MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

Perú

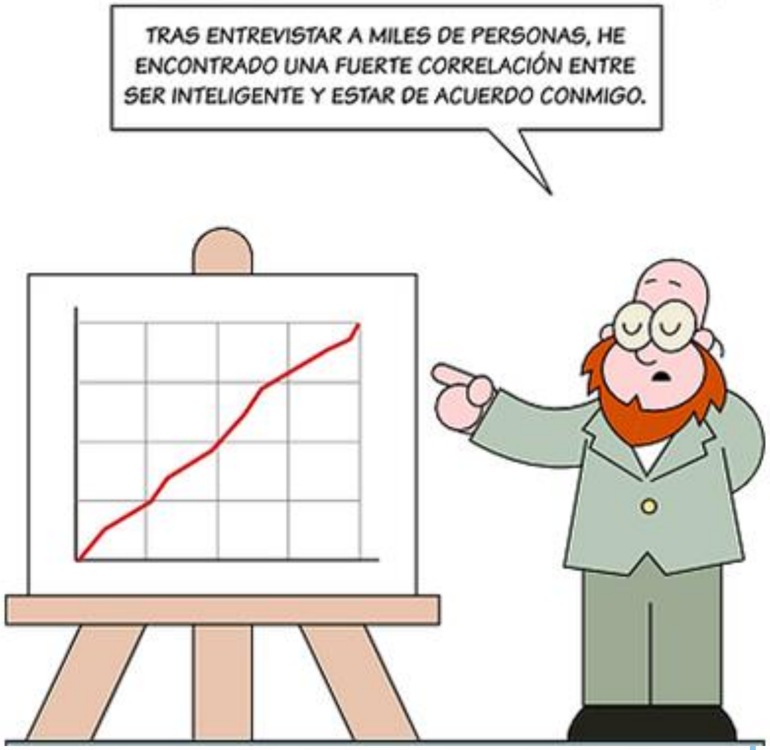
« Un esfuerzo mas y lo que iba a ser un fracaso se convierte en un éxito; no existe el fracaso si nos esforzamos cada vez mas y mas»

Marat

AGENDA

- Introducción a la Regresión Lineal.
- Diagrama de Dispersión.
- Especificación e Interpretación del Modelo de Regresión Lineal.
- Bondad de Ajuste
- Validación del Modelo
- Significancia de Parámetros del Modelo
- Modelos de Regresión Penalizados
- Regresión Ridge
- Regresión Lasso

Regresión Lineal Simple y Múltiple



INTRODUCCIÓN

- Determinar la ecuación de regresión sirve para:
  - Describir de manera concisa la relación entre variables.
  - **<u>Predecir los valores de una variable en función de la otra.</u>**
- Veremos EXCLUSIVAMENTE relaciones lineales.
- La regresión lineal simple estudia la relación entre sólo dos variables (el caso de relación más sencillo posible).

INTRODUCCIÓN

DENOMINACIÓN DE LAS VARIABLES	
X	Y
Predictora, regresor	Criterio
Explicativa	Explicada
Predeterminada	Respuesta
Independiente	Dependiente
Exógena	Endógena
(Explica la variabilidad de otra variable)	(Su variabilidad es explicada por otra variable)

DIAGRAMA DE DISPERSIÓN

- A grandes rasgos, como paso previo, el diagrama de dispersión permite vislumbrar si:
  - Existe relación entre variables.
  - La relación es lineal o de otro tipo.
  - Intensidad de la relación (por la estrechez de la nube de puntos).
  - Valores anómalos (Outliers) distorsionan la relación.
  - La dispersión de los datos es o no uniforme (homocedasticidad vs. heterocedasticidad).

ESPECIFICACIÓN DEL MODELO DE REGRESIÓN LINEAL

$$Y = \alpha + \beta X + \varepsilon$$

$$Y = \hat{Y} + \varepsilon$$

$$\hat{Y} = \alpha + \beta X$$

$$\varepsilon = Y - \hat{Y}$$

$$\varepsilon$$

- Puede denominarse:
  - Error
  - Perturbación
  - Residual
- Se debe fundamentalmente a:
  - Medición incorrecta de la variable.
  - Influencia de otras variables no incluidas en el modelo.
  - Variabilidad inherente a la conducta humana.

