Regresión Lineal

Contenido

- Introducción
- Rectas: Concepto de recta y pensar en cómo encontrar los parámetros
- Mínimos Cuadrados
- Ecuación Normal: Derivando la ecuación, convexidad del error cuadrático, mostrando el resultado del ajuste y los parámetros
- Múltiples regresiones lineales: Extender la idea al plano

Reserva si no se alcanzan los 10 minutos

- Normalización de los datos
- Interpretación de la regresión: Qué significan los parámetros
- Motivación para el iterativo: En la vida real se pueden tener datos masivos, producto vector matriz es factible
- Descenso del gradiente: Ejemplo resolviendo una ecuación no lineal, Definición para Reg Lineal
- Ajuste: Visualizando parámetros conforme ajustax

Para el sgte vid: Ajuste polinomial, prueba de hipótesis..., Lasso, Ridge, Otros métodos iterativos

Introducción:

> Cambiar dataset por P 35 tabla 2.1 una muestra aleatoria

Cuando se tienen datos de la siguiente forma,
a veces es necesario predecir valores fuera del rango conocido

Por ejemplo, si decimos que x representa el ingreso semanal de una persona

y y el gasto de consumo semanal

Sería útil poder conocer el gasto semanal de la persona dado su ingreso

pero cómo podemos estimar el mejor valor en base a la información disponible?

La idea consiste en encontrar la línea de mejor ajuste a los puntos,

y cómo lo conseguimos habiendo tantas rectas posibles?

Rectas 1

Recordemos que una recta está definida por su pendiente m e intercepto b
`ecuación de la recta`
donde la pendiente indica la inclinación de la recta como la proporción del cambio vertical con respecto al horizontal, y el intercepto en qué punto toca el eje y

En la literatura estadística se les llama β_1 al intercepto y β_2 a la pendiente.

variando estos dos valores podemos construir infinitas rectas.
`Mostrando rectas con parámetros aleatorios`

Rectas 2

Ya conocemos la entrada y respuesta de algunos puntos

modelamos esta información asumiento que tienen una relación lineal y pueden ser explicados por una recta `line_eq1`, a esto se le agrega un término de error `lineas azules`, el cual representa variables no consideradas o ruido `line_eq2` en nuestro ejemplo el gasto mensual no solamente depende del ingreso, existen variables del diario vivir que no se están considerando. A esto se le suma el hecho que las observaciones pueden ser sólo una muestra del total de datos.

Ahora queremos predecir el consumo para valores desconocidos pero que provienen de la misma distribución de datos.

Para estas observaciones nuevas no podemos calcular el término de error, por lo que lo omitimos, lo cual hará que nuestra estimación no sea perfecta, pero si lo más cercana al valor real posible

A esta estimación la llamaremos $\hat{y}$, y debido a que asumimos que nuestros puntos eran una muestra, renombraremos los parámetros con el sombrero para indicar que son una estimación de los parámetros de la población real.

Podemos definir una forma de medir, dados β_1 y β_2 cuán distinta es la estimación del valor $\hat{y}$ al valor original y_i para los puntos (x_i, y_i) conocidos.

Promediando estos valores para todos los puntos obtenemos una medida de error total.

Esta medida de error se llama error cuadrático medio: $E(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$, y es una de las funciones de error más utilizadas.

Ahora sabemos que una línea ajusta mejor mientras menor el valor de $E(\beta)$

por lo que necesitamos una manera automática de dado un conjunto de datos X y Y, obtener los valores de β que produzcan el menor valor posible de $J(\beta)$

Ecuación Normal 1 (Notación)

Notemos que todo sistema de ecuaciones

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases}$$

se puede reescribir en función de matrices.

$$\begin{bmatrix} a_{11}, a_{12}, \ldots, a_{1n} \\ a_{21}, a_{22}, \ldots, a_{2n} \\ \vdots \\ a_{n1}, a_{n2}, \ldots, a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$A\mathbf{x} = \mathbf{b}$

Así para simplificar los cálculos es conveniente expresarlo de esta manera.

