Pronóstico de datos numéricos: Métodos de regresión II

Modelos lineales generalizados y regresión logística

Como se demostró en el análisis de los datos del lanzamiento del transbordador espacial Challenger, la regresión lineal estándar es un método útil para modelar la relación entre un resultado numérico y uno o más predictores. No es de extrañar que la regresión haya resistido la prueba del tiempo. Incluso después de más de cien años, sigue siendo una de las técnicas más importantes de nuestro conjunto de herramientas, aunque no es más sofisticada que encontrar la mejor línea recta que se ajuste a los datos.

Sin embargo, no todos los problemas son adecuados para ser modelados por una línea y, además, las suposiciones estadísticas realizadas por los modelos de regresión se violan en muchas tareas del mundo real. Incluso los datos del Challenger son menos que ideales para la regresión lineal, ya que violan el supuesto de regresión de que la variable objetivo se mide en una escala continua. Como el número de fallas de juntas tóricas solo puede tomar valores contables, no tiene sentido que el modelo pueda predecir exactamente 2.21 eventos de avería en lugar de dos o tres.

Para modelar valores de conteo, para resultados categóricos o binarios, así como otros casos en los que el objetivo no es una variable continúa distribuida normalmente, la regresión lineal estándar no es la mejor herramienta para el trabajo, aunque muchos todavía la aplican a este tipo de problemas, y a menudo funciona sorprendentemente bien.

Para abordar estas deficiencias, la regresión lineal se puede adaptar a otros casos de uso utilizando el GLM, que fue descrito por primera vez en 1972 por los estadísticos John Nelder y Robert Wedderburn. El GLM flexibiliza dos supuestos del modelado de regresión tradicional.

Primero, permite que el objetivo sea una variable no distribuida normalmente y no continua. Segundo, permite que la varianza de la variable objetivo se relacione con su media. La primera propiedad abre la puerta a la modelización de datos categóricos o de conteo, o incluso casos en los que existe un rango limitado de valores para predecir, como valores de probabilidad que caen en un rango entre 0 y 1. La segunda propiedad permite que el modelo se ajuste mejor a los casos en los que los predictores se relacionan con las predicciones de una manera no lineal, como el crecimiento exponencial en el que un aumento de una unidad de tiempo conduce a aumentos cada vez mayores en el resultado.

Estas dos generalizaciones de la regresión lineal se reflejan en los dos componentes clave de cualquier GLM:

1. La **familia** se refiere a la distribución de la característica objetivo, que debe elegirse entre los miembros de la **familia exponencial** de distribuciones, que incluye la distribución Gaussiana normal, así como otras como Poisson, Binomial y Gamma. La distribución elegida puede ser discreta o continua, y puede abarcar diferentes rangos de valores, como solo valores positivos o solo valores entre cero y uno.

2. La **función de enlace** (*link function*) transforma la relación entre los predictores y el objetivo de tal manera que se pueda modelar mediante una ecuación lineal, a pesar de que la relación original sea no lineal. Siempre hay una función de enlace canónica, que está determinada por la familia elegida y se utiliza de forma predeterminada, pero en algunos casos, se puede elegir un enlace diferente para variar la forma en que se interpreta el modelo o para obtener un mejor ajuste del modelo.

La variación de las funciones de familia y de enlace le otorga al enfoque GLM una enorme flexibilidad para adaptarse a muchos casos de uso diferentes del mundo real y para ajustarse a la distribución natural de la variable objetivo. Saber qué combinación usar depende tanto de cómo se aplicará el modelo como de la distribución teórica del objetivo. Comprender estos factores en detalle requiere conocimiento de las diversas distribuciones en la familia exponencial y una formación en teoría estadística. Afortunadamente, la mayoría de los casos de uso de GLM se ajustan a algunas combinaciones comunes de familia y enlace, que se enumeran en la siguiente tabla:

Familia	**Función de enlace canónico**	**Rango objetivo**	**Notas y Aplicaciones**
Gaussiana (normal)	Identidad	$-\infty$ a ∞	Se utiliza para el modelado de respuesta lineal; reduce el GLM a una regresión lineal estándar.
Poisson	Log	Enteros 0 a ∞	Conocida como regresión de Poisson; modela el recuento de un evento que ocurre (como el número total de fallas de juntas tóricas) estimando la frecuencia con la que ocurre el evento.
Binomial	Logit	0 a 1	Conocida como regresión logística; se utiliza para modelar un resultado binario (como si alguna junta tórica falló) estimando la probabilidad de que ocurra el resultado.
Gamma	Inverso negativo	0 a ∞	Una de las muchas posibilidades para modelar

			datos sesgados a la derecha; se podría utilizar para modelar el tiempo hasta un evento (como los segundos hasta la falla de una junta tórica) o datos de costos (como los costos de reclamos de seguros por un accidente automovilístico).
Multinomial	Logit	1 de K categorías	Conocida como regresión logística multinomial; se utiliza para modelar un resultado categórico (como un lanzamiento exitoso, fallido o abortado de un transbordador) al estimar la probabilidad de que el ejemplo caiga en cada una de las categorías. Generalmente utiliza paquetes especializados en lugar de una función GLM para ayudar a la interpretación.

Debido a los matices de la interpretación de los GLM, se necesita mucha práctica y un estudio cuidadoso para ser experto en la aplicación de solo uno, y pocas personas pueden afirmar ser expertos en el uso de todos ellos. Se dedican libros de texto completos a cada variante de GLM. Afortunadamente, en el dominio del aprendizaje automático, la interpretación y la comprensión son menos importantes que poder aplicar la forma GLM correcta a problemas prácticos y producir predicciones útiles. Si bien este tema del curso no puede cubrir cada uno de los métodos enumerados, una introducción a los detalles clave te permitirá más adelante buscar las variantes del GLM más relevantes para tu propio trabajo.

Comenzando con la variante más simple que se enumera en la tabla, la regresión lineal estándar puede considerarse como un tipo especial de GLM que utiliza la familia gaussiana y la función de enlace de identidad. El **enlace de identidad** implica que la relación entre el objetivo y y un predictor x_i no se transforma de ninguna manera. Por lo tanto, al igual que con la regresión estándar, un parámetro de regresión estimado β_i puede interpretarse de manera bastante simple como el aumento en y dado un aumento de una unidad en x_i, suponiendo que todos los demás factores se mantienen iguales.

Las formas de GLM que utilizan otras funciones de enlace no son tan simples de interpretar y comprender completamente el impacto de los predictores individuales requiere un análisis mucho más cuidadoso. Esto se debe a que los parámetros de regresión deben interpretarse como el aumento aditivo en y para un aumento de una unidad en x_i, pero solo después de ser transformados a través de la función de enlace.

Por ejemplo, la familia Poisson que utiliza la función de enlace logarítmico para modelar el recuento esperado de eventos relaciona *y* con los predictores x_i a través del logaritmo natural; en consecuencia, el efecto aditivo de $\log(\beta_1 x_1) + \log(\beta_2 x_2)$ sobre *y* se vuelve multiplicativo en la escala original de la variable de respuesta. Esto se debe a que, al utilizar las propiedades de los logaritmos, sabemos que $\log(\beta_1 x_1) + \log(\beta_2 x_2) = \log(\beta_1 x_1 * \beta_2 x_2)$, y esto se convierte en $\beta_1 x_1 * \beta_2 x_2$ después de exponenciar para eliminar el logaritmo.

Debido a este impacto multiplicativo, las estimaciones de los parámetros se entienden como tasas relativas de aumento en lugar del aumento absoluto de *y* como en la regresión lineal.

Para ver esto en la práctica, supongamos que construimos un modelo de regresión de Poisson del recuento de fallas de juntas tóricas en función de la temperatura de lanzamiento. Si x_1 es la temperatura y el β_1 estimado = -0.103, entonces podemos determinar que hay alrededor de un 9.8 por ciento menos de fallas de juntas tóricas en promedio por cada grado adicional de temperatura en el lanzamiento.

Esto se debe a que exp(-0.103) = 0.902, o el 90.2 por ciento de las fallas por grado, lo que implica que esperaríamos un 9.8 por ciento menos de fallas por cada grado de aumento. Aplicando esto a la temperatura de lanzamiento del transbordador Challenger a 36 grados Fahrenheit, podemos extrapolar que un lanzamiento 17 grados más cálido (53 grados Fahrenheit fue el lanzamiento más frío anterior) habría tenido aproximadamente (0.902)^17 = 17.2 por ciento del número esperado de fallas, equivalente a una caída del 82.8 por ciento.