SUPUESTOS DEL MODELO

- Características estadísticas:
  - Linealidad.
  - Homocedasticidad: las varianzas de Y para cada valor de X son todas iguales.
  - Ausencia de autocorrelación: las variables Y son independientes entre sí (problema en estudios longitudinales).
  - Normalidad.
- Características como modelo descriptivo:
  - El modelo ha de estar correctamente especificado:
    - No se excluyen variables independientes relevantes.
    - No se incluyen variables independientes irrelevantes.
  - La variable independiente ha de haber sido medida sin error.

ESTIMACIÓN DE PARÁMETROS

- Para estimar los parámetros a y β usamos: **mínimos cuadrados.**
- En puntuaciones directas:

$$\hat{Y} = \alpha + \beta X$$

$$a = \overline{Y} - b\overline{X}$$

$$b = r_{XY} \frac{S_Y}{S_X}$$

INTERPRETACIÓN DEL MODELO DE REGRESIÓN LINEAL

En el modelo teórico de regresión lineal

$$Y = a + bX + e$$

distinguimos los siguientes elementos:

- e → error de estimación o puntuaciones residuales, parte aleatoria; aquello no explicado por el modelo.

$$\hat{Y} = a + bX$$

- $\hat{Y}$ → puntuación estimada: valor promedio previsto para todos los sujetos que han obtenido en la variable X un valor de X_i.
- b → pendiente de la recta: cambio en Y por cada unidad de cambio en X.
- a → ordenada en el origen: valor medio de Y cuando X=0.

INTERPRETACIÓN DEL MODELO DE REGRESIÓN LINEAL

$$\hat{Y} = 600 + 300X$$

- Supongamos que tenemos la ecuación de regresión, donde X es el número de años de experiencia profesional, e Y es el sueldo mensual.

  - ✓ Interpreta a y b.
  - ✓ Una persona con 3 años de experiencia laboral, ¿qué sueldo mensual tendrá? Interpreta el resultado.
  - ✓ Si una persona con 3 años de experiencia laboral tiene un sueldo mensual de 1700 €, ¿cuál será su error asociado? Interpreta el resultado.

INTERPRETACIÓN DEL MODELO DE REGRESIÓN LINEAL

$$\hat{Y} = 600 + 300X$$

- ✓ b=300 → Cambio en Y por cada unidad de cambio en X. Por cada año de experiencia laboral, el sueldo mensual aumenta 300 €.

- ✓ a=600 → Valor medio de Y cuando X=0. Sueldo medio de aquellas personas sin experiencia laboral.

- ✓ Una persona con 3 años de experiencia laboral, ¿qué sueldo mensual tendrá? Interpreta el resultado.

$$X = 3 \Rightarrow \hat{Y} = 600 + 300 * 3 = 1500$$

- ✓ $\hat{Y} = 1500$ → Valor promedio previsto para todos los sujetos que han obtenido en la variable X un valor de X_i. Las personas con 3 años de experiencia tienen un sueldo promedio de 1500 €.

INTERPRETACIÓN DEL MODELO DE REGRESIÓN LINEAL

- Si una persona con 3 años de experiencia laboral tiene un sueldo mensual de 1700 €, ¿cuál será su error asociado? Interpreta el resultado.

$$e = Y - \hat{Y} = 1700 - 1500 = 200$$

✓ El modelo estimó un sueldo de 1500 € para una persona con 3 años de experiencia laboral. Si esta persona concreta tiene un sueldo de 1700 €, esta diferencia de 200 € es el error; aquello que el modelo no explica.

