UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Selección de un modelo de regresión lineal múltiple para el cálculo de la precipitación media en verano.

Apellidos, nombre	Balaguer Beser, Angel[1] (abalague@mat.upv.es) Ruiz Fernández, Luis Ángel[2] (laruiz@cgf.upv.es)
Departamento	[1]Departamento de Matemática Aplicada [2]Departamento de Ingeniería Cartográfica, Geodesia y Fotogrametría
Centro	E.T.S.I. Geodésica, Cartográfica y Topográfica

1 Resumen de las ideas clave

En este artículo se comparan distintos procedimientos para la selección de modelos de regresión lineal múltiple usando datos reales. Se aplican los métodos de selección paso a paso hacia adelante y selección paso a paso hacia atrás. Se trabaja con modelos cuyas variables son estadísticamente significativas con un Valor-P inferior a 0.05. Se selecciona el mejor modelo en función del coeficiente R-cuadrado ajustado, la raíz del error cuadrático medio (RMSE en las siglas en inglés) y el error absoluto medio (MAE en siglas en inglés), el criterio de información de Akaike, el criterio Bayesiano de Schwarz-Bayesian y el criterio de Hannan-Quinn. También se analizan los residuos de cada modelo para verificar si se cumplen las hipótesis de linealidad, homocedasticidad, independencia y normalidad de los residuos. Esta metodología se aplica para obtener modelos de regresión en la predicción de la precipitación media durante los meses del verano meteorológico en el territorio de la Comunitat Valenciana (España) y áreas adyacentes, usando algunas variables de carácter geográfico y topográfico descritas en Portalés et al. (2010). Para ello se utiliza el programa Statgraphics Centurion XVII.

2 Objetivos

Después de leer con detenimiento este documento, el lector será capaz de:

- Proponer modelos de regresión lineal múltiple para analizar de manera conjunta la influencia de varias variables cuantitativas sobre un fenómeno a estudiar.
- Partiendo de un conjunto de variables continuas, analizar cuáles son estadísticamente significativas para explicar la variable respuesta mediante un modelo de regresión lineal múltiple.
- Calcular los modelos de regresión lineal múltiple usando el método de selección paso a paso hacia adelante, la selección paso a paso hacia atrás, el criterio de información de Akaike, el criterio Bayesiano de Schwarz-Bayesian y el criterio de Hannan-Quinn.
- Obtener el mejor modelo de regresión lineal múltiple con variables estadísticamente significativas, teniendo en cuenta el número de variables explicativas y los valores de la R-cuadrado ajustada, RMSE y MAE.
- Comprobar si los residuos del modelo obtenido cumplen las hipótesis de normalidad, linealidad, homocedasticidad e independencia.
- Estimar el valor esperado de la respuesta para unos valores prefijados de las variables explicativas.
- Obtener un modelo explicativo de la precipitación media usando los datos registrados en dicha variable en un conjunto de observatorios meteorológicos durante un periodo de tiempo, junto con un conjunto de variables de carácter geográfico y topográfico para usar como variables explicativas.
- Usar el programa Statgraphics Centurion XVII para realizar los cálculos necesarios para alcanzar los objetivos anteriores.

3 Introducción

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes o explicativas $(X_1, X_2, \cdots, X_k)$ usando la siguiente ecuación (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

siendo β_i los parámetros que se tendrán que estimar. Para calcular dichos coeficientes se usa una tabla de datos formada por observaciones de la variable respuesta y todas las variables explicativas en un conjunto de "n" individuos o lugares. De esta forma, se dispone de una matriz de datos formada por "n" filas y "k+1" columnas. El requisito para poder hacer los cálculos es que el número de datos tiene que ser igual o mayor que "k+2" (número de variables explicativas + 2) y ninguna de las variables explicativas puede ser combinación lineal exacta de las restantes.

Los modelos de regresión lineal múltiple que usan variables explicativas que están muy correlacionadas pueden conducir a cálculos imprecisos de los parámetros a estimar en el modelo (1), siendo el caso extremo cuando una variable es combinación lineal de las otras, el cual recibe el nombre de **multicolinealidad**. Para determinar si nuestro modelo sufre de multicolinealidad podemos construir una matriz donde se muestren los coeficientes de correlación, de unas variables con otras. En aquellos casos en los que observemos valores de correlación altos, podremos sospechar que existe multicolinealidad.

