suavizadores lineales son diferentes a la regresión lineal.

Los pesos en todos los suavizadores, tienen la propiedad:

$$\forall x, \quad \sum_{i=1}^{n} l_i(x) = 1 \tag{4}$$

Tomado de la sección de regresión no paramétrica de Wassermann (2006).

4.1. Regresión

Una de las aproximaciones a la regresión no paramétrica, es a la vez una aproximación no paramétrica y robusta de la siguiente forma:

- Vamos a asumir que la función de suavización $r(X) = X\beta$. La aproximación robusta y no paramétrica de cada β esta dada por:

$$\beta_i = \rho_k \frac{Mad(ProvinciaBarcelona)}{Mad(Provincia_i)}$$

  - Donde: ρ_k es el coeficiente de correlación de Kendall y $Mad(X)$ se define como:

$$Mad(X) = Median(X - Median(X))$$

- Utilizar la técnica de Boostrap para encontrar un intervalo de confianza no paramétrico para cada β_i y así determinar si la temperatura de las otras provincias explican la provincia elegida.

Cabe aclarar que para el caso trabajado en el taller, consideramos 1000 repeticiones del proceso para utilizar la técnica de Bootstrap, además de recordar los datos obtenidos en los percentiles 2.75 y 97.5 respectivamente para poder construir un intervalo de confianza al con una confianza del 95 %.

Cuadro 4: Regresión no paramétrica, con Intervalos de confianza no Paramétrico (Intervalo Boostrap usando 1000 repeticiones)

Provincia	Coeficiente	Limite Inferior I.C	Limite Superior I.C	Provincia	Coeficiente	Limite Inferior I.C	Limite Superior I.C
albacete6	0.456392193	0.456345072	0.459595067	lleida32	0.494019188	0.493971829	0.49409136
alicante7	0.783981377	0.783922801	0.784093888	madrid33	0.455501112	0.455445928	0.455587012
almeria8	0.765098577	0.765013073	0.765239101	malaga34	0.744085416	0.744007268	0.744241956
avila9	0.499289894	0.499227455	0.499385701	melilla35	0.814415403	0.814330861	0.814570392
badajoz10	0.487312282	0.479757547	0.487380201	murcia36	0.610403358	0.610352444	0.610504666
baleares11	0.8271998	0.818084417	0.827269696	orense37	0.513822003	0.513726332	0.513945031
burgos13	0.48930587	0.489241342	0.489402548	asturias38	0.696928195	0.696791678	0.697122744
caceres14	0.480621646	0.48055816	0.480719181	palencia39	0.481375059	0.481310536	0.481474388
cadiz15	0.863785448	0.863672486	0.863956706	laspalmas40	1.156976936	1.156734263	1.157432414
castellon16	0.722576961	0.722526163	0.722670304	navarra41	0.525312865	0.525245408	0.525411522
ceuta17	0.982982078	0.982889687	0.983163164	pontevedra42	0.650299979	0.650167903	0.65047545
cordoba18	0.441284385	0.441231233	0.441372285	salamanca43	0.474288139	0.474224861	0.474384047
coruña19	0.913994611	0.913838824	0.914247976	tenerife44	1.029043152	1.028878542	1.029324947
creal20	0.432157811	0.432102877	0.432238403	cantabria45	0.794224379	0.783510791	0.794419942
cuenca21	0.445657678	0.445598219	0.445744481	segovia46	0.479945927	0.479883268	0.480046527
gerona22	0.689596756	0.689545255	0.689686113	sevilla47	0.502027825	0.501967362	0.502119632
granada23	0.462675469	0.462622994	0.462764101	soria48	0.474354541	0.47428835	0.474445351
guadalajara24	0.451244923	0.451184857	0.451330959	tarragona49	0.617398363	0.617343833	0.617486763
guipuzcoa25	0.706184736	0.697433932	0.706338056	teruel50	0.477794173	0.477739997	0.481344683
huelva26	0.619168122	0.619086488	0.619284774	toledo51	0.445223627	0.445169344	0.445299871
huesca27	0.493812425	0.491335516	0.498597586	valladolid52	0.457785428	0.45772913	0.45786964
jaen28	0.475902175	0.475844464	0.475990607	valencia53	0.738853721	0.73878851	0.738969802
leon29	0.489640435	0.489565146	0.48975204	vizcaya54	0.658003752	0.657877265	0.658181456
rioja30	0.495941107	0.495885438	0.496033204	zamora55	0.477658151	0.477596484	0.4777611
lugo31	0.612237414	0.612123693	0.61239428	zaragoza56	0.494823622	0.494773946	0.494903471
alava5	0.555060978	0.554976698	0.555167259				