La variante GLM que utiliza una distribución de familia binomial con una función de enlace logit se conoce como **regresión logística**, y es quizás la forma más importante, ya que permite adaptar la regresión a tareas de clasificación binaria. El **enlace logit** es una función en la forma $\log(p / (1 - p))$, donde *p* es una probabilidad; la parte interna $(p/(1 - p))$ expresa la probabilidad como probabilidades, exactamente como las probabilidades utilizadas en los juegos de azar y las apuestas deportivas en frases como “el equipo tiene una probabilidad de 2:1 de ganar”.

Después de tomar el logaritmo natural, los coeficientes de regresión estimados se interpretan como probabilidades logarítmicas. Como entendemos las probabilidades de manera más intuitiva que las probabilidades logarítmicas, generalmente exponenciamos los coeficientes de regresión logística estimados para convertir las probabilidades logarítmicas en probabilidades para interpretación. Sin embargo, como los coeficientes de regresión logística indican la diferencia en las probabilidades de *y* debido a un aumento de una unidad en *x*, las probabilidades exponencializadas se convierten en **razones de probabilidades** (*odds ratios*), que expresan el aumento o disminución relativa en las probabilidades de que *y* suceda.

En el contexto de los datos del transbordador espacial, supongamos que construimos un modelo de regresión logística para la tarea de clasificación binaria de predecir si un lanzamiento tendría o no una o más fallas de junta tórica. Un factor que no cambia la probabilidad de una falla de junta tórica mantendría las probabilidades equilibradas en 1:1 (probabilidad 50-50), lo que se traduce en probabilidades logarítmicas de log(0.5/(1 – 0.5)) = 0 y el coeficiente de regresión estimado $\beta = 0$ para esta característica. Si se obtiene la razón de probabilidades como exp(0) = 1, se observa que las probabilidades permanecen invariables independientemente del valor de este factor.

Ahora bien, supongamos que un factor como la temperatura reduce la probabilidad de que se produzca el resultado, y que en el modelo de regresión logística con x_1 como temperatura, entonces el β_1 estimado = -0.232. Al exponenciar esto, obtenemos la razón de probabilidades exp(-0.232) = 0.793, lo que significa que las probabilidades de un fallo caen aproximadamente un 20 por ciento por un aumento de un grado en la temperatura, suponiendo que todo lo demás se mantiene igual. Es muy importante señalar que esto no implica que la probabilidad de un fallo caiga un 20 por ciento por cada aumento de un grado.

Como la relación entre probabilidades y posibilidades no es lineal, el impacto de un cambio de temperatura en la probabilidad de falla depende del contexto en el que se produce el cambio de temperatura.

Las probabilidades y posibilidades están relacionadas a través de la conexión inversa entre las funciones logit y logística. La función logística tiene la propiedad conveniente de que para cualquier valor de entrada *x*, la salida es un valor en el rango entre 0 y 1, exactamente el mismo rango que una probabilidad. Además, la función logística crea una curva en forma de S cuando se grafica, como se ilustra en la figura 6, que muestra un modelo de regresión logística hipotético de la probabilidad de falla de la junta tórica en función de la temperatura de lanzamiento. La probabilidad de falla en el eje *y* cambia más fuertemente en la mitad del rango de temperatura; en temperaturas extremas, la probabilidad de falla predicha cambia muy poco por cada grado adicional de temperatura agregado o eliminado.

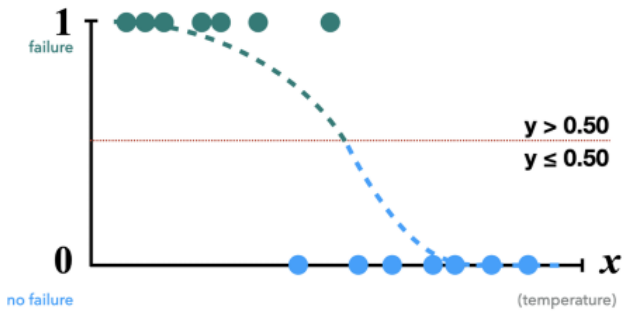

Figura 6: Curva de regresión logística hipotética que representa los datos del lanzamiento del transbordador espacial.