Los modelos de regresión múltiple pueden emplearse para estimar el valor de la variable respuesta conocidos los valores de las variables explicativas en otra localización o individuo diferente. Un caso de aplicación se da en el del cálculo de la precipitación media en una localización espacial, la cual depende en gran medida de factores geográficos y topográficos (véase Marquínez et al. (2003)). Dado que la precipitación generalmente se conoce sólo en ciertos lugares, se necesitan procedimientos para estimar esta variable en otros puntos. En este trabajo se muestra el procedimiento de obtención de un modelo de regresión lineal múltiple para este tipo de variables respuesta.

4 Desarrollo

Para entender la metodología que se expone a continuación se necesitan algunos conocimientos básicos de estadística descriptiva y de álgebra lineal sobre el método de mínimos cuadrados, los cuales se pueden adquirir a través de las referencias Balaguer-Beser et al (2014) y Marín-Molina et al. (2012).

En este trabajo se explica el proceso de obtención de un modelo de regresión lineal múltiple utilizando el programa Statgraphics Centurion XVII. Como observaciones de la variable respuesta del modelo de regresión se han usado datos de precipitación media durante los meses de junio, julio y agosto, en 212 observatorios meteorológicos de la Comunitat Valenciana (España) y territorios adyacentes. Dichas medias han sido calculadas a partir de datos de precipitación obtenidos durante 35 años, desde 1960 hasta 2005, los cuales fueron suministrados a los autores de este trabajo por la Agencia Estatal de Meteorología (AEMET). Las variables explicativas serán las variables de carácter geográfico y topográfico que aparecen en la tabla 1, en la cual puede verse una breve descripción de las mismas. El proceso de obtención de dichas variables está explicado con detalle en la referencia Portalés et al. (2010).

Nombre	Significado
XUTM	Longitud en el sistema de coordenadas UTM
YUTM	Latitud en el sistema de coordenadas UTM
COAST	Distancia mínima al mar Mediterráneo
Z5	Elevación media dentro de un área circular de 5 km de radio
Z10	Elevación media dentro de un área circular de 10 km de radio
D5	Diferencia de altura entre los puntos más altos y más bajos dentro de un área circular de 5 km de radio.
D10	Diferencia de altura entre los puntos más altos y más bajos dentro de un área circular de 10 km de radio.
VS_N	(Producto escalar normalizado del vector que apunta a la dirección norte y el vector en la dirección del flujo sinóptico en la superficie + 1) / 2.
ALTDIF	Diferencia de elevación entre el punto más alto de una ladera y el punto más alto dentro de un área orientada dentro de la dirección del flujo sinóptico en 850 hPa

Tabla 1. Nombre de las variables explicativas usadas en los modelos de regresión junto con una explicación de su significado.

El modelo de regresión obtenido a partir de las variables dadas en la tabla 1 se usará para obtener una estimación de dicha precipitación media en verano en otros lugares donde se conozca el valor de dichas variables explicativas. Los pasos que se siguen para ello se recogen en los siguientes apartados.

4.1 Estudio de la relación lineal entre las variables. Variables estadísticamente significativas.

Antes de proceder al cálculo de los modelos de regresión lineal múltiple es importante analizar si existe relación lineal entre las variables de la tabla 1 (variables explicativas) y la variable precipitación media en verano (variable respuesta). También habrá que analizar la relación entre las variables explicativas para evitar problemas de multicolinealidad tal y como se ha comentado en la introducción.

Con Statgraphics Centurion XVII se puede usar el procedimiento: **Describir + Datos Numéricos + Análisis multivariado**. Luego se elige el gráfico de correlación y obtenemos la imagen que aparece en la tabla 2. En ella aparecen los coeficientes de correlación de Pearson entre las variables. En rojo se han señalado aquellos pares de variables que tienen un índice de correlación más cercano a 1 (correlación lineal positiva) y en azul los que tienen un índice más cercano a -1 (correlación lineal negativa). Vemos que las variables que presentan una correlación más alta con la precipitación media en verano son YUTM, Z5 y Z10. Sin embargo, la relación lineal entre estas dos últimas variables es casi perfecta y positiva (correlación de Pearson igual a 0.99) razón por la cual no podremos incorporar las Z5 y Z10, al mismo tiempo, como variables explicativas en el modelo de regresión pues tendremos multicolinealidad. También debemos tener cierta precaución al considerar las variables Coast y Z5 (o Z10) en el mismo modelo pues tienen un índice de correlación de Pearson igual a 0.83 (0.84 con Z10).

