Introducción al Modelado con Regresión Lineal Múltiple

Dr. Andrés Farall

September 29, 2021

El objetivo de este tutorial es motivar el modelado de datos, sin el uso de nociones formales de probabilidad ni de estadística. El enfoque se centra en los datos, y se introduce la necesidad de modelado de los mismos a partir de un análisis exploratorio (EDA) aplicado a un dataset real.

Si bien el cuerpo teórico estadístico es, quiza, el mejor contexto para definir y formalizar el modelado de datos, este documento busca brindar una aproximación intuitiva al problema del modelado basado en datos, siendo la posterior profundoización en el conocimiento estadístico el complemento esencial de las ideas presentadas aquí. Asimismo, varias técnicas de modelado podrían haber sido utilizadas a los fines de motivar el modelado de datos, sin embargo la Regresión Lineal Múltiple es ideal, no sólo por su simplicidad, sino por sus capacidades Descriptivas, Explicativas y Predictivas.

Con qué datos vamos a trabajar?

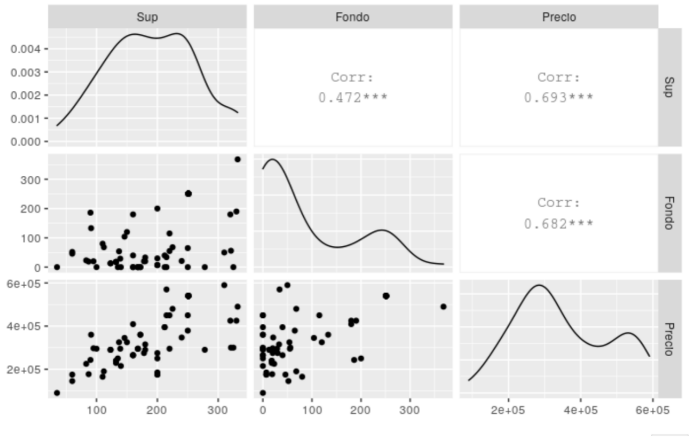
A los efectos de ejemplificar los conceptos de este tutorial, utilizaremos un dataset real de publicaciones digitales de inmuebles, provisto en forma libre por Properati (https://www.properati.com.ar/). Con este dataset trataremos de modelar los precios de las casas de CABA en función de algunas variables relevantes, como son la superficie cubierta (llamada Sup en el dataset) y no cubierta (llamada Fondo) de las mismas. Una primera visualización de los datos se muestra a continuación

Casas en el Barrio de Boedo

Sup <dbl>	Fondo <dbl>	Precio <dbl>
140	0	215000
35	0	90000
132	18	239000
212	0	394900
225	68	480000
60	46	175000
135	0	250000
200	30	275000
110	80	165000
200	7	175000

1-10 of 68 rows Previous 1 2 3 4 5 6 7 Next

En este caso contamos con 68 casas del barrio de Boedo, con precios en dólares que van desde los u$s 90000 hasta los u$s 600000. Un resumen básico de estos datos se presenta en el próximo gráfico.



Puede verse que ambas variables, Sup y Fondo, se hallan positivamente correlacionadas con Precio; Asimismo, también se aprecia una correlación positiva entre ellas.

En lo que sigue, definiremos los conceptos básicos del modelado, aplicando en cada caso estos conceptos al dataset presentado.

Qué es un Modelo para la Ciencia de Datos ?

> Para un problema de interés cualquiera, un Modelo es una representación simplificada de las relaciones existentes entre las variables relevantes a ese problema. Desde la perspectiva de los datos, **Un Modelo es una explicitación simplificada de algunas relaciones potencialmente existentes entre las variables disponibles en los datos**

Toda tarea de modelado consta de tres elementos o etapas fundamentales:

- **Representación**
- **Evaluación**
- **Optimización**

Para cada uno de estos tres elementos existen muchisimas alternativas de elección posibles. En algunas técnicas de la estadistica y del machine learning, los tres elementos vienen pre-determinados en la técnica en cuestión. Sin embargo, esencialmente, estos elementos son independientes, pudiedo combinarse a voluntad las distintas opciones para cada uno de ellos.

Veremos en lo que sigue cada uno de estos elementos en detalle.

Representación

La repesentación, que puede pensarse como el modelo en si mismo, es la definición y explicitación de las relaciones, de las infinitas posibles, que vinculan a las variables involucradas en el análisis. Esta etapa define que será posible captar de los datos, y que no será posible. Esta elección de las variables y sus relaciones se conoce como **especificación** del modelo.

¿ Porqué entonces no incluir siempre una inmensa cantidad de relaciones entre las variables ? Básicamente porque cada relación que pongamos va a "consumir" una cierta cantidad de información de los datos, y los datos son limitados. Por ende, tenemos que incluir las relaciones más importantes, si queremos utilizar la información de la mejor manera posible.

Ejemplo: Pensemos en el sigueinte problema. Tenemos los salarios de 50 empleados de una empresa, y podríamos intentar modelar estos salarios en función del nivel educativo (primario, secundario y universitario) de los empleados. Esto nos llevaría a un modelo en el que cada una de las tres categpriás del nivel educativo se vería representada por los salarios de los empleados que caen en las mismas, pongamos por caso 16 con nivel primario, 20 con nivel secundario y 14 universitarios. Con este modelo podríamos creer razonable que los promedios de los 16, 20 y 14 salarios nos aporten información válida sobre la relación entre el nivel educativo y el salario. Sin embargo, si quiesieramos agregar a este modelo el efecto potencial del barrio en el que vive cada empleado, muy probablemente la mayor parte de los empleados queden representados por una única combinación de nivel educativo y barrio. En este último caso, cualquier diferencia observada entre los salarios, dificilmente pueda ser confiablemente atribuida a los efectos del nivel educativo y del barrio. Acá se ve claramente que la cantidad de datos habilitará ciertas representaciones (modelos) y no otras.

Ejemplo de Representación

Usando nuestro ejemplo de aplicación, podríamos plantear como representación del precio de las casas de Boedo, la siguinte relación:

$$Precio^{M2} = \mu + \alpha * Sup + \beta * Fondo$$

El superíndice M2 significa que estamos definiendo el precio según el modelo (el 2 es por la cantidad de variables), no el precio real de las casas. En esta representación, a μ, α, β se los denomina *parámetros*, a Sup y Fondo *predictoras*, y a Precio la variable a ser predicha (o *target*). Nótese que esta representación seguramente es erronea, pues el precio de las casas debe depender de otras variables importantes que no figuran en la representación, como ser la antiguedad, la ubicación y el estado del inmueble. Más aún, muy posiblemente la relación entre el precio y la superficie cubierta (Sup) podría no ser lineal. De hecho, sería razonable que a partir de un cierto valor de superficie cubierta, el aumneto de precio sea cada vez menor por cada metro cuadrado adicional.

Otro supuesto muy fuerte de esta representación, es el efecto **aditivo** que ambas variables ejercen en la formación del precio, según el modelo. Es decir, el incremento del precio por cada metro de Fondo que se le agrega a una casa, el modelo supone que se le "suma" al precio que le correponde por los metros cubiertos de superficie cubierta que la casa posee. Esto podría ser falso. Quizá, el incremento de precio que cada metro de Fondo produce sea mayor en casas con pocos metros cubiertos, o al revés.

El modelo recién planteado puede pensarse como una función, que dados los parámetros $\boldsymbol{\theta} = (\mu, \alpha, \beta)$, asigna un precio a una combinación cualquiera de valores de Sup y Fondo. O sea

$$Precio^{M2} = F_{\mu,\alpha,\beta}\left(Sup, Fondo\right)$$

Es así evidente que el modelo no queda completamente deperminado hasta que no se establezcan los valores de los parámetros $\boldsymbol{\theta} = (\mu, \alpha, \beta)$. Para que el modelo sea de utilidad, tanto la especificación del mismo (que variables y que relaciones) debe ser buena, como así también la determinación de los valores de los parámetros. Por ejemplo, la misma especificación del modelo (representación) podría ser aplicada a dos barrios distintos, pero los valores de parámetros propios de cada barrio podrían ser muy distintos. Puntualmente, es de esperar que en barrios caros el parámetro α sea mayor al de los barrios baratos.

