Introducción al aprendizaje estadístico

Índice

El aprendizaje estadístico se refiere a una conjunto de herramientas y entendimiento de conjuntos de datos complejos. Se ha desarrollado recientemente en el area de estadística y se mezcla con el desarrollo en pararelo de las ciencias computaciones, en particular, con aprendizaje de máquina.

1. Introducción

Aprendizaje estadístico se refiere a un vasto conjunto de herramientas para entender los datos. Estas herramientas pueden ser clasificadas como con supervisión o sin supervisión. Amplimente hablando, aprendizaje estadístico con supervisión se refiere a construir un modelo estadístico para predecir, o estimar, un valor de salida dado uno o mas valores de entrada. En el aprendizaje estadístico sin supervisión, hay valores de entrada pero no hay valores de salida; aún así podemos aprender las relaciones y estructura de los datos.

En este curso no se discuten los detalles técnicos relacionados con el ajuste de los procedimientos y propiedades teóricas.

2. Aprendizaje estadístico

En escencia, al hablar de aprendizaje estadístico nos refererimos a un conjunto de enfoques para estimar f. Empezamos con los conceptos teóricos clave para estimar f y con las herramientas para evaluar los estimadores que obtengamos. Nos interesa estimar f para dos cosas: inferencia y prediccion.

2.1. ¿Qué es el aprendizaje estadístico ? En general, supongamos que observamos una respuesta cuantitativa Y y que $X_1, X_2, ..., X_p$ son las predictoras. Asumimos que hay alguna relación entre Y y $X = (X_1, ...X_p)$, la cual puede ser escrita en la forma general

$$Y = f(X) + \epsilon$$

Aquí f es alguna función de $X_1, ..., X_p$ fija pero desconocida, y ϵ es un término de error aleatorio, el cual es independiente de X y tiene media cero. En esta formulación, f representa la información sistemática que X provee acerca de Y.

En general, la función f podría involucrar más de una variable de entrada. Esta funcion f debe ser estimada basada en los datos observados.

En esencia, el aprendizaje estadístico se refiere a una conjunto de enfoques para estimar f. En este capítulo indicaremos algunos de los conceptos teoréticos clave para estimar a f, así como herramientas para evaluar los estimados obtenidos.

2.2. Por qué estimar f. Hay dos razones principales por la que queremos estimar a f: predicción e inferencia. Discutiremos cada uno.

Predicción

En muchas situaciones, el conjunto de varibales de entrada X es fácil de obtener pero la respuesta Y es más complicada. En esta configuración, como los errores promedian cero, podemos predecir Y usando

$$\tilde{Y} = \tilde{f}(X),$$

donde $\tilde{f}$ representa nuestra estimación para f, y $\tilde{Y}$ representa la predicción para Y. En esta configuración, $\tilde{f}$ es tratada como una caja negra en el sentido de que uno no está típicamente preocupado con la forma exacta de $\tilde{f}$, mientras esta mantenga predicciones precisas para Y.

Como ejemplo, supongamos que $X_1, ..., X_p$ son las características de la sangre de un paciente y Y es la variable que describe el riesgo a una reacción adversa del paciente al tomar cierto medicamento. Es natural quere predecir Y usando X para evitar una reacción adversa en un paciente.

La precisión de $\tilde{Y}$ como predictor para Y depende de dos cantidades, las cuales llamaremos el **error reducible** y el **error irreducible**. En general $\tilde{f}$ no será un estimador perfecto de f y esta impresición introduce un errror. Este error es reducible pués podemos escoger la técnica de aprendizaje estadístico más apropiada. Aún si es posible encontrar la forma perfecta de f y nuestra respuesta estimada es

$$\tilde{Y} = f(X)$$

nuestra predicción todavía tendría algo de error ahi pués Y también es una función de ϵ. Así la variabilidad asociada con ϵ afecta la precisión de nuestras predicciones.

Esto es lo que conocemos como el error irreducible.

El error irreducible es positivo pués ϵ podría tener variables no medibles que son útiles para predecir Y. Además también podría contener una variación no medible. Por ejemplo, el riesgo de una reacción adversa puede variable en cada paciente dependiendo del día, dependiendo en la variación de la producción de la medicina o en el estado de ánimo del paciente.

Considere una estimación $\tilde{f}$ y un conjunto de predictoras X, y que obtenemos la predicción $\tilde{Y} = \tilde{f}(X)$. Asuma por un momento que tanto $\tilde{f}$ como X son fijas. Entonces, es fácil mostrar que

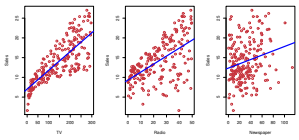
$$E(Y - \tilde{Y})^2 = E(f(X) + \epsilon - \tilde{f}(X))^2 = (f(X) - \tilde{f}(X))^2 + Var(\epsilon),$$

donde $E(Y - \tilde{Y})^2$ representa el primedio, or el valor esperado, de la distance al cuadrado entre la prediccin y el valor actual de Y, y $Var(\epsilon)$ representa la varianza asociada con el término de error ϵ. Nos enfocaremos en este curso en minimzar el error reducible $(\tilde{f}(X))^2 = (f(X) - \tilde{f}(X))^2$). El error irreducible siempre proveerá una cota superior para la precisión de nuestra predicción para Y. Esta cota es casi siempre desconocida en la práctica.

Inferencia

Frecuentemente estamos interesados en entender la forma en que Y es afectada por cambios en $X_1, ..., X_p$. En esta situación, queremos estimar f pero nuestro objetivo no es hacer predicciones para Y. Lo que queremos es entender la relación entre Y y X o más específicamente, como Y cambia siendo una función de $X_1, ..., X_p$. Ahora $\tilde{f}$ no puede ser tratada como una caja negra por que lo que necesitamos saber en su forma exacta. En esta configuración puede estar interesado en responder las siguientes preguntas.