El modelo de regresión logística ajustado crea una curva que representa una estimación de probabilidad en una escala continua en cualquier punto del rango entre 0 y 1, a pesar de que el resultado objetivo (representado por círculos en la figura) solo tome el valor $y = 0$ o $y = 1$. Para obtener la predicción binaria, simplemente define el umbral de probabilidad dentro del cual se debe predecir el resultado objetivo. Por ejemplo, si la probabilidad predicha de una falla de junta tórica es mayor que 0.50, entonces predice "falla" y, de lo contrario, predice "no falla". El uso de un umbral del 50 por ciento es común, pero se puede usar un umbral más alto o más bajo para ajustar la sensibilidad del modelo a los costos.

El examen de la curva logística en la figura 6 conduce a otra pregunta: ¿cómo determina el algoritmo de modelado la curva que mejor se ajusta a los datos? Después de todo, dado que no se trata de una línea recta, el algoritmo OLS utilizado en la regresión lineal estándar ya no parece aplicarse.

De hecho, los modelos lineales generalizados utilizan una técnica diferente llamada estimación de máxima verosimilitud (MLE, **Maximum likelihood estimation**), que encuentra los valores de los parámetros para la distribución especificada que tienen más probabilidades de haber generado los datos observados.

Debido a que la estimación de OLS es un caso especial de estimación de máxima verosimilitud, el uso de OLS o MLE para un modelo lineal no hace ninguna diferencia siempre que se cumplan los supuestos del modelado de OLS. Para aplicaciones fuera del modelado lineal, la técnica MLE producirá resultados diferentes y debe usarse en lugar de OLS. La técnica MLE está incorporada en el software de modelado GLM y, por lo general, aplica técnicas analíticas para identificar los parámetros óptimos del modelo iterando repetidamente sobre los datos en lugar de encontrar la solución correcta directamente. Afortunadamente, como veremos en breve, construir un GLM en R no es más desafiante que entrenar un modelo lineal más simple.

Esta introducción solo arañó la superficie de lo que es posible con la regresión lineal y el GLM.
Aunque la teoría y los ejemplos simples como el conjunto de datos Challenger son útiles para comprender cómo funcionan los modelos de regresión, hay más involucrado en la construcción de un modelo útil de lo que hemos visto hasta ahora. Las funciones de regresión integradas de R incluyen la funcionalidad adicional necesaria para ajustar modelos más sofisticados y, al mismo tiempo, brindan una salida de diagnóstico adicional para ayudar a la interpretación del modelo y evaluar el ajuste. Apliquemos estas funciones y ampliemos nuestro conocimiento de la regresión al intentar realizar una tarea de aprendizaje del mundo real.

Realizaremos dos ejemplos de regresión: lineal y logística.

Comprensión de los árboles de regresión y los árboles modelo

Si recuerdas anteriormente en el tema de divide y conquista: Clasificación mediante árboles de decisión y reglas, un árbol de decisión crea un modelo, muy parecido a un diagrama de flujo, en el que los nodos de decisión, los nodos de hoja y las ramas definen una serie de decisiones que se utilizan para clasificar ejemplos.

Dichos árboles también se pueden utilizar para la predicción numérica haciendo solo pequeños ajustes al algoritmo de crecimiento del árbol.

En esta sección, consideraremos las formas en que los árboles para la predicción numérica difieren de los árboles utilizados para la clasificación. Los árboles para la predicción numérica se dividen en dos categorías.

La primera, conocida como **árboles de regresión**, se introdujo en la década de 1980 como parte del algoritmo seminal de árboles de clasificación y regresión (CART, **Classification and regression tree**).

A pesar del nombre, los árboles de regresión no utilizan métodos de regresión lineal como se describió anteriormente en este tema; en cambio, hacen predicciones basadas en el valor promedio de los ejemplos que llegan a una hoja.

El segundo tipo de árbol para la predicción numérica se conoce como **árbol modelo**. Introducidos varios años después que los árboles de regresión, son menos conocidos, pero quizás más potentes. Los árboles modelo se desarrollan de forma muy similar a los árboles de regresión, pero en cada hoja se construye un modelo de regresión lineal múltiple a partir de los ejemplos que llegan a ese nodo. Según la cantidad de nodos de hoja, un árbol modelo puede construir decenas o incluso cientos de estos modelos.