Un concepto fundamental a destacar a esta altura es el de *predicción*, que proviene de aplicar la función $F_{\mu,\alpha,\beta}\left(Sup, Fondo\right)$ a una combinación de las variables Sup y Fondo.

Dada la representación y los datos, la determinación de los parámetros se logrará en la etapa de Optimización.

Evaluación

La etapa de evaluación consiste en establecer una métrica o medida de falta de ajuste a los datos, de una determinación cualquiera de los parámetros del modelo. Con "determinación cualquiera de los parámetros" queremos decir alguna elección particular de parámetros del modelo elegido.

La definición de la métrica en este etapa de evaluación es tan importante como la elección del modelo (representación), pues la métrica define en que aspecto particular nos interesa que el modelo se parezca a los datos. A la medida de falta de ajuste elegida, suele denominarsela *función de pérdida* (Loss function).

Ejemplo de Evaluación

En este caso, dada la representación y los datos, tenemos que generar la métrica de falta de ajuste del modelo a los datos. Una posibilidd muy usual es considerar el promedio de las discrepancias al cuadrado entre los precios reales (observados) y los predichos por el modelo. Es decir

$$\begin{aligned} L(\mu, \alpha, \beta) &= \frac{1}{68} \sum_{i=1}^{68} \left(Precio_i - Precio_i^{M2} \right)^2 \\ &= \frac{1}{68} \sum_{i=1}^{68} \left(Precio_i - (\mu + \alpha * Sup_i + \beta * Fondo_i) \right)^2 \end{aligned}$$

Claramente, la falta de ajuste o función de pérdida (L) depende de los valores de los parámetros (μ, α, β) que determinan el modelo. Dados los datos y el modelo, se buscarán aquellos valores de los parámetros que minimicen la pérdida.

Obviamente exusten infinitas elecciones posibles de funciones de pérdida. La pérdida recién definida se conoce como pérdida cuadrática, y tiene, entre otras, la característica de castigar de la misma forma las sobreestimaciones y las subestimaciones del modelo. En algunas situaciones, podría ser razonable utilizar una pérdida que castigue más fuertemente la discrepancia en un sentido. Esto podría logarse facilmente modificando la función de pérdida. Otra característica importante de la pérdida cuadrática es la "desporporcionada" importancia (por el cuadrado !) que le dá a las discrepancias grandes (en valor absoluto).

Optimización

La tarea de optimización consiste en, dadas la representación y la evaluación elegidas, definir un procedimiento efectivo que pueda hallar la determinación óptima de los parámetros del modelo, es decir, la determinación de parámetros que produzacan la menor falta de ajuste posible. Esta tarea es conocida también como la etapa de *entrenamiento* o *ajuste* del modelo.

Ejemplo de Optimización

El objetivo ahora es plantear algún procedimiento que dados los datos, la representación y la métrica de evaluación, permita escoger los valores de parámetros más apropiados. Más específicamente, necesitamos hallar los valores de los parámetros (μ, α, β) que minimicen la pérdida L. Es decir

$$Min_{\mu,\alpha,\beta}\, L(\mu,\alpha,\beta) = Min_{\mu,\alpha,\beta}\, \frac{1}{68} \sum_{i=1}^{68} (Precio_i - (\mu + \alpha * Sup_i + \beta * Fondo_i))^2$$

Otro concepto fundamental que aparece en esta etapa es el de *estimación*, que es el nombre que se le da a estos valores de parámetros que minimizan la función de pérdida, o sea

$$ArgMin_{\mu,\alpha,\beta}\, L(\mu,\alpha,\beta) = \left(\widehat{\mu}, \widehat{\alpha}, \widehat{\beta}\right)$$

El agregado del "sombrerito" a los nombres de los parámetros indica que son valuaciones provenientes del procedimiento de optimización.

En este caso particular, el problema de optimización planteado puede resolverse analíticamente. Sólo debemos derivar la función de pérdida L con respecto a los tres parámetros (μ, α, β), igualar a cero, y despejar para hallar los valores de los parámetros que satisfacen la condición necesaria de punto crítico, pues esta función de pérdida es convexa y tiene un sólo punto crítico que es el mínimo absoluto de la función. Sin embargo, inicialmente no vamos a seguir este camino.

Vamos a aplicar dos mecanismos muy básicos de optimización.

El primero, y más sencillo, es un método de optimización aleatoria llamado **Luus-Jaakola**, que sólo presupone que el lector sepa generar números aleatorios con distribución uniforme.

El segundo, apenas más complejo, es un método de optimización por **gradiente descendente**, que presupone que el lector posea conocimientos básicos de análisis matemático.

Para seguir con la lectura de este tutorial, el lector puede elegir cualquiera de los dos métodos de optimización. Una vez terminada la lectura del método elegido, se podrá continuar a partir de la sección Modelo Entrenado.

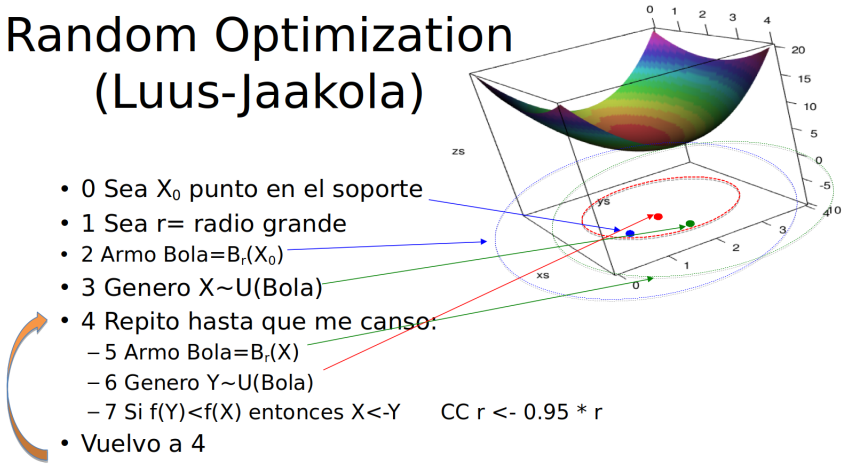
Para ambos enfoques nos convendrá simplificar los nombres de las variables. Así, con la simplificación de los nombres, buscaremos resolver

$$Min_{\mu,\alpha,\beta} \sum_{i=1}^{68} (P_i - (\mu + \alpha * S_i + \beta * F_i))^2$$

Optimización Aleatoria o Luus-Jaakola

El método consiste en recorrer iterativamente el espacio paramétrico mediante la generación aleatoria de valores. En nuestro caso el espacio paramétrico es $\mathbb{R}^3$, ya que tenemos 3 parámetros que potencialmente podrían tomar cualquier valor. En cada iteración la región sobre la cual se generan aleatoriamente los valores se centra en los mejores valores de los parámetros hallados hasta ese momento. De esa forma, cada vez que se encuentra un valor mejor, con menor valor de pérdida, se produce una **actualización** y se mueve la región de búsqueda hacia ese punto. Para asegurar la convergencia, cada vez que en un nuevo intento aleatorio no se consigue una mejora, se reduce levemente el tamaño de la región de búsqueda. Esto produce que alrededor de zonas "malas" (probablemente al inicio del algoritmo) se busque con mucha libertad, mientras que una vez que se alcanza una zona "buena" el entorno de búsqueda se reduce, acelerando la convergencia hacia el óptimo.

