Máster en Big Data y Data Science

Estadística Avanzada

Actividad 1: Análisis de datos

Laura Molinos Mayo

Febrero 2022

Índice

1. Introducción al dataset y objetivo **3**

2. Preprocesado de datos **6**

2.1. Missing values 6

2.1.1. Variables numéricas de oleaje 7

2.1.2. Variables numéricas oceanográficas y meteorológicas 7

2.2. Datos aberrantes y outliers 7

2.3. Variables transformadas 7

3. Regresión lineal **8**

4. Regresión multilineal **14**

5. Regresión logística **20**

6. Regresión bayesiana **24**

1. Introducción al dataset y objetivo

El dataset está formado por datos medidos por tres boyas de la Red de Boyas de Aguas Profundas de Puertos del Estado[1], solicitados mediante formulario a la institución: Cabo Silleiro, Estaca de Bares y Vilán-Sisargas. El objetivo es analizar la evolución del cambio climático en la zona atlántica gallega[2] a partir de series temporales de temperatura del aire y de temperatura y salinidad marinas obtenidas en las últimas décadas[3], mediante el cálculo de los índices concebidos por el Equipo Experto en la Detección del Cambio Climático (ETCCDI)[4]:

- FD (*frost days*): recuento de días al año en los que la temperatura mínima (T_{min}) es menor que 0 °C.
- SD (*summer days*): recuento de días al año en los que la temperatura máxima (T_{max}) supera los 25 °C.
- ID (*icing days*): recuento de días al año en los que la T_{max} es bajo cero.
- TR (*tropical nights*): recuento de días al año en los que T_{min} supera los 20 °C.
- GSL (*growing season length*): número de días que pasan entre los seis primeros días consecutivos del año con temperatura media mayor que 5 °C y los seis primeros días consecutivos, después del 1 de julio, con temperatura media menor que 5 °C.
- TX_x: máximo mensual de la temperatura máxima diaria, T_{max}.
- TN_x: máximo mensual de la temperatura mínima diaria, T_{min}.
- TX_n: mínimo mensual de la temperatura máxima diaria, T_{max}.
- TN_n: mínimo mensual de la temperatura mínima diaria, T_{min}.

Una vez calculados todos estos índices a partir de las series diarias de temperatura, se aplicarán modelos de regresión sobre ellos, para deducir posibles tendencias.

Se hará lo mismo con los valores diarios de la temperatura del agua y la salinidad. Esto permitirá confirmar (o no) indicios de cambio climático en la zona.

Para ello se ha realizado un preprocesado de los datos, dentro de la asignatura de Minería de Datos, y se necesita completar en próximas asignaturas las series de interés (temperaturas de aire y agua y salinidad) utilizando una red neuronal.

[1] Estas boyas, al estar fondeadas a más de 200 m de profundidad, evitan efectos climáticos locales y por lo tanto son representativas de una gran zona litoral.

[2] La alteración del medio marino en dicha zona supondría un gran impacto, debido al peso del sector pesquero y el marisqueo en la economía y gastronomía gallegas.

[3] Uno de los principales efectos del cambio climático es el calentamiento global, evidenciado en el aumento de las temperaturas de aire y océanos y mares, pero también se puede ver reflejado en los valores de salinidad, debido a la redistribución de las masas de agua, mayor evaporación y otros fenómenos derivados del incremento de temperatura.

[4] https://www.wcrp-climate.org/etccdi. Existen más índices, pero estos son los más ampliamente usados.

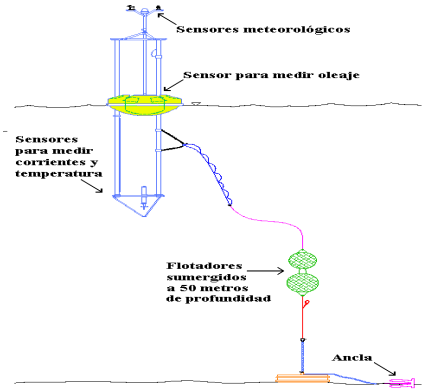

Figura 1: Localización y esquema de las boyas SeaWatch

	Estaca de Bares	**Vilán-Sisargas**	**Cabo Silleiro**
Código de boya	2244	2246	2248
Profundidad de fondeo (m)	1800	386	600
Coordenadas	7.68° O, 44.12° N	9.21° O, 43.50° N	9.43° O, 42.12° N
Fecha de activación	19/7/1996 a las 15:00	13/5/1998 a las 00:00	6/7/1998 a las 20:00

Cuadro 1: Información de las boyas

Inicialmente se contaba con tres archivos csv (uno para cada boya), con datos enviados a cada hora por vía satélite a Puertos del Estado (series horarias). Los datos han pasado por un sencillo control de calidad a su recepción, de manera que las medidas tomadas bajo condiciones desfavorables de los sensores constaban en los archivos como -9999.9.

Además, aunque todas las magnitudes son enviadas a cada hora, no todas se miden de la misma forma: algunas son instantáneas, tomándose ese valor como representativo de toda la hora; otras se miden durante un periodo y el procesador de la boya realiza el promedio u otro cálculo estadístico antes de emitir el dato.

Magnitudes	**Duración de la medida**
Magnitudes del oleaje	30 min
Velocidad del viento	10 min
Velocidad de la corriente	10 min
Temperatura del aire	Instantánea
Temperatura del agua	Instantánea
Presión atmosférica	Instantánea
Salinidad	Instantánea

Cuadro 2: Medición de las magnitudes en el tiempo

A continuación se enumeran todas las variables que constaban en los archivos iniciales:

- **Fecha (GMT)**: la fecha y hora de la medida, en el formato *año mes día hora*, según el estándar de tiempo medio de Greenwich (GMT).

Las fechas se encuentran comprendidas entre la activación de la boya y la fecha tomada como final: el 1/12/2021 a las 23:00[5]. Tipo **datetime**.

- **Variables numéricas de oleaje:**
  - **Altura Signif. del Oleaje (m)**: la altura significante del oleaje en metros, es decir, la altura promedio del 33 % de olas más altas registradas durante esa hora. Es un parámetro estadístico muy usado para describir la distribución de altura de las olas, y por lo tanto es calculado por el procesador de las boyas antes de la emisión. Posee dos decimales de precisión. Tipo **float**.
  - **Altura Maxima del Oleaje (m)**: la altura en metros de la ola más alta registrada en esa hora. Dos decimales de precisión. Tipo **float**.
  - **Periodo Medio Tm02 (s)**: tiempo promedio en segundos entre dos olas consecutivas cualesquiera. También es un parámetro calculado por la boya. Dos decimales de precisión. Tipo **float**.
  - **Periodo de Pico (s)**: tiempo promedio en segundos entre dos olas dentro del 33 % de las más altas. Calculado por la boya. Dos decimales de precisión. Tipo **float**.
  - **Periodo de la Ola Maxima (s)**: tiempo en segundos que separa la ola más alta con la anterior. Dos decimales de precisión. Tipo **float**.
  - **Direcc. Media de Proced. ($0 = N$,$90 = E$)**: dirección promedio de todas las olas registradas en esa hora. Calculada por la boya. Se da en grados y va de 0 a 360 ° (sin decimales). Tipo **int**.
  - **Direcc. de pico de proced. ($0 = N$,$90 = E$)**: dirección promedio del 33 % de las olas más altas. Calculada por la boya. Se da en grados y va de 0 a 360 ° (sin decimales). Tipo **int**.
  - **Direcc. Media de Proced. en el pico (grados)**: desviación típica de la dirección de las olas más altas respecto a la dirección media. Calculada por la boya. Se da en grados y va de 0 a 90 ° (sin decimales). Tipo **int**.
- **Variables numéricas oceanográficas:**
  - **Dir. de prop. de la Corriente ($0 = N$,$90 = E$)**: dirección promedio de la corriente marítima durante esa hora. Calculada por la boya. Se da en grados y va de 0 a 360 ° (sin decimales). Tipo **int**.
  - **Velocidad media de Corriente** (cm/s): velocidad media de la corriente marítima durante esa hora en centímetros por segundo. Calculada por la boya. Un decimal de precisión. Tipo **float**.
  - **Temperatura del Agua (celsius)**: temperatura del agua en grados centígrados, medida a 3 metros de profundidad. Dos decimales de precisión. Tipo **float**.
  - **Salinidad (psu)**: concentración de sal en unidades prácticas de salinidad[6], obtenidas a partir de la medida de la conductividad[7]. Dos decimales de precisión. Tipo **float**.

[5]Esta fecha fue seleccionada a propósito, para las tres boyas, pues a partir de ella no se podía garantizar que los datos hubiesen pasado ya por el control de calidad inicial.

[6]Es la unidad más utilizada hoy en día para la salinidad. Relaciona la conductividad de la muestra de agua de mar con respecto a la de una disolución de KCl estándar (32,4356 g de KCl por kg de disolución, a 15 °C).

[7]La concentración de sales disueltas en el agua está muy ligada a la conductividad, puesto que estas suponen un libre tránsito de iones. A mayor salinidad, mayor conductividad.