COMPONENTES DE VARIACIÓN

$$\sum_{i=1}^{N}\left(Y-\overline{Y}\right)^{2}=\sum_{i=1}^{N}(\hat{Y}-\overline{Y})^{2}+\sum_{i=1}^{N}(Y-\hat{Y})^{2}$$

- ✓ Suma de cuadrados total = suma de cuadrados explicada + suma de cuadrados no explicada
- ✓ Variación Total = Variación Explicada + Variación No Explicada

BONDAD DE AJUSTE

$$R^2 = r_{XY}^2 = \frac{SC_{\exp}}{SC_t} = \frac{\sum\left(\hat{Y} - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

- Coincide con el coeficiente de determinación.
- La proporción de variabilidad no explicada= 1-R^2

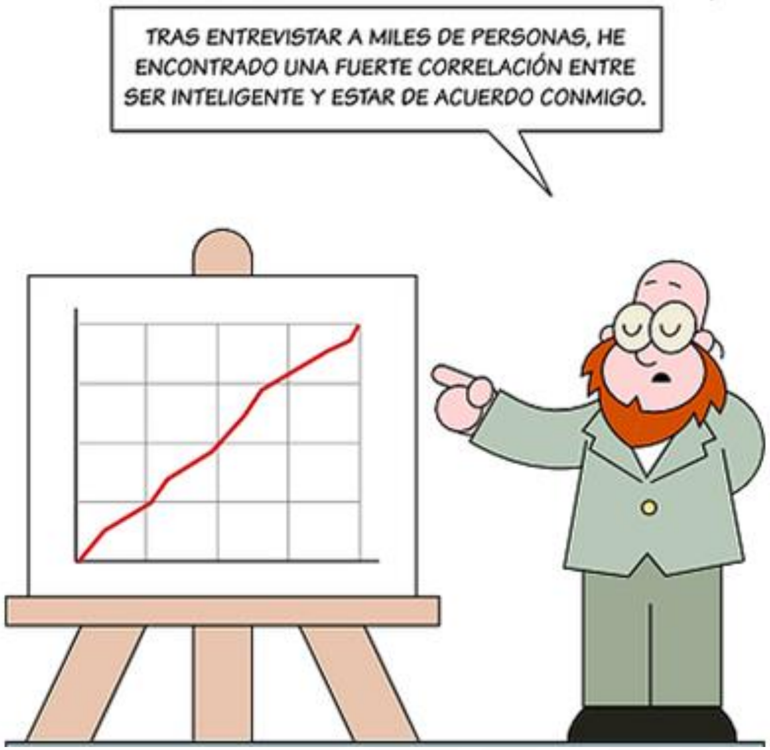
$$\overline{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{k-1}$$

VALIDACIÓN DEL MODELO

Fuentes de variación	Sumas de cuadrados	gl	Varianza	F
Regresión o explicada	$\sum\left(\hat{Y}-\bar{Y}\right)^2$	k	$S_{\exp}^2=\frac{SC_{\exp}}{k}$	$\frac{S_{\exp}^2}{S_{res}^2}=\frac{\frac{R_{XY}^2}{k}}{\frac{1-R_{XY}^2}{N-k-1}}$
Residual o no explicada	$\sum\left(Y-\hat{Y}\right)^2$	N-k-1	$S_{res}^2=\frac{SC_{res}}{N-k-1}$	
Total	$\sum\left(Y-\bar{Y}\right)^2$	N-1	$S_t^2=\frac{SC_t}{N-1}$	

- ANOVA
  - $F > F_{(\alpha,k,N-k-1)}$ → Se rechaza la Hipótesis nula. Las variables están relacionadas. El modelo es válido.
  - $F \leq F_{(\alpha,k,N-k-1)}$ → Se acepta la Hipótesis nula. Las variables no están relacionadas. El modelo no es válido. (k = número de variables independientes)

Regresión Lineal Múltiple



REGRESIÓN LINEAL MÚLTIPLE

- Se ha visto el tema del análisis de regresión simple:

$$\text{Precio de la casa} = \beta_0 + \beta_1(\text{Área de la casa}) + \varepsilon$$

- Pero en general, una variable dependiente depende de más de una variable independiente:

- Precio de la casa puede depender de:
  - Área
  - Antigüedad
  - Número de baños
  - Área del garaje
  - Etc.