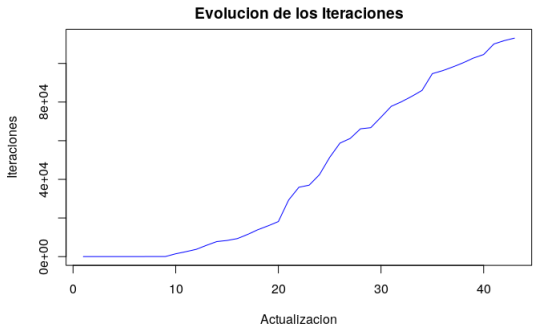
El próximo gráfico muestra el algoritmo junto a una ilustración de funcionamiento del miemo.



Veamos la aplicación de este método a nuestro problema. Comenzamos inicializando arbitrariamente los valores de los parámetros

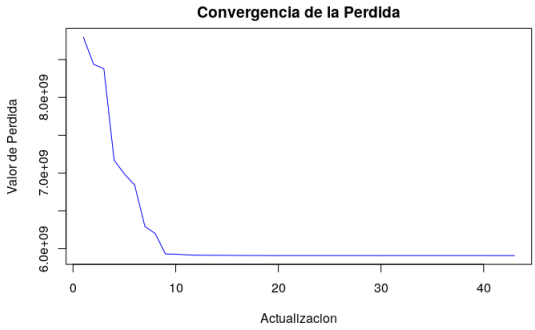
μ_0=100000
α_0=1000
β_0=1000

Empecemos viendo la evolución de la cantidad de iteraciones en función de la cantidad de actualizaciones (modificación del óptimo hasta ese momento) a lo largo del proceso.

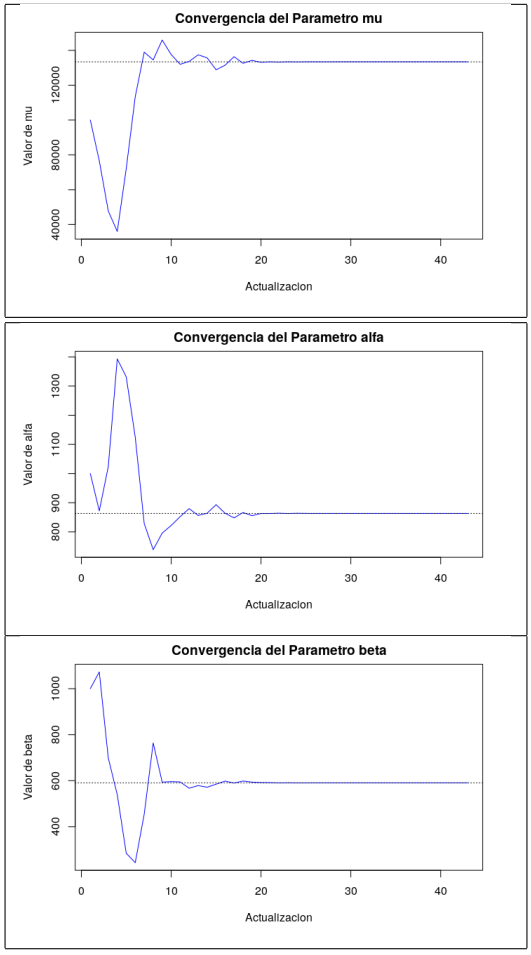


Como es de esperar, al inicio del algoritmo la cantidad de iteraciones necesarias para hallar una mejora en la pérdida (actualización) es baja, mientras que conforme se va acercando a los valores óptimos de los paramétros, al método le cuesta más iteraciones para encontrar una mejora en la pérdida.

El siguiente gráfico muestra la evolución de las mejores pérdidas que se van obteniendo a lo largo del proceso.



Pareciera que a partir de la décima actualización la pérdida no mejora. En realidad la pérdida sigue disminuyendo, por eso se siguen produciendo actualizaciones. Sin embargo las mejoras son insignificantes. Veamos entonces el comportamiento de la evolución de los parámetros a lo largo del proceso.



Se ve claramente que la evolución de cada uno de los tres parámetros convergen a un valor final (linea negro de puntos), a los que podemos considerar como las estimaciones. Estos valores finales son:

$\widehat{\mu}$=133401
$\widehat{\alpha}$=863
$\widehat{\beta}$=591

Optimización por Gradiente Descendente

La solución que propondremos en esta instancia consiste en determinar un procedimiento muy general, llamado método del **gradiente descendente**. El método consiste en comenzar con valores iniciales de parámetros aleatorios, y calcular (o aproximar numéricamente) el gradiente (o sea el vector de derivadas parciales), para moverse en la dirección que disminuya la pérdida lo más posible. Una vez dado el primer paso, se repite iterativamente este procedimiento, hasta alcanzar valores de parámetros para los cuales no exista dirección en la que pueda achicarse la périda. Para realizar este procedimiento deberemos definir de antemano el tamaño del paso que vamos a dar, al cual llamaremos λ (learning rate). Suponiendo que ya tenemos para el paso t algún valor de los parámetros $\boldsymbol{\theta}_t = (\mu, \alpha, \beta)$, la sencilla fórmiula de este procedimiento la podemos enunciar así

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta}_t - \lambda \nabla L$$

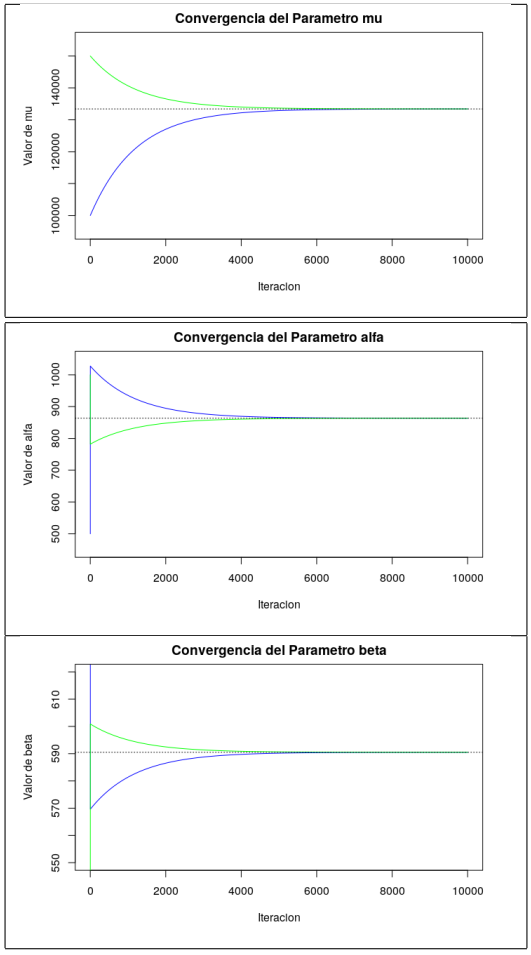
donde ∇L representa al gradiente de la función de pérdida.

Este procedimiento tiene la ventaja de poder aplicarse en situaciones en las cuales no es trivial igualar a cero las derivadas parciales y despejar. Incluso en situaciones en la cuales las derivadas parciales sólo pueden aproximarse numéricamente, puede seguirse este procedimiento. Por supuesto, los valores a los cuales converja (si sucede) dependerán de la función a minimizar, pudiendo caer en puntos silla o mínimos locales, en algunos casos.

En nuestro ejemplo particular, la expresión analítica del gradiente se calcula muy facilmente de la siguiente manera

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial \mu} \\ \frac{\partial L}{\partial \alpha} \\ \frac{\partial L}{\partial \beta} \end{pmatrix} = \begin{pmatrix} \frac{1}{68} \sum_{i=1}^{68} (-2) \left(P_i - \mu - \alpha * S_i - \beta * F_i\right) \\ \frac{1}{68} \sum_{i=1}^{68} (-2) S_i \left(P_i - \mu - \alpha * S_i - \beta * F_i\right) \\ \frac{1}{68} \sum_{i=1}^{68} (-2) F_i \left(P_i - \mu - \alpha * S_i - \beta * F_i\right) \end{pmatrix}$$

Vamos ahora a optimizar los tres parámetros del modelo propuesto. Para ejemplificar el proceso de convergencia, mostrado en el próximo gráfico, proponemos dos conjuntos de paámetros iniciales distintos. El primer conjunto, representado en color azul en el gráfico, será $\mu_0 = 100000, \alpha_0 = 500, \beta_0 = 1000$. El segundo conjunto, representado en color verde, será $\mu_0 = 150000, \alpha_0 = 1000, \beta_0 = 500$.