- **Variables numéricas meteorológicas:**
  - **Direc. de proced. del Viento** ($0 = N$,$90 = E$): dirección media de procedencia del viento durante esa hora. Calculada por la boya. Se da en grados y va de 0 a 360 ° (sin decimales). Tipo **int**.
  - **Velocidad media del viento** (m/s): velocidad media del viento durante esa hora en metros por segundo. Calculada por la boya. Un decimal de precisión. Tipo **float**.
  - **Presion atmosferica (hPa)**: presion atmosférica en hectopascales, medida a 3 metros de la superficie. Tipo **int**.
  - **Temperatura del Aire (celsius)**: temperatura del aire en grados centígrados, medida a 3 metros de la superficie. Dos decimales de precisión. Tipo **float**.
- **Variables categóricas (category):** los archivos contienen cuatro campos referentes al canal de obtención de las medidas. En las boyas existen 3 canales de medición, de manera que cada medida de una magnitud podrá ser tomada por uno de ellos. Cada una de las columnas **Canal de obtencion de los datos** hace referencia a las magnitudes de su izquierda.

Los tres archivos fueron integrados en un único dataset. Para ello se creó un nuevo campo, con el nombre de la boya, y se introdujeron las fechas que faltaban en los csv originales (debido a periodos de mantenimiento de las boyas o errores) para contar con series temporales completas. Se crearon a mayores campos para el año, mes y día, para facilitar las tareas de limpieza de datos.

El dataset utilizado para esta actividad de Estadística Avanzada se trata del final, después del preprocesado, comentado en el siguiente apartado.

2. Preprocesado de datos

A pesar de que las magnitudes de interés de cara el objetivo final se tratan de la temperatura del aire, temperatura del agua y salinidad, el preprocesado se aplicó a la totalidad de campos del dataset. Fue realizado en Python, concretamente en Jupyter Notebook.

2.1. Missing values

Todas las variables contaban con *missing values*, excepto la fecha (una vez completada), los campos derivados de ella (año, mes, día) y el campo con los nombres de las boyas. Por preferencia se cambiaron los valores indeterminados por defecto (-9999.9) por el genérico NaN.

Pero no todos los *missing values* eran causados por el mismo motivo. Algunos se debían a que las boyas no contaban todavía con medidor para esa magnitud; otros estaban causados por medidas incorrectas puntuales (una, dos o tres medidas horarias seguidas) por algún fallo en el sistema o alteración del entorno y otros abarcaban un gran periodo de tiempo (más de dos días) y se debían a la revisión o mantenimiento de las boyas.

Para el tratamiento de los dos últimos casos (medidas erróneas puntuales o prolongadas en el tiempo) cambiamos el NaN por un valor numérico razonable, usando un método que dependía de la variable en cuestión. En cambio, para el caso de los *missing values* debidos a que ninguna boya medía aún la magnitud, los mantuvimos como valores indeterminados, concretamente como -1, que carecía de sentido físico para todas las magnitudes.

En ningún caso se eliminaron instancias, aunque todos los atributos fuesen NaN, porque esto provocaría discontinuidades en las series temporales.

2.1.1. Variables numéricas de oleaje

- Los valores faltantes de la altura significante y máxima de las olas se imputaron con los valores medios mensuales. Esto es más razonable que tomar la media de todas las instancias, pues el oleaje en los meses de invierno difiere mucho del de los meses de verano, debido a la mayor presencia de temporales y fuertes rachas de viento.
- En los periodos (medio, de pico y de ola máxima) se imputaron mediante interpolación, concretamente con el método *time*.
- Por último, para las direcciones (media, de pico y media en el pico), se rellenaron los *missing values* con el valor más frecuente mensual (moda mensual). Esto se aplicó también a la dirección de la corriente y del viento, al comportarse de manera análoga.

2.1.2. Variables numéricas oceanográficas y meteorológicas

La velocidad de corriente era bastante variable de una hora a otra, por lo que se optó por imputar los valores faltantes con la mediana mensual. Para la velocidad del viento y la presión atmosférica, mucho más lineales, se aplicó interpolación, con el método *linear*.

2.2. Datos aberrantes y outliers

- Para las magnitudes secundarias (las que no son relevantes para el objetivo final) se localizaron los posibles *outliers* con el uso de diagramas de caja (*boxplot*). Los valores que salientaban eran estudiados (observando las instancias más cercanas por ejemplo) para constatar si eran válidos; en caso de no serlo se corregían con la media de los dos datos más cercanos.
- En las magnitudes de interés (temperatura del aire, temperatura del agua y salinidad) se establecieron unos límites[8], que en caso de ser sobrepasados implicaban automáticamente un valor incorrecto. Después se localizaban valores fuera de rango (aquellos que se alejaban más de cuatro desviaciones típicas de la media) y se comprobaba su validez. Los datos tomados como incorrectos eran sustituídos por *missing values*, para ser completados en un futuro con la red neuronal junto con los ya existentes.

2.3. Variables transformadas

Una vez imputados los *missing values* de las variables secundarias y corregidos los datos aberrantes de las secundarias y las de interés, se obtuvieron las series diarias para las temperaturas del aire y agua y salinidad, pues sobre estas se aplicará la red neuronal. Para eso, para cada una de las boyas, se recorrieron todos los días registrados, calculando la media de cada variable.

[8]Para la temperatura del aire se tomaron como límites infranqueables -35,6 $^{\circ}C$ y 47,4 $^{\circ}C$ (mínimo y máximo históricos de temperatura en España, ambos alcanzados en 2021). Para la temperatura del agua se tomaron -5$^{\circ}C$ y 25$^{\circ}C$. Para la salinidad, 12 psu y 40 psu (salinidad del salobre Mar Caspio y el bastante salino Mar Rojo).

En los días en los que faltaban todas las medidas horarias, la media sería también NaN; en los que faltaba un porcentaje razonable de los datos la media se calculaba con los datos completos; y en los días en los que había demasiados NaN la media se consideraba también *missing value*[9].

El dataset que utilizamos en esta actividad cuenta entonces, a mayores de las variables ya vistas, con las medias diarias de temperatura del aire y agua y salinidad. El dataset tiene en total **634069** instancias y **29** columnas.

3. Regresión lineal

Para la regresión lineal de dos variables utilizaremos la **altura significante de ola** y la **altura máxima de ola**.

La altura de las olas sigue una distribución de Rayleigh en condiciones ideales, que se aproxima de manera razonable al comportamiento real cuando no se da viento extremo o muy variable, pues en ese caso dicha distribución se vuelve bastante más compleja. Esto quiere decir que habrá muy pocas olas que alcancen alturas próximas a la máxima registrada y muy pocas olas cercanas a la mínima. En base a este modelo ideal, la altura significante se corresponde con la mitad de la altura máxima.

El objetivo de esta primera regresión es estudiar si existe empíricamente una relación lineal entre ambas magnitudes y en caso afirmativo, si se aproxima o no al valor teórico.

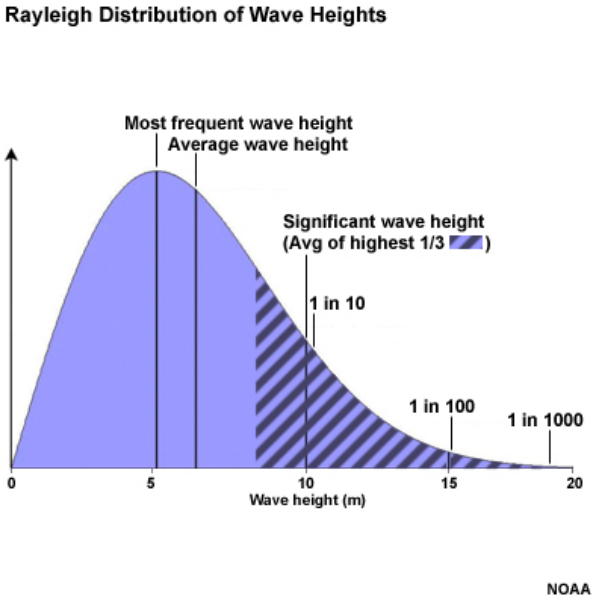

Figura 2: Distribución teórica de la altura de las olas

[9]Si en un día tenemos sólo valores para las horas 12:00, 13:00 y 14:00 y el resto son NaN, por ejemplo, no es razonable tomar el dato diario como la media de sólo esas tres medidas. En ese caso es mejor considerar el dato diario como NaN.

Primero, realizamos un análisis exploratorio a ambas magnitudes.

	Mínimo	**Máximo**	**Media**	**Desviación típica**	**Q1**	**Mediana**	**Q3**
Altura significante (m)	0,010	13,460	2,487	1,232	1,620	2,240	3,090
Altura máxima (m)	0,180	27,810	3,805	1,913	2,460	3,430	4,750

Cuadro 3: Descripción estadística de las magnitudes

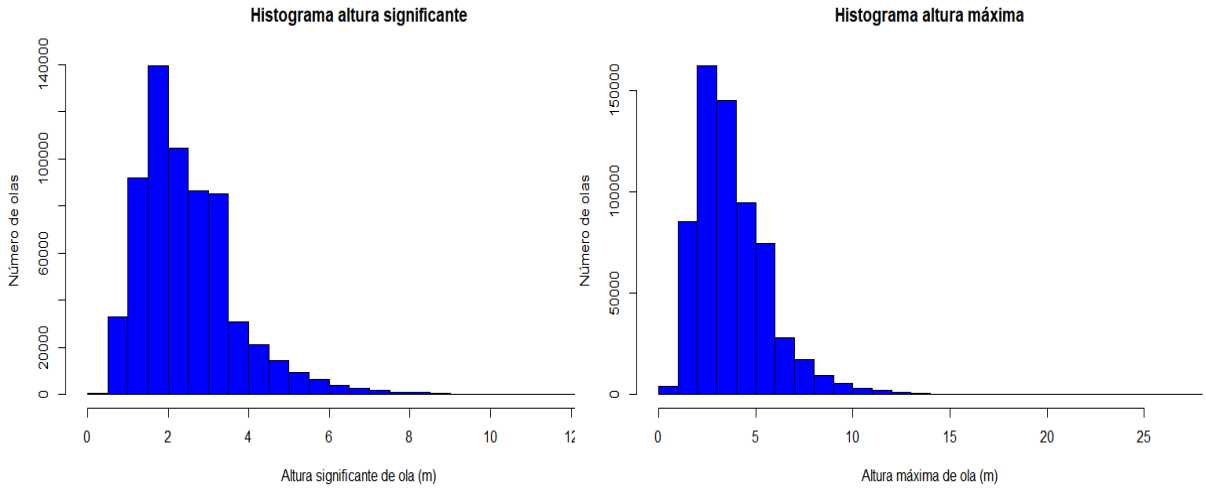

Figura 3: Histogramas de las magnitudes

Como vemos la altura significante se sitúa mayoritariamente entre el metro y medio y los tres metros, y en pocas ocasiones alcanza valores mayores a 6 m. Asimismo, las olas más altas rara vez superan los 10 m.