Según los intervalos de confianza de cada coeficiente en la regresión no paramétrica, es posible observar que en ningún intervalo de confianza al 95 % hay un cambio de signo y el valor 0 no está incluido, por lo que podríamos decir que, con esta aproximación de regresión no paramétrica, las temperaturas de todas las provincias, explican la temperatura de Barcelona

5. Identificar mediante regresión robusta, qué temperaturas explican la elegida

La regresión robusta es una alternativa en vez del método de mínimos cuadrados, cuando dentro de nuestro conjunto de datos tenemos presencia de datos atípicos o outliers[3]. Una aproximación robusta a la regresión, es por el método de Estimación-M. Este método defina una función de peso tal que la ecuación a estimar es:

$$\sum_{i=1}^{n} w_i(y_i - x'b)x_i' = 0 \tag{5}$$

Acá los pesos dependen de los residuales, y los residuales dependen de los pesos. La ecuación es

[3]https://stats.idre.ucla.edu/r/dae/robust-regression/

resuelta por el método de mínimos cuadrados reutilizados, donde para la iteración j, el resultado de la estimación es:

$$B_j = (X'W_jX)^{-1}X'W_jY \tag{6}$$

Cuadro 5: Regresión robusta explicando la temperatura de España utilizando conjuntamente la información de las demás provincias

Variable	**Value**	**Std. Error**	**t-value**	**Variable**	**Value**	**Std. Error**	**t-value**
(Intercept)	-31.4346	2.2129	-14.2051	lleida32	0.0521	0.0123	4.252
alava5	-0.0033	0.0158	-0.2066	madrid33	0.012	0.0198	0.6062
albacete6	-0.089	0.0158	-5.6409	malaga34	-0.0017	0.0102	-0.1674
alicante7	0.1004	0.0168	5.9732	melilla35	0.0691	0.0123	5.6037
almeria8	0.0275	0.0102	2.7008	murcia36	-0.0436	0.0148	-2.9528
avila9	0.0357	0.0174	2.0504	orense37	0.0148	0.0131	1.1288
badajoz10	-0.0782	0.018	-4.3339	asturias38	-0.0551	0.0132	-4.178
baleares11	0.2141	0.0121	17.6981	palencia39	-0.036	0.0104	-3.4732
burgos13	0.015	0.0141	1.0604	laspalmas40	0.0039	0.0146	0.2687
caceres14	0.0715	0.0215	3.33	navarra41	0.0073	0.0149	0.4916
cadiz15	0.0052	0.0133	0.392	pontevedra42	-0.0086	0.0118	-0.7235
castellon16	0.2856	0.017	16.8043	salamanca43	-0.0837	0.0188	-4.4575
ceuta17	0.0618	0.0178	3.4678	tenerife44	0.0596	0.0141	4.2174
cordoba18	0.0236	0.0196	1.2048	cantabria45	0.0397	0.0152	2.6168
coruña19	0.0455	0.0134	3.3873	segovia46	0.0024	0.0191	0.1278
creal20	-0.0138	0.0169	-0.8169	sevilla47	0.0046	0.0206	0.223
cuenca21	0.0049	0.0168	0.2945	soria48	-0.0221	0.0154	-1.4335
gerona22	0.36	0.0101	35.5165	tarragona49	0.003	0.0126	0.2394
granada23	0.0293	0.015	1.9592	teruel50	0.0015	0.0143	0.102
guadalajara24	-0.0267	0.0166	-1.6157	toledo51	0.0458	0.0197	2.3197
guipuzcoa25	-0.0622	0.0138	-4.4889	valladolid52	0.0353	0.0203	1.7406
huelva26	0.018	0.0146	1.23	valencia53	-0.0568	0.0155	-3.6672
huesca27	0.006	0.0107	0.5594	vizcaya54	0.0886	0.0154	5.7692
jaen28	-0.0742	0.0178	-4.1775	zamora55	0.0353	0.0172	2.0557
leon29	-0.0016	0.0127	-0.1234	zaragoza56	-0.0349	0.0147	-2.3673
rioja30	-0.0363	0.0142	-2.5602	zaragoza56	-0.0349	0.0147	-2.3673
lugo31	0.0012	0.013	0.0886				

De los resultados de la regresión robusta, sabemos que valores t grandes, están asociados a variables estadisticamente significativas a nuestro modelo, por lo tanto podemos ver que las temperaturas de las provincias: **lleida32, albacete6, alicante7, melilla35, murcia36, badajoz10, asturias38, baleares11, palencia39, caceres14, castellon16, salamanca43, ceuta17, tenerife44, coruña19, gerona22, guipuzcoa25, valencia53, vizcaya54, jaen28**, son las que explican la distribución de temperatura en Barcelona.