Esto hace que los árboles modelo sean más difíciles de entender que el árbol de regresión equivalente, con el beneficio de que pueden dar como resultado un modelo más preciso.

Incorporación de regresión a los árboles

Los árboles que pueden realizar predicciones numéricas ofrecen una alternativa convincente, aunque a menudo pasada por alto, al modelado de regresión. Las fortalezas y debilidades de los árboles de regresión y los árboles de modelos en relación con los métodos de regresión más comunes se enumeran en la siguiente tabla:

Fortalezas	Debilidades
• Combina las fortalezas de los árboles de decisión con la capacidad de modelar datos numéricos • No requiere que el usuario especifique el modelo por adelantado • Utiliza la selección automática de características, lo que permite que el enfoque se utilice con una gran cantidad de características • Puede ajustar algunos tipos de datos mucho mejor que la regresión lineal • No requiere conocimientos de estadística para interpretar el modelo	• No es tan conocido como la regresión lineal • Requiere una gran cantidad de datos de entrenamiento • Es difícil determinar el efecto neto general de las características individuales en el resultado • Los árboles grandes pueden volverse más difíciles de interpretar que un modelo de regresión

Aunque los métodos de regresión tradicionales suelen ser la primera opción para las tareas de predicción numérica, en algunos casos, los árboles de decisión numéricos ofrecen ventajas distintivas. Por ejemplo, los árboles de decisión pueden ser más adecuados para tareas con muchas características o muchas relaciones complejas y no lineales entre las características y el resultado; estas situaciones presentan desafíos para la regresión. El modelado de regresión también hace suposiciones sobre los datos que a menudo se violan en los datos del mundo real; este no es el caso de los árboles.

Los árboles para la predicción numérica se construyen de la misma manera que para la clasificación. Comenzando en el nodo raíz, los datos se dividen utilizando una estrategia de dividir y conquistar de acuerdo con la característica que dará como resultado el mayor aumento en la homogeneidad en el resultado después de realizar una división.

En los árboles de clasificación, recordarás que la homogeneidad se mide por la entropía. Esto no está definido para los datos numéricos. En cambio, para los árboles de decisión numéricos, la homogeneidad se mide por estadísticas como la varianza, la desviación estándar o la desviación absoluta de la media.

Un criterio de división común se denomina reducción de la desviación estándar (SDR, **Standard deviation reduction**). Se define mediante la siguiente fórmula:

$$\text{SDR} = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

En esta fórmula, la función *sd*(*T*) se refiere a la desviación estándar de los valores del conjunto *T*, mientras que T_1, T_2, ..., T_n son conjuntos de valores resultantes de una división en una característica. El término |T| se refiere al número de observaciones del conjunto T. Básicamente, la fórmula mide la reducción de la desviación estándar comparando la

desviación estándar anterior a la división con la desviación estándar ponderada posterior a la división.

Por ejemplo, considera el siguiente caso en el que un árbol decide si realizar una división en la característica binaria A o una división en la característica binaria B:

original data	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature A	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
split on feature B	1	1	1	2	2	3	4	5	5	6	6	7	7	7	7
				T_1								T_2			

Figura 7: El algoritmo considera divisiones en las características A y B, lo que crea diferentes grupos T_1 y T_2.

Usando los grupos que resultarían de las divisiones propuestas, podemos calcular la SDR para A y B de la siguiente manera. La función length() utilizada aquí devuelve la cantidad de elementos en un vector. Ten en cuenta que el grupo general T se llama tee para evitar sobrescribir las funciones T() y t() integradas de R.

```
> tee <- c(1, 1, 1, 2, 2, 3, 4, 5, 5, 6, 6, 7, 7, 7, 7)
> at1 <- c(1, 1, 1, 2, 2, 3, 4, 5, 5)
> at2 <- c(6, 6, 7, 7, 7, 7)
> bt1 <- c(1, 1, 1, 2, 2, 3, 4)
> bt2 <- c(5, 5, 6, 6, 7, 7, 7, 7)
> sdr_a <- sd(tee) - (length(at1) / length(tee) * sd(at1) +
+   length(at2) / length(tee) * sd(at2))
> sdr_b <- sd(tee) - (length(bt1) / length(tee) * sd(bt1) +
+   length(bt2) / length(tee) * sd(bt2))
```

Comparemos la SDR de A con la SDR de B:

```
> sdr_a
[1] 1.202815
>  sdr_b
[1] 1.392751
```

La SDR para la división en la característica A fue de aproximadamente 1.2 frente a 1.4 para la división en la característica B. Dado que la desviación estándar se redujo más para la división en B, el árbol de decisión utilizaría B primero. Esto da como resultado conjuntos ligeramente más homogéneos que A.

