

Introducción al aprendizaje estadístico

ÍNDICE

1.	Introducción	3
2.	Aprendizaje estadístico	4
2.1.	¿Qué es el aprendizaje estadístico ?	4
2.2.	Por qué estimar f	4
2.3.	Evaluación de la precisión del modelo	9
2.4.	Precisión en clasificación	12
3.	Regresión Lineal	16
3.1.	Regresión lineal simple	16
3.2.	Regresión Lineal Múltiple	21
3.3.	Otra consideraciones el modelo de regresión	26
3.4.	El ejemplo del problema de publicidad	31

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 paralelo 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. Amplamente 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 esencia, al hablar de aprendizaje estadístico nos referiremos 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 predicción.

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

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

Aquí f es alguna función de X_1, \dots, 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 función f debe ser estimada basada en los datos observados.

En esencia, el aprendizaje estadístico se refiere a un conjunto de enfoques para estimar f . En este capítulo indicaremos algunos de los conceptos teóricos 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 variables 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, \dots, 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 que 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 imprecisión introduce un error. Este error es reducible pues 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 ahí pues 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 pues 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 variar 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

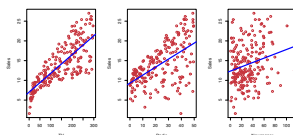
$$E(Y - \tilde{Y})^2 = E(f(X) + \epsilon - \tilde{f}(X))^2 = (f(X) - \tilde{f}(X))^2 + Var(\epsilon),$$

donde $E(Y - \tilde{Y})^2$ representa el promedio, o el valor esperado, de la distancia al cuadrado entre la predicción 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 minimizar 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, \dots, 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, \dots, 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 están 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 pues 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 $i = 1, \dots, p$ y $j = 1, \dots, p$. Correspondientemente, sea y_i la variable respuesta para la observación i . Entonces nuestro conjunto de entrenamiento consiste en los puntos $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ donde $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$.

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 paramétricos o no paramétricos.

Métodos Paramétricos

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

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

Éste es un modelo lineal, el cual lo analizaremos en el capítulo 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 $\beta_0, \beta_1, \dots, \beta_p$ tal que

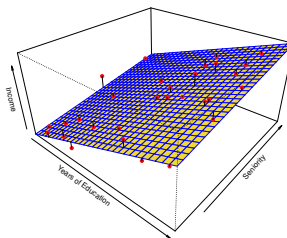
$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

El método más común para ajustar el modelo en 1 es mínimos cuadrados aunque hay otras maneras las cuales las discutiremos en el capítulo 6.

Una de las desventajas de usar este enfoque paramétrico es que el modelo que escogamos 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, o 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

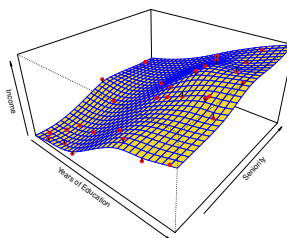
$$income \approx \beta_0 + \beta_1 education + \beta_2 seniority$$



Métodos No Paramétricos

Estos métodos no hacen supuesto explícitos 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.



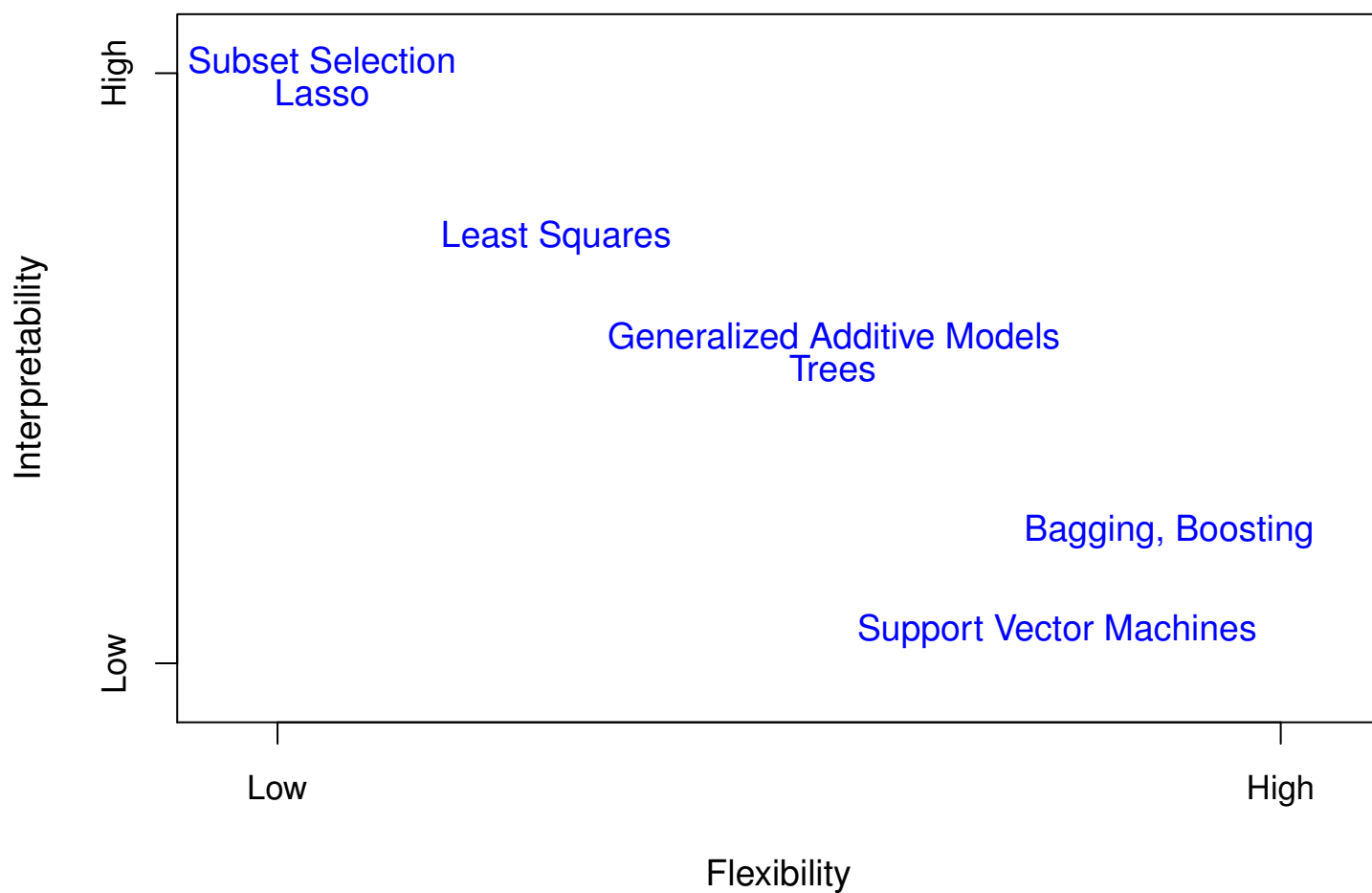
Hay ventajas y desventajas en los métodos paramétricos y no paramétricos para el aprendizaje estadístico. Exploraremos ambos tipos de métodos a lo largo del curso.