REGRESIÓN LINEAL MÚLTIPLE

- Para tratar este tipo de problemas se requiere expandir el análisis de regresión:

Regresión Lineal Simple

↓

Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

↓

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\dots\dots + \beta_p x_p + \varepsilon$$

MODELO DE REGRESIÓN MÚLTIPLE

Objetivo: Examinar la relación lineal entre
una variable dependiente (y) y
dos o más variables independientes (x_i)

Modelo poblacional:

Y-intercepto

Pendientes

Error aleatorio

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

Modelo de regresión múltiple estimado:

Valor estimado o predecido de $\hat{y}$

y-intercepto estimado

Pendientes estimadas

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

METODOLOGÍA PARA LA CONSTRUCCIÓN DE MODELOS

Las 3 etapas:

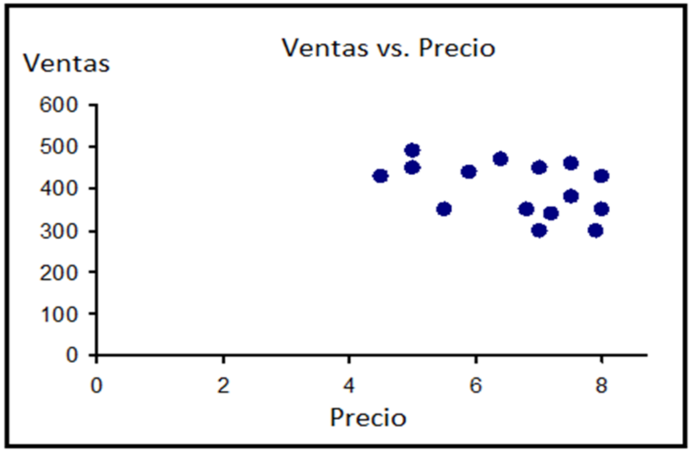
- **Especificación del modelo**
  - Especificación del modelo de regresión poblacional.
  - Recolección de la data muestral.
- **Formulación o construcción del modelo**
  - Cálculo de los coeficientes de correlación entre las distintas variables, dependientes e independientes.
  - Ajuste del modelo a la data. Estimación de la ecuación de regresión múltiple.
- **Diagnóstico del modelo**
  - Pruebas estadísticas para determinar la bondad de ajuste del modelo a la data.
  - Verificación de los supuestos de regresión múltiple.

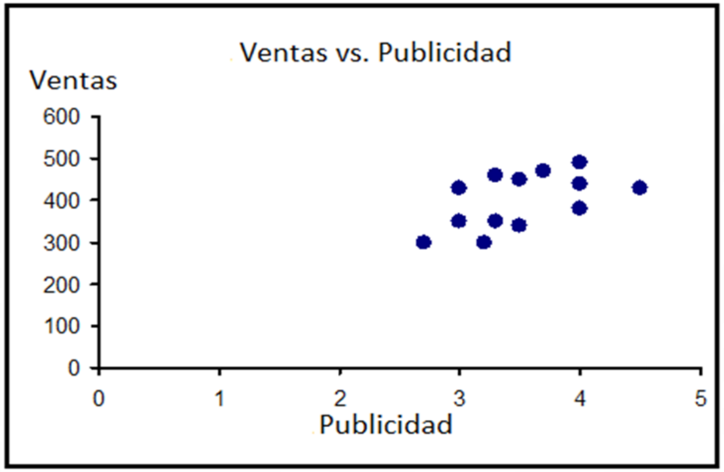
EJEMPLO DE APLICACIÓN

- Un distribuidor de pies (postres) desea evaluar los factores que se cree influyen en la demanda de sus productos, y si existen éstos factores cuantificar cual de éstos es el más importante.

EJEMPLO DE APLICACIÓN

- Diagrama de Dispersión.





EJEMPLO DE APLICACIÓN

- Especificación del Modelo

Un distribuidor de pies (postres) desea evaluar los factores que se cree influyen en la demanda

➢ Variable **Dependiente**: Ventas (unidades / semana)

➢ Variables **Independientes**: Precio ($) y Publicidad ($100)

Modelo de Regresión múltiple Poblacional:

$$\textbf{Ventas} = \beta_0 + \beta_1(\textbf{Precio}) + \beta_2(\textbf{Publicidad}) + \varepsilon$$

EJEMPLO DE APLICACIÓN

- Interpretación de los Coeficientes Estimados

- Pendientes (B_i)

  - Estiman el cambio en el valor promedio de "y" como b_i unidades por cada unidad de incremento en x_i manteniendo las otras variables constantes.