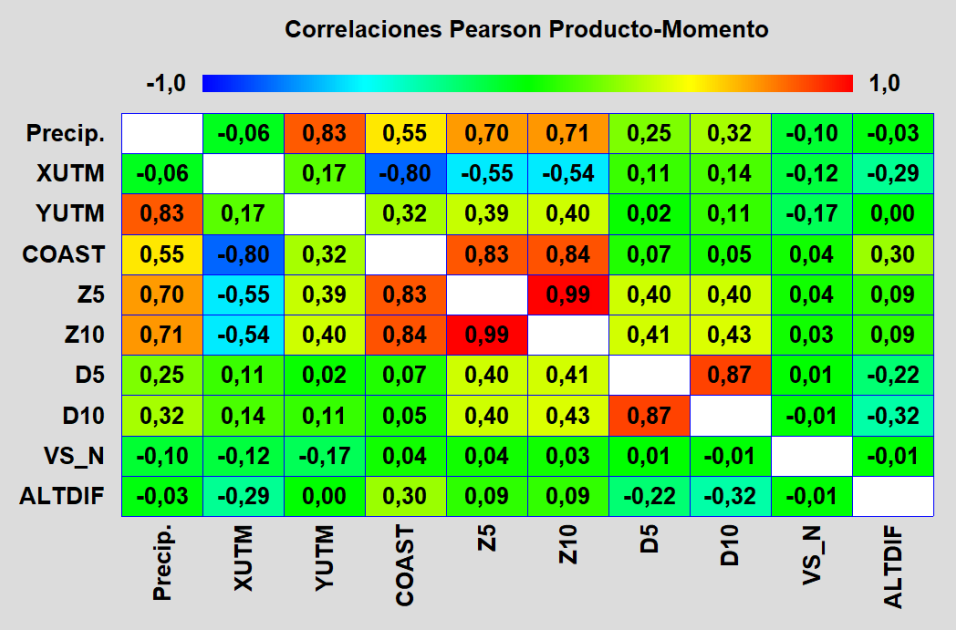


Tabla 2: Valores de los índices de correlación de Pearson entre todas las variables. Precip. indica la precipitación media en los meses de junio, julio y agosto. La descripción del resto de variables se muestra en la tabla 1.

El procedimiento **Relacionar + Varios factores + Regresión Multiple** de Statgraphics Centurion permite calcular el modelo de regresión mediante el método de mínimos cuadrados ordinarios, los cuales calculan los coeficientes del modelo (1) minimizando la suma de cuadrados de los errores entre los valores observados de precipitación y los estimados por dicho modelo. Como variables explicativas eliminamos la variable Z10 debido a su alta correlación con Z5. Después de introducir dichas variables explicativas en el modelo de regresión de la Precipitación media en verano y aplicar el método de mínimos cuadrados ordinarios, se obtienen la estimación de los coeficientes que aparece en la tabla 3. Dicha estimación, junto con el error estándar de cada coeficiente permite calcular el estadístico T, el cual tiene asociado el Valor-P que aparece en la tabla 3. Aquellos coeficientes con un Valor-P inferior a 0,05 son estadísticamente significativos con un 95% de confianza. El que tiene un Valor-P más alto es VS_N y al ser superior a 0,05 deberíamos eliminar esta variable del modelo.

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	-795,062	42,2769	-18,8061	0,0000
XUTM	0,000158879	0,0000496081	3,20269	0,0016
YUTM	0,000166446	0,00001483	11,2236	0,0000
COAST	0,000185876	0,0000792126	2,34655	0,0199
Z5	0,0300282	0,00461042	6,51311	0,0000
D5	-0,00682133	0,00592263	-1,15174	0,2508
D10	0,00795474	0,00518911	1,53297	0,1268
VS_N	-0,388288	2,1133	-0,183735	0,8544
ALTDIF	-0,00475853	0,00293364	-1,62206	0,1063

Tabla 3: Estimación de los coeficientes del modelo de regresión (1) para la variable precipitación media en verano, junto con el error estándar, el estadístico T y el Valor-P, usando el método de mínimos cuadrados ordinarios.

Sin embargo, después de eliminar VS_N nos quedan otras variables que no son estadísticamente significativas (D5, D10 y ALTDIF) y el proceso de eliminar una a una de forma manual puede ser largo si tenemos muchas variables. Para evitar esto podemos usar alguno de los métodos de selección de variables explicativas que se explican en el siguiente apartado.