Para ambos conjuntos de parámetros, los valores a los cuales converge son muy cercanos a:

$\widehat{\mu}$=133401
$\widehat{\alpha}$=863
$\widehat{\beta}$=591

Estos valores son los que efectivamente anulan el gradiente, y se muestran en los gráficos anteriores como lineas horizontales negras punteadas.

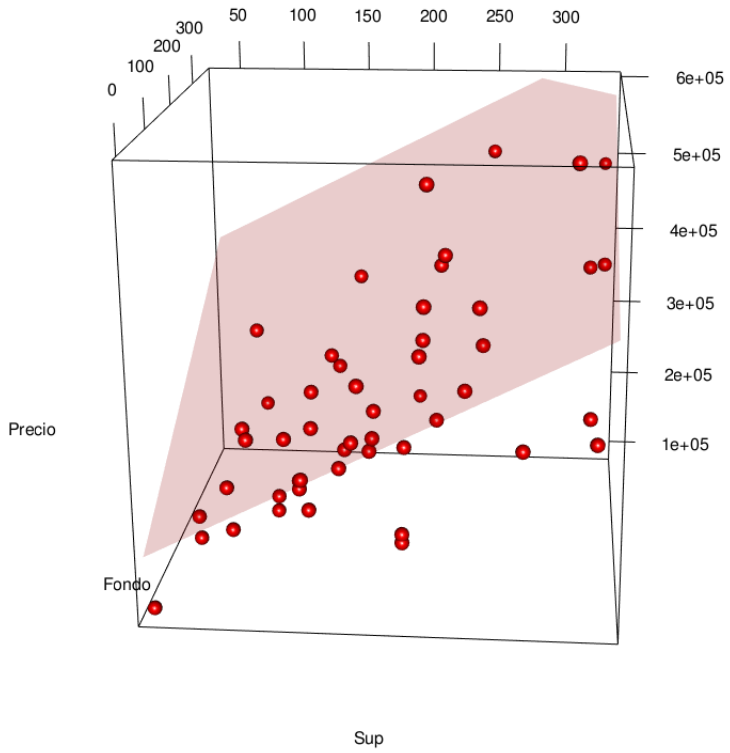
Modelo Entrenado

Ambos enfoques de estimación arrojaron los mismos resultados. Como es de esperar, la estimación de los parámetros correspondientes a las variables Sup y Fondo son positivos, indicando que un aumento de ambas variables redundará en un incremento del precio de las casas. La estimación positiva de $\widehat{\mu}$ estaría indicando que las casas mas grandes tendrían un precio por metro cuadrado menor que las casas más chicas. Lo cual parece razonable. A su vez, el mayor valor de $\widehat{\alpha}$ comparado con $\widehat{\beta}$ podría estar indicando que el metro cuadrado cubierto aporta más al precio que el descubierto; razonable.

Pese a la utilidad de los métodos iterativos de estimación usados previamente, de ahora en más, por simplicidad, vamos a hacer uso de la solución analítica que nos provee la función *lm* del R. En efecto, la estimación analítica de los tres parámetros arroja los mismos valores mencionados anteriormente.

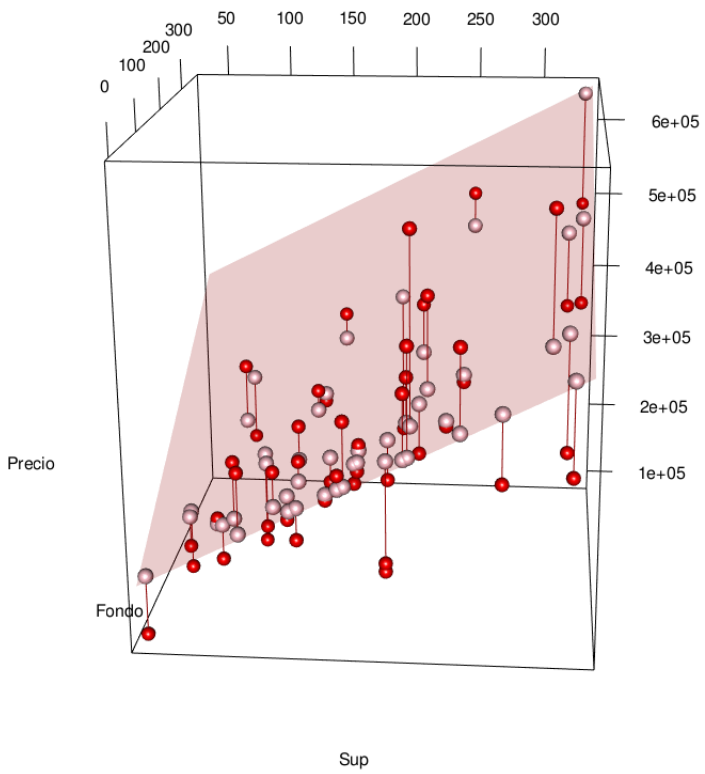
```
> ajus<-lm(Precio~Sup+Fondo,data=datos1d)
> coe<-coef(ajus)
> coe
(Intercept)         Sup       Fondo
133400.9620    863.4824    590.5066
```

Una vez determinados los valores paramétricos óptimos del modelo, podemos visualizar el ajuste a los datos.



Se aprecia que el modelo estimado, que en este caso es un plano, acompaña las observaciones, prediciendo aumentos de precios en las direcciones en las que ambas variables crecen.

En el próximo gráfico agregamos las predicciones (como puntos más claros) que realiza el modelo para las 68 combinaciones de Sup y Fondo observadas en la base de datos. Para resaltar las discrepancias entre los precios observados y los predichos, unimos ambos elementos con segmentos.



Es claro que las predicciones no son demasiado precisas.

A continuación plantearemos una serie de preguntas relevantes que se originan en un análisis como este. En cada caso, trateremos de responder estas preguntas usando procediemintos basados en los datos que creemos son relativamente intuitivos.

¿ Cuán bueno es el ajuste del modelo a los datos ?

Si lo que queremos decir con "cuan bueno es el ajuste" es cuan bien predice el modelo ajustado/entrenado los precios de las casas de Boedo, la respuesta deberá involucrar alguna medida de discrepancia entre los precios reales y los que predice el modelo. Podríamos usar la misma función de pérdida que optimizamos para entrenar el modelo. El problema de esta medida es su poca interpretabilidad (diferencias al cuadrado !). Una medida más amigable consiste en calcular el promedio de las distancias absolutas entre los precios observados y los predichos (Mean Absolute Error, o MAE), o sea

$$\begin{aligned} MAE &= \frac{1}{68}\sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M2} \right\| \\ &= 60757 \end{aligned}$$

Es decir que en promedio erramos por u$s 60757. Esto ya es más interpretable. Si quisieramos una medida relativa del error, podríamos calcular el MAE Proporcional, o PMAE

$$\begin{aligned} PMAE &= \frac{1}{\overline{Precio}}\frac{1}{68}\sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M2} \right\| \\ &= \frac{60757}{347997} = 0.1746 \end{aligned}$$

Esto indica que le estamos errando por un poco más del 17% del valor promedio de las casas. Entonces, para una nueva casa de Boedo ¿ puedo esperar que ese sea el error de mi modelo ?

CUIDADO ! El error fue medido con los mismos datos con los que ajustamos el modelo. Este es un **GRAN problema**. De hecho si hubieramos trabajado de entrada sólo con 3 casas, el error hubiera sido CERO, ya que siempre existe un plano que pasa por 3 puntos en el espacio. Este último ejemplo extremo nos muestra que debemos tener cuidado en la medición del error. La medición del error del modelo hecha con los mismos datos con los que se entrena se conoce como medición ingenua (o naive) del error. Esta medición tiende a subestimar sistemáticamente al error real, es decir, el error que obtendríamos al aplicar las predicciones del modelo a nuevas (y muchas) casas de Boedo.