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 líneas o planos que vimos en las figuras anteriores. Los métodos como plate splines de las figuras que hemos vistas son mas flexibles pues 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 compensación entre flexibilidad e interpretabilidad usando diferentes métodos de aprendizaje estadístico.



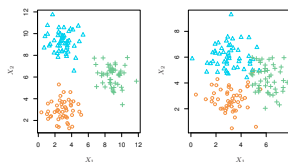
Uno esperaría que si solo nos interesa la predicción usaremos un método que sea muy flexible, pero esto no siempre es el caso! Esto tiene que ver con el fenómeno de overfitting.

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 supervisión. la imagen de la izquierda 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: género (masc o fem), Marca de carro (BMW, Mercedes, GM), etc. Tenedemos a referirnos a los problemas con valores de respuesta cuantitativos 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 pues 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.

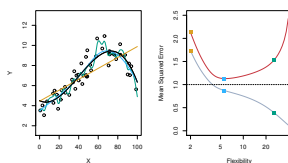
2.3. Evaluación de la precisión del modelo. En este curso veremos muchos métodos de aprendizaje estadístico que van más allá de la regresión lineal estándar. En estadística no existen un método que domine a todos los demás en todos los posibles conjuntos de datos. Es una tarea importante decidir para 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 desempeño 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 lo 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 garantía de que el método con el mas chico training MSE tambien tenga 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 anaranjada), 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 línea punteada el mínimo posible test MSE sobre todos los métodos ($\text{Var}(\epsilon)$). Los cuadraditos representan el training y test MSE para los tres ajustes mencionados.

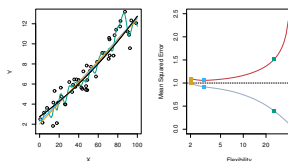
Podemos ver que entre mas flexibilidad, la curva ajusta mejor a los datos. También vemos que la curva verde es la mas flexible y ajusta mejor los datos pero difiere de la curva original f en negro por que ondula mucho. (Podemos ajustar los niveles de flexibilidad en el smoothing spline para producir 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 estadístico que es válida para cualquier conjunto de datos y método estadístico.

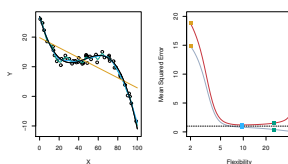
Si aumenta la flexibilidad, el training MSE se reducirá pero el test MSE puede que no. Cuando un modelo tenga training MSE chico pero un test MSE grande entonces estamos overfitting los datos.

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 monótonamente 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 la función original f no es altamente 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 training 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

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

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 $\text{Var}(\epsilon)$, 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 pequeños 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 dato solo moverá un poco la línea.

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, \dots, 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 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 precisión del modelo cuando tenemos un problema de regresión. Cuando estamos interesados en la clasificación, el enfoque más común para cuantificar la precisión 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 interesados 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

$$Prom((y_o \neq \tilde{y}_o).)$$

Clasificador de Bayes

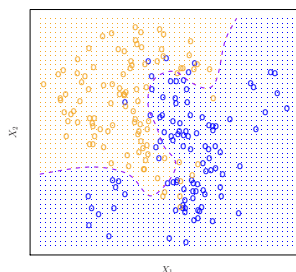
Veremos una cota de error para los métodos de clasificación, que está dada por el clasificador de Bayes. Veamos.

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

$$Pr(Y = j|X = x_0)$$

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

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

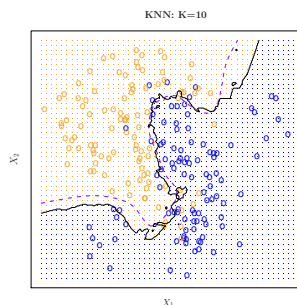
Finalmente, KNN aplica la regla de Bayes y clasifica la observación de prueba a la clase con la más alta probabilidad.

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

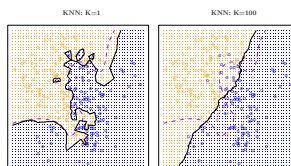
[width=4cm, center]Fig2/214.pdf

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

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



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



Veamos la relación entre el error de entrenamiento y el error de prueba.

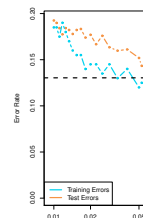
El error de entrenamiento fue calculado con 200 observaciones y el error de prueba con 5,000. El nivel de flexibilidad crece con la función $1/K$ (eq. el número K decrece). La línea punteada indica el error de Bayes. Los brinco en las líneas se deben a los pocos datos de entrenamiento.

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.



Repuestas:

1. a) El método flexible sería mejor por que con muchos datos, la función ajustaría bien la forma de la función desconocida f y como son pocas variables predictoras no dificultaría su interpretación, que es una desventaja.

b. El flexible sería peor por que son pocos datos, quizás la función no se ajuste bien y como son muchas variables predictoras, dificultaría su interpretación.

c. Flexible es mejor pues ajustaría mejor la forma de los datos no lineales.

d. El método flexible intentará ajustarse más a los datos y como estos presentan mucha variación, la varianza será grande.

2.

3. Regresión Lineal

Recordemos los datos de gasto en publicidad que hemos visto antes. Para un artículo 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 próximo año. 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 buena 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 contribuye 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 dólar que se gasta en un 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 qué tan precisa es esta predicción.