Como tenemos m puntos tambien llamados observaciones en nuestro conjunto de datos

Podemos reordenar la respuestas conocidos como un vector de resultados $\hat{\mathbf{y}}$ donde cada elemento o "fila" corresponde a una observación.

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 x_1 \\ \beta_1 + \beta_2 x_2 \\ \vdots \\ \beta_1 + \beta_2 x_m \end{bmatrix}$$

Ahora podemos reescribir la suma de esta manera $\beta_1 \cdot 1 + \beta_2 \cdot x$ escribiendo de manera explícita el uno por β_1

`Mover a la izquierda y empezar a escribir a la derecha` recordando la definición del producto punto entre dos vectores.

Si tenemos $\mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$ y $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, su producto punto es $\langle \mathbf{x}_i, \beta \rangle = 1 \cdot \beta_1 + x_i \cdot \beta_2$.

`De nuevo a la izquierda` entonces cada elemento del vector es el producto punto entre los vectores

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \langle \begin{bmatrix} 1 \\ x_1 \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \rangle \\ \langle \begin{bmatrix} 1 \\ x_2 \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \rangle \\ \vdots \\ \langle \begin{bmatrix} 1 \\ x_m \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \rangle \end{bmatrix}$$

`A la derecha` Pero podemos reescribir el producto punto como el vector $\mathbf{x}_i$ traspuesto volviendose un vector fila, por el vector β así. $\mathbf{x}_i^T \beta = [\,1, x_i\,] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

Esta notación se basa en considerarlos directamente como matrices de $v^T : 1 \times 2$ y $u : 2 \times 1$ respectivamente.

`Izquierda` Reescribiendo así el vector $\hat{\mathbf{y}}$:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \beta \\ \mathbf{x}_2^T \beta \\ \vdots \\ \mathbf{x}_m^T \beta \end{bmatrix}$$

Hasta este punto ya simplificamos mucho la notación, pero notemos que β se repite muchas veces, podemos simplificarla más considerando cada vector $\mathbf{x}_i^T$ como fila de una matriz.

`Borrar la derecha` De modo que tenemos la matriz $X = \begin{bmatrix} 1, x_1 \\ 1, x_2 \\ \vdots \\ 1, x_m \end{bmatrix}$ llamada matriz de diseño. Al multiplicarla por el vector β obtenemos `multiplicación larga` que se escribe matricialmente como $\hat{\mathbf{y}} = \mathbf{X}\beta$.

Esta notación es muy util para simplificar los cálculos y escribir programas que lo resuelvan.

Mínimos Cuadrados

Existen casos donde el sistema no tiene solución, sin embargo existirá un vector $\mathbf{x}^*$ que al multiplicarse por A da como resultado un vector b^* lo más cercano posible a b.

Al vector $\mathbf{x}^*$ se le llama solución por mínimos cuadrados.

Esta formulación es equivalente a encontrar el $\mathbf{x}^*$ que minimice $(\underset{\mathbf{x}}{\operatorname{argmin}} ||A\mathbf{x} - b||^2)$

Es decir la distancia entre b^* y b

La regresión lineal es uno de estos sistemas, por lo que se debe resolver, entre comillas, ya que sabemos que no encontraremos una solución exacta, mediante mínimos cuadrados

Ecuación Normal 2

Se puede reescribir el error cuadrático medio como `ms1 a mse2` de manera matricial, para enccontrar el mínimo de esta función podemos prescindir el término $\frac{1}{m}$ simplificandose a: `mse3`

Recordemos que $\hat{\mathbf{y}} = X\beta$

Esta formulación es una reescritura de la norma descrita en mínimos cuadrados (Escribir la ecuación de mínimos cuadrados con β

`Mover arriba y empezar el desarrollo` Desarrollando esta expresión tenemos.

La función del error cuadrático medio forma una superficie convexa, es decir que tiene un único punto mínimo. `Graficar el MSE`
Se deriva la expresión con respecto a β

Y se la iguala a 0, técnica común para encontrar mínimos.