4.2 Obtención de los modelos de regresión con una selección de variables explicativas.

En el procedimiento **Relacionar + Varios factores + Regresión Multiple** de Statgraphics Centurion se puede elegir una de estas dos Opciones de Análisis.

• **Selección paso a paso hacia atrás.** Comienza con un modelo que involucra todas las variables y elimina del modelo la variable que es estadísticamente menos significativa. El proceso elimina una variable en cada paso hasta que los Valores-P de todas las variables que quedan son inferiores al Valor-P especificado para "quitar". Además, las variables eliminadas del modelo con anterioridad mediante este procedimiento pueden ser reingresadas más tarde si sus Valores-P son inferiores al valor especificado por el usuario para "agregar". La tabla 4 muestra las variables que quedan en el modelo dado en la tabla 3 después de aplicar este criterio, usando el criterio del Valor-P con el mismo α=0,05 para quitar y para agregar. Con este criterio, las variables VS_N, D5 y D10 han sido eliminadas del modelo (en ese orden) y todas las variables que quedan son estadísticamente significativas.

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	-802,101	39,1347	-20,4959	0,0000
XUTM	0,000157041	0,0000462488	3,39556	0,0008
YUTM	0,000168918	0,0000134199	12,5871	0,0000
COAST	0,000170243	0,0000783867	2,17183	0,0310
Z5	0,0314981	0,00371926	8,46893	0,0000
ALTDIF	-0,00579415	0,00282509	-2,05096	0,0415

Tabla 4. Estimación de los coeficientes del modelo de regresión (1) para la variable precipitación media en verano, junto con el error estándar, el estadístico T y el Valor-P, usando el método de selección paso a paso hacia atrás.

• **Selección paso a paso hacia adelante.** Comienza con un modelo que involucra un solo término constante y en cada paso, el algoritmo introduce en el modelo la variable que será estadísticamente la más significativa si se ingresa. La variable más significativa será introducida dentro del modelo mientras que tenga un Valor-P para "agregar" menor o igual al especificado en el cuadro de diálogo. Además, las variables introducidas en el modelo pueden ser eliminadas más tarde si sus Valores-P están por encima del criterio especificado para "quitar". La tabla 5 muestra los resultados obtenidos con este procedimiento y α=0,05.

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	-819,47	38,6971	-21,1765	0,0000
XUTM	0,0000831609	0,0000235596	3,52981	0,0005
YUTM	0,00018579	0,0000102707	18,0894	0,0000
Z5	0,0376395	0,00270316	13,9243	0,0000

Tabla 5. Estimación de los coeficientes del modelo de regresión (1) para la variable precipitación media en verano, junto con el error estándar, el estadístico T y el Valor-P, usando el método de selección paso a paso hacia adelante.

El método de selección paso a paso hacia adelante ha elegido un conjunto de variables explicativas más pequeño que el de la selección hacia atrás, aunque los dos modelos (tablas 4 y 5) usan un conjunto de variables que son estadísticamente significativas. ¿Qué modelo elegimos para estimar la precipitación media en verano en otro punto de la Comunitat Valenciana? Necesitaremos algunas herramientas estadísticas para evaluar los dos ajustes de regresión.

4.3 Evaluación del ajuste de regresión.

Conocida una estimación de los parámetros del modelo de regresión (1), $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$, se pueden calcular los errores (residuos) cometidos en cada observación usando la fórmula (2),

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \tag{2}$$

siendo y_i los valores observados de la variable respuesta y $x_{i1}, \cdots, x_{ik}$ los valores de las variables explicativas del modelo en la localización número "i". Con ello se puede calcular:

- **La raíz del error cuadrático medio** (RMSE por sus siglas en inglés). Se calcula con la fórmula: RMSE=$\sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-k-1}}$, siendo "n" el número de datos y "k" el número de variables explicativas.
- **El error medio absoluto** (MAE por sus siglas en inglés), el cual indica el error en promedio en la predicción de la variable respuesta. Su fórmula es: MAE=$\frac{\sum_{i=1}^{n} |\hat{e}_i|}{n}$.
- **R-cuadrado ajustada por los grados de libertad,** que representa el porcentaje de variabilidad de la respuesta que se ha explicado mediante el modelo ajustado de regresión, teniendo en cuenta el número de variables explicativas. La fórmula para calcularla es: $\bar{R}^2 = 100 \cdot \left(1 - \frac{(RMSE)^2}{\hat{S}_y{}^2}\right)$, siendo $\hat{S}_y{}^2$ el estimador insesgado de la varianza de los valores respuesta. A diferencia de los anteriores, este parámetro no depende de las unidades de medida de la variable respuesta.