6. La relación es lineal ?

Hay una relación más o menos de una línea recta entre el gasto en publicidad en los diversos medios y las ventas ? Si no es así, quizás podemos transformar las predictoras o las respuestas para que la regresión lineal pueda ser usada. 7. Hay sinergia (interacción) entre los medios ?

Quizás tenemos más ventas invirtiendo 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:

$$(2) \quad Y \approx \beta_0 + \beta_1 X$$

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

$$(3) \quad \text{ventas} \approx \beta_0 + \beta_1 TV$$

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:

$$(4) \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

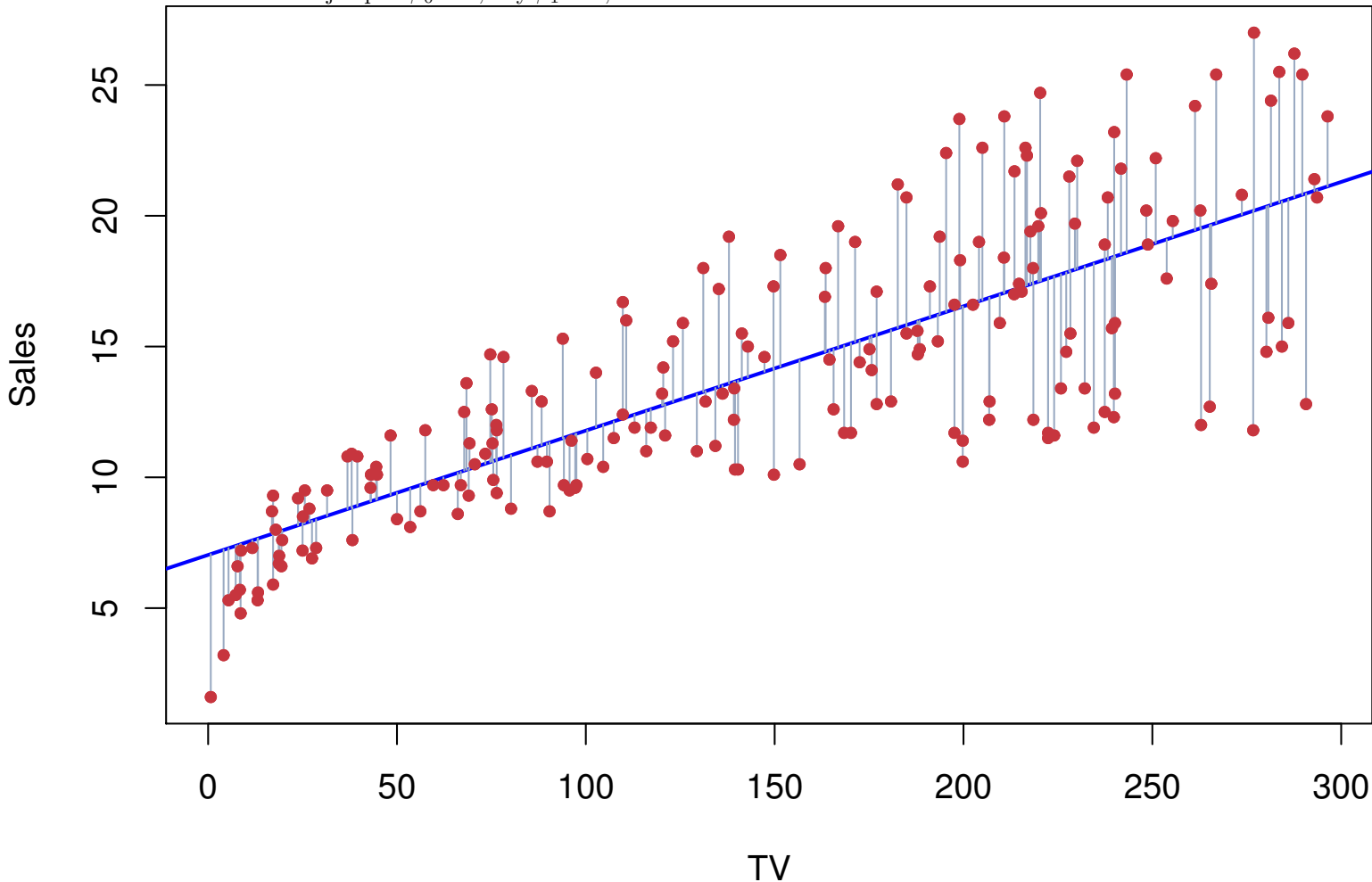
3.1.1. Estimación de los Coeficientes. En realidad no conocemos los parámetros β_0 y β_1 . Nuestra tarea es encontrar los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor ajustan a los datos. Esto es, dados los datos de entrenamiento $(x_1, y_1), \dots, (x_n, y_n)$ podemos predecir $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. 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 la suma de los cuadrados residuales RSS

$$RSS = e_1^2 + \dots + e_n^2$$

El método de mínimos cuadrados provee de estos estimadores:

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

Donde \bar{x} y \bar{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

$$(6) \quad Y = \beta_0 + \beta_1 X + \epsilon$$

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 línea 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 :

$$(7) \quad SE(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \text{ y } SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

(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 $SE(\hat{\beta}_1)$ para indicar que se ha hecho una estimación pero por simplicidad, no usaremos el sobrerito en nuestra notación.

Los errores estándar nos sirven para encontrar 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 ! Y podemos decir que por haber rechazado $\beta_0 = 0$ es la ausencia de inversión en publicidad en TV la ventas serán mayores que cero y por haber rechazado $\beta_1 = 0$ hay una relación entre X e Y .

3.1.3. Evaluando la precisión del modelo. Una vez que hemos demostrado que hay una relación entre X e Y quisiéramos saber que tan bueno el modelo ajusta a los datos. La calidad del ajuste de la regresión lineal típicamente se evalúa usando dos cantidades relacionadas: el error residual estándar (RSE) y el estadístico R^2 . Hemos visto como R nos da estos valores.