    - Ejemplo: Si b_1 = -20, entonces se espera que las ventas promedio (y) se reduzcan en 20 pies por semana por cada $1 en que se incremente el precio (x_1), manteniendo constante la variable publicidad (x_2).

- Intercepto (B_0)

  - Estima el valor promedio de y cuando todas las variables x_i son iguales a cero (suponiendo que el valor cero está dentro de los rangos de valores que pueden tomar los x_i).

EJEMPLO DE APLICACIÓN

Sema-na	Venta de pies	Precio ($)	Publicidad ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Modelo de Regresión Múltiple:

$$\text{Ventas} = b_0 + b_1\,(\text{Precio}) + b_2\,(\text{Publicidad})$$

Matriz de correlación:

	Venta de pies	Precio	Publicidad
Venta de Pies	1		
Precio	-0.44327	1	
Publicidad	0.55632	0.03044	1

EJEMPLO DE APLICACIÓN

	Ventas de pies	Precio	Publicidad
Ventas de pies	1		
Precio	-0.44327	1	
Publicidad	0.55632	0.03044	1

- Ventas vs. Precio : r = -0.44327
  - Hay una asociación lineal **negativa** entre las ventas y el precio
- Ventas vs. Publicidad : r = 0.55632
  - Hay una asociación lineal **positiva** entre las ventas y la publicidad

EJEMPLO DE APLICACIÓN

$$Ventas = 306.526 - 24.975(Precio) + 74.131(Publicidad)$$

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = 0.52148$$

Estadísticas de la regresión	
correlación múltiple	0.7221343
Coeficiente de determinación R^2	0.5214779
R^2 ajustado	0.4417243
Error típico	47.463413
Observaciones	15

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	29460.02687	14730.01343	6.53860679	0.012006372
Residuos	12	27033.30647	2252.775539		
Total	14	56493.33333			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	306.52619	114.2538935	2.682851182	0.01993159	57.58834426	555.464042
Precio	-24.97509	10.83212512	-2.305650022	0.03978846	-48.5762627	-1.3739163
Publicidad	74.130957	25.96731792	2.854779139	0.01449363	17.55303206	130.708883

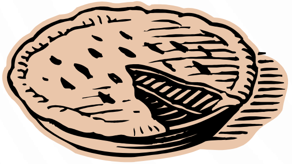
EJEMPLO DE APLICACIÓN

Ecuación estimada de regresión múltiple:

$$\text{Ventas} = 306.526 - 24.975(\text{Precio}) + 74.131(\text{Publicidad})$$

Donde:

Ventas (número de pies por semana)
Precio ($)
Publicidad ($100's)



b_1 = -24.975: Las ventas decrecerán en promedio 24.975 pies por semana por cada $1 incrementado en el precio, manteniendo constante la publicidad

b_2 = 74.131: Las ventas crecerán en promedio 74.131 pies por semana por cada $100 incrementado en publicidad, manteniendo constante el precio.

Ejemplo de aplicación

DIAGNÓSTICO DEL MODELO: PRUEBA F (SIGNIFICANCIA GENERAL)

- Prueba F para la significancia del modelo (General)
- Muestra si hay una relación lineal entre todas las variables x (consideradas en forma conjunta) e y
- Usa el estadístico de prueba F
- Hipótesis:
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (No hay relación lineal)
  - H_A: Al menos un $\beta_i \neq 0$ (Existe relación lineal entre (y) y al menos un x_i)

Ejemplo de aplicación

DIAGNÓSTICO DEL MODELO: PRUEBA F (SIGNIFICANCIA GENERAL)

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.72213429
Coeficiente de determinación R^2	0.52147794
R^2 ajustado	0.44172426
Error típico	47.4634126
Observaciones	15

$$F = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	29460.02687	14730.01343	6.538606789	0.012006372
Residuos	12	27033.30647	2252.775539		
Total	14	56493.33333			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	306.526193	114.2538935	2.682851182	0.019931591	57.58834426	555.4640423
Precio	-24.9750895	10.83212512	-2.305650022	0.039788461	-48.5762627	-1.373916335
Publicidad	74.1309575	25.96731792	2.854779139	0.014493627	17.55303206	130.7088829

Ejemplo de aplicación

PREDICCIONES

Predecir las ventas de una semana en la cual el precio es $5.50 y la publicidad es $350.