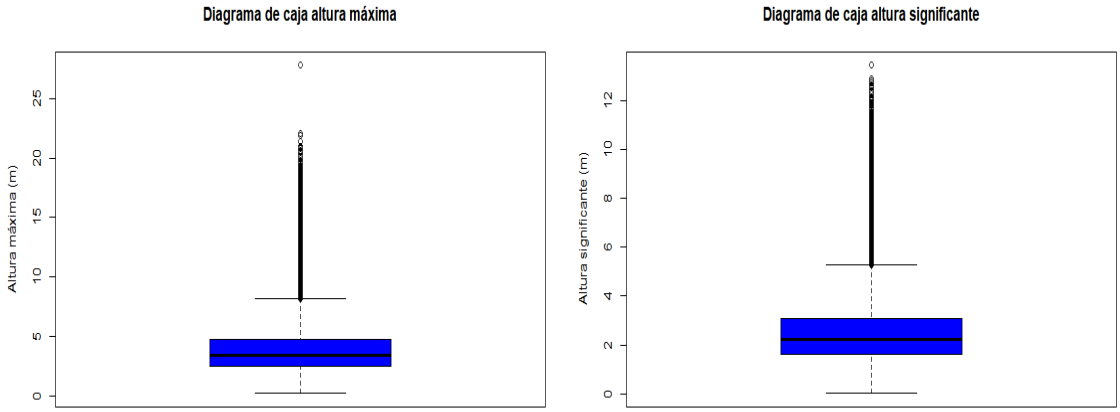

Figura 4: Boxplot de las magnitudes

En ambos diagramas de caja destaca un punto, pero no es un error, se trata de la ola más alta registrada en aguas españolas: 27,81 m, registrada por Vilán-Sisargas en enero de 2014.

Graficamos la altura máxima frente a la significante para tener una noción inicial de su comportamiento.

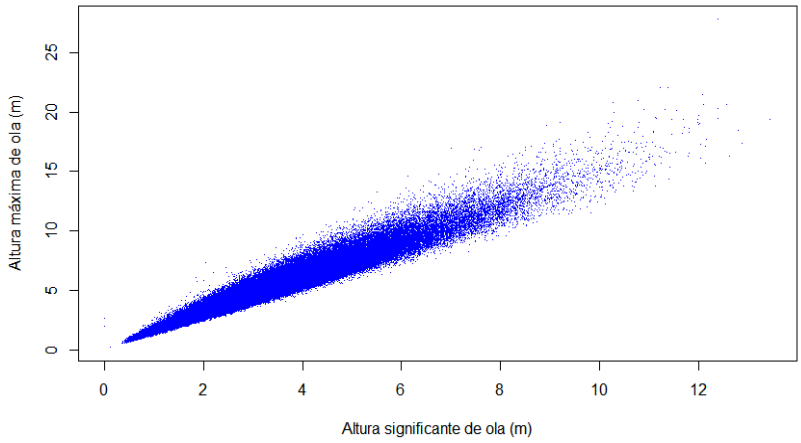

Figura 5: Altura máxima *vs* altura significante

Como vemos la nube de puntos es adecuada para el análisis: si un modelo lineal es apropiado o no lo deduciremos de los parámetros obtenidos de la regresión. La regresión lineal se lleva a cabo con la función *lm* de R, y se aplica a un conjunto de entrenamiento, que suponemos del 80 % del original[10].

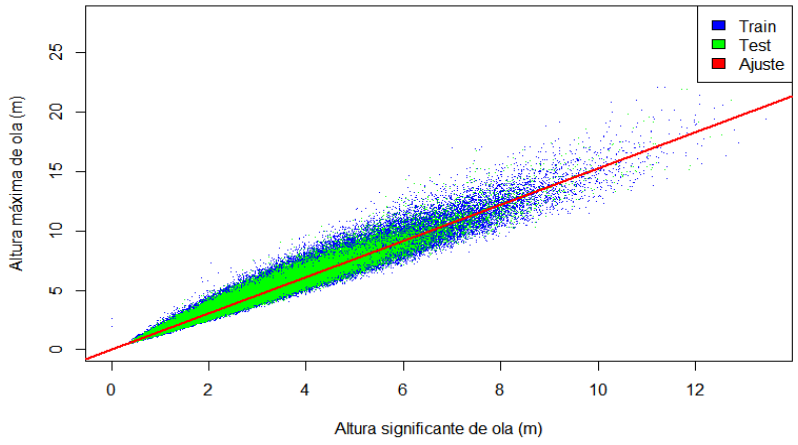

Figura 6: Recta de ajuste y datos de entrenamiento y test

Visualmente la recta parece ajustarse fielmente a los datos, pero hay que analizar los parámetros para poder constatar su validez.

[10]Se ha comprobado que la salida de la función no cambiase de manera importante al ajustar con todos los datos, para asegurarse de que el conjunto train está balanceado.

```
> summary(fit1)

Call:
lm(formula = Altura.Maxima.del.Oleaje..m. ~ Altura.Signif..del.Oleaje..m.,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5155 -0.1522 -0.0017  0.0916  8.8980

Coefficients:
                             Estimate Std. Error
(Intercept)                 0.0153534  0.0011403
Altura.Signif..del.Oleaje..m. 1.5239260  0.0004107
                            t value Pr(>|t|)
(Intercept)                   13.46   <2e-16 ***
Altura.Signif..del.Oleaje..m. 3710.12   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3609 on 507253 degrees of freedom
Multiple R-squared:  0.9645,    Adjusted R-squared:  0.9645
F-statistic: 1.376e+07 on 1 and 507253 DF,  p-value: < 2.2e-16

>
```

Figura 7: Salida de la función *lm*

- La desviación típica de los parámetros del modelo es mucho menor que sus valores, concretamente del 7,4 % para la ordenada en el origen y del 0,027 % para la pendiente. Sus valores t (*t-values*) son bastante mayores de la unidad y grandes en comparación con las desviaciones típicas, pudiéndose afirmar con un nivel de significación prácticamente de 0 (<2e-16), que ambos parámetros no son nulos.
- El coeficiente de determinación (R^2), se acerca mucho a 1, de lo que se deduce que el modelo explica realmente bien la variabilidad de la variable dependiente (altura máxima).
- El valor p del test F (*p-value*) es prácticamente 0, lo que indica que se puede rechazar la hipótesis nula[11], es decir, que existe una relación entre las dos magnitudes. Se puede llegar a la misma conclusión observando el gran valor de F.
- El error estándar residual (RSE), que supone el error que tendrán asociado las predicciones de la variable dependiente por el modelo, ya se tiene en cuenta en el cálculo de los t-valores. Sería muy útil para comparar el modelo con otros.
- De los datos estadísticos de los residuos (diferencia entre los valores dependientes reales y predichos por el modelo) se intuye que su distribución no es totalmente simétrica, pero no obstante, la mediana se acerca bastante al cero, lo que es positivo para el modelo. Analicemos los residuos detenidamente con los siguientes gráficos.

Uno de los supuestos de los modelos lineales es que los residuos siguen una distribución normal. Se puede observar fácilmente si esto se cumple graficando la función de densidad junto a la normal que correspondería para el punto medio y la desviación de los residuos.

[11] La hipótesis nula consiste en afirmar que la pendiente de la recta es 0; dicho de otra forma, que no hay relación alguna entre las dos variables.

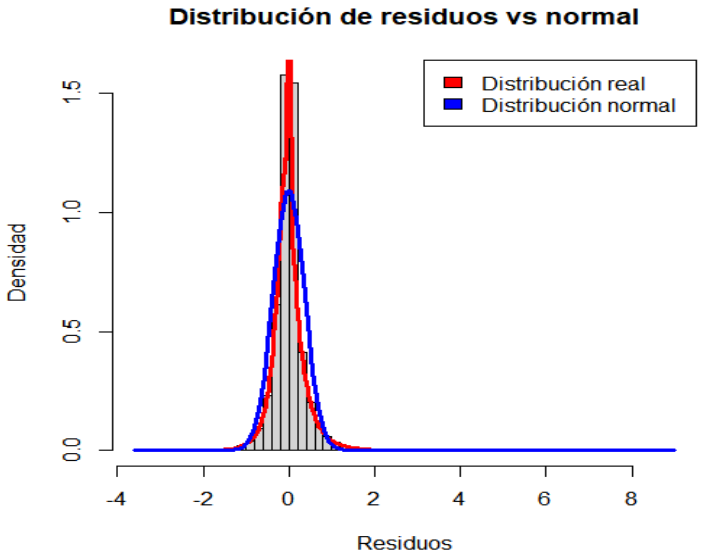

Figura 8: Distribución de los residuos

Aunque la distribución de los residuos presenta una simetría decente, se diferencia de la normal al ser bastante más picuda.