Una posible solución a este problema consiste en ajustar el modelo con datos distintos a los que se usarán para la medición del error. La forma de hacer esto, maximizando la cantidad de datos con los que hacemos el ajuste del modelo, es separar una observación de las 68, y ajustar el modelo con las 67 que quedan, para luego medir el error con la observación que dejé afuera. Repitiendo esto para cada observación, obtenddremos 68 mediciones de error que pieden ser pormediadas para cuantificar una medida más fiable de dicho error de predicción. Nótese que de esta forma van a ser entrenados 68 modelos, que arrojarán 68

errores, y en ningún caso la observación con la que se mide el error será usada en el entrenamiento del modelo con el que se calcula ese error.

Este procedimiento es conocido como Validación Cruzada o Leave One Out Cross Validation (LOOCV).

En nuestro caso el MAE y PMAE por validación cruzada dan 64087 y 0.1842, respectivamente. Nótese que ambos mediciones son mayores a la primer evaluación (la medición ingenua).

La capacidad que tiene un modelo de predecir correctamente los valores de la target, en nuestro caso, predecir bien los valores de los precios, es considerada como la **capacidad predictiva** de un modelo.

¿ Cuán certeras (creibles/estables/repetibles) son las relaciones estimadas por el ajuste del modelo ?

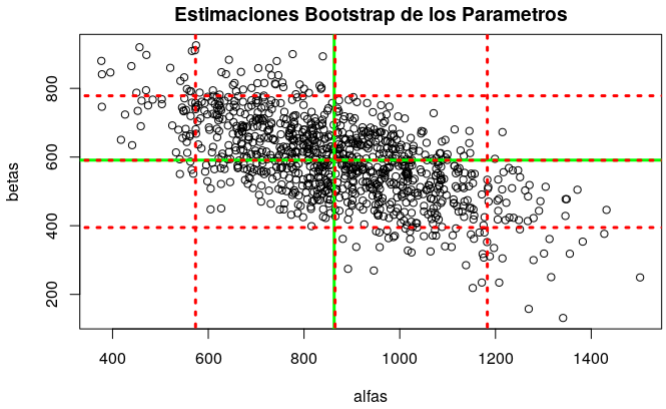
La pregunta que nos hacemos ahora es: si hubieramos ajustado el modelo con **TODAS** las casas del barrio de Boedo, ¿ los valores de los tres paramétros hubieran sido los mismos ? Seguramente no. A los valores de los parámetros que hubieramos obtenido de haber aplicado el ajuste a todas las casas de Boedo, los llamaremos *parámetros poblacionales*. ¿ Podemos aproximar, o estimar, que tan lejos están los valores de los parámetros estimados de los valores de los parámetros poblacionales ? La respuesta es sí. Pero esto ya no es tan facil de resolver.

La idea que está detrás de la incerteza de la estimación hecha por el ajuste del modelo, deriva, en gran parte, de la aleatoriedad con la que creemos se obtuvieron las 68 casas del total de casas existentes en Boedo. ¿ Como podemos considerar la incerteza que esta aleatoriedad produce ?

Una posibilidad muy util, es la de recrear sintéticamente (artificialmente) esa incerteza en una situación conocida por nosotros y parecida al problema original. La idea es pensar a la muestra de 68 casas como si fuera nuestra población conocida, y generar muestrs aleatorias con reposición de tamaño 68 de la muestra original. El supuesto inherente a esta heurística es el de pensar que la incerteza en la estimación de los parámetros que observaremos producto de nuestra aleatorización de la muestra, es equivalente a la producida por la aleatorización que generó a la muestra original a partir de la población de casas de Boedo. A esta heurística se la conoce como **Bootstrap**.

Apliquemos este procedimiento a nuestros datos. Generaremos $B = 1000$ muestras con reposición de la muestra original. Para cada muestra ajustaremos el modelo, y guardaremos las estimaciones de los parámetros. Cada estimación será (muy probablemente) distinta, pues provienen de ajustes hechos con muestras distintas.

En el siguiente gráfico vemos las $B = 1000$ estimaciones de los parámetros α y β.



Las estimaciones originales de ambos parámetros se representan con lineas en color verde, en tanto que los cuantiles 0.05, 0.50 y 0.95 de las estimaciones bbotstrap se representan con lineas de puntos rojas. Varias conclusiones pueden extraerse de este gráfico:

- Las medianas de las estimaciones bootstrap coinciden con las estimaciones originales.¿ Podría no ocurrir esto ?
- Existe una correlación negativa (-0.62) entre las estimaciones bootstrap de ambos parámetros.¿ Porqué pasa esto ?
- El 90% más central de las estimaciones bootstrap de α se hallan en el intervalo $(573, 1183)$. Considerando esto último, ¿ Cuán seguros estamos de que el precio se relaciona positivamente con esta variable ?
- El 90% más central de las estimaciones bootstrap de β se hallan en el intervalo $(395, 779)$ Considerando esto último y el punto anterior,. ¿ Cuán seguros estamos que la variable Fondo contribuye menos al precio que Sup ?

La capacidad que tiene un modelo de captar certeramente las relaciones existentes entre las variables, en nuestro caso, estimar bien los valores de los parámetros, es considerada como la **capacidad explicativa** de un modelo.

¿ Qué realación existe entre la capacidad predictiva de un modelo y su capacidad explicativa ?

Ambas capacidades de un modelo pueden ser muy distintas. Veamos dos ejemplos en los cuales estas características difieren esencialmente.

Ejemplo 1: Pensemos en un modelo que al ser entrenado estima perfectamente las relaciones entre el precio de las casas y la superficie cubierta y descubierta. O sea, supongamos que en nuestro caso, en Boedo, pudieramos contar con todas las casas y estimar así de forma muy confiable la relación entre el precio y la superficie cubierta (el parámetro α), y la relación entre el precio y la superficie no cubierta (el parámetro β). Aún así, si los precios de las casas dependieran de muchos otros atributos no considerados en el modelo (i.e. estado del inmueble o ubicación) es esperable que el error de predicción sea grande, pues casas con valores de Sup y Fondo similares podrían presentarjar situaciones muy distintas en términos de los atributos no considerados. Quiza, el "modelo verdadero" sea el siguiente:

$$\begin{aligned} Precio^{Verdadero} &= \mu + \alpha * Sup + \beta * Fondo + \rho * Estado + \tau * Ubicacion \\ &= \mu + \alpha * Sup + \beta * Fondo + Error \end{aligned}$$

Si en nuestro modelo sólo etnemos en cuenta los primeros dos factores, los atributos Estado (estado del inmueble) y Ubicación (ubicación del inmueble) se cuelan en las predicciones como un error de predicción. Cuanto más importante sea el efecto de estos atributos en el precio (ρ y τ) y más variabilidad tengan estos atributos, mayor será el error de predicción.

Ejemplo 2: Pensemos ahora en una situación hipotética en la cual los valores de los atributos Sup y Fondo de las casas en Boedo sea el mismo (aproximadamente), por ejemplo una casa podría tener $Sup = 251$ y $Fondo = 249$, en tanto que otra casa podría tener $Sup = 136$ y $Fondo = 138$. Supongamos también que los únicos atributos que afectan el precio de una casa son Sup y Fondo. Si los verdaderos valores de los parámetros (los poblacionales) son $\alpha = 900$ y $\beta = 500$,y por algún motivo (pocos datos o mala suerte en el muestreo de las casas) obtuvieramos estimaciones de parámetros $\widehat{\alpha} = 500$ y $\widehat{\beta} = 900$, las prediciones serían casi perfectas. Por ejemplo, para la primer casa mencionada tendríamos que el precio real sería $\alpha * 251 + \beta * 249 = 900 * 251 + 500 * 249 \approx 500 * 251 + 900 * 249 = \widehat{\alpha} * 251 + \widehat{\beta} * 249$. Sin embargo los efectos estarían pesimamente estimados, invirtiendo la importancia de los mismos.