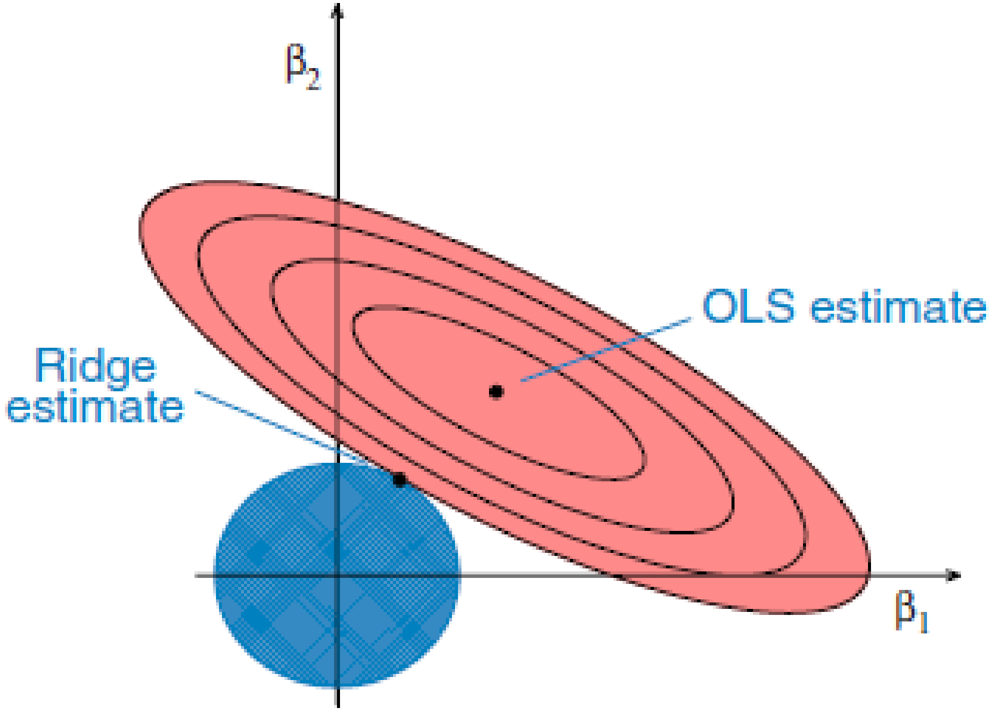
$$\begin{aligned} \text{Ventas} &= 306.526 - 24.975(\text{Precio}) + 74.131(\text{Publicidad}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62 \end{aligned}$$

La venta predecida es 428.62 pies

Nota: La publicidad está en $100's, entonces $x_2 = 3.5$ significa $350

Regresión Avanzada

Ridge y Lasso



ALGUNOS PROBLEMAS CON LOS MODELOS DE REGRESIÓN

- Falta de interpretabilidad; si se utiliza un gran número de predictores, sería deseable determinar un pequeño subconjunto de éstos con fuerte poder explicativo y predictivo.

- Si existen predictores o features fuertemente correlacionados(Multicolinealidad).

- Existencia del caso p>= n (Número de variables mayor al número de observaciones).

SOLUCIONANDO LOS PROBLEMAS DE LOS MODELOS DE REGRESIÓN

- Una solución habitual implica hacer selección de variables para obtener un modelo más parsimonioso y estable (George, 2000).

- Necesitamos de alguna medida que tome en cuenta el ajuste a los datos de entrenamiento pero que penalice la complejidad del modelo de forma tal que posea buen poder predictivo.

- Las técnicas más conocidas en estos casos son los métodos secuenciales o de a pasos , en los cuales en el pasaje de un modelo a otro se agregan o eliminan variables de a una por vez (Forward Selection, Backward Elimination o Forward-Backward).