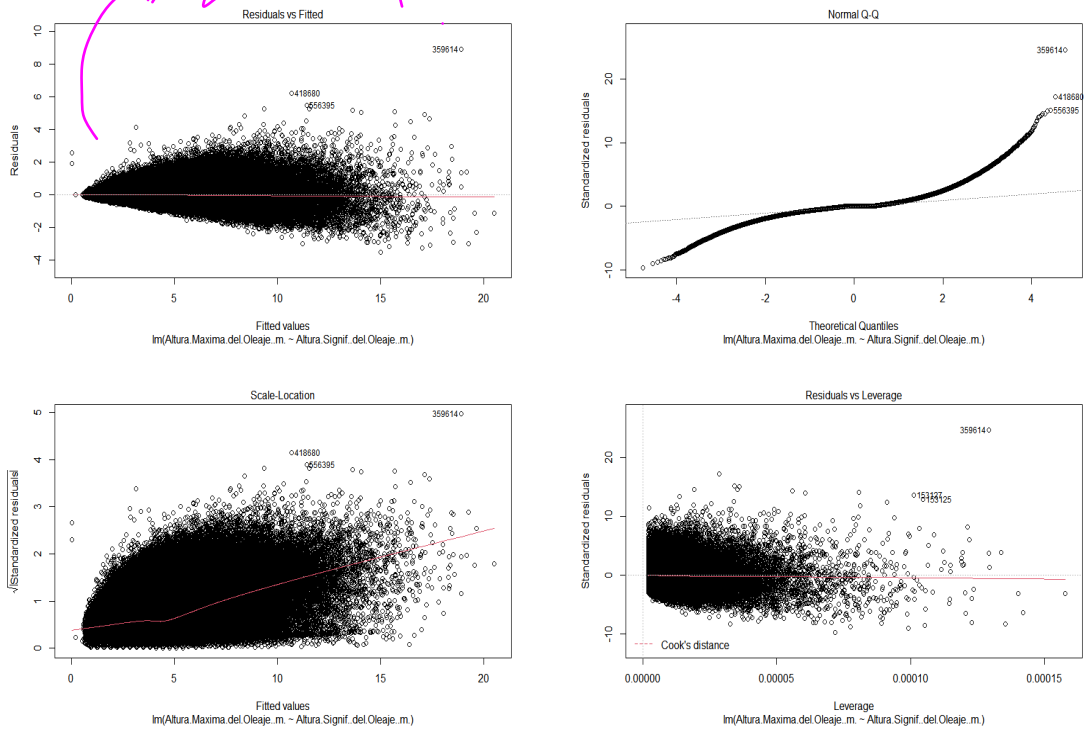

Figura 9: Gráficos de residuos

- Del primer gráfico ($residuals\ vs\ fitted-values$) se deduce claramente que no existen efectos no lineales entre las variables, lo que confirma la adecuación de un modelo lineal a los datos.

- En el gráfico Q-Q se manifiestan las diferencias con la distribución normal vistas en la función de densidad.
- Del tercer gráfico ($scale - location$), es evidente que la magnitud media de los residuos no es independiente de la variable predecida, sino que aumenta al incrementarse esta (la recta roja debería ser horizontal idealmente pero tiene pendiente positiva). Además la dispersión de los puntos no es homógenea a lo largo de la línea. Entonces tampoco se cumple el supuesto de varianza constante.
- Del último gráfico ($residuals\ vs\ leverage$), al no aparecer líneas de Cook, se entiende que no existen valores influenciables, es decir, no hay puntos que de ser eliminados cambiarían apreciablemente los parámetros del modelo. Por lo tanto no hay porque estudiar ningún punto problemático.
- En cuanto a los *outliers*, el que más salienta es el comentado anteriormente. Los demás han sido estudiados también en el preprocesado y se tratan de datos válidos.

Por último graficamos los datos test con su predicción y con el intervalo de predicción, que tiene en cuenta el ruido y la magnitud de los residuos. Las líneas de puntos rojas se tratan de los límites inferior y superior de dicho intervalo.

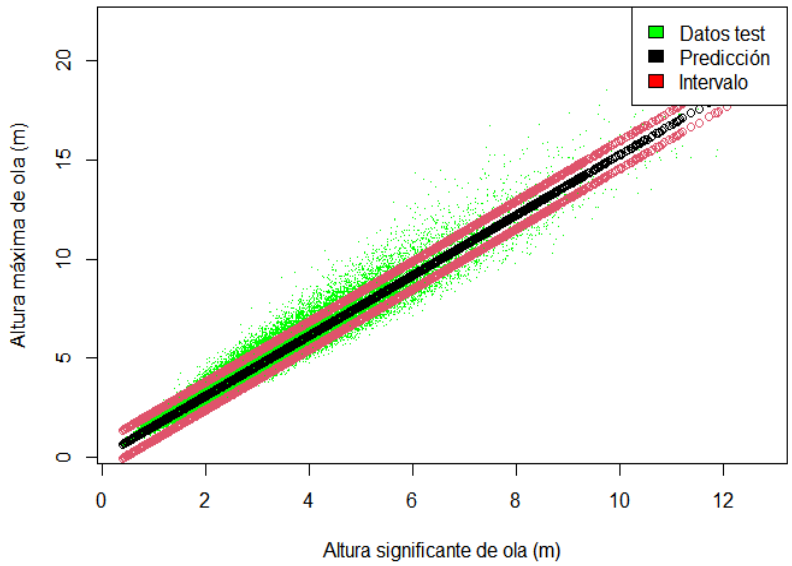

Figura 10: Intervalos de predicción

Conclusión:

Con los resultados de la regresión y gráficos asociados se puede confirmar que existe una relación lineal entre las dos variables, y que los parámetros obtenidos son los de mejor ajuste para el modelo lineal.

En la teoría la altura máxima se corresponde con el doble de la altura significante, lo que se traduce en una relación lineal con pendiente 2 y ordenada en el origen nula.

La pendiente empírica es de aproximadamente 1,52, lo que supone una diferencia del 24 % respecto al valor teórico; además, la ordenada en el origen empírica no es nula, pero tiene un valor muy cercano a cero: 0,015. Es lógico que la relación real difiera de la teórica, puesto que la distribución ideal ignora muchas variables existentes en el oleaje, y no tiene en cuenta los días con condiciones climatológicas desfavorables.

Aunque el modelo lineal explica acertadamente la variación de la variable dependiente, no acaba de modelar satisfactoriamente los residuos. En este sentido, sería adecuado probar con otros modelos[12], aunque podrían no mejorar al actual.

4. Regresión multilineal

En este apartado realizaremos una regresión lineal con más de una variable predictora. Para ello, utilizando la matriz de correlaciones, observamos qué magnitud es más adecuada para actuar de dependiente y cuales deben ser sus variables predictoras asociadas.

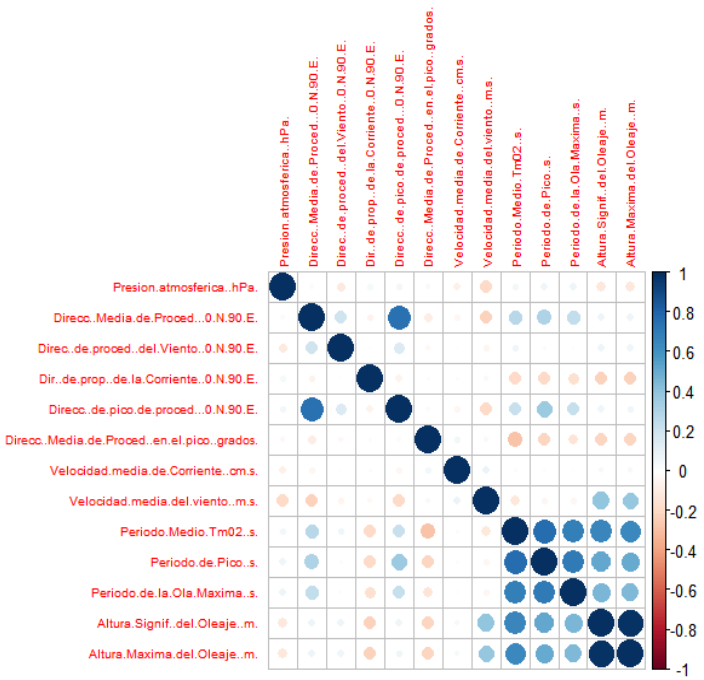

Figura 11: Matriz de correlación

De la matriz se percibe que muy pocas magnitudes están notablemente correlacionadas. La dirección media del oleaje y la dirección de pico están bastante correlacionadas, pero al estarlo sólo entre ellas, se podrían utilizar para un ajuste lineal normal pero no uno multilineal. Entonces nos centramos en la esquina inferior derecha, donde se concentran la mayoría de correlaciones.

[12] En principio los modelos polinómicos no serían adecuados, al no observarse tendencia no-lineal en los gráficos de residuos.