1. Qué variables predictoras están asociadas con la respuesta? A veces solo una parte de todas las variables estan sustancialmente asociadas a Y.
2. Cuál es la relación entre la respuesta y cada una de las predictoras ? Algunas variables tienen una relación positiva y algunas otras el sentido opuesto.
3. Puede la relación entre Y y cada una de las predictoras ser apropiadamente resumida usando una ecuación lineal, o es la relación más complicada ? Históricamente la mayoría de los métodos para estimar f han tenido una forma lineal.



Veremos ejemplos de predicción, inferencia y una combinación de ambos.
Cómo estimaremos a f?

En este curso exploraremos varios enfoques lineales y no lineales para estimar f. Estos métodos en general comparten algunas características. Siempre asumiremos que hemos observado un conjunto de n puntos diferentes. A este conjunto le

llamaremos los datos de entrenamiento pués usaremos estas observaciones para entrenar, o estimar a nuestro método cómo estimar f. Sea x_{ij} el valor que representa la observación i en el predictor j para $i = 1, ..., p$ y $j = 1, ...p$. Correspondientemente, sea y_i la variable respuesta para la observación i. Entonces nuestro conjunto de entrenamiento consiste en los puntos $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ donde $x_i = (x_{i1}, x_{i2}, ...x_{ip})^T$.

Nuestro objetivo es aplicar un método de aprendizaje estadístico a los datos de entrenamiento para estimar la función desconocida f. La mayoría de los métodos para estimar f están en la categorías de parámetricos o no paramétricos.

Métodos Paramétricos

1. Asumimos una forma de la funcionón de f. Por ejemplo, supongamos que f es lineal en X:

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Éste es un modelo lineal, el cual lo analizaremos en el cap
itulo 3. Solo es necesario encontrar los valores de los coeficientes β.

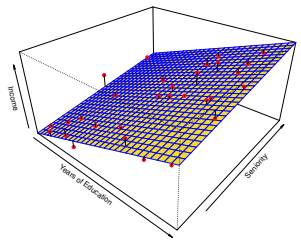
2. Ya que seleccionamos el modelo usaremos los datos de entrenamiento para ajustar el (o entrenar) modelo. Necesitamos encontrar $\beta_0, \beta_1, ..., \beta_p$ tal que

$$Y \approx \beta_0 + \beta_1 X_1 + ... + \beta_p X_p.$$

El método más común para ajustar el modelo en 1 es mínimos cuadrados aunque hay otras maneras las cuales las dicutiremos en el cap
'itulo 6.
Una de las desventajas de usar este enfoque paramétrico es que el modelo que escojamos podría no concordar con la verdadera función f. Y así nuestras estimaciones no serían buenas. Podríamos evitar este problema escogiendo un modelo que sea más flexible y que pueda ajustar muchas formas diferentes para f. Pero en general los modelos más flexibles necesitan más parámetros, los cuales hay que encontrar. Estos modelos más complicados pueden llevar al fenómeno de sobreajustar los datos, lo cual significa que siguen los errores, or ruido. Discutiremos este fenómeno a lo largo del curso.

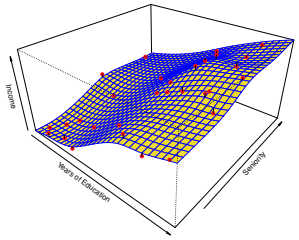
En la figura 2.4 vemos un ejemplo de un método paramétrico aplicado a los datos de income. Se ha ajustado modelo lineal de la forma

$$income \approx \beta_0 + \beta_1 education + \beta_2 seniority$$



Métodos No Paramétricos
Estos métodos no hacen supuesto explícitios acerca de la forma de f. Buscan un estimado de f cerca de los datos. Tienen la ventaja de poder ajustar una gran

rango de formas de f. Una gran desventaja es que se necesita una gran cantidad de datos para poder estimar con precisión a f. En la figura 2.5 vemos el método thin-plate spline para estimar f. La superficie amarilla aproxima bastante bien f. Pero el spline que se observa en la figura 2.6 tiene mucha mas variacion que la funcion original f. Este es un ejemplo de overfitting. El problema es que las respuestas obtenidas con nuevos datos no serán muy precisas. Veremos mas ejemplos a detalle en los capítulo 5 y 7.