El error residual estándar (RSE)

Como hemos mencionado nuestro modelo poblacional es $Y = \beta_0 + \beta_1 X + \epsilon$ y el RSE es un estimado de la desviación estándar de ϵ . Es la cantidad promedio que la

respuesta se desviará de la línea de regresión verdadera. Su fórmula

$$RSE = \sqrt{RSS/(n-2)} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Para el problema de la publicidad vimos que el RSE es 3.26. Esto quiere decir que las ventas reales en cada mercado están alejadas de la línea de regresión poblacional aproximadamente en 3,260 unidades en promedio. Otra forma de interpretar es que aún si conociéramos los valores de β_0 y β_1 , cualquier predicción de las ventas basados en TV estará desviada por aprox 3,260 unidades en promedio. Si es bueno o malo depende del contexto. En nuestro problema el promedio de las ventas sobre todos los mercados es de aprox 14,000 unidades y así el porcentaje de error es de $3,260/14,000 = 23\%$.

Al RSE lo consideramos una medida de la falta de ajuste del modelo poblacional a los datos. Es claro que si $\hat{y}_i \approx y_i$ el RSE chico y podemos concluir que el modelo ajusta a los datos muy bien. Por otro lado, si los \hat{y}_i están lejos de y_i , el RSE será grande e indicará que el modelo no ajusto a los datos muy bien. (No es claro que es un buen RSE, pues está medido en las unidades de Y).

El estadístico R^2

Este estadístico es una medida alternativa de ajuste. Toma la forma de una proporción (la proporción de la varianza explicada) y así su valor está entre 0 y 1 y es independiente de la escala de Y .

Para calcular R^2 usamos la fórmula

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

donde $TSS = \sum (y_i - \bar{y})^2$ es la suma total de los cuadrados. Esta cantidad mide la varianza total en la respuesta Y y puede ser interpretada como la cantidad de la variabilidad inherente a la respuesta antes de aplicar la regresión. En contraste, el RSS mide la cantidad de variabilidad que deja sin explicar después de aplicar la regresión. De aquí que $TSS - RSS$ mide la cantidad de variabilidad en la respuesta que es explicada (o removida) por aplicar la regresión, y R^2 mide la proporción de la variabilidad en Y que puede ser explicada usando X . Cuando el R^2 es cercano a 1 indica que una gran proporción de la variabilidad en la respuesta ha sido explicada por la regresión. Cuando está cercana a 0 indica que la regresión no explicó mucha de la variabilidad en la respuesta; es puede ocurrir por que el modelo lineal está mal, o el error inherente σ^2 es alto, o ambos. En nuestro problema $R^2 = 0.61$, así que solo un poco menos de dos terceras partes de la variabilidad en las ventas es explicada por la regresión lineal en TV.

El estadístico R^2 tiene una ventaja en la interpretación sobre el RSE pero aún así puede ser difícil determinar que valor es un buen R^2 , en general esto depende de la aplicación. Por ejemplo, en física sabemos que algunos datos vienen de un modelo lineal con error residual chico. En este caso, esperaríamos ver un R^2 muy cercano a 1, y un valor substancialmente mas chico indicaría que hubo un problema serio con el experimento del que provienen los datos. Por otro lado, en otras áreas como

biología, psicología, etc, el modelo lineal es a lo más una quasi aproximación a los datos y los errores residuales debido a factores no medibles son por lo general muy grandes. En este caso, esperaríamos que solo una pequeña proporción de la varianza en la respuesta es explicada por la predictora y así un R^2 abajo de 0.1 sería más realista.

El estadístico R^2 es una medida de la relación lineal entre X e Y . Recuerde que la correlación está definida como

$$Cor(X, Y) = \frac{\sum_{i=1}^n (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^n (x_i - \hat{x})^2} \sqrt{\sum_{i=1}^n (y_i - \hat{y})^2}}$$

es también una medida de la relación lineal entre X e Y . Esto sugiere que podríamos usar $r = Cor(X, Y)$ en lugar de R^2 para evaluar el ajuste de un modelo lineal. De hecho, se puede mostrar que la regresión lineal simple, $R^2 = r^2$. En la regresión lineal múltiple veremos la diferencia entre estos conceptos.

3.2. Regresión Lineal Múltiple. El modelo de regresión lineal múltiple está dado por

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n + \epsilon$$

donde X_j representa la predictora j y β_j cuantifica la asociación entre la variable y la respuesta. Interpretamos β_j como el efecto promedio de Y por cada unidad de incremento en X_j mientras todas las demás predictoras están fijas.

Para nuestro ejemplo, el modelo es

$$Y = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Periodico + \epsilon$$

3.2.1. Estimación de los coeficientes. De forma análoga al caso de la regresión lineal simple, los coeficientes poblacionales $\beta_0, \beta_1, \dots, \beta_n$ son desconocidos y debemos estimarlos. Dados los estimadores $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_n$ podemos hacer predicciones usando la fórmula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n$$

Las fórmulas de estos parámetros son encontradas utilizando mínimos cuadrados. Se busca minimizar

$$(8) \quad RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_p x_{ip})^2$$

en este curso no daremos esta fórmula pues es complicada y tediosa. El programa R encontrara estos valores por nosotros.

Si aplicamos la regresión lineal simple a cada una de las predictoras y comparamos los valores de los estimadores con la regresión lineal múltiple podemos ver que los coeficientes de TV y Radio son similares. Pero el coeficiente del periódico en la regresión lineal múltiple es casi cero y el p-valor es grande! Esto muestra que los coeficientes de la regresión lineal simple y la múltiple pueden ser muy diferentes. Y también la podemos explicar de la siguiente manera. En la regresión lineal simple, la pendiente representa el efecto promedio de incrementar \$1000 en la publicidad del periódico ignorando las predictoras TV y Radio. En contraste, en la regresión múltiple, el coeficiente representa el efecto promedio de incrementar \$1000 en la publicidad del periódico mientras el gasto en TV y en la Radio está fijo.