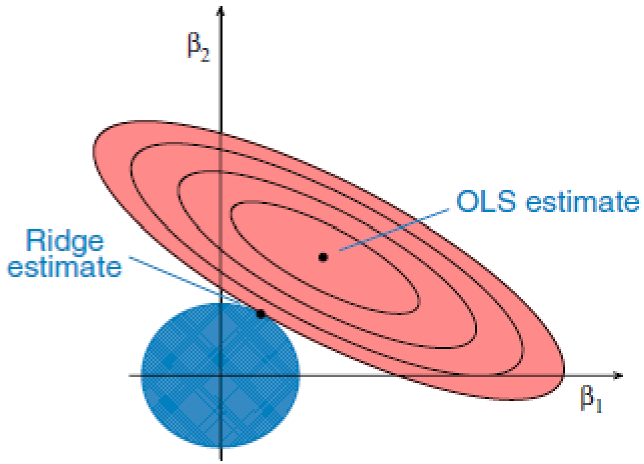
SOLUCIONANDO LOS PROBLEMAS DE LOS MODELOS DE REGRESIÓN

- Estos son métodos **greedy** que reemplazan la búsqueda de un óptimo global por la consideración sucesiva de óptimos locales, con lo cual no garantizan la mejor solución y ni siquiera la misma entre sus distintas variantes.

- Sin embargo, la mayor desventaja que poseen es su fuerte inestabilidad en el sentido de que pequeños cambios en el conjunto de datos pueden producir grandes modificaciones en los resultados (**Breiman**, 1996).

- Esto se debe principalmente a que realizan un proceso discreto de exploración del espacio de modelos (**cada variable es seleccionada o descartada**).

MODELO DE REGRESIÓN PENALIZADA RIDGE



$$Cost\,(W) = RSS(W) + \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

- El principal problema a resolver en la aplicación de **Regresión Ridge** es la determinación del valor de k más adecuado.

- La elección de este parámetro involucra un balance entre los componentes de sesgo y variancia del error cuadrático medio al estimar β.

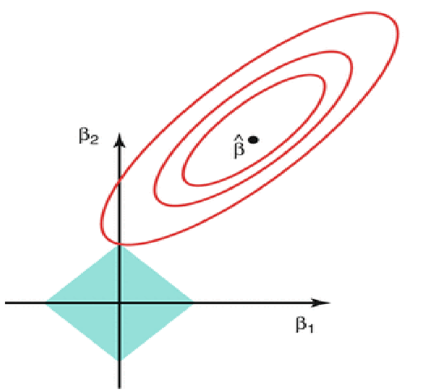
CONCLUSIONES DEL MODELO DE REGRESIÓN PENALIZADA RIDGE

- En general Regresión Ridge produce predicciones más precisas que los modelos obtenidos por MCO + selección "clásica" de variables.

- Sin embargo, si bien al aumentar λ (mayor penalización) los coeficientes estimados se contraen hacia cero, ninguno de ellos vale exactamente cero por lo cual no se produce selección de variables.

- Es recomendable estandarizar variables y validar su aporte al modelo final.

MODELO DE REGRESIÓN PENALIZADA LASSO



$$L(b) = \sum_{i=1}^{n} (y_i - x_i\, b)^2 + \lambda\, |b|$$

- También motivado por el objetivo de encontrar una técnica de regresión lineal que fuera estable pero que realizara selección de variables, **Tibshirani** (1996) propuso Lasso (Least Absolute Shrinkage and Selection Operator).
- Lasso es una técnica de regresión lineal regularizada, como Ridge, con una leve diferencia en la penalización (norma **L1** en lugar de L2.

Modelo de Regresión Penalizada Lasso

- Para valores crecientes de λ , los coeficientes βj se contraen hacia cero como en Ridge (shrinkage), con la diferencia de que algunos de ellos se anulan.

- Esto es, Lasso produce estimación y selección de variables en forma continua y simultánea, siendo especialmente útil en el caso $p>=n$.

- En los últimos años se han presentado algunas generalizaciones y extensiones de las técnicas presentadas anteriormente, especialmente diseñadas para ciertas situaciones particulares.

GRACIAS
POR SU ATENCIÓN