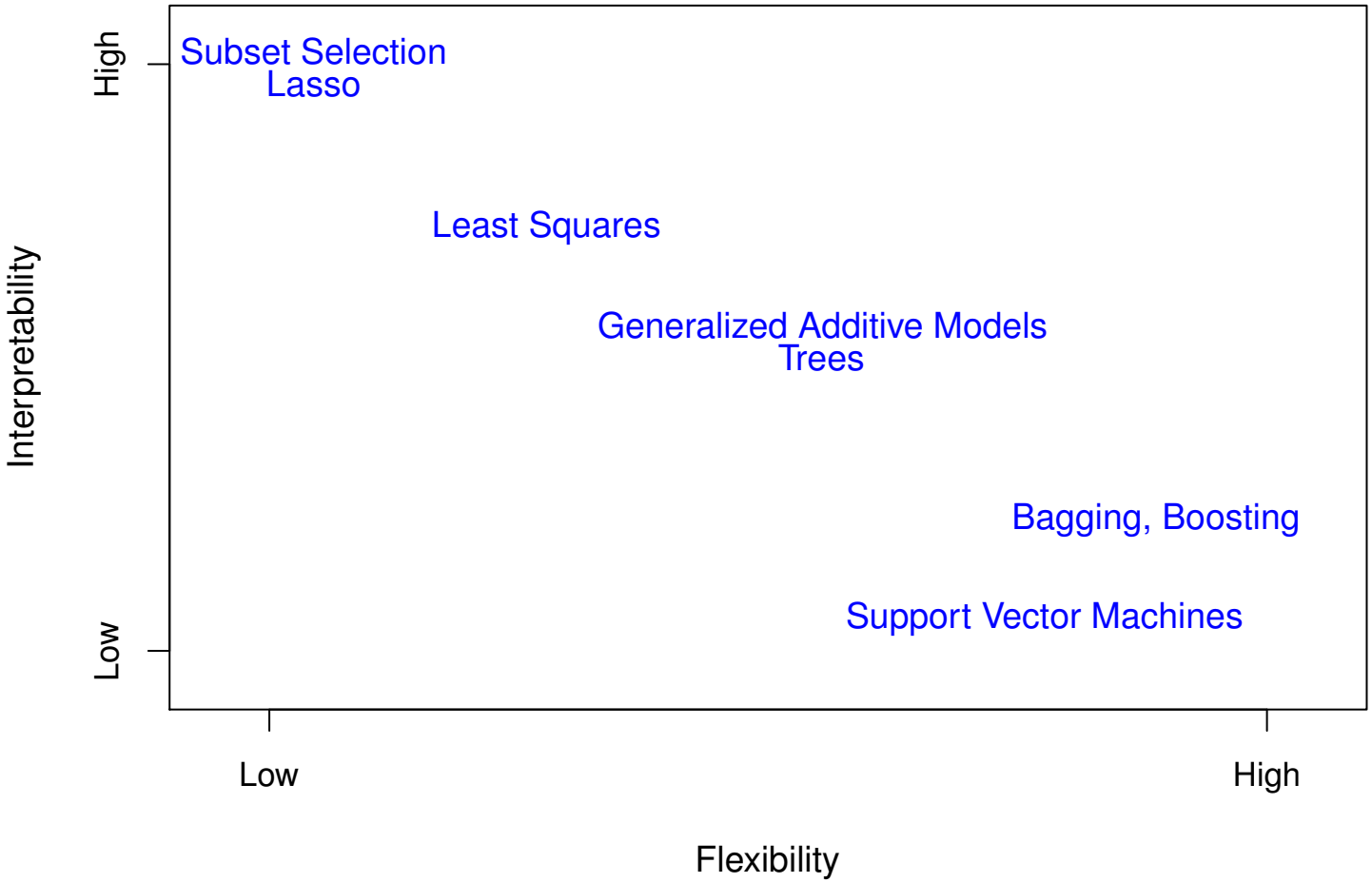
Hay ventajas y desventajas en los métodos paramétricos y no paramétricos para el aprendizaje estadístico. Exploraremos ambos tipos de métodos a lo largo del curos.

La compensación entre precisión de predicción y la interpretación del modelo.

De los muchos métodos que veremos en este curso, algunos son poco flexibles (o mas restrictivos), en el sentido de que solo pueden producir un conjunto chico de formas para estimar f. Por ejemplo, regresión lineal puede generar solo funciones lineales como lineas o planos que vumos en las figuras anteriores. Los métodos como plate splines de las figuras que hemos vistas son mas flexibles pués pueden tomar muchas formas para estimar a f.

Si estamos interesados en la inferencia, los métodos restrictivos son más interpretables. En los métodos más flexibles es más difícil entender como cada predictor es asociado con la respuesta.

En la siguiente figura vemos un representación de la compensació entre flexibilidad e interpretabilidad usando diferentes métodos de aprendizaje estadístico.



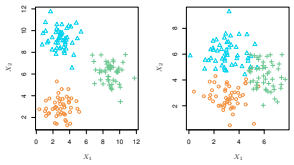
Uno esperaría que si solo nos interesa la predicción usaremos un método que sea muy flexible, pero esto no simpre es el caso! Esto tiene que ver con el fenómeno de overfitting.

Supervision vs no supervision

La mayoría de los problemas de aprendizaje estadístico caen en una de dos categorías: supervisión o sin supervisión. Los ejemplos que hemos visto son problemas de supervisión. La mayor parte de este curso está dedicada a este tipo de problemas.

Cuando tenemos observaciones y no hay una variable de respuesta estamos frente a un problema de aprendizaje sin supervisión. Qué tipo de análisis estadístico

podemos llevar a cabo ? Buscamos entender las relaciones entre las variables o entre las observaciones. Una herramienta que utilizaremos será el cluster analysis. El objetivo del cluster analysis es en que grupo colocar una observación.



En la figura anterior podemos ver que hay observaciones con dos variables X_1 y X_2. Cada observación corresponder a uno de los tres grupos distintos. En realidad uno no sabe cuantos grupos hay en un problema sin supervision. la imagen de la izquiera es fácil distinguir los grupos, en la de la derecha no lo tan fácil. En el capítulo 10 examinaremos este tipo de problemas.

Muchos problemas pueden ser tratados fácilmente como con supervisión o sin supervisión. Pero en algunos casos no está claro si un problema debe tratarse como alguno de estas dos opciones.

Regresión vs Clasificación

Las variables pueden ser clasificadas como cuantitativas o cualitativas (categóricas). Las cuantitativas toman valores numéricos mientras que la cualitativas toman un valor de alguna clase. Ejemplos de cualitativas son: genéro (masc o fem), Marca de carro (BMW, Mercedez, GM), etc. Tenedemos a referirnos a los problemas con valores de respuesta quantitativos como problemas de regresión mientras que los que involucran una respuesta cualitativa son referidos como problemas de clasificación. Es claro que la regresión lineal es un problema de regresión. Mientras la regresión logística normalmente es tratada como un problema de clasificación (respuesta binaria). Pero también puede ser considerada como de regresión pués estima probabilidades.

Escogemos un método de aprendizaje estadístico basados en el tipo de variable de respuesta. Es menos importante saber si las predictoras son cuantitativas o cualitativas siempre y cuando las cualitativas estén propiamente clasificadas.

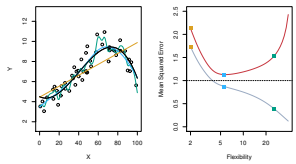
2.3. Evaluación de la precisión del modelo. En este curso veremos muchos métodos de aprendizaje estadístico que van más allá de la regresión linel estándar. En estadística no existen un método que domine a todos los demás en todos los posibles conjuntos de datos. Es una tarea importante decidir para un conjunto de datos qué método produce los mejores resultados. Seleccionar el mejor enfoque es uno de los problemas más complejos en la práctica.
En esta sección discutiremos algunos de los conceptos más importantes para seleccionar un procedimiento de aprendizaje estadístico dado un conjunto de datos.