Entonces la regresión lineal múltiple nos sugiere que no hay relación entre las ventas y el periódico mientras que la regresión lineal simple nos dice lo contrario. Veamos que pasa. Observemos la matriz de correlación, vemos que la correlación entre radio y el periódico es de 0.35. Sabemos que esto significa que tenemos una tendencia a gastar mas en publicidad en periodico en mercados donde se gasta en radio. Bien, ahora supongamos que la regresión lineal múltiple es la correcta y que la publicidad en el periódico no tiene un impacto directo en las ventas, y que la publicidad en el radio si incrementa las ventas. Entonces en los mercados donde gastamos mas en el radio las ventas tenderán a ser más altas y como la matriz de correlación lo indica, también gastaremos más en el periódico en esos mismo mercados. Asi que cuando usamos la regresión lineal simple de ventas vs periódico, observaremos que valores más altos en el periódico tienden a asociar valores más altos en ventas, aún cuando la publicidad en el periódico no afecta la ventas. De esta manera, el periódico se lleva el crédito del efecto de la radio en la ventas.

Veamos algunas preguntas importantes

Al efectuar la regresión lineal múltiple estamos interesados en responder algunas preguntas importantes.

1. Es alguna de las predictoras X_1, \dots, X_p útil en la predicción de la respuesta ?
2. Todas las predictoras ayudan a explicar Y , o solo algunas ?
3. Qué tan bien el modelo ajusta los datos ?
4. Dado un conjunto de entrenamiento, qué valor de respuesta debemos predecir y qué tan precisa es nuestra predicción ?

Responderemos estas preguntas:

1. De manera análoga a la regresión simple, haremos una prueba de hipótesis de la siguiente manera:

$$H_0 : \beta_1 = \dots = \beta_p = 0$$

vs

$$H_a : \text{al menos uno de los } \beta_j \text{ es no cero}$$

y usaremos el estadístico F que tiene la forma

$F = ((TSS - RSS)/p) / (RSS/(n - p - 1))$ donde TSS y RSS son las mismas fórmulas que en la regresión simple. Si los supuestos del modelo lineal son correctos, entonces uno puede mostrar que

$$E(RSS/(n - p - 1)) = \sigma^2$$

y si H_0 es verdadera

$$E(TSS - RSS/p) = \sigma^2$$

Así que cuando no hay relación entre la respuesta y las predictoras, podemos esperar que el estadístico F tome un valor cercano a 1. Por otro lado, si H_a es verdadera, entonces $E(TSS - RSS/p) > \sigma^2$, así que esperaríamos que F sea mayor que 1.

En nuestro ejemplo el estadístico F es 570. Esto sugiere que al menos una de las predictoras está relacionada con las ventas. Pero que pasaría si el estadístico estuviera cercano a 1, que tanto es cercano? Esto depende de los valores de n y p . Cuando n es grande, si F es un poco más grande que 1, esto sería suficiente evidencia contra H_0 . Y cuando n es chica, necesitamos un F mas grande para rechazar H_0 . Cuando H_0 es verdadera, los errores ϵ_i tienen una distribución normal y F tienen una distribución F (aún si los errores no son normales, F tiene aprox una distribución F) y así podemos encontrar un p-valor. Para nuestro ejemplo el p-valor es $2e^{-16}$, casi cero! Así que tenemos suficiente evidencia para rechazar H_0 y encontrarnos que alguna de las predictoras está asociada con el incremento en las ventas.

(Rechazamos cuando F es muy grande. Si n es muy grande, es suficiente que F sea ligeramente mayor a 1, si n es chica F debe ser grande. Claro, si p-valor es chico, tambien rechazamos).

Para concluir, a pesar de que podemos ver algunos p-valores casi cero en algunos de los coeficientes, nos basaremos en el p-valor del estadístico F para rechazar la prueba de hipótesis nula y aceptar que alguna de las predictoras está asociada con la respuesta.

Dejaremos para otra sección la situación cuando queramos hacer una prueba de hipótesis con un subconjunto de las variables predictoras.

Selección de variables

Una vez que encontramos la asociación entre la respuesta y las predictoras, quisiéramos encontrar cuales son las que si están relacionadas. En el capítulo 6 se ve un estudio más profundo de este tema. Para escoger el subconjunto de predictoras que nos dan el mejor modelo existen varios métodos que pueden medir la calidad del modelo. Algunos de estos son Mallows's C_p , Akaike information criterion (AIC), Bayesian information criterion (BIC), and R^2 ajustada. También se puede determinar cual modelo es el mejor graficando varios modelos, como el de residuos, para buscar patrones.

Existen 2^p modelos, lo cual hace difícil evaluar cada uno. Existen 3 formas clásicas de hacer esta tarea de manera eficiente y automatizada:

Selección para adelante. Empezamos con un modelo que solo tiene el intercepto pero no predictoras. Hacemos regresiones lineales simples y vamos agregando al modelo inicial la variable que tenga el menor RSS. Y así sucesivamente vamos agregando variables hasta que un criterio de parada se satisface.

Selección hacia atrás. Se inicia con todas las variables en el modelo. Y se remueve la variable con el p valor mas alto. Se hace otro modelo con las $p-1$ variables

y se remueve la que tenga el p-valor mas alto. Se sigue este procedimiento hasta que un criterio se satisface.

Selección mixta. Es una combinación de las anteriores dos estrategias.

La selección hacia atrás no puede ser usada si $p > n$, mientras que la selección hacia adelante siempre puede ser usada. La selección hacia adelante puede incluir variables en una etapa temprana que más tarde pueden ser redundantes.