`Desarrollar`

Obtenemos la siguiente expresión, que nos da una fórmula exacta para los β que producen el menor error.

Es decir el punto más cercano al mínimo. `Punto amarillo`

Ecuación Normal 3

`Graficando` Regresando al conjunto de puntos del inicio

`Hacer pequeña la gráfica y mandarla a la izquierda`

`mostrar el matrices con pocos datos`

Si se resuelve la ecuación anterior `mostrar ecuación`

Se obtiene el vector de parámetros

Graficando la recta obtenida por estos parámetros podemos observar el ajuste obtenido.

Interpretación de la regresión

> Esperanza condicional

`Llevar la gráfica al centro nuevamente de la anterior sección`

Podemos preguntarnos entonces, si la regresión es útil para predecir puntos fuera de nuestro conjunto de datos, por qué esta predicción se ve distante a los valores reales de los puntos conocidos

> Usar dataset P 35 tabla 2.1

Simplificando nuestro ejemplo con fines inlustrativos, en la realidad el grupo de personas con un sueldo x_i no van a gastar todos lo mismo, es por esto que si tuvieramos la información completa de toda una población, tendríamos la siguiente grafica: `grafica con varios puntos por x_i` con distintos consumos correspondiendo al mismo ingreso.

Si se calcula el promedio de cada grupo, se tiene una media de gasto para cada consumo denotada por los puntos naranjas, `Puntos naranjas` a este promedio por grupos se le llama la esperanza de y_i dado x_i, `Escribir E(y_i|x_i)`, Se puede trazar una línea que pase por cada promedio `Dibujar la línea`, a esa línea se le llama la regresión lineal poblacional.

Sin embargo en problemas reales generalmente se tiene sólo un punto correspondiendo a cada x_i, como en el ejemplo visto hasta ahora.

Al ajustar la regresión lineal estamos tratando de estimar esta esperanza condicional para cada valor de x, en base a un fragmento o muestra de la población `Graficar la regresión poblacional vs la regresión con la muestra`, e inevitablemente la estimación no será perfecta por la pérdida de información al trabajar con un subconjunto de puntos.

Es por esto que se hace la distinción de los parámetros y la variable dependiente estimados con un sombrero `reescribir la ecuación`.

Por todo esto, el interés está en predecir un valor lo más cercano posible al **promedio real** de la variable dependiente, *en nuestro ejemplo el gasto semanal* dado un valor de la variable dependiente llamada regresor, en base a una muestra.

> Parámetros

Sumado a esto, notemos que $\hat{\beta}_2$ es la pendiente, `agregar la pendiente como en Rectas1` y nos dice cuánto cambia en el gasto semanal al incrementar el ingreso en 1.

Como cada predicción es una estimación del valor esperado de y dado x, cuando $x = 0$, $y = \hat{\beta}_1$ `Desarrollarlo`, por esta razón el intercepto sólo tendrá sentido interpretarlo si en algún momento se puede tener que x vale cero, lo cual en nuesto ejemplo no sirve ya que nadie puede gastar si no recibe dinero

> Notar que correlación no implica causalidad, además que en el análisis de correlación no se hace distinción de las variables, ambas se asumen como aleatorias, mientras que en regresión se asume que sólo la dependiente es aleatoria y las explicativas son fijas, no estocásticas.

Ejemplo (TODO)

DESARROLLAR EL EJEMPLO DE LA PAG 80

Regresión Lineal Múltiple

Volviendo al modelo, consideremos ahora una variable explicativa más. La cantidad de años de estudio, de manera general, mientras más especializada la persona, mejor sueldo podría tener.