2.3.1. Midiendo la calidad del ajuste. Para evaluar el desempeno de un método de aprendizaje estadístico en un conjunto de datos, necesitamos medir que tan buenas son las predicciones con los valores observados. En regresión la medida más común es el "mean squared error"(MSE), dado por

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{f}(x_i))^2,$$

donde $\tilde{f}(x_i)$ es el valor de la predicción de la función $\tilde{f}$. Claramente, el MSE será chico si las predicciones están cercanas a los valores observados y será grande de los contrario. Pero en realidad no estamos tan interesados en saber el MSE sobre los datos que ya conocemos, sino que queremos minimizar el MSE para datos que no conocemos y que no fueron utilizados para entrenar el modelo, el cual llamaremos test MSE. Queremos escoger el método con el menor test MSE.

No hay garatía de que el método con el mas chico training MSE tambien tenga el el test MSE mas chico. Muchos métodos estadísticos estiman los coeficientes para minimizar el training MSE. En estos métodos el training MSE sera chico pero el test MSE será por lo general mucho mas grande.

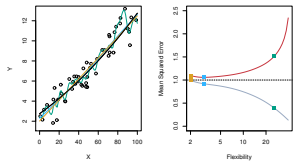


En la figura de la izquierda vemos los datos simulados por f. Y tenemos tres estimaciones para f, la regresión lineal (curva anarajada), y dos smoothing spline (curvas verde y azul). En la derecha tenemos la curva gris que denota el training MSE , la curva roja para el test MSE y la linea punteada el minimo posible test MSE sobre todos los métodos ($\text{Var}(\epsilon)$). Los cuadraditos representa el training y test MSE para los tres ajustes mencionados.
Podemos ver que entre mas flexibilidad, la curva ajusta mejor a los datos. También vemos que la curva verde es la mas flexible y ajusta mejor los datos pero difiere de la curva original f en negro por que ondula mucho. (Podemos ajustar los niveles de flexibilidad en el smoothing spline para producir muchos ajustes diferentes a estos datos).

Podemos ver que la función original f es no lineal, por eso el método lineal no es tan bueno para estimar. Vemos que los métodos flexibles tienen mejor training MSE.

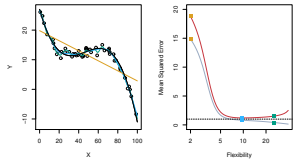
También en la imagen del lado derecho podemos ver que cuando aumenta la flexibilidad del método estadístico, el training MSE decrece monótonamente y vemos una forma de U en el test MSE. Esta es una propiedad fundamental del aprendizaje estadítico que es válida para cualquier conjunto de datos y método estadístico.

Si aumenta la flexibilidad, el training MSE se reducirá pero el test MSE puede que no. Cuando un modelo tenga traininig MSE chico pero un test MSE grande entonces estamos overfitting los datos.
Veamos otro ejemplo con datos con un compartamiento mas lineal.



Podemos ver que la función original f es aproximadamente lineal. También vemos que el training MSE decrece monótomamente cuando la flexibilidad aumenta y que el test MSE toma forma de U. Note que como la función es lineal, si se aumenta la flexibilidad también se aumenta el test MSE.

Y también tenemos otro ejemplo donde las función original f no altamente no lineal.



Vemos que el training MSE y el test MSE tienen un comportamiento muy similar. Pero cuando la flexibilidad aumenta el test MSE empieza a aumentar.

En la práctica es fácil calcular el training MSE, en el capítulo 5 veremos como estimar el test MSE usando el trainning data.

2.3.2. Intercambio Bias-variance. La forma de U que podemos ver en las curvas del test MSE son el resultado de dos propiedades de los métodos de aprendizaje estadístico.

Ahora, sea (x_0, y_0) una observación en el conjunto de prueba. Es posible mostrar que

$$E(y_0 - \tilde{f}(x_0))^2 = Var(\tilde{f}(x_0)) + (Bias(\tilde{f}(x_0)))^2 + Var(\epsilon)$$

Donde $E(y_0 - \tilde{f}(x_0))^2$ define el valor esperado del test MSE y se refiere al promedio test MSE que obtendríamos si repetidamente estimamos f utilizando una gran número de conjuntos de entrenamiento y evaluamos en x_0 cada uno. Bien, ahora si consideramos todos los x_0 del conjunto de prueba podemos encontrar el valor esperado del test MSE general si promediamos $E(y_0 - \tilde{f}(x_0))^2$ sobre todos los x_0 del conjunto de prueba.

La ecuación anterior nos dice que para minimizar el valor esperado del test MSE, necesitamos seleccionar un método que tenga poca varianza y poco bias. Dado que estos dos valores son positivos es claro que el valor esperado del test MSE nunca puede estar por debajo de $Var(\epsilon)$, que es el error irreducible.

¿Qué significa varianza y bias en el aprendizaje estadístico ?
Con la varianza nos referimos a la cantidad en que la estimación $\tilde{f}$ cambiaría si la estimamos usando otro conjunto de datos de entrenamiento. Es claro que cada

conjunto nos proporcionará una $\tilde{f}$ diferente, idealmente quisieramos que nuestra función no cambiara mucho con diferentes conjuntos de datos de entrenamiento. Sin embargo, si el método tiene mucha varianza entonces pequenos cambios en los datos de entrenamiento pueden generar grandes cambios en $\tilde{f}$. Ver la curva verde que sigue a los datos, si cambiamos cualquier dato, $\tilde{f}$ cambiará considerablemente. En el caso de la regresión lineal como es relativamente inflexible, cambiar cualquier datos solo moverá un poco la linea.

Por otro lado, bias se refiere al error que es introducido por la aproximación simplista que hacemos a un problema de la vida real que puede ser muy complicado. Por ejemplo, en la regresión lineal asumimos que hay una relación lineal entre Y y $X_1, ..., X_p$. En la vida real esto es poco probable, así que sin lugar a dudas usar la regresión lineal nos dará un resultado sesgado en el estimado de f. Ver los ejemplos donde f es no lineal.