Este ejmplo puede parecer muy forzado, pero cuando se consideran muchos atributos en un modelo, estas situaciones se vuelven más probables. Es decir, es más facil hallar conjuntos de predictoras fuertemente relacionadas que no dañen la capacidad predictiva, pero si la capacidad explicativa, a fuerza de confundir el efecto de un atributo con el de otro.

¿ Qué sucede si modelamos el Precio sólo con la variable Sup ?

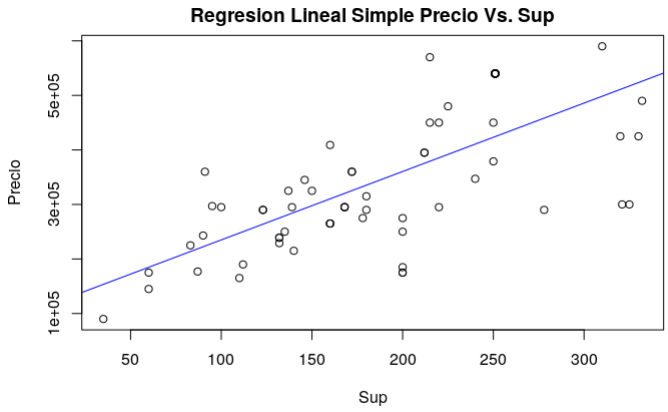
Supongamos que sólo nos interesa conocer el efecto de la superficie cubierta (Sup) en el precio de las casas. Parece razonable proponer un modelo más simple aún, con una única predictora, la variable Sup. La representación sería:

$$Precio^{M1} = \gamma + \delta * Sup$$

Tuve la precaución de darles nombres distintos a los parámetros, debido a que son modelos distintos. Las estimaciones de los parámetros de este modelo son:

$$\boxed{\widehat{\gamma} = 109601}$$
$$\boxed{\widehat{\delta} = 1254}$$

El gráfico del ajuste del nuevo modelo a los datos es



Se ve que, al igual que en el ajuste del modelo múltiple anterior, un aumento de Sup se relaciona con un incremento del Precio. Calculemos ahora el error de predicción ingenuo de este modelo

$$\begin{aligned} MAE &= \frac{1}{68}\sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M1} \right\| \\ &= 76741 \end{aligned}$$

$$PMAE = \frac{1}{\overline{Precio}} \frac{1}{68} \sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M1} \right\|$$
$$= \frac{76741}{347997} = 0.22$$

Vemos así que el error es mayor al del modelo múltiple. Esto se debe a que en la Regresión Lineal Múltiple, como en casi todas las técnicas de modelado, el agregado de predictores (relevantes) en el modelo mejora la evaluación (disminuye la pérdida). Este es un resultado general del modelado de datos:

Los modelos más complejos se ajustan mejor a los datos

El modelo inicial múltiple es más complejo que el último modelo simple propuesto. En el caso particular de la técnica de Regresión Lineal Múltiple con pérdida cuadrática puede demostrase que siempre el agregado de variables disminuye el error. Seguiremos esta discusión un poco más adelante.

Pasando ahora a analizar la relación del Precio con Sup en ambos modelos veremos que

$$\widehat{\alpha} = 863 < \widehat{\delta} = 1254$$

¿ Por qué son tan distintos ? Los datos son los mismos, así que la explicación tiene que recaer en el hecho de haber eliminado del modelo la variable Fondo. Recordemos que ambas predictoras Sup y Fondo se hallan positivamente correlacionadas ($cor(Sup, Fondo) = 0.47$), por lo que podríamos pensar en "representar" a la variable Fondo (la que eliminamos en este nuevo modelo más simple) en función de la variable Sup, de la siguiente manera

$$Fondo = a + b * Sup$$

Si esta relación fuera cierta, al modelo múltiple inicial lo podríamos plantear:

$$\begin{aligned} Precio^{M2} &= \mu + \alpha * Sup + \beta * Fondo = \\ &= \mu + \alpha * Sup + \beta * (a + b * Sup) = \\ &= (\mu + \beta * a) + (\alpha + \beta * b) * Sup = \\ &= \widetilde{a} + \widetilde{b} * Sup \end{aligned}$$

Así, cunado proponemos un modelo para Precio sólo dependiente de la variable Sup

$$Precio^{M1} = \gamma + \delta * Sup$$

deberíamos esperar que el parámetro δ que capta el efecto de la variable Sup "absorba" también la componente que le corresponde a la variable Fondo, pues $\delta = \widetilde{b} = \alpha + \beta * b$. Es decir que al efecto puro de Sup (medido por α) se le agrega una parte (b) del efecto que Fondo tiene con Precio (β).

Otra forma de entender esto es pensar que cada vez que el modelo simple "observa" casas con valores de Sup grandes, en cierta medida también está observado casas con Fondo grande, por lo que observa Precios más grandes que los que les corresponden sólo por el efecto Sup.

Entonces, ¿ Es incorrecto plantear el modelo más simple ? No, no lo es. De hecho podría ser un buen modelo predictivo, bastante sencillo. Pero si queremos enteder o explicar el efecto que la variable Sup genera en el precio, por sobre el efecto de (dejando constantes) las demás variable, entonces ese sí es un modelo incorrecto. Este modelo no permite aislar el efecto de Sup del efecto de Fondo. Este modelo confunde eses efectos.

¿ Siempre es mejor incluir más variables en el modelo ?

Pasemos ahora a proponer un modelo más complejo, con más variables que el original. Puesto que es razonable considerar a la ubicación como una variable relevante al problema, vamos a incluir las variables latitud (Lat) y longitud (Lon) como predictoras. El modelo que proponemos es:

$$Precio^{M4} = \beta_0 + \beta_{Sup} * Sup + \beta_{Fondo} * Fondo + \beta_{Lon} * Lon + \beta_{Lat} * Lat$$

Este nuevo modelo posee 5 parámetros (4 variables) a ajustar. Las estimaciones se muestran a continuación:

$\beta_0 = 226437200$
$\beta_{Sup} = 829$
$\beta_{Fondo} = 113$
$\beta_{Lon} = 2523830$
$\beta_{Lat} = 2276729$

Salvo por los coeficientes de Sup y Fondo, no es necesario interpretar el resto de los coeficientes, ya que no son muy interpretables ni interesantes. Lo que sí es interesante es observar, nuevamente, que el agregado de otras variables (Lon y Lat) modifican las estimaciones de los efectos de Sup y Fondo.

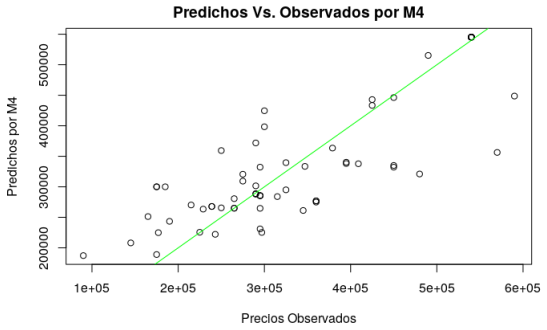
Los errores ingenuos de predicción del modelo son

$$
\begin{aligned}
MAE &= \frac{1}{68}\sum_{i=1}^{68}\left\|Precio_i - Precio_i^{M4}\right\| \\
&= 46397
\end{aligned}
$$

$$
\begin{aligned}
PMAE &= \frac{1}{\overline{Precio}}\frac{1}{68}\sum_{i=1}^{68}\left\|Precio_i - Precio_i^{M4}\right\| \\
&= \frac{46397}{347997} = 0.1333
\end{aligned}
$$

Este es el menor error ingenuo que tenemos hasta ahora. Pero que pasa con el error por validación cruzada para este modelo ? En este caso el MAE y PMAE por validación cruzada dan 49931 y 0.1435, respectivamente. Nótese que ambos mediciones son mayores a los ingenuos. Este modelo tiene los menores errores CV vistos hasta ahora.