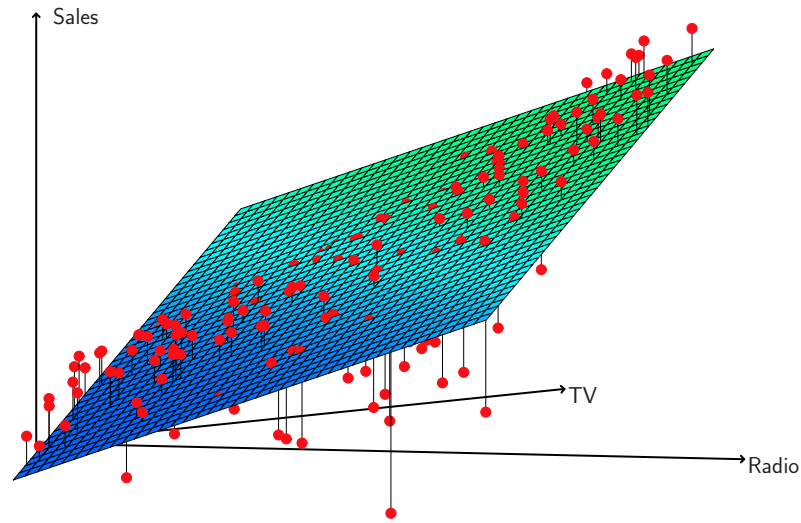
Ajuste del modelo

Las medidas numéricas más comunes son el RSE y el R^2 . Un valor de R^2 cercano 1, indica que el modelo explicar una gran proporción de la varianza en la variable respuesta. En el ejemplo de publicidad, el modelo con las 3 variables tiene un R^2 de 0,8972. Si se usan las dos variables TV y Radio el R^2 es de 0,89719. Esto muestra que siempre que incluimos una variable al modelo el R^2 aumenta, aún cuando el p valor no sea significativo. En este caso de la publicidad, el agregar el periódico al modelo solo aumenta muy poco el R^2 y como su p-valor es chico, podemos rescindir de esta variable. El incluirla podría llevar a resultados pobres debido al overfitting.

Por otro lado, el modelo que solo tiene a la TV como predictora, tiene un R^2 de 0.61 y si incluimos al Radio, el valor aumenta considerablemente. Esto significa que es mejor el modelo con estas dos variables.

Con respecto a al RSE, el modelo con radio y TV, su valor es de 1.681, y con las 3 variables es de 1.686. Y con solo la variable TV el RSE es de 326. Esto corrobora que la conclusión de que el modelo con la TV y el radio es mas preciso (con los datos de entrenamiento) que solo con la TV. Como observacion, de las formula del RSE, los modelos con mas variables pueden tener un RSE mas grande si el decrecimiento en el RSS es chico comparado con el incremento en p.

Adicionalmente al RSE y al R^2 podemos echarle un ojo a la gráfica de los datos.



Del patrón de los residuos, podemos ver que hay una relación no lineal en los datos. Los residuales positivos (sobrestimados) tienden a estar a lo largo de una línea de 45 grados, donde los presupuestos de la TV y Radio se dividen uniformemente. Los residuales negativos (subestimados) casi no se ven, tienden a estar lejos de esta línea, donde los presupuestos están cargados hacia un lado. Esto sugiere una sinergia, combinados las dos variables tienen más efecto que en una sola.

Predicción

Tenemos 3 tipos de incertidumbres al hacer las predicciones. 1. Recordando que

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

es solo un estimado del plano de regresión verdadero

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

y la imprecisión en los coeficientes estimados es debida al error reducible. Podemos calcular un intervalo de confianza para determinar que tan cerca estará \hat{Y} de $f(X)$.

2. En la práctica, asumir un modelo lineal para $f(X)$ es algo irrealista, y esto es una fuente de error reducible que llamamos el sesgo del modelo (bias). Entonces cuando usamos el modelo lineal, ignoraremos este hecho y pensaremos que el modelo lineal es correcto.

3. Aún si conociéramos los valores verdaderos de los coeficientes, las predicciones serían predichas perfectamente pues tenemos el error aleatorio ϵ , que llamamos error irreducible. Usaremos los intervalos de predicción para saber que tanto varía Y de \hat{Y} . Estos intervalos son más grandes que los de confianza pues incorporan tanto el error reducible para $f(X)$ como la incertidumbre de que tanto difiere un solo

punto del plano de regresión poblacional error irreducible).

Veamos un poco más acerca de los intervalos mencionados en 3 para nuestro ejemplo. Un intervalo de confianza cuantifica la incertidumbre sobre el promedio de ventas sobre una gran cantidad de ciudades. Ejemplo, dado que gastamos \$100,000 en publicidad de TV y \$20,000 de Radio en cada ciudad, un intervalo del 95 % de confianza es $[10,985, 11,528]$. Esto significa que 95 % de los intervalos de esta forma tendrán el valor verdadero $f(X)$. Por otro lado, un intervalo de predicción puede ser usado para cuantificar la incertidumbre sobre las ventas en una ciudad en particular. Con los mismo datos anteriores, obtenemos el intervalo $[7,930, 14,580]$. Esto significa que 95 % de los intervalos de esta forma tendrán el valor verdadero de Y para esta ciudad.

3.3. Otra consideraciones el modelo de regresión. Veamos la base datos crediticia.

3.3.1. Predictores cualitativos. Predictores con solo dos niveles

Cuando una variable predictora tenga dos posibles valores cualitativos, como sería el caso de género, creamos una variable artificial de la siguiente manera. Asignamos 1 a la si la persona x_i es mujer y 0 si x_i es hombre y tendremos el siguiente modelo. Si la persona i es mujer

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

y si la persona i es masculina

$$y_i = \beta_0 + \epsilon_i$$

Interpretamos a β_0 como el promedio del credito en el saldo para hombres, y $\beta_0 + \beta_1$ el analogo para mujeres. β_1 es la diferencia promedio entre hombres y mujeres.

De la regresión lineal podemos ver que $\beta_0 = 509,8$ y $\beta_1 = 19,73$, así que las mujeres tiene una deuda promedio de 529,53. En este caso, el p-valor asociado a nuestra variable artificial es muy alto, 0.669. Esto indica que no hay evidencia estadística de una diferencia entre el promedio de deuda entre los géneros.

Podríamos alternar los valores que asignamos al género y los resultados de la regresión serían lo mismo, lo que cambiaría sería la interpretación. Otra forma de definir la variable artificial sería hacer $x_i = 1$ para mujer y $x_i = -1$ para hombre. Si la persona i es mujer

$$y_i = \beta_0 + \beta_1 + \epsilon_i$$

y si la persona i es masculina

$$y_i = \beta_0 - \beta_1 + \epsilon_i$$

β_0 se interpreta como el promedio general de deuda (ignora el efecto de género), y β_1 es la cantidad que las mujeres están por arriba del promedio y por abajo del promedio en hombres. Para nuestro ejemplo. $\beta_0 = 519,665$, que es la mitad de los promedios anteriores y $\beta_1 = 9,865$ que es la mitad de la diferencia entre hombres y mujeres. Los resultados de las predicciones serán los mismos, lo que varía es la interpretación.