Como regla general, entre más flexible es el método, la varianza se incrementará pero el bias decrecerá. La tasa relativa de cambio de estas dos cantidades determina si el test MSE crece o decrece. Cuando incrementamos la flexibilidad en una clase de métodos, el bias tiende a decrecer más rápido que el incremento de la varianza. Claro, existe un punto donde al incrementar la flexibilidad tiene poco impacto en el bias pero un impacto significativo en el incremento de la varianza. Cuando esto pasa, el test MSE se incrementa. Ver ejemplos.

La relación que vimos entre el bias, varianza y el test MSE se refieren al trade-off (intercambio) de bias y la varianza. Claramente el reto es escoger un método que tenga poca varianza y poco bias. Es fácil escoger casos extremos donde el sesgo es muy bajo pero la varianza muy alta (ejemplo?) o donde el la varianza es muy baja pero el sesgo muy alto (ejemplo?).

En la realidad cuando no conocemos f no podemos explícitamente calcular el bias, la varianza o el test MSE, pero tenemos que tener en mente ese intercambio. En el capítulo 5 veremos como calcular el test MSE usando los datos de entrenamiento.

2.4. Precisión en clasificación. Hasta ahora solo hemos discutido la precición del modelo cuando tenemos un problema de regresión. Cuando estamos interesados en la clasificación, el enfoque mas común para cuantificar la precición de nuestro estimado $\tilde{f}$ es la tasa de error de entrenamiento:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \tilde{y}_i).$$

Esta ecuación calcula la fracción de clasificaciones erróneas. De igual manera que en la sección anterior, estamos interesado en minimizar la tasa de error de prueba. Sea (x_o, y_o) un punto del conjunto de prueba. Entonces un buen método para clasificar es el que minimiza la siguiente ecuación

$$Prom\left((y_0 \neq \tilde{y}_0).\right)$$

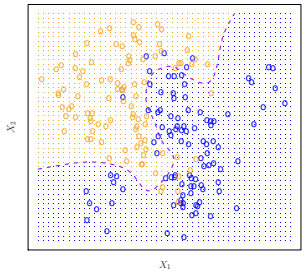
Clasificador de Bayes

Veremos una cota de error para los métodos de clasificación, que está dada por el clasificador de Bayes. Veamos.

Este clasificador asigna a cada observación la clase más probable dados sus valores de predicción. Esto es, simplemente asignamos a una observación de prueba con predictor x_0 la clase j para la cual

$$Pr\left(Y = j|X = x_0\right)$$

es más grande. Este es el clasificador de Bayes, el cual a su vez minimiza $Prom\left((y_0 \neq \tilde{y}_0).\right)$ (la demostración está fuera del alcanze de este curso). Claramente si tenemos dos clases entonces a la observación con la predictora x_0 le asignaremos la clase 1 si $Pr\left((Y = 1|X = x_0\right) > 0{,}5$, y la clase 2 de lo contrario.
En el siguiente ejemplo



Los círculos son los elementos del conjunto de entrenamiento. La línea punteada en lila representa la frontera de decisión de Bayes. Una observación de prueba en la región anaranjada pertenece a la clase anaranjada. De manera similar para la región azul.
El clasificador de Bayes produce el error de prueba más bajo posible, llamado la tasa de error de Bayes. La tasa de error en $X = x_0$ será $1 - \text{máx}_j\, Pr(Y = j|X = x_0)$. La tasa de error de Bayes en general está dada por

$$1 - E\left(\max_j Pr(Y = j|X = x_0)\right)$$

donde el valor esperado promedia la probabilidad sobre todos los posibles valores de X. El error de Bayes es análogo al error irreducible que discutimos anteriormente.

K-Vecinos más cercanos (KNN) En teoría siempre quisiera predecir respuestas cualitativas usando el clasificador de Bayes. Pero con datos reales no conocemos la distribución de Y dado X. Este método (asi como otros) estima esta distribución y luego clasifica una observación con la probabilidad estimada más alta. Dado un entero positivo K y una observación x_0, KNN primero identifica los primeros K puntos más cercanos a x_0, representados por N_0. Entonces estima la probabilidad condicional para la clase j como una fracción de los puntos en N_0 cuyo valor de respuesta equivale j:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j). \tag{1}$$

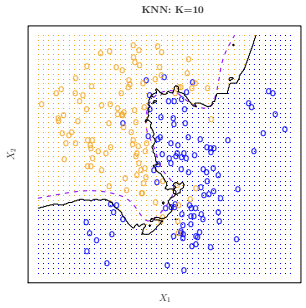
Finalmente, KNN aplica la regla de Bayes y clasifica la observaci
on de prueba a la clase con la más alta probabilidad.

En la imagen anterior, en el panel izquierdo $K = 3$. A la observacion x le asignamos la clase azul. En la imagen de la izquierda vemos la linea de la frontera de la decision de las clases.

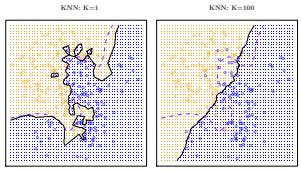
[width=4cm, center]Fig2/214.pdf

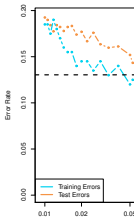
El parámetro K tiene un gran efecto al momento de clasificar. Cuando K crece el método se hace menoss flexible y produce que la linea de la frontera de la decision sea cercana a una línea. Esto correponde a un método de poca varianza y mucho bias.

En la siguiente imagen $K = 10$ y las lineas de frontera de decision de KNN y Bayes son muy parecidas.



En la siguiente imagen $K = 1$ y $K = 100$. Las linea de frontera de decision Bayes es punteada.





Veamos la relacion entre el error de entrenamiento y el error de prueba.
El error de entrenamiento fue calculado con 200 observaciones y el error de prueba con $5,000$. El nivel de flexibilidad crece con la funcion $1/K$ (eq. el numero K decrece). La linea punteda indica el error de Bayes. Los brincos en las lineas se deben a los pocos datos de entrenamiento.