En la ventana de Statgraphics, "resumen del análisis" obtenida con el procedimiento: **Relacionar + Varios factores + Regresión Múltiple**, se pueden encontrar los valores de la R-cuadrado ajustada, RMSE y MAE. Statgraphics denomina el RMSE como "error estándar del estadístico". La tabla 6 muestra los resultados obtenidos para los modelos de las tablas 4 y 5. En ella se observa que el modelo definido por los coeficientes de la tabla 4 tiene unos errores medios de RMSE y MAE más pequeños, junto con un mayor valor de R-cuadrado ajustada, aunque en el mismo intervienen dos variables más que en modelo definido por la tabla 5. De todas formas, la diferencia entre los resultados obtenidos con ambos modelos no es muy alta. ¿Qué modelo elegimos, el definido por la tabla 4 que tiene errores medios más pequeños y $\bar{R}^2$ mayor, pero con mayor número de variables? O tal vez, ¿es mejor elegir el modelo de la tabla 5 pues consigue una precisión similar con sólo tres variables explicativas?

Modelo de regresión definido en la tabla 4	**Modelo de regresión definido en la tabla 5**
RMSE=9,94027, MAE=7,36157, $\bar{R}^2 = 85{,}94\%$	RMSE =10,0567, MAE=7,38283, $\bar{R}^2 = 85{,}61\%$

Tabla 6. Valores de RMSE, MAE y R-cuadrado ajustada, obtenidos en los modelos de regresión definidos en las tablas 4 y 5. RMSE y MAE están obtenidos en unidades de (l/m^2).

Los valores de RMSE y MAE tienen que ser lo más pequeños posibles. Sin embargo, el valor de la R-cuadrado ajustada tiene que ser lo más cercano posible al 100%. El programa Statgraphics Centurion dispone de un procedimiento: **Relacionar + Varios factores + Selección de modelos de regresión,** mediante el cual se pueden calcular los valores de la R-cuadrado ajustada, para modelos con distinto número de variables explicativas. La gráfica de la figura 1 muestra cómo varía dicho valor cuando aumentamos el número de coeficientes en el modelo de regresión. El uso de las variables YUTM+Z5 proporciona un modelo con una $\bar{R}^2 = 84{,}82\%$, cercana a los valores mostrados en la tabla 6, pero con sólo tres coeficientes en el modelo.

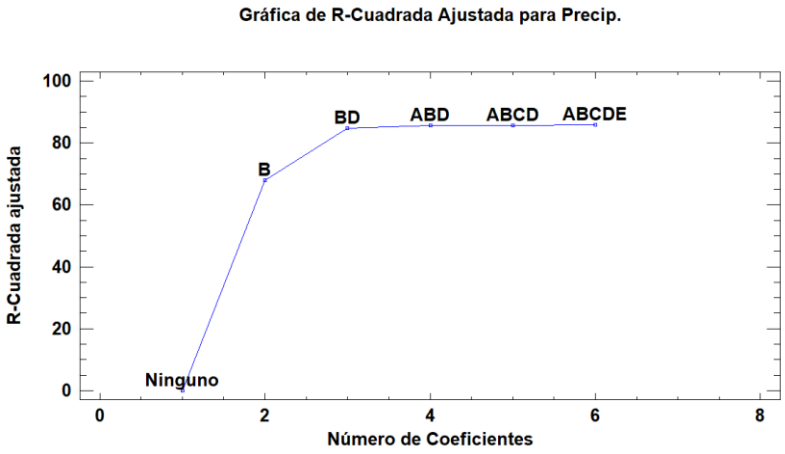


Figura 1: Valores de la R-cuadrado ajustada en función del número de coeficientes en el modelo de regresión, teniendo en cuenta la constante. Las letras significan estas variables: A=XUTM, B=YUTM, C=COAST, D=Z5, E=ALTDIF

Las fórmulas de la R-cuadrado ajustada y el RMSE tienen en cuenta el número de variables en el modelo. En general se busca elegir un modelo de regresión con el menor número de variables explicativas que tenga un buen nivel de precisión (R-cuadrado ajustada alta y RMSE bajo). Para ello también se pueden utilizar otros criterios de información cuya meta es seleccionar un modelo con el mínimo error residual y con tan pocos coeficientes como sea posible, relativo a la cantidad de datos disponibles en la muestra (véase Vrieze (2012)). El programa Statgraphics Centurion proporciona los resultados obtenidos con estos tres criterios de información:

- **El criterio de información de Akaike (AIC):** es una medida de la calidad relativa de un modelo estadístico que usa este indicador: *AIC*=$2\ Ln(RMSE) + \frac{2(k+1)}{n}$.
- **Hannan-Quinn Criterion (HQC):** $HQC = 2\ Ln(RMSE) + \frac{2(k+1)Ln(n)}{n}$
- **Schwarz-Bayesian Information Criterion (SBIC):** $SBIC = 2\ Ln(RMSE) + \frac{(k+1)Ln(n)}{n}$

La diferencia entre estos criterios se encuentra en el uso de una penalización diferente para el número de parámetros estimados *(k+1)* teniendo en cuenta el número de datos *(n).*

Siguiendo este procedimiento de Statgraphics: **Relacionar + Varios Factores + Selección de modelos de regresión**, en la tabla denominada "Mejor Criterio de información" se encuentran los resultados obtenidos para modelos obtenidos con diferente número de variables. El valor más bajo obtenido con el criterio AIC se obtiene con el modelo descrito en la tabla 4. En cambio, los valores menores con los otros dos criterios se consiguen con el modelo de la tabla 5. Estos dos últimos criterios tienden a elegir modelos con menos variables explicativas.

$(RMSE)^2$	*Coeficientes*	*AIC*	*HQC*	*SBIC*	*Variables incluidas*
98,809	6	4,64979	4,68819	4,74479	ABCDE
101,137	4	4,65422	4,67981	4,71755	ABD

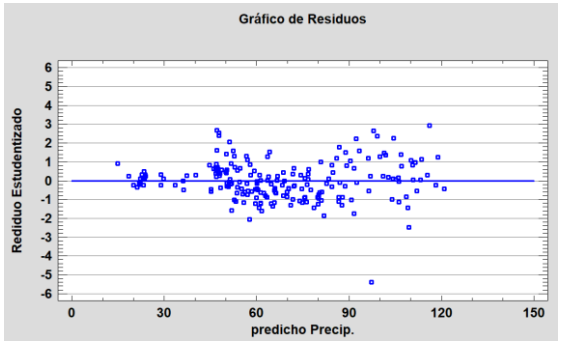
Tabla 7. Modelos con los criterios de información más bajos. El número de coeficientes incluye la constante. Las letras significan estas variables: A=XUTM, B=YUTM, C=COAST, D=Z5, E=ALTDIF. En rojo están señalados los valores más pequeños de cada criterio.

4.4 Verificación de las hipótesis de regresión.

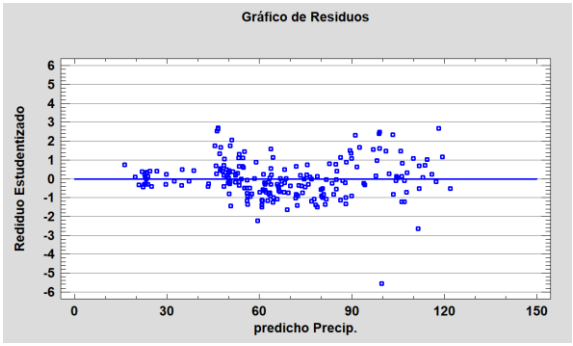
Los errores del modelo de regresión (residuos) tienen que verificar las siguientes hipótesis:

1. **Linealidad**: su media cero, es decir, $E(e_i) = 0$, para cada $i = 1, \cdots, n$
2. **Homocedasticidad**: Varianza constante para todo i, $\text{Var}(e_i) = \sigma^2$, $\forall i = 1, \cdots, n$
3. **Independencia:** $E(e_i e_j) = 0$, para todo $i \neq j$
4. **Normalidad:** $e_i \sim \text{Normal}(0, \sigma^2)$, para todo $i = 1, \cdots, n$.