Si una variable tiene mas de dos niveles, podríamos crear más de una variable artificial.

3.3.2. Extensiones del modelo lineal. El modelo de regresión lineal que hemos visto, nos arroja resultados facil de interpretar y se desempeña bien en muchos problemas del mundo real. Sin embargo hace dos supuestos muy restrictivos entre la variable de respueta y la predictora: aditividad y linealidad. La primera nos dice que el efecto de los cambios en la predictora j sobre la respuesta es independiente de los valores de las otras predictoras. El supuesto de linealidad es que el cambio en la respuesta dado el cambio de una unidad en la predictora j es constante, sin importar el valor la predictora j . En este curso veremos varios métodos que relajan estos supuestos. Ahora veremos brevemente algunas formas classicas de extender le modelo lineal.

Quitanto el supuesto de aditividad

Como vimos en el ejemplo de la publicidad, existe una sinergia entre las variables radio y TV. Cuando alguno de estos dos valores es bajo, los valores verdaderos de las ventas están por debajo de lo que predice el modelo. Y cuando las dos predictoras tienen valores muy parecidos, el modelo tiende a subestimar las ventas. Nuestro modelo está dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Para extender nuestro modelo, podemos incluir una tercera variable predictora que se construye a partir de X_1 y X_2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Y podemos reescribir

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$

o

$$Y = \beta_0 + \hat{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$

Para nuestro ejemplo de publicidad, el modelo toma la forma

$$sales = \beta_0 + (\beta_1 + \beta_3 radio) TV + \beta_2 radio + \epsilon.$$

Podemos interpretar a β_3 como el incremento de la efectividad de la Tv por una unidad de aumento en la publicidad en radio (o viceversa).

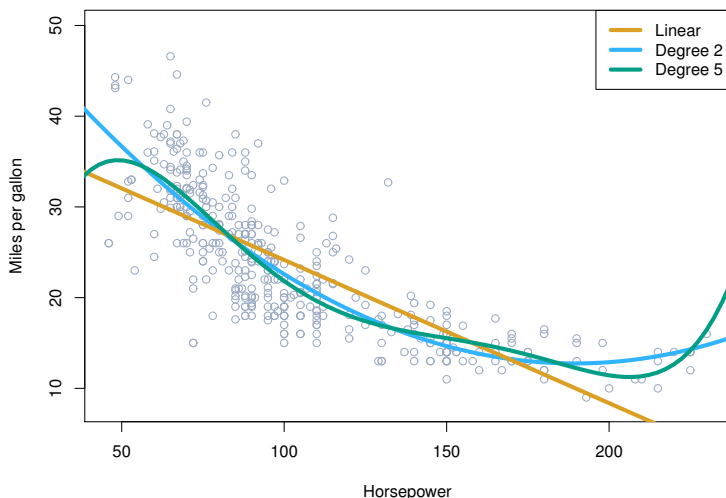
Usando la regresión lineal, los resultados indican que utilizar el término de interacción arroja resultados superiores contra el modelo que contiene solo los efectos principales. El p-valor de la interaccion es extremadamente bajo, lo que indica que hay mucha evidencia para $\beta_3 \neq 0$. En otras palabras, es claro que la relación es no lineal. El R^2 es de 96.8% compardo con solo el 89.7%. Esto muestra que el $(96.8-89.7)/89.7=69\%$ de la variabilidad en las ventas que permanece después de ajustar el modelo con el término de interacción ha sido explicado por el término de interacción.

En este ejemplo, los p-valores asocidos son estadísticamente significativos, asi que incluimos las tres variables en el modelo. Podría darse el caso el p-valor del término de interacción sea muy chico y que el de las variables principales no lo sea. En este caso también incluimos todas las variables.

Estas interacciones también aplican para variables cualitativas.

Relaciones no lineales

Cuando hay evidencia que hay una relación no lineal entre la variable de respuesta y las predictoras podemos usar otros métodos. En el capítulo 6 veremos más a detalle esto. Una forma sencilla de extender el modelo lineal es a través de la regresión polynomial.



En la figura tenemos el modelos lineal, el modelo de grado 2

$$(9) \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

y el de grado 5. A pesar de que la ecuación anterior incluye un término cuadrático, todavía estamos hablando de un modelo lineal. El nombre de regresión lineal proviene de los coeficientes β s.

3.3.3. Problemas potenciales. Al aplicar la regresión lineal a un conjunto, los problemas más comunes son

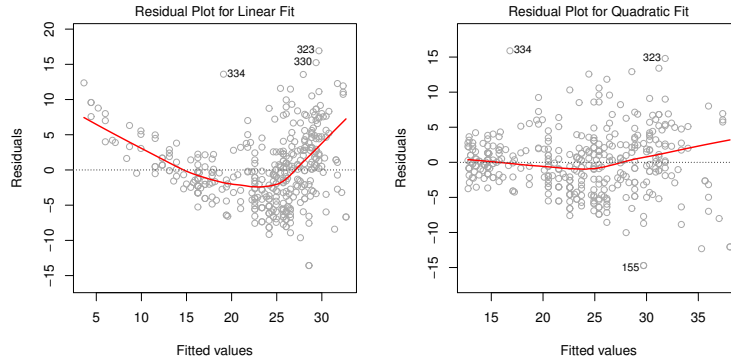
1. No linealidad entre la respuesta y las predictoras
2. Correlación en los términos de error
3. Varianza de los errores no constante
4. Outliers
5. Puntos con gran peso
6. Colinealidad.

En la práctica, identificar y sobreponerse a estos problemas es más arte que ciencia. Existen muchos libros y artículos al respecto. Dado que la regresión lineal no es nuestro objetivo principal, aquí veremos un breve resumen de los puntos más importantes.

1. No linealidad entre la respuesta y las predictoras

Si la relación no es lineal entonces todos nuestros resultados de la regresión estarán en duda.

Podemos usar la gráfica de los residuos $R_i = y_i - \hat{y}_i$ vs la predictora x_i . En la regresión lineal múltiple graficaremos los residuos vs los valores predecidos. Si existe algún patrón entonces hay una relación no lineal y podríamos hacerle alguna transformación a X como la raíz cuadrada, logaritmo o elevar al cuadrado. Entre otras.