Escoger el correcto nivel de flexibilidad es crítico para tener éxito en el aprendizaje estadístico, tanto para la regresión como para la clasificación. El intercambio Bias-Varianza y la forma de U en el error de prueba hacen esta tarea difícil. En el capítulo 5 veremos este tema para escoger el nivel óptimo de flexibilidad.

Tarea 1: Pag 52
2.4: 1- 7 y 9.
Ejercicio en clase 8 y si hay tiempo 10.

3. Regresión Lineal

Recordemos los datos de gasto en publicidad que hemos visto antes. Para un artı'uculo que se publicita, la variable de respuesta es ventas y las predictoras son el dinero invertido en publicidad de TV, Radio y Periódico. Supongamos que somos unos consultores en estadística y nos piden que hagamos un plan de publicidad basado en los datos que resulte en altas ventas para el proximo ano. Aquí están las siguientes preguntas que queremos responder:

1. Hay alguna relación entre la inversión en publicidad y las ventas ?
Queremos buscar evidencia de una asociación entre el gasto en publicidad y las ventas.
2. Qué tan fuerte es esta relación ? Dado un presupuesto para la publicidad si podemos predecir las ventas con una buen precisión entonces podemos decir que hay una relación fuerte. Si nuestras predicciones son un poco mejor que una adivinanza entonces esta relación es débil.
3. Qué medio contrinuye a las ventas ?
Los tres medios contribuyen ? solo uno o dos ? Buscaremos la forma de separar el efecto individual de cada medio cuando se invierte en los tres medios.
4.Qué tan preciso podemos estimar el efecto de cada medio en las ventas ?
Por cada dolar que se gasta en una medio, qué tanto se incrementan las ventas ? Qué tan preciso podemos predecir este incremento ?
5. Qué tan preciso podemos predecir las ventas futuras ?
Para cada nivel de gasto en tv, radio o periódico, cuáles son las predicciones para las ventas y que tan precisa es esta predicción.
6. La relación es lineal ?
Hay una relacion mas o menos de una linea recta entre el gasto en publicidad en los diversos medios y las ventas ? Si no es así, quizas podemos transformar las predictoras o las respuestas para que la regresiń lineal pueda ser usada. 7. Hay sinergía (intereacción) entre los medios ?
Quizás tenemos más ventas inviertiendo 50,000 en tv y 50,000 en radio que 100,000 en solo uno de ellos.

Responderemos estas preguntas de manera general y más adelante de manera específica.

3.1. Regresión lineal simple. Empezamos con la regresión más sencilla. Tenemos una respuesta cuantitativa Y con solo una predictora X y asumimos que hay una relación aproximadamente lineal entre ellas. Esta relación la podemos escribir de la siguiente manera:

$$Y \approx \beta_0 + \beta_1 X \tag{2}$$

Para nuestro problema de ventas, diremos que X representa el gasto en publicidad en la TV, de esta forma tenemos que ajustar el modelo

$$ventas \approx \beta_0 + \beta_1 TV \tag{3}$$

A β_0 y β_1 les llamaremos coeficientes o parámetros del modelo lineal. Usaremos el conjunto de entrenamiento para producir los estimadores $\hat{\beta_0}$ y $\hat{\beta_1}$ y así podremos predecir las ventas futuras dado un valor de TV, esto es:

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x \tag{4}$$

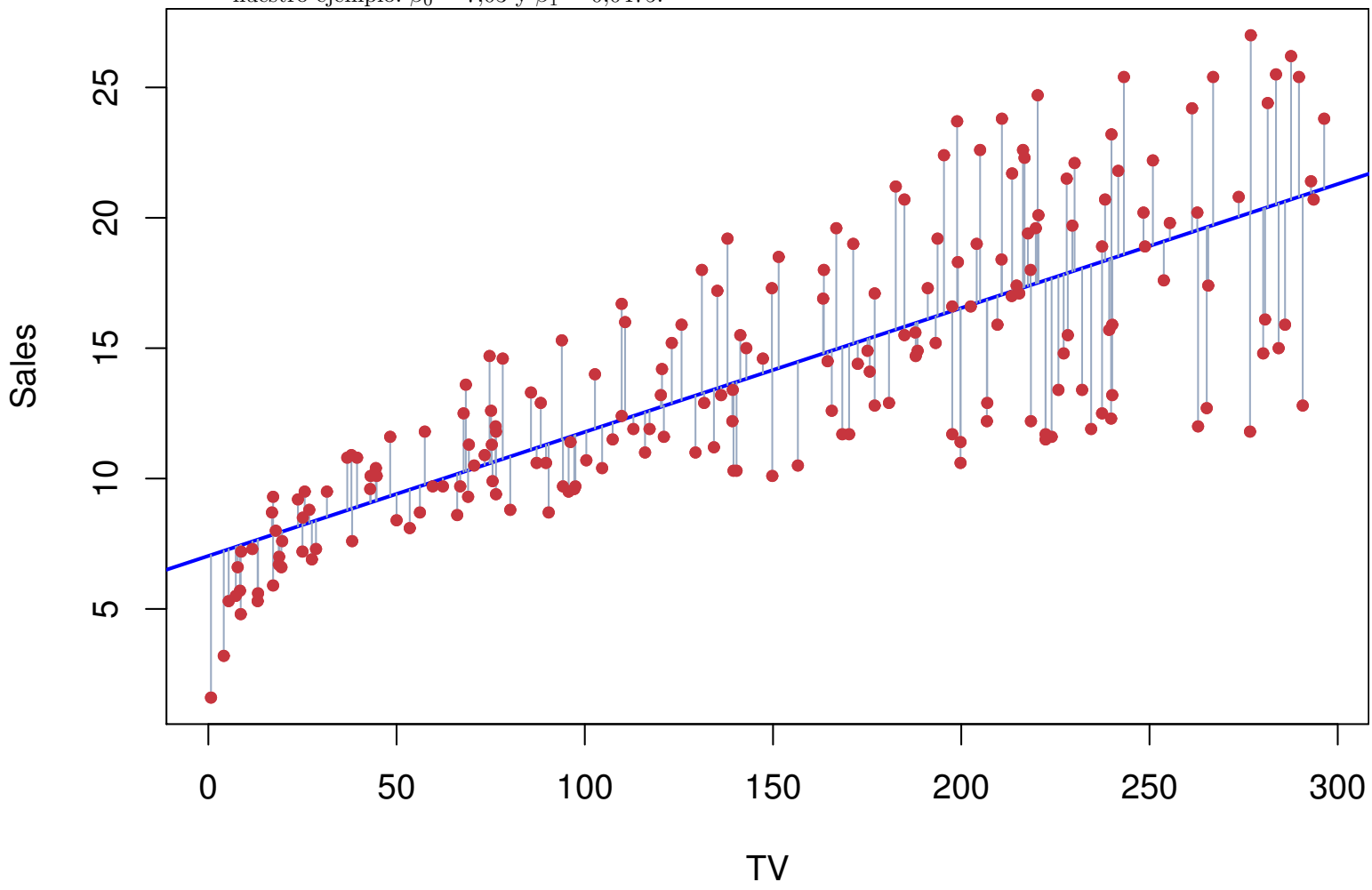
3.1.1. Estimación de los Coeficientes. En realidad no conocemos los parámetros β_0 y β_1. Nuestra tarea es encontrar los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor ajustan a los datos. Esto es, dados los datos de entrenamiento $(x_1, y_1), ..., (x_n, y_n)$ podemos predecir $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Y así podemos encontrar la diferencia (residuo) entre la predicción y el valor real $e_i = y_i - \hat{y}_i$. Y lo que nos interesa es minimizar el la suma de los cuadrados residuales RSS