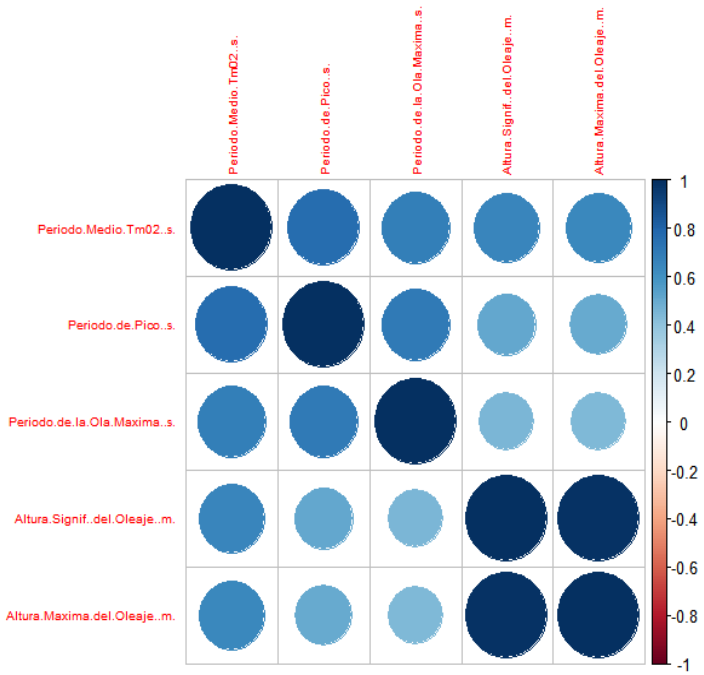

Figura 12: Magnitudes más correlacionadas

El periodo medio de ola está bastante correlacionado con las otras cuatro magnitudes (periodo de pico, periodo de ola máxima y las alturas de ola usadas en la primera regresión), por lo que se sitúa como el mejor candidato a variable dependiente. Es evidente que la altura significante de ola y la altura máxima no pueden usarse simultáneamente como variables predictoras puesto que están fuertemente correlacionadas como se puede ver en la imagen y como se demostró en la regresión lineal. Por lo tanto optamos por una de ellas, por ejemplo, la altura máxima. A mayores se selecciona el periodo de ola máxima, por estar menos correlacionada con la altura máxima que el periodo de pico. Así garantizamos la menor correlación entre las variables predictoras, para evitar multicolinealidad y sus efectos perjudiciales sobre los resultados.

Antes de realizar la regresión se presenta una descripción estadística de las variables que intervienen (la de la altura máxima aparece anteriormente, en la tabla 3 y la figura 14).

	Mínimo	**Máximo**	**Media**	**Desviación típica**	**Q1**	**Mediana**	**Q3**
Periodo medio (s)	2,970	14,790	6,321	1,491	5,170	6,110	7,270
Periodo de ola máxima (s)	2,660	25,540	9,622	2,802	7,460	9,370	11,420

Cuadro 4: Descripción estadística de las magnitudes

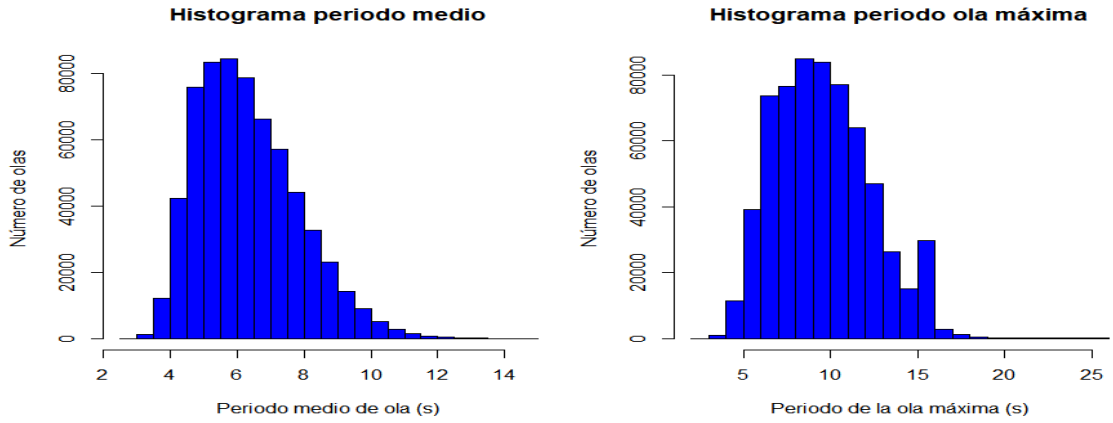

Figura 13: Histogramas de las magnitudes

Del resumen estadístico e histogramas se puede deducir que las olás más altas están asociadas con periodos más largos, pues el periodo de ola máxima alcanza valores más grandes que el periodo medio.

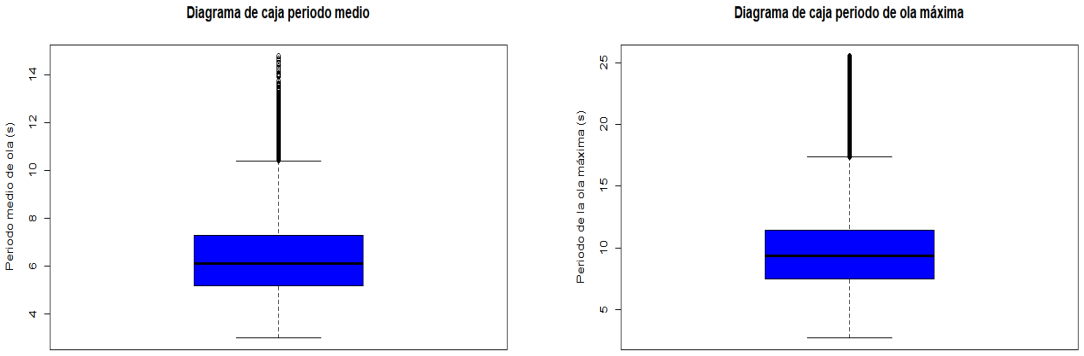

Figura 14: Boxplot de las magnitudes

De los diagramas de caja se observa que no hay valores que destaquen especialmente, pues los *outliers* han sido tratados durante el preprocesado.

Tras realizar la regresión multilineal, con el **periodo medio** como variable **dependiente** y la **altura máxima de ola** y el **periodo de ola máxima** como variables **predictoras**, obtenemos la siguiente salida:

```
> summary(multiAll)

Call:
lm(formula = Periodo.Medio.Tm02..s. ~ ., data = trainAjuste)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2672 -0.5559 -0.1130  0.5100  9.0195

Coefficients:
                              Estimate Std. Error
(Intercept)                  2.5263217  0.0046853
Periodo.de.la.Ola.Maxima..s. 0.2624435  0.0005177
Altura.Maxima.del.Oleaje..m. 0.3337481  0.0007577
                             t value Pr(>|t|)
(Intercept)                    539.2   <2e-16 ***
Periodo.de.la.Ola.Maxima..s.   507.0   <2e-16 ***
Altura.Maxima.del.Oleaje..m.   440.5   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9237 on 507252 degrees of freedom
Multiple R-squared:  0.6162,    Adjusted R-squared:  0.6162
F-statistic: 4.073e+05 on 2 and 507252 DF,  p-value: < 2.2e-16

> vif(multiAll)
Periodo.de.la.Ola.Maxima..s.
                    1.251029
Altura.Maxima.del.Oleaje..m.
                    1.251029
```

Figura 15: Salida de la función *lm*

- La desviación típica de los parámetros del modelo es muy notablemente menor que sus valores: 0,18 % para la ordenada en el origen, 0,20 % para el parámetro que multiplica al periodo de ola máxima y 0,23 % para el de la altura máxima. Además, los *t-values* son mucho mayores que la unidad y las desviaciones típicas. Se puede rechazar la hipótesis nula con un nivel de significación prácticamente de 0 (<2e-16), es decir, con una gran probabilidad los tres parámetros son distintos de cero.

- El coeficiente de determinación (R^2) no cambia respecto al ajustado, lo que se debe a que no hemos seleccionado variables predictoras a mayores innecesarias ni inoportunas. Su valor indica que en torno al 60 % de la variabilidad del periodo medio se puede explicar con el modelo. Este valor es mejorable, pero era esperable dado el nivel correlación observado en la matriz.

- El gran valor del test F y el diminuto valor de su *p-value* permiten concluir que existe relación entre la variable dependiente con cada una de las dos predictoras.

- El VIF (*variance inflation factor*) manifiesta que existe una correlación leve entre las dos variables predictoras (ya se observaba en 12), que no genera multicolinealidad, lo que es muy positivo.

- Los residuos cuentan con una mediana cercana al cero y unos primer y tercer cuartiles simétricos. Veámoslo con detalle con los gráficos de residuos:

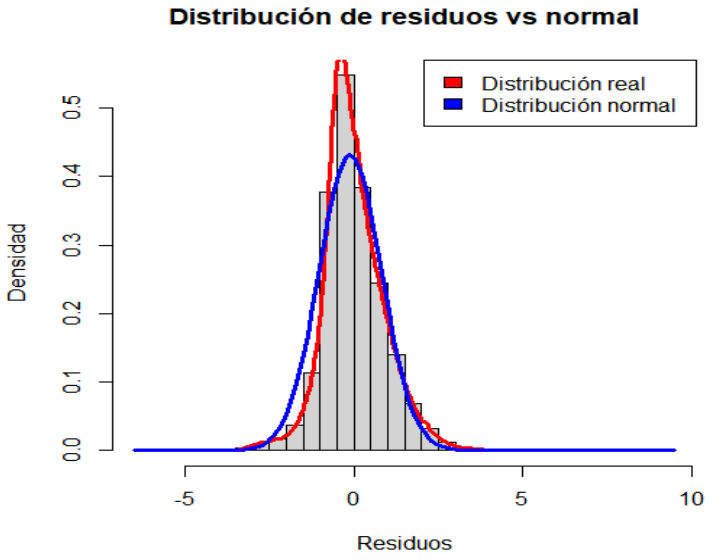

Figura 16: Distribución de los residuos

Se aprecia una simetría razonable en la función de densidad de los residuos, pero decrece más rápidamente que la distribución normal correspondiente a la media y desviación de los residuos.