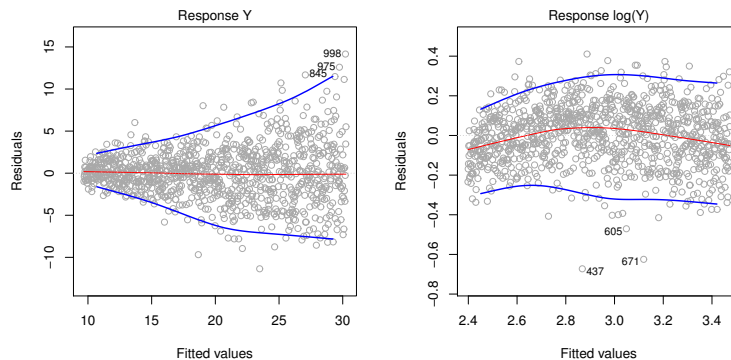


2. Correlación en los términos de error

Un importante supuesto es que los errores $\epsilon_1, \dots, \epsilon_n$ no están correlacionados. Esto significa que el signo de ϵ_i no provee ninguna información sobre el signo de ϵ_{i+1} . En caso de haber correlación, esta generaría un sentido de confianza sobre nuestro modelo, esto es, los estimadores de los errores estándares subestiman los valores reales y con ello obtengamos intervalos de confianza y predicción con una probabilidad menor que la que les corresponde. Además que los p-valores sean mas chicos de lo que deberían ser.

3. Varianza de los errores no constante

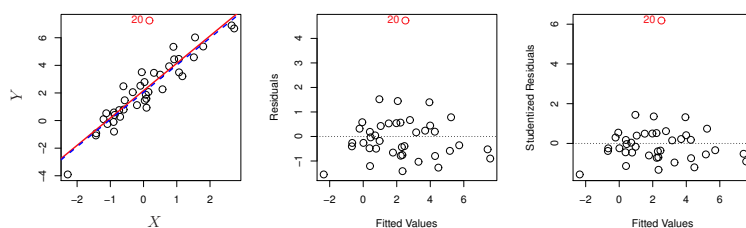
Otro supuesto de la regresión lineal es los términos de error tienen varianza constante, $Var(\epsilon_i) = \sigma^2$. Esto se puede ver cuando la varianza de los errores crece más con los valores de la respuesta.



Una solución a este problema es aplicar una transformación cóncava a Y como la raíz cuadrada o el logaritmo.

4. Outliers

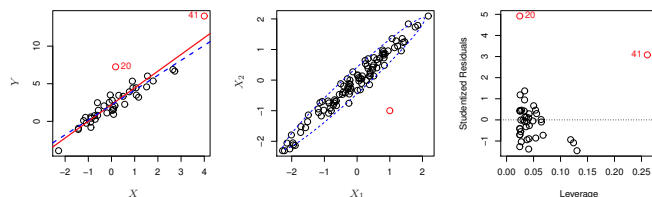
Un outlier es un punto para el cual y_i está lejos del valor predicho por el modelo. En la figura de la izquierda está la línea roja que el incluye el outlier y la punteada azul no lo incluye. Vemos que no hay mucha diferencia en las líneas. En algunos casos el outlier puede impactar de manera negativa el RSE y por tanto los intervalos de confianza y los p-valores. Inclusive puede hacer que el R^2 disminuya.



La gráfica de los residuos no puede ayudar a identificar outliers. Existe un manera más metódica de identificarlos. Se studentizan los residuos y los que estén fuera del intervalo $-3,3$ serán outliers. En algunos casos los outliers son causados por algún error en la recolección de los datos y esos se pueden remover. En otros casos pueden senalar una deficiencia en el modelo.

5. Puntos con gran apalancamiento

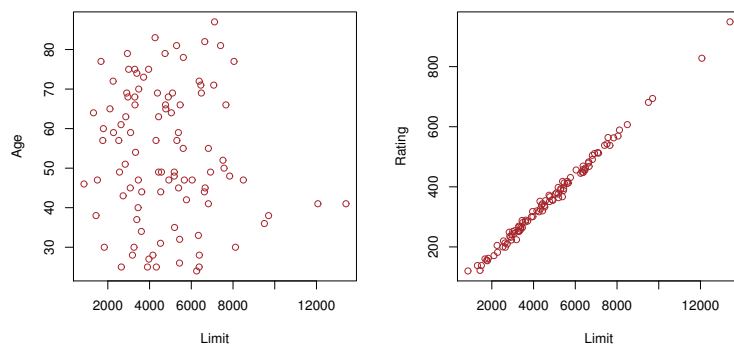
Tienen un peso fuerte en el modelo. Existe una fórmula del estadístico del apalancamiento.



El punto 41 de la figura de la izquierda tiene un apalancamiento grande, mientras que el 20 no. Si removemos el punto 41, la línea de regresión se mueve un poco ! En la imagen del medio, el punto roja tiene un apalancamiento alto, está lejos de la multitud de los demás puntos a pesar de no estar tan lejos con respecto a cada una de la componentes. En la última imagen, el punto 41 tiene un alto apalancamiento y un alto residual.

6. Colinealidad.

Es cuando dos o más variables predictoras están muy relacionadas entre ellas. La colinealidad de una variable se puede medir con el coeficiente VIF (Varianza inflation factor), si es muy grande entonces hay colinealidad entre las variables.



Limit y rating están altamente correlacionadas, decimos que son colineales. No hay ninguna relación en las variables de la izquierda.

3.4. El ejemplo del problema de publicidad. 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 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 periódico, 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 amplios que los intervalos de confianza pues estos consideran la incertidumbre asociada con ϵ , el error irreducible.

6. La relación es lineal ?

El plot de los residuos no ayuda a identificar no linealidad. Si la relación 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 sinergia (interacción) 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 para acomodar las relaciones no aditivas. Un valor chico del p-valor indica la presencia de esta relación. Los resultados muestran que los datos de la publicidad pueden ser no aditivos. Si incluimos un término de interacción en el modelo, el R^2 se incrementa del 90 % a casi el 97 %.