Con tantas variables, ya no es factible graficar los precios en función de las predictoras. Una buena forma de visualizar el ajuste de un modelo es hacer un gráfico de dispersión entre los valores observados de la target, y los valores predichos según el modelo.



Otra medida de ajuste de un modelo entrenado es la proporción de variabilidad explicada por el mismo. La idea es simple. Empecemos por calcular la variabilidad de la target en relación a la prediccion del modelo más simple que podemos plantear. El modeo más simple es

$$Precio^{M0} = \lambda$$

Este es un modelo que no usa ni una sóla predictora. Se modela el precio sólo con una cosntante. Es facil ver que bajo la pérdida cuadrática, la estimación óptima del parámetro λ es $\widehat{\lambda} = \overline{Precio}$. El precio promedio.

Así la variabilidad de la target (o error cuadrático) en relación a la prediccion del modelo más simple es

$$\begin{aligned} SSTo &= \sum_{i=1}^{68} \left(Precio_i - Precio_i^{M0}\right)^2 \\ &= \sum_{i=1}^{68} \left(Precio_i - \overline{Precio}\right)^2 \end{aligned}$$

A esta magnitud se la llama suma de cuadrados totales. Calculemos ahora la variabilidad de la target, pero en relaciuón al modelo de interés que queremos evaluar ($Precio_i^{M4}$)

$$SSRes = \sum_{i=1}^{68} \left(Precio_i - Precio_i^{M4}\right)^2$$

A esta magnitud se la llama la suma de cuadrados de los residuos, pues a la discrepancia entre lo observado y lo predicho por el modelo se lo llama residuo.

Parece ahora razonable definir, como medida de bondad de ajuste de un modelo, a la proporción de reducción de variabilidad producida por el modelo de interés, en relación al modelo más simple, es decir:

$$R^2 = \frac{SSTo - SSRes}{SSTo} = 1 - \frac{\sum_{i=1}^{68} \left(Precio_i - Precio_i^{M4}\right)^2}{\sum_{i=1}^{68} \left(Precio_i - \overline{Precio}\right)^2} = 0.7351$$

A esta medida se la conoce como coeficiente de determinación o R^2. Dado que se satisface que $SSRes \leq SSTo$ (al menos en la Regresión Lineal Múltiple, y para la mayor parte de los métodos de predicción), se cumple que $0 \leq R^2 \leq 1$. En el caso de este modelo, el coeficiente arroja un valor de 0.7351, implicando que casi tres cuartas partes de la variabilidad total estaría siendo explicada por el modelo.

Obviamente, valores cercanos a uno(cero) implicarán ajustes muy buenos (malos).

Vamos por mas .....

Dado que hemos ido mejorando el ajuste cuanto más variables agregamos, ajustemos un último modelo, sumando efectos cuadráticos y cúbicos a la latitud y la longitud. Este agregado permitiría captar comportamientos no lineales en el espacio, por ejemplo si la zona más cara de Boedo se hallará en un círculo en el centro del barrio, el modelo anterior no podrá captarlo, pero este nuevo modelo sí. El nodelo así propuesto sería

$$\begin{aligned} Precio^{M8} &= \beta_0 + \beta_{Sup} * Sup + \beta_{Fondo} * Fondo + \beta_{Lon} * Lon + \beta_{Lat} * Lat + \\ &+ \beta_{Lon2} * Lon^2 + \beta_{Lat2} * Lat^2 + \\ &+ \beta_{Lon3} * Lon^3 + \beta_{Lat3} * Lat^3 \end{aligned}$$

Es muy importante notar que incluso con la inclusión de estos efectos no lineales, el modelo sigue siendo lineal ! ¿ Por qué ? Porque lo relevante en términos de entrenamiento de un modelo son los parámetros, no el comportamiento o transformaciones que le hagamos a las variables. Fijese que la expresión anterior es lineal en todos los betas.

Una vez entrenado el modelo, los errores ingenuos son

$$\begin{aligned} MAE &= \frac{1}{68} \sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M8} \right\| \\ &= 45137 \end{aligned}$$

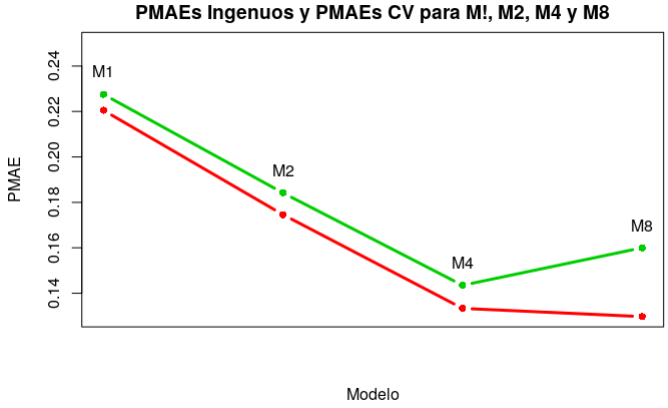
$$\begin{aligned} PMAE &= \frac{1}{\overline{Precio}} \frac{1}{68} \sum_{i=1}^{68} \left\| Precio_i - Precio_i^{M8} \right\| \\ &= \frac{46397}{347997} = 0.1297 \end{aligned}$$

Este es el menor error ingenuo que tenemos hasta ahora. Pero que pasa con el error por validación cruzada para este modelo ? En este caso el MAE y PMAE por validación cruzada dan 55684 y 0.16, respectivamente. Nótese que ambos mediciones no sólo son mayores a los ingenuos, sino que son mayores a los errores por validación cruzada del modelo $M4$. Lo que parece ser una mejora, en realidad no lo es.

¿ Cómo elegimos el nivel de complejidd óptima ?

A esta altura debiera quedar claro que no debieramos guiarnos por el error ingenuo. Vimos que en el último modelo el error ingenuo disminuyó, pero el error por CV aumentó.

El próximo gráfico nos muestra la evolución de los PMAEs (ingenuos en color rojo y por CV en color vede) de los cuatro modelos entrenados hasta el momento.



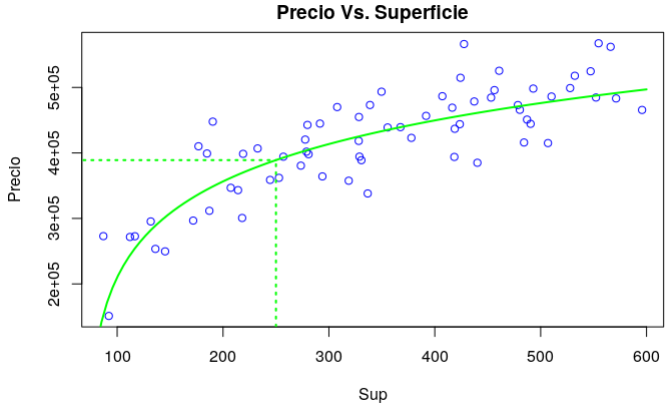
Varias características son destacables:

- El eje horizontal del gráfico puede pensarse como un eje de complejidad, ya que el mismo fue ordenado desde el modelo más simple (M1) hasta el más complejo (M8)
- Los errores ingenuos son siempre menores a los de validación cruzada
- Los errores ingenuos son siempre decrecientes con la complejidad del modelo
- Los errores por CV tienen un mínimo en un nivel de complejidad que NO es el máximo. En este caso el modelo M4.

El último punto refleja un concepto **FUNDAMENTAL** en el modelado de datos, pues muestra que modelos demasiado sencillos poseen mucho error, pero, modelos demasiado complejos pueden generar también errores grandes. Este fenómeno es conocido como **Trade-off Sesgo-Varianza**.