Reescribiendo como $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$ (se considerará $x_{i1} = 1$ constante) ahora contiene una nueva variable x_3 con su respectivo parámetro $\hat{\beta}_3$, entonces la esperanza condicional $E(y_i|x_{i2}, x_{i3})$ de y_i depende ahora de los valores de x_{i2} y x_{i3}, es decir, si antes se consideraba un grupo a todos los valores de y_i en la línea vertical correspondiente a x_i `Graficar la línea vertical en ejemplo ilustrativo`, ahora el grupo está dado por todos los valores y_i en los planos verticales que forman 90 grados y se cruzan en el punto dado por el valor de ambas variables. `Mientras se explica pasar al 3d, ya se tiene un plano, mostrar el segundo plano`

Al tener dos variables independientes y una dependiente, la gráfica ahora se debe realizar en 3 dimensiones.

Repitiendo el proceso de graficar la población `Borrar los planos y poner los puntos poblacionales` observamos que ahora tenemos variación también en la nueva dimensión.

Graficando las esperanzas condicionales para cada grupo, se puede trazar un plano que pase por todos los puntos medios, esta es la línea entre comillas de regresión, ya que en realidad es la extensión a 2D de la recta, un plano

De manera similar al caso con una variable, se puede re escribir como matrices, `mostrar desarrollo`, así, el desarrollo para encontrar el vector de betas es idéntico y la ecuación normal `Mostrar eq` sirve para una cantidad de variables arbitraria, consideranda cada una, una colúmna de la matriz X

Si se considera la muestra del conjunto de puntos, se tiene el plano ajustado:

> Estandarización se deja para el método iterativo

Normalización de datos

Como las variables en la vida real tienen distintas escalas, esto puede hacer dificil de interpretar los parámetros, para evitar esto, y simplificar la función de error se estandarizan los datos.

Vamos a denotar la segunda colúmna de la matriz X como $X_{:,2}$

Para la segunda colúmna de X se calcula la media y su desviación estándar,

y se realiza el cálculo elemento a elemento en la colúmna, que se guarda en la segunda colúmna de la matriz estandarizada X tilde. Esta matriz también tiene unos en su primera colúmna.

Se realiza el mismo procedimiento para $\mathbf{y}$

`Graficar puntos en escala real y llevarlos a la normal` aplicandolo, observamos como se desplazan los puntos al centro

`Zoom al grafico`

`Graficar Superficie de error` la superficie de error es el siguiente paraboloide:

El punto $(\beta_1, \beta_2, E(\beta))$ está lo más cerca posible al mínimo de la superficie de error como se puede observar `Graficar el punto mínimo`.

Con los valores de los betas y el error calculados `Evaluar el punto mínimo`

Tenemos la línea de mejor ajuste en los datos estandarizados

Nótese que al estandarizar, el intercepto se vuelve cero.

Cómo obtener los parámetros no estandarizados desde los estandarizados

> How do we interpret the beta coefficients? The interpretation is that if the (standardized)
> regressor increases by one standard deviation, on average, the (standardized) regressand
> increases by β_2^* standard deviation units. Thus, unlike the traditional model in Eq. (6.3.3), we
> measure the effect not in terms of the original units in which Y and X are expressed, but in
> standard deviation units.

Motivación

Revisar

En la vida real, la matriz X puede llegar a ser gigante con una gran cantidad de datos.

Veamos una vez más la fórmula de la ecuación normal $\beta = (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$, notemos que tenemos que multiplicar $X^T X$ si suponemos que X tiene dimensiones $m \times n$, una estimación de la cantidad de operaciones para calcularlo es $O(m^2 n)$ **notación O indica el peor caso**

Una computadora moderna puede realizar unas 10^8 operaciones por segundo en un lenguaje veloz como **c++** y **Fortran** que es en lo que están escritas las librerías de algebra lineal computacional

Si tuvieramos una matriz de 300000×1000, realizar la operación tomaría unas 10^{13} operaciones

Es plausible esperar unos segundos para esto, sin embargo con 300000×10000 un caso que se puede dar en información geoespacial, ya no cabe en la memoria de una computadora común, es por estas limitaciones de tiempo y memoria que se plantea encontrar los parámetros de manera iterativa.