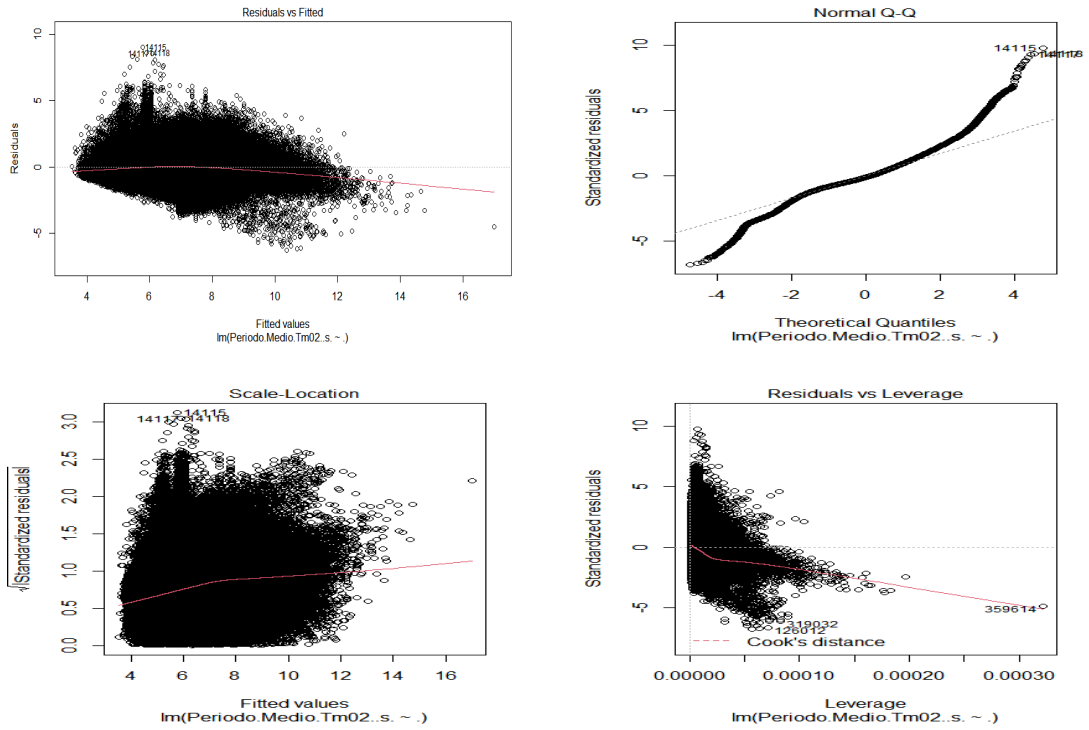

Figura 17: Gráficos de residuos

- En el primer gráfico ($residuals\ vs\ fitted - values$) se observa una desviación de la línea roja respecto a la central, lo que puede deberse a efectos no lineales.

- En el gráfico Q-Q se aprecia la discrepancia con la normal vista en la función de densidad: las colas, muy simétricas, se alejan de la línea.
- En el tercer gráfico ($scale - location$) volvemos a ver una línea roja con pendiente, lo que manifiesta que los residuos no son independientes de la variable dependiente, y se incumple el supuesto de varianza constante.
- En el último gráfico (*residuals vs leverage*), al no aparecer la línea de Cook, se concluye que no existen puntos influenciables que deban ser tomados en consideración.

Comparamos ahora los resultados obtenidos de la regresión multilineal con los que se obtendrían con una regresión lineal con cada una de las variables predictoras por separado.

```
> fit2=lm(Periodo.Medio.Tm02..s.~ Periodo.de.la.Ola.Maxima..s.,data=trainAjuste)
> summary(fit2)

Call:
lm(formula = Periodo.Medio.Tm02..s. ~ Periodo.de.la.Ola.Maxima..s.,
    data = trainAjuste)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5685 -0.5872 -0.0245  0.6228  9.6909

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.8131931  0.0054555   515.7   <2e-16 ***
Periodo.de.la.Ola.Maxima..s. 0.3645821  0.0005442   670.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.086 on 507253 degrees of freedom
Multiple R-squared:  0.4695,    Adjusted R-squared:  0.4695
F-statistic: 4.489e+05 on 1 and 507253 DF,  p-value: < 2.2e-16

> fit3=lm(Periodo.Medio.Tm02..s.~ Altura.Maxima.del.Oleaje..m.,data=trainAjuste)
> summary(fit3)

Call:
lm(formula = Periodo.Medio.Tm02..s. ~ Altura.Maxima.del.Oleaje..m.,
    data = trainAjuste)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9647 -0.8445 -0.1803  0.6632  7.9694

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.3976987  0.0035422  1241.5   <2e-16 ***
Altura.Maxima.del.Oleaje..m. 0.5058242  0.0008315   608.3   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.134 on 507253 degrees of freedom
Multiple R-squared:  0.4218,    Adjusted R-squared:  0.4218
F-statistic: 3.7e+05 on 1 and 507253 DF,  p-value: < 2.2e-16
```

Figura 18: Salida de la función *lm* para cada variable predictora

La desviación típica de los parámetros y su nivel de significación para rechazar la hipótesis nula son también muy correctos en las regresiones lineales, así como los valores del test F. Donde se evidencia que el multilineal es un mejor modelo es en el coeficiente de determinación y en el RSE: en ambas regresiones lineales el R^2 se sitúa en torno al 0,4, por lo que solo un 40 % de la variabilidad del periodo medio podría explicarse con ambos modelos, frente al 60 % del multilineal. El RSE es de 0,9237 para el multilineal pero sobrepasa la unidad en los lineales, por lo que el error cometido en la predicción será mayor en estos.

Si hubiese que elegir entre una de las dos regresiones lineales, en base a estos dos coeficientes, la del periodo de ola máxima sería mejor por poca diferencia.

Se grafica también la recta de mejor ajuste de la regresión lineal para cada una de las variables predictoras junto a los datos predichos por el modelo multilineal (en ambos para los datos de entrenamiento). Como es lógico la multilineal no se trata de una recta, pues es la proyección de un plano. Visualmente es también notable la mejoría del modelo multilineal.

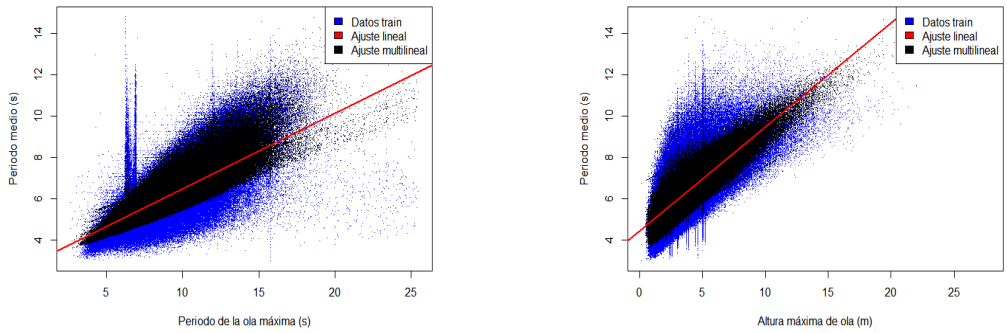

Figura 19: Regresión multilineal vs regresión lineal

Conclusión:

Se ha comprobado que el modelo multilineal es una alternativa a la regresión lineal convencional, que en muchas ocasiones supone una mejora, como es el caso. En otras, sin embargo, puede empeorar los resultados, debido a multicolinealidad o el uso de variables que no aportan información valiosa al modelo. A mayores, sería adecuado probar con modelos polinómicos, también en varias variables, debido a la aparente, pero no exagerada, presencia de efectos no lineales. De esta forma, quizás también se modelarían mejor los residuos.

5. Regresión logística

Para la regresión logística se ha transformado una variable discreta del dataset en una binaria[13]: se ha creado un nuevo campo (*Binaria*) que vale 1 para las medidas tomadas en primavera y verano y 0 para las tomadas en otoño e invierno[14]. Es una transformación lógica, puesto que las magnitudes climatológicas y oceánicas tienen comportamientos muy diferenciados según la época del año. Concretamente realizamos la regresión con la **temperatura del aire** al ser de entre todas la que más varía con la estacionalidad. Como ya se ha comentado, esta magnitud aún contiene *mising values*, pues aún no han sido completados con la red neuronal, por lo que la regresión se realiza con los datos disponibles: 509989 temperaturas.

Se presenta el resumen estadístico de esta magnitud.

	Mínimo	**Máximo**	**Media**	**Desviación típica**	**Q1**	**Mediana**	**Q3**
Temperatura del aire (ºC)	1,50	26,80	15,11	2,78	13	15	17,40

Cuadro 5: Descripción estadística de la temperatura del aire

[13]Las variables binarias que ya contiene el dataset, correspondientes a los canales de medición, no guardan relación alguna con los valores de las magnitudes medidas, por lo que no serían adecuadas para una regresión.

[14]Para eso se ha utilizado el campo *Mes*: a las instancias en las que este valía 1,2,3,10,11 o 12 se ha atribuido un 0 y en las demás un 1.