Supongamos que el árbol dejó de crecer aquí utilizando esta única división. El trabajo de un árbol de regresión está hecho. Puede hacer predicciones para nuevos ejemplos dependiendo de si el valor del ejemplo en la característica B coloca al ejemplo en el grupo T_1 o T_2. Si el ejemplo termina en T_1, el modelo predeciría mean(bt1) = 2, de lo contrario predeciría mean(bt2) = 6.25.

En cambio, un árbol modelo iría un paso más allá. Usando los siete ejemplos de entrenamiento que caen en el grupo T_1 y los ocho en T_2, el árbol modelo podría construir un modelo de regresión lineal del resultado versus la característica A. Ten en cuenta que la característica B no es de ayuda para construir el modelo de regresión porque todos los ejemplos en la hoja tienen el mismo valor de B: se colocaron en T_1 o T_2 de acuerdo con su valor de B. El árbol modelo puede entonces hacer predicciones para nuevos ejemplos usando cualquiera de los dos modelos lineales.

Para ilustrar aún más las diferencias entre estos dos enfoques, trabajemos con un ejemplo del mundo real (Ejemplo de la calidad de los vinos).

Resumen

En este tema, estudiamos dos métodos para modelar datos numéricos. El primer método, la regresión lineal, implica ajustar líneas rectas a los datos, pero una técnica llamada modelado lineal generalizado también puede adaptar la regresión a otros contextos. El segundo método utiliza árboles de decisión para la predicción numérica. Este último se presenta en dos formas: árboles de regresión, que utilizan el valor promedio de los ejemplos en los nodos de hoja para hacer predicciones numéricas, y árboles de modelo, que construyen un modelo de regresión en cada nodo de hoja en un enfoque híbrido que es, en cierto modo, lo mejor de ambos mundos.

Empezamos a comprender la utilidad del modelado de regresión al usarlo para investigar las causas del desastre del transbordador espacial Challenger. Luego usamos el modelado de regresión lineal para calcular los costos esperados de reclamos de seguros para varios segmentos de conductores de automóviles.

Debido a que la relación entre las características y la variable objetivo está bien documentada por el modelo de regresión estimado, pudimos identificar ciertos grupos demográficos, como los conductores de alto kilometraje y los que trabajan a altas horas de la noche, a quienes se les puede cobrar tarifas de seguro más altas para cubrir sus costos de reclamos superiores al promedio. Luego aplicamos la regresión logística, una variante de la regresión utilizada para la clasificación binaria, a la tarea de modelar la retención de clientes de seguros. Estos

ejemplos demostraron la capacidad de la regresión para adaptarse de manera flexible a muchos tipos de problemas del mundo real.

En una aplicación algo menos empresarial del aprendizaje automático, se utilizaron árboles de regresión y árboles de modelos para modelar la calidad subjetiva de los vinos a partir de características mensurables. Al hacerlo, aprendimos cómo los árboles de regresión ofrecen una forma sencilla de explicar la relación entre las características y un resultado numérico, pero los árboles de modelos más complejos pueden ser más precisos.

En el camino, aprendimos nuevos métodos para evaluar el desempeño de los modelos numéricos. En marcado contraste con este tema, que abordó los métodos de aprendizaje automático que dan como resultado una comprensión clara de las relaciones entre la entrada y la salida, el siguiente tema cubre métodos que dan como resultado modelos casi incomprensibles (Redes neuronales que no revisaremos y máquinas de soporte vectorial). La ventaja es que son técnicas extremadamente poderosas (entre los clasificadores de acciones más poderosos) que se pueden aplicar tanto a problemas de clasificación como de predicción numérica.