Para analizar la **linealidad y homocedasticidad**, Statgraphics Centurion permite obtener la gráfica de los residuos versus el valor predicho por el modelo de regresión dentro del procedimiento **Relacionar + Varios factores + Regresión Múltiple**. Los residuos deberían estar dispuestos en una nube sin forma alrededor de la línea y=0. Además, la amplitud de los valores positivos y negativos debería mantenerse constante a medida que varía el valor de la predicción por el modelo. En el eje de las Y se puede elegir entre los residuos (sin modificar) y los residuos estudentizados, los cuales están en una escala diferente que permite analizar si existen valores anómalos. Residuos estudentizados superiores a 2 en módulo se consideran residuos atípicos. Pero, sobre todo conviene examinar las observaciones con residuos mayores a 3 en módulo para determinar si son valores anómalos que debieran ser eliminados. La figura 2 muestra los residuos estudentizados para los dos modelos calculados en apartados precedentes. En ambos modelos existen 12 residuos con valores mayores que 2 en módulo, pero sólo uno de ellos tiene un valor inferior a -3. Se trata de un lugar en zona de montaña en latitudes superiores, en el cual la precipitación observada ha sido muy inferior a la calculada por los modelos. Recordemos que en los dos modelos de regresión los coeficientes de YUTM y Z5 son positivos, indicando una mayor precipitación en lugares con mayor latitud y altitud.



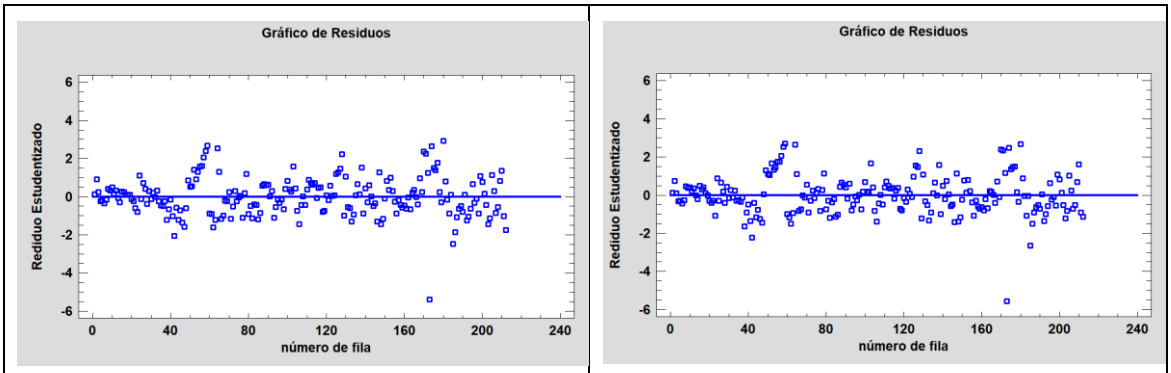
Modelo de regresión definido en la tabla 4



Modelo de regresión definido en la tabla 5

Figura 2. Residuos estudentizados en función de los valores predichos de precipitación media en verano para los modelos con los coeficientes descritos en las tablas 4 y 5, respectivamente.

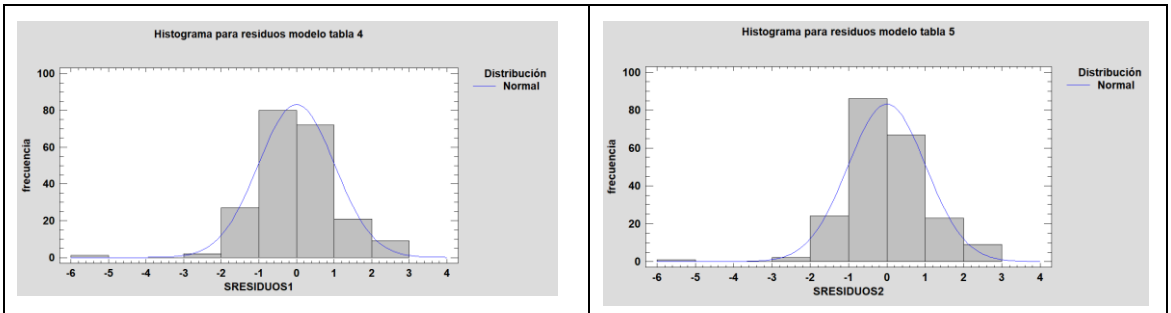
La independencia de los residuos se puede analizar observando el gráfico de los residuos estudentizados versus el número de fila, en la cual debería aparecer una nube de puntos sin forma. La figura 3 muestra estos resultados observando algún tipo de tendencia en los residuos. Statgraphics también nos ofrece el estadístico de Durbin-Watson (DW), el cual también indica una posible correlación serial en los residuos de ambos modelos con un nivel de confianza del 95,0%.



Modelo de regresión definido en la tabla 4	Modelo de regresión definido en la tabla 5

Figura 3. Residuos estudentizados en función del número de fila para los modelos de regresión con los coeficientes descritos en las tablas 4 y 5, respectivamente.