6. Punto adicional

Sean $X_1, X_2, ..., X_n$ una muestra de cierta distribución. Podemos denotar $X_{[1]}, X_{[2]}, ..., X_{[n]}$ donde $X_{[i]}$ se conoce como el i-ésimo estadístico ordenado. Como casos particulares, podemos ver que

- $X_{[1]} = min(X_1, X_2, ..., X_n)$
- $X_{[n]} = max(X_1, X_2, ..., X_n)$

Sabemos que, dada una muestra aleatoria $X_1, X_2, ..., X_n$ de una misma distribución F, su función de distribución conjunta está dada por:

$$f(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} f(x_i) \tag{7}$$

Donde $f(x_i)$ es la función de densidad de X_i.Ahora, todas las posibles combinaciones ordenadas (permutaciones) de $X_1, X_2, ..., X_n$ puede darse $n!$ veces. Es decir, la función de distribución de todos los estadísticos de orden está dada por:

$$f(X_{[1]}, X_{[2]}, ..., X_{[n]}) = n! \prod_{i=1}^{n} f(x_i) \tag{8}$$

Ahora, con base en la distribución conjunta de todos los estadísticos de orden, la función de distribución conjunta de r estadísticos de orden está dada por:

$$f(X_{[1]}, X_{[2]}, ..., X_{[r]}) = [1 - F(x_r)]^{n-r} \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(x_i) \tag{9}$$

Finalmente, queriendo expresar la función de distribución conjunta de 2 estadísticos ordenados. Sean entonces $X_{[r]}, X_{[s]}$ dos estadísticos ordenados tal que $r < s$. Su distribución conjunta está dada por:

$$P(X_{[r]} \leq x \wedge X_{[s]} \leq y) = \int_0^{F_r(x)} \int_x^{F_s(y)} \frac{n! x^{r-1} (y-x)^{s-r-1} (1-y)^{n-s}}{(r-1)!(s-r-1)!(n-s)!} dxdy \tag{10}$$

Teoría tomada del capitulo de estadísticos de orden de Shahbaz *et al.* (2016).

7. Conclusiones

- Existen diferentes métodos para la estimación de densidades de un conjunto de datos. Para el caso particular de la distribución de temperatura en Barcelona, es posible concluir que tanto la aproximación sencilla de un histograma de frecuencias como la estimación de la densidad a través de kernels, dan un buen desempeño para ajustar la densidad de los datos.
- Al utilizar técnicas no paramétricas de tests de homogeneidad como lo son el Kolmogorov-Smirnov y el método de rangos, es posible concluir que la distribución de las temperaturas en la provincia de Barcelona, se asemeja a la distribución de temperaturas de las demás provincias consideradas.

- Es interesante considerar diferentes técnicas de regresión tales como versiones robustas y no paramétricas para tratar de explicar variables, sin embargo, en el caso que trabajamos en este taller, la aproximación robusta nos plantea que las variables significativas en el modelo de regresión son lleida32, albacete6, alicante7, melilla35, murcia36, badajoz10, asturias38, baleares11, palencia39, caceres14, castellon16, salamanca43, ceuta17, tenerife44, coruña19, gerona22, guipuzcoa25, valencia53, vizcaya54, jaen28, mientras que en la versión no paramétrica, encontramos que todas las variables consideradas son significativas para el modelo. Por lo que sería interesante proponer un método para determinar de una manera más precisa la significancia estadística de las variables para el método no paramétrico, así se poder contrastar si todas las variables significativas de la regresión robusta también lo son en la no paramétrica.

Referencias

Shahbaz, Muhammad Qaiser, Ahsanullah, Mohammad, Shahbaz, Saman Hanif, & Al-Zahrani, Bander M. 2016. *Ordered Random Variables: Theory and Applications*. Springer.

Wackerly, Dennis D, Muñoz, Romo, Humbertotr, Jorge, *et al.* . 2010. *Estadística matemática con aplicaciones*.

Wassermann, Larry. 2006. All of nonparametric statistics. *New York*.