$$RSS = e_1^2 + ... + e_n^2$$

El método de mínimos cuadrados provee de estos estimadores:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \text{ y } \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{5}$$

Donde $\overline{x}$ y $\overline{y}$ son los promedios respectivos del conjunto de entrenamiento. Para nuestro ejemplo: $\hat{\beta}_0 = 7{,}03$ y $\hat{\beta}_1 = 0{,}0475$.



3.1.2. Precisión en los coeficientes. Recordemos que asumimos la forma general de la verdadera relación entre X e Y tiene la forma $Y = f(X) + \epsilon$ donde f es una

función desconocida y ϵ es el término de error con media cero. Si f es aproximada por una función lineal entonces podemos escribir esta relación como

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{6}$$

donde

- β_0 es el intercepto, el valor esperado de Y cuando $X = 0$.
- β_1 es la pendiente, el aumento promedio en Y asociado con una una unidad de incremento en X.
- ϵ es el término que captura lo que se pierde en este modelo sencillo: la verdadera relación podría ser no lineal, podría haber otras variables que generan variación en Y y podrían haber errores en la medición. Por lo general asumimos que este error es independiente de X.

El modelo dado por la ecuación anterior define la linea de regresión de la población, la cual es la mejor aproximación lineal a la verdadera relación entre X e Y.

Vimos las fórmulas que para estimar β_0 y β_1 usamos las formulas de la ecuación anterior: $\hat{\beta_0}$ y $\hat{[}\beta_1]$. Estos estimadores son insesgados, es decir, sistemáticamente no subestiman ni subestiman a los parámetros poblacionales β_0 y β_1 sino que si obtenemos $\hat{[}\beta_0]$ y $\hat{[}\beta_1]$ de muchos conjuntos de datos y los promediamos obtenemos los parámetros poblacionales ! Pero podríamos preguntarnos que tan lejos está un solo estimado $\hat{[}\beta]$ del parámetro poblacional $\hat{[}\beta_0]$ y β_0? Esto se reponden con el error estándar $SE(\hat{\beta})$. El error estándar nos dice la cantidad promedio que difiere el estimado $\hat{[}\beta_0]$ del valor real β_0. Sabemos que entre más observaciones tengamos el error estándar disminuye. Las fórmulas del error estándar para nuestro estimadores son :

$$SE(\hat{\beta_0}) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right), \text{ y } SE(\hat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{7}$$

(Válidas si las n observaciones están no correlacionadas.) donde $\sigma^2 = Var(\epsilon)$. Para que estas fórmulas sean válidas necesitamos asumir que los errores ϵ_i para cada observación no están correlacionadas con la varianza en común σ^2. Esto claramente no es cierto si vemos la gráfica, pero aún así la fórmula es una buen aproximación. Note que en la fórmula $SE(\hat{\beta_1})$ es mas chica cuando los x_i están más dispersos; intuitivamente tenemos más apalancamiento para estimar la pendiente en este caso. Note que σ^2 es en general desconocida pero la podemos estimar de los datos. El estimador de σ es conocido como el error residual estándar y está dado por $RSE = \sqrt{RSS/(n-2)}$. Siendo estrictos cuando σ^2 sea estimada de los datos deberíamos escribir $\hat{SE}(\hat{\beta_1})$ para indicar que se ha hecho una estimación pero por simplicidad, no usaremo el sobrerito en nuestra notación.

Los errores estándar nos sirven para enconrar intervalos de confianza. Un intervalo de confianza del 95 % es definido como un rango de valores tales que con un 95 % de probabilidad, el rango contendrá el valor verdadero el parámetro desconocido. Para la regresión lineal, el intervalo de confianza del 95 % para β_1 toma la forma

$$\hat{\beta_1} \pm 2SE(\hat{\beta_1}).$$

(el número 2 es aproximado) Esto es, hay aproximadamente un 95 % de probabilidad de que el intervalo

$$(\hat{\beta_1} - 2SE(\hat{\beta_1}), \hat{\beta_1} + 2SE(\hat{\beta_1}))$$

contiene el valor verdadero de β_1. Similarmente para β_0

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0).$$

Para el problema de la publicidad, los intervalos de confianza del 95 % para β_0 y β_1 son [6,13, 7,935] y [0,042, 0,053] respectivamente. Esto significa que en la ausencia de publicidad en TV, las ventas estarán entre 6103 y 7935 unidades. Y por cada aumento de inversión en publicidad en la TV, las ventas aumentarán entre 42 y 53 unidades. Los errores estándares también serán usados para efectuar las pruebas de hipótesis en los coeficientes. Las más frecuentes serán
H_0 :No hay relación entre X e Y vs H_1: Hay alguna relación entre X e Y.