La normalidad de los residuos estudentizados puede ser analizada con distintas pruebas en la opción: **describir + ajuste de distribuciones + ajuste de datos no censurados.** También disponemos de algunas opciones gráficas a través de dicha ventana. Una de ellas es la del histograma de los residuos junto con la curva de densidad de la distribución normal, la cual se muestra en la figura 4. Algunas pruebas de normalidad como el Estadístico W de Shapiro-Wilk, indican que los residuos no siguen una distribución normal.



Modelo de regresión definido en la tabla 4	Modelo de regresión definido en la tabla 5

Figura 4. Histograma de residuos estudentizados comparado con el ajuste a la función de densidad de la distribución normal, usando los coeficientes descritos en las tablas 4 y 5, respectivamente.

Deberíamos mejorar los modelos anteriores para que se cumplan todas las hipótesis de la regresión. Para ello, podemos probar con alguna transformación de la variable respuesta. El programa Statgraphics Centurion nos ofrece la Transformación Box-Cox que puede ser útil en casos con heterocedasticidad. También dispone de la transformación de Cochrane-Orcutt que provee un mecanismo para manejar situaciones en las que los residuos no son independientes. Aunque el uso de estas transformaciones se escapa del objetivo de este artículo docente.

5 Conclusiones

Si buscamos un modelo suficientemente preciso con el menor número de variables podemos elegir el modelo (1) con la estimación de los coeficientes para las tres variables explicativas indicadas en la tabla 5. Los estadísticos de los residuos de dicho modelo (RMSE, MAE y $\bar{R}^2$) son similares al modelo descrito en la tabla 4, aunque en la ecuación de este último intervienen dos variables más. Con el modelo de la tabla 5 se puede calcular el valor de la precipitación media en verano en otra localización, conociendo sus coordenadas UTM (XUTM≡longitud, YUTM≡latitud) junto con la elevación media dentro de un área circular de 5 km de radio (Z5). A modo de ejemplo, hemos probado, el cálculo de la precipitación media en verano en la localidad de Puçol (Valencia, España), obteniendo un valor estimado de 58,9194 l/m^2. En cambio, el uso del modelo descrito en la tabla 4 requiere el cálculo de variables más complejas, especialmente la variable ALTDIF que se calcula como la diferencia de elevación entre el punto más alto de la ladera del nuevo punto y el punto más alto dentro de un área orientada dentro de la dirección del flujo sinóptico en 850 hPa. Detalles del cálculo de esta variable pueden consultarse en Portalés et al. (2010).

Para validar el modelo de regresión elegido deberíamos de conocer el valor de la precipitación media en verano en lugares diferentes a los usados para construir el modelo. Con ello se podrían calcular los errores cometidos en dichos lugares como diferencia entre el valor observado y el valor predicho.

Más detalles sobre el uso del programa Statgraphics Centurion para la selección de variables en los modelos de regresión pueden verse en los siguientes videos:

https://media.upv.es/player/?id=c8e91d70-a4e8-11ea-a55e-2790bf869373

https://media.upv.es/player/?id=5f262940-abdd-11ea-888c-63e032ecdc13.

6 Bibliografía

[1] Balaguer Beser, A.; Capilla Roma, MT.; Felipe Román, MJ.; Marín Molina, J.; Monreal Mengual, L. (2014). Métodos matemáticos. Editorial Universitat Politècnica de València. http://hdl.handle.net/10251/70684

[2] Marín Molina, J.; Balaguer Beser, A.; Felipe Román, MJ.; Capilla Roma, MT. (2012). Álgebra lineal. Editorial Universitat Politècnica de València. http://hdl.handle.net/10251/72444

[3] Marquínez, J., Lastra, J., & García, P. (2003). Estimation models for precipitation in mountainous regions: the use of GIS and multivariate analysis. *Journal of hydrology*, *270*(1-2), 1-11. https://doi.org/10.1016/S0022-1694(02)00110-5

[4] Portalés, C., Boronat, N., Pardo-Pascual, J. E., Balaguer-Beser, A. (2010). Seasonal precipitation interpolation at the Valencia region with multivariate methods using geographic and topographic information. International journal of climatology, 30(10), 1547-1563. https://doi.org/10.1002/joc.1988

[5] Vrieze, S.L. (2012). Model Selection and Psychological Theory: A Discussion of the Differences Between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Psychological Methods, 17(2), 228–243. https://doi.org/10.1037/a0027127