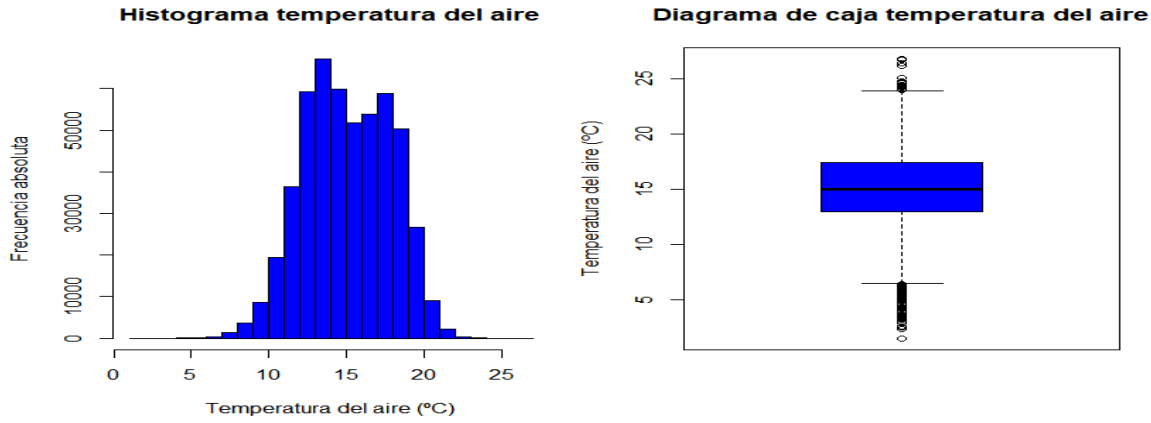

Figura 20: Histograma y diagrama de caja de la temperatura del aire

Como vemos, la mayoría de temperaturas se encuentran entre los 10 ° C y los 20 ° C. Los *outliers* han sido minuciosamente estudiados en el preprocesado y ya no hay valores erróneos.

Representamos el diagrama de caja, pero ahora agrupando las temperaturas según su categoría (0: otoño-invierno, 1: primavera-verano)[15]. Se aprecia una diferencia en la mediana para ambas categorías; si esto no ocurriese no sería una magnitud adecuada para una regresión logística.

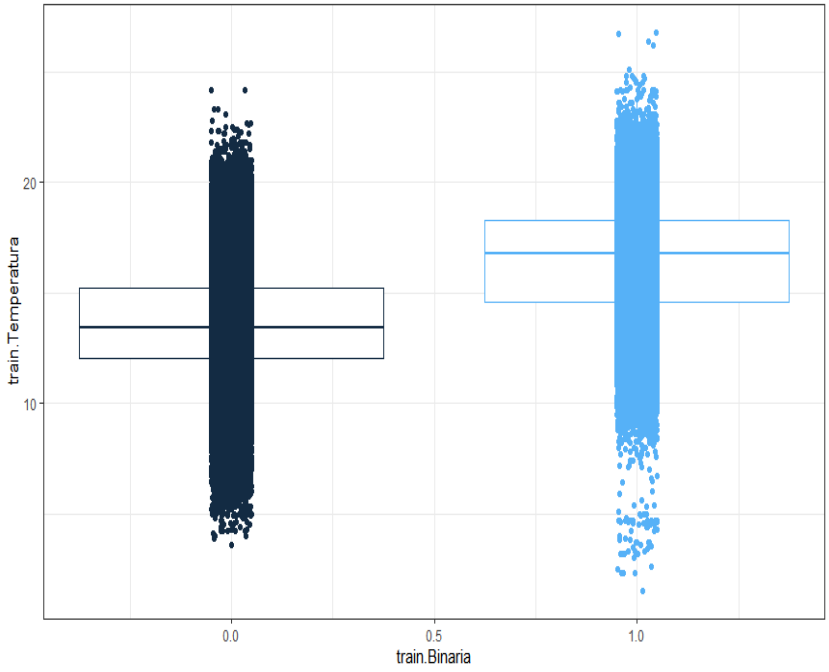

Figura 21: Diagrama de caja según categoría

Comprobamos que las dos clases (0 y 1) están balanceadas, es decir, que cuentan más o menos con el mismo porcentaje de instancias, y que los datos de entrenamiento representan correctamente el porcentaje de la muestra total.

[15]Se ha aplicado *geom_jitter* para añadir un pequeño ruido a los puntos, y evitar su solapamiento, para visualizarlos mejor.

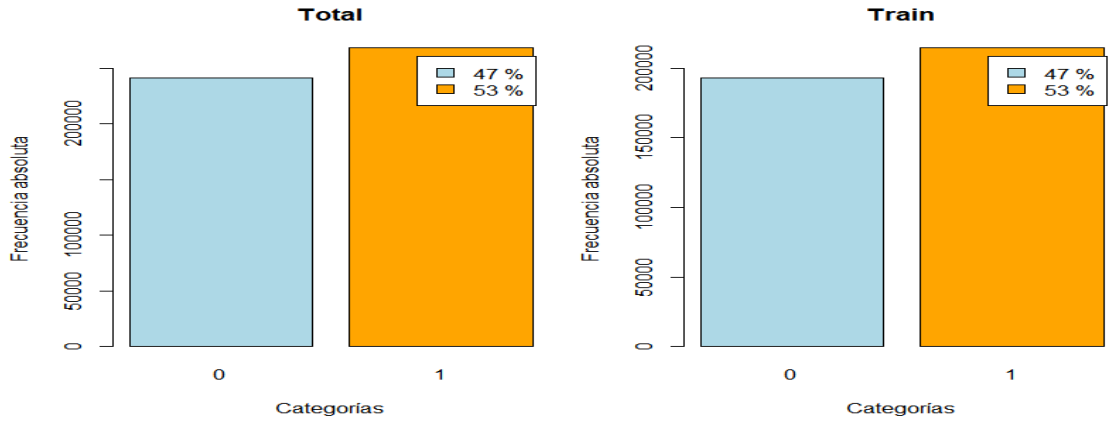

Figura 22: Las dos clases están suficientemente balanceadas

Para graficar la variable binaria frente a la temperatura del aire se utiliza la función *sunflowerplot*, pues evita el solapamiento de puntos: si el par de datos aparece una sola vez se representa como un punto normal, si aparece dos veces contiene una aspa, si aparece tres, dos aspas, y así consecutivamente. Esto permite visualizar las zonas de puntos con mayor ocurrencia.

Se muestra dicha gráfica incluyendo la sigmoide de mejor ajuste:

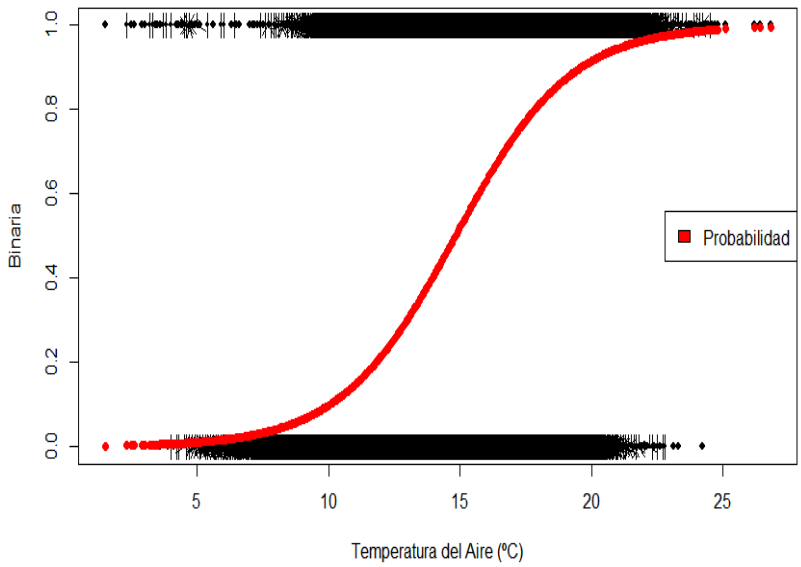

Figura 23: Regresión logística

```
> summary(logist)

Call:
glm(formula = Binaria ~ Temperatura, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9445  -0.8992   0.4285   0.8534   3.5018

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.819873   0.024181  -282.0   <2e-16 ***
Temperatura  0.460398   0.001596   288.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 564466  on 407990  degrees of freedom
Residual deviance: 448383  on 407989  degrees of freedom
AIC: 448387

Number of Fisher Scoring iterations: 4

> anova(logist, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Binaria

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       407990     564466
Temperatura  1   116084    407989     448383 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 24: Salida de la función *glm*

De los resultados del modelo, sacamos las siguientes conclusiones:

- Los parámetros poseen una desviación típica pequeña en comparación con su valor, los *z-value* superan con creces las desviaciones típicas, y se puede afirmar, con un nivel de significación cercano al cero, que ambos parámetros son no-nulos.
- Del test ANOVA se confirma (con un nivel de significación también <2e-16) que el modelo es más válido que el modelo nulo (sin predictores).
- El AIC (criterio de información de Akaike) no es informativo por si solo; es útil para comparar su valor entre varios modelos.

Comprobemos por último, como de bueno sería el modelo si lo transformamos en uno de clasificación, usando los datos de test: si la probabilidad de que un valor pertenezca a 1 es mayor de 0,5 se le atribuye esta categoría; si es menor, se le atribuye la categoría 0.

		Predicciones		
		0	1	
Observaciones	0	34431	13896	**48327**
	1	14762	38909	**53671**
		49193	**52805**	

Cuadro 6: Matriz de confusión

Los valores en verde se tratan de los verdaderos 0 y los verdaderos 1, es decir, los correctamente predichos por el modelo; en rojo, los falsos 0 y los falsos 1.

Se calcula también la exactitud, que indica el porcentaje de datos predichos correctamente (verdaderos 0 y verdaderos 1) respecto al total; la sensibilidad, que supone el porcentaje de verdaderos 1 respecto al total de observaciones de esa categoría (verdaderos 1 y falsos 0) y la precisión, que es el porcentaje de verdaderos 1 respecto al total de valores 1 predichos (verdaderos 1 y falsos 1).

Exactitud (accuracy)	**Sensibilidad (recall)**	**Precisión (precision)**
71,90 %	72,50 %	73,68 %

Cuadro 7: Coeficientes del modelo de clasificación

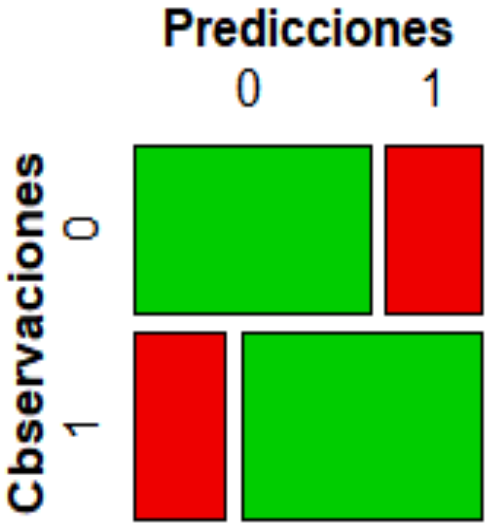

Figura 25: Representación gráfica de la matriz de confusión

Por lo tanto, el modelo de regresión logística da lugar a un modelo de clasificación razonablemente bueno.

6. Regresión bayesiana

Repetimos el ajuste lineal para la **altura significante** y la **altura máxima de ola** nuevamente, pero ahora desde el enfoque bayesiano, por lo que no obtendremos un valor medio y una desviación típica para los dos parámetros (pendiente y ordenada en el origen), sino una distribución de probabilidad.

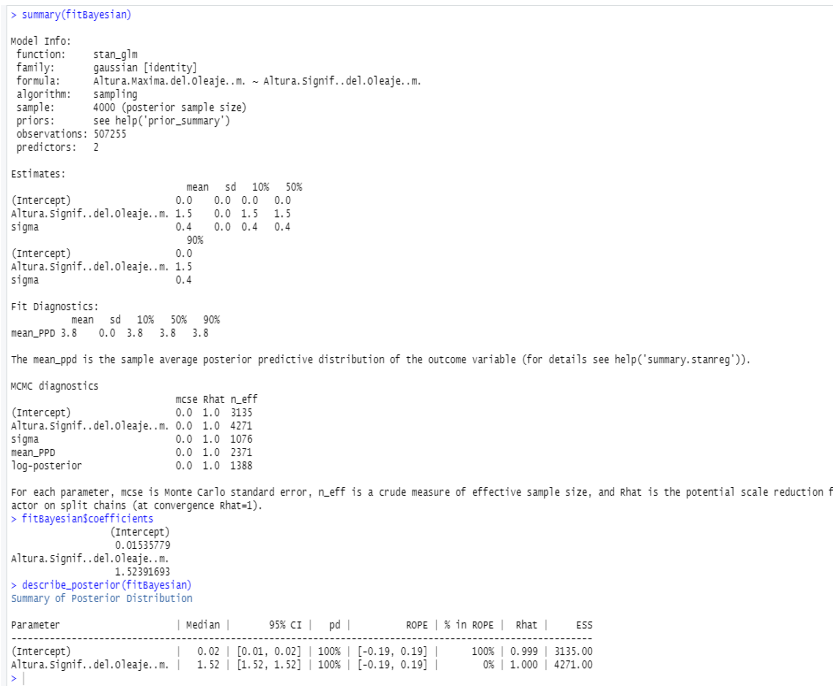

Figura 26: Salida de la función *stan_glm*

De los datos obtenidos de la regresión lineal bayesiana se observa lo siguiente:

- La media de los parámetros bayesianos se acerca a los parámetros frecuentistas salvo una pequeña diferencia lógica.
- *mean_PDD* se parece a la media de la magnitud dependiente (altura máxima), como podemos ver en la tabla 3. Esto no significa que el modelo sea bueno necesariamente, sólo que puede predecir bien la media, pero si ambos valores divergiesen supondría un problema.
- Los valores de *Rhat* son la unidad, lo que indica una convergencia correcta de MonteCarlo.
- El intervalo de confianza para los parámetros del modelo tiene una anchura razonable para ambos.
- El *pd* es similar al *p-value* frecuentista y es correcto que se acerque al 100 % (equivaldría a un *p-value* prácticamente nulo).
- El *ROPE* (*Region of Practical Equivalence*) es el análogo al test de hipótesis frecuentista. No debería ser el mismo para ambos parámetros, pero se trata de una simplificación que considera que las dos magnitudes varían en el mismo rango. El *% in ROPE* indica qué porcentaje de elementos de la muestra de cada parámetro cae en dicho intervalo, por lo que debido a la magnitud de *Intercept* (ordenada en el origen) es lógico que sus valores se encuentren totalmente en el intervalo. Entonces, no podríamos aplicar el test de hipótesis y afirmar que *Intercept* es distinto de cero, dado la simplicidad de cálculo del *ROPE*; pero esto no es problema porque ya ha sido rechazada la hipótesis nula en la regresión frecuentista.
- El *Rhat* de los parámetros es también muy cercano a 1.

Por lo tanto no hay ningún indicio preocupante en la regresión bayesiana. Presentamos las distribuciones para los dos parámetros y para el error estándar residual del modelo, que en la regresión frecuentista también rondaba ese valor, 0,3609 (ver fig.7).

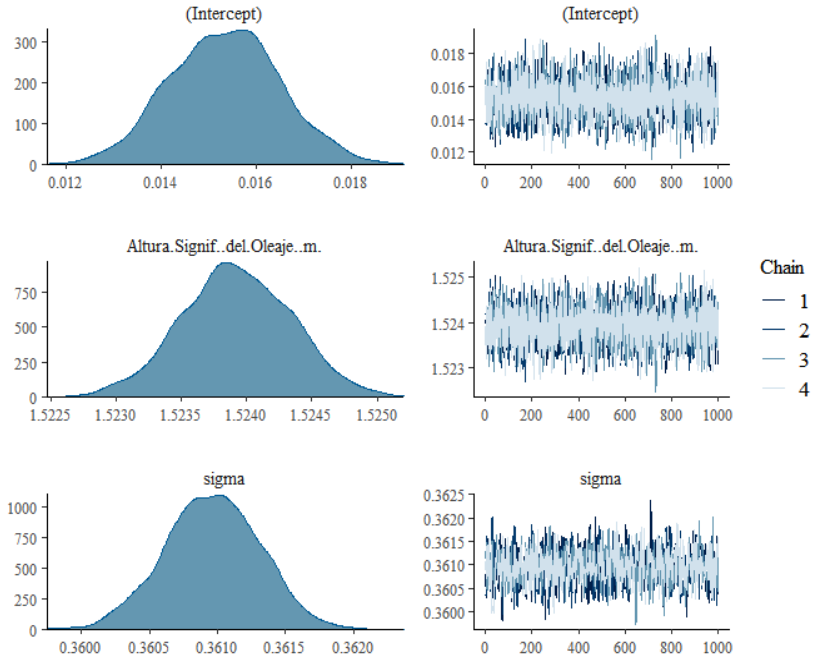

Figura 27: Distribuciones de los parámetros y cadenas de Markov

Las cadenas de Markov presentan también un comportamiento aceptable: varían de manera más o menos azarosa en torno a un valor medio.

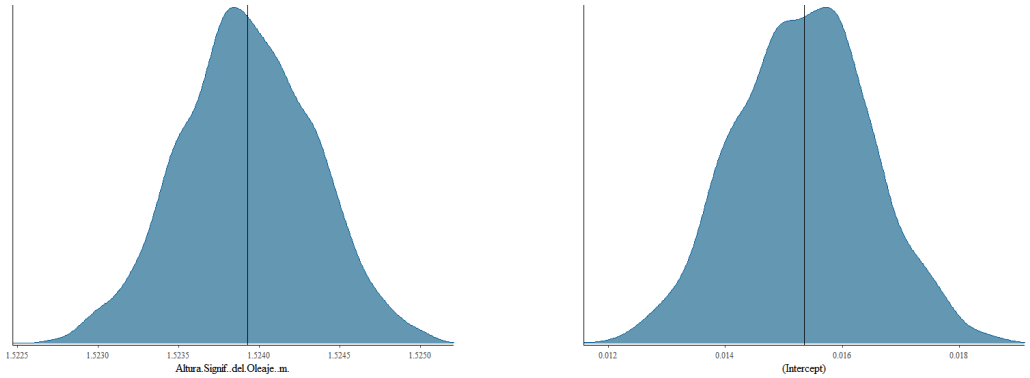

Figura 28: Distribuciones bayesianas con el valor frecuentista

Como era de esperar, los valores frecuentistas de los parámetros se encuentran centrados en las distribuciones bayesianas. Se ha demostrado, por lo tanto, que la regresión bayesiana arroja resultados acordes a los frecuentistas[16].

[16]Sería interesante también obtener la predicción bayesiana de los datos test, con la función *posterior_predict*, pero la memoria de R no lo permitía, dado el tamaño de la muestra de datos.