Trade-off Sesgo-Varianza

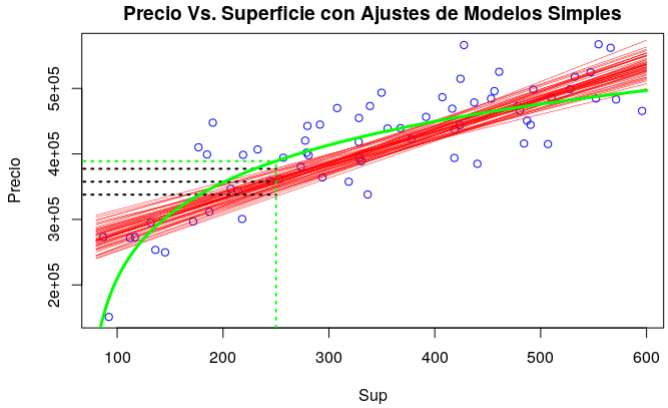
Para discutir este concepto, siguiendo con nuestro ejemplo de precios de casas en Boedo, vamos a considerar un modelo de Regresión Lineal que explique el precio (Precio) en función sólo de la superficie cubierta (Sup). Pero ahora, en lugar de trabajar con los 68 datos observados en la base, vamos a generar una muestra artificial de superficies y precios que siga una relación conocida por nosotros, y que supondremos no lineal. La idea de proponer una relación conocida entre precio y superficie es la de poder comparar las estimaciones que obtengamos con la verdadera relación, la propuesta por nosotros. La relación que supondremos será la que se muestra en color verde en el siguiente gráfico



En el gráfico también se observan en color azul unas 68 observaciones obtenidas aleatorizando valores de superficies entre 80 y 600, y asignandoles precios según la relación, pero sumando un ruido aleatorio, centrado en el precio predicho por la relación, con desvio de valor 50000. De no ser por el ruido agregado, los puntos azules caerían sobre la curva verde. El agregado del ruido simula la existencia de factores adicionales no contemplados en el modelo, que afectan al precio, más allá de la superficie cubierta. La justificación de una relación no lineal podría obedecer al hecho que un aumento de metros en casas chicas es de esperar sea mayor que en casas grandes. Marcamos en el gráfico (lineas verdes de puntos) un caso particular de interés de la predicción del precio de una casa de 250 metros cuadrados.

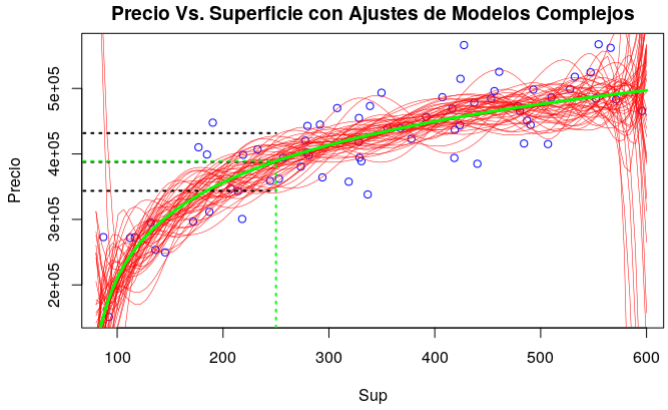
De ahora en más, para cada modelo que postulemos, realizaremos 50 entrenamientos del modelo, basado en 50 muestras de 68 observaciones provenientes de la relación propuesta (curva verde), pero generando ruidos aleatorios diferentes en cada simulación.

Empecemos proponiendo un modelo muy simple, en el que el precio dependa linealmente de la superficie.



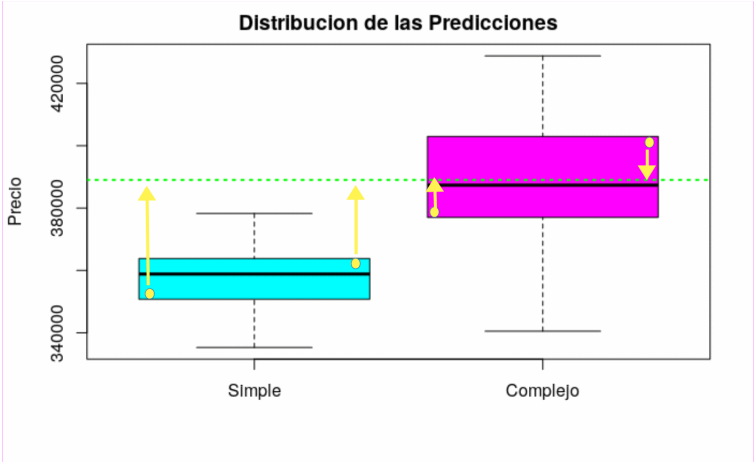
Se observa que todas las estimaciones sobreestiman el precio de casas chicas y grandes, subestimando las de tamaño intermedio. Este comportamiento se debe a la falta de flexibilidad del modelo, el que no puede adapterse a la no lienalidad del fenómeno. Al mismo tiempo se aprecia la relativamente baja variabilidad de las estimaciones (las rectas rojas están bastante pegadas). En particular, la predicción del precio de la casa de 250 metros cuadrados daría una clara subestimación. Esto puede verse claramente mediante el cálculo del promedio de las estimaciones, sumado y restado 2 desvios estandar (lineas negras puntedas horizontales).

Probemos ahora un modelo mucho más complejo, un polinomio de grado 12.

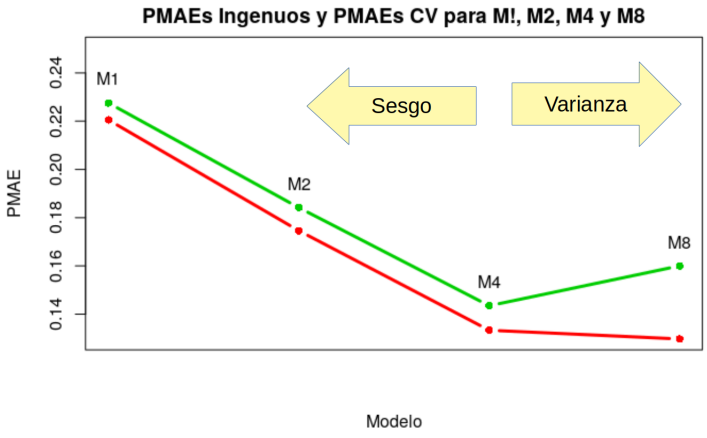


Se observa ahora un comportamiento inverso al anterior, con estimaciones muy variables, pero sin errores sistemáticos (sesgos) en el precio. En particular, la predicción del precio de la casa de 250 metros cuadrados sería muy variable. pero sin sesgo (error sistemático).

Podemos comparar directamente los dos conjuntos de estimaciones del precio de la casa de 250 metros, como se presenta en el siguiente gráfico.



En color cyan (magenta) podemos ver la distribución de las estimaciones del modelo simple (complejo). La linea horizontal verde de puntos representa el verdadero precio. Esta situación es una característica general del modelado de datos, y se conoce como **Trade-off Sesgo-Varianza**. Los modelos simples poseen mucho sesgo, pero poca varianza. En tanto que los modelos complejos poseen poco sesgo y mucha varianza. El error de predicción (como el PMAE) puede pensarse como la distancia promedio de cada estimaión al valor verdadero. En el caso del modelo simple, las distancias son grandes (flechas amarillas del boxplot de la izquierda) pues todas las estimaciones son sistemáticamente más bajas (**Sesgo**). En el caso del modelo más complejo, las distancias son grandes (flechas amarillas del boxplot de la derecha) porque se mueven mucho alrededor del verdadero valor (**Varianza**). Por ende, el error se comporta como una suma del Sesgo y de la Varianza. Eso explica la razón de los mayores errores de los modelos muy simples (M1) y muy complejos (M8) en la predicción del precio de la base de Properati, como se explicita en el próximo gráfico.



Por supuesto que si calculamos el error de forma ingenua, no notaremos el incremento del error al aumentar la complejidad del modelo. Sólo veremos un descenso sistemático del mismo, como en la curva roja del gráfico.