Matemáticamente $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$, claramente si $\beta_1 = 0$, X no está relacionada con Y. Para probar la prueba de hipótesis nula necesitamos determinar si $\hat{\beta}_1$, nuestro estimador de β_1 está lo suficientemente lejos de cero y podemos estar seguros de que β_1 no es cero. Qué es suficientemente lejos ? Esto depende de la precisión de $\hat{\beta}_1$, esto es, depende de $SE(\hat{\beta}_1)$. Si $SE(\hat{\beta}_1)$ es chica, entonces aún valores relativamente chicos de $\hat{\beta}_1$ pueden dar fuerte evidencia de que $\beta_1 \neq 0$, y así que existe una relación entre X e Y. En contraste, si $SE(\hat{\beta}_1)$ es grande, entonces $\hat{\beta}_1$ debe de ser grande en valor absoluto para poder rechazar la hipótesis nula. En la práctica, calculamos el estadístico t dado por

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

que mide el número de desviaciones estándar que $\hat{\beta}_1$ está del número cero. Si en realidad no hay relación entre X e Y entonces esperamos que t tendrá una distribución t con $n-2$ grados de libertad. Consecuentemente, es fácil calcular la probabilidad de observar cualquier número mayor o igual a $|t|$ asumiendo que $\beta_1 = 0$. A esta probabilidad la llamamos $p-$valor. Y lo interpretamos de la siguiente manera: Un p-valor chico indica es muy poco probable observar tal asociación entre la predictora y la respuesta debido a la casualidad, en la ausencia de una asociación real entre la predictora y la respuesta. Así que si vemos un p-valor chico podemos inferir que hay una asociación entre la predictora y la respuesta!. Esto es, rechazamos la hipótesis nula y aceptamos la hipótesis alterna. Normalmente esocgermos un p-valor de 5 o 1 %.
En nuestro problema rechazaremos ambas pruebas de hipótesis, ambos p-valores serán muy chicos. Esto es, la probabilidad de observar los valores que obtuvimos dado que H_0 es verdad es casi cero !

Regresemos al problema de la inversión en publicidad en los medios de comunicación para incrementar las ventas de un producto.

1. Hay alguna relación entre la inversión en publicidad y las ventas ?

Si, ajustamos el modelo de regresión múltiple de las ventas con la TV, radio y el periódico. Hacemos la prueba de hipótesis $H_0 : \beta_{TV} = \beta_{radio} = \beta_{periodico} = 0$. El estadístico F es muy chico, indicando que hay una clara evidencia de la relación entre la publicidad y las ventas.

2. Qué tan fuerte es esta relación ?

El RSE estima la desviación estándar de la respuesta de la linea de regresión poblacional. Para nuestro problema el RSE es de 1,681 unidades mientras que la valor medio de las respuestas es de 14,022 indicando error de porcentaje de alrededor de 12 %. Ahora, el estadístico R^2 guarda la el porcentaje de la variabilidad en la respuesta que es explicada por las predictoras. Las predictoras explican casi el 90 % de la varianza en las ventas.

3. Qué medio contribuye a las ventas ?

Podemos ver que el $p - valor$ del estadístico t asociado con cada predictor es chico para para TV y radio pero no para el periódico. Esto sugiere que solo los dos primeros están relacionados con las ventas.

4. Qué tan grande es el efector de cada medio en las ventas ?

El error estándar de $\hat{\beta}_j$ puede ser usado para construir intervalos de confianza para β_j. Los intervalos de confianza de TV y radio son chicos y alejados de cero, lo que indica evidencia de que estos medios si están relacionados con las ventas. El intervalo para el periódico incluye el cero, esto indica que esta variable no estadísticamente significativa dados los valores de TV y radio.

Con respecto a la colinealidad(colinealidad puede generar errores estándares grandes), los resultados de VIF son 1,005, 1,145 y 1,145 para la TV, radio y el peridódico, lo que sugiere que no hay evidencia de colinealidad.

Si efectuamos la regresión lineal para cada una de variables predictoras por separado, veremos que hay evidencia de una asociación extremadamente fuerte entre la TV y las ventas y también entre el radio y las ventas. Hay una ligera asociación entre el periódico y las ventas.

5. Qué tan preciso podemos predecir las ventas futuras ?

Si queremos predecir (i) una respuesta individual o $Y = f(X) + \epsilon$ utilizaremos un intervalo de predicción y para (ii) el promedio de las respuestas $f(X)$ un intervalo de confianza. Los intervalos de predicción siempre serán mas amplíos que los intervalos de confianza pués estos consideras la incertidumbre asocidada con ϵ, el error irreducible.

6. La relación es lineal ?

El plot de los residuos no ayuda a indentificar no linealidad. Si la relacion es lineal entonces esta gráfica no debe tener ningún patrón. Para nuestro problema vemos un efecto no lineal. Hemos visto que se pueden aplicar transformaciones a las predictoras para acomodar estas relaciones no lineales.

7. Hay sinergía (intereacción) entre los medios ?

El modelo de regresión estándar asume una relación aditiva entre las predictoras y las respuestas. Un modelo aditivo es fácil de interpretar por que el efecto de cada predictora con la respuesta no está relacionada con los valores de las otras predictoras. Pero para algunos conjuntos de datos esto no es realista. Hemos visto como incluir un término de interacción en el modelo de regresión oara acomodar las relciones no aditivas. Un valor chico del p-valor indica la precensia de esta relación. Los resultados muestra que los datos de la publicidad pueden ser no aditivos. Si incluimos un término de interación en el modelo, el R^2 se incrementa del 90 % a casi el 97 %.