Universidad de San Andrés
Maestría en Inteligencia Artificial
Modelado Estocástico

Solución a la Ejercitación 4

El siguiente ejercicio consiste en la valuación de una casa según sus atributos en función a la base de datos que vimos en clase de Gary Koop. La misma tiene información sobre 546 casas vendidas en Windsor, Canadá. La variable dependiente (Y) es el precio de venta en dólares canadienses. Las variables explicativas (las x´s) son:

- LOTE: el tamaño del lote, medido en pies al cuadrado (*square feet*)
- CUARTOS: número de dormitorios
- BANOS: número de baños
- PISOS: número de pisos, excluido el sótano
- ENTRADA: Dummy = 1 si la casa tiene una entrada para el auto (*driveway*) (= 0 si no tiene)
- REC: Dummy = 1 si la casa tiene un cuarto de recreación (*rec room*) (= 0 si no tiene)
- SOTANO: Dummy = 1 si la casa tiene un sótano (*basement*) (= 0 si no tiene)
- CALEF: Dummy = 1 si la casa tiene calefacción central (gas) (= 0 si no tiene)
- AIRE: Dummy = 1 si la casa tiene aire acondicionado (*air cond*) (= 0 si no tiene)
- GARAGE: mide la cantidad de autos que entran en el *garage* (= 0 si no tiene)
- NBHD: Dummy = 1 si la casa está en un barrio agradable (= 0 si no lo está)
- DB2: Dummy = 1 si la casa tiene exactamente 2 baños (= 0 sino)
- DB3: Dummy = 1 si la casa tiene exactamente 3 baños (= 0 sino)

En este ejercicio las casas tienen 1, 2 ó 3 baños. En las regresiones que se presentan a continuación, se omitió una observación de la base de datos original que satisfacía que tenía 4 baños (la única), de modo que las observaciones usadas en las regresiones a continuación se basan en las restantes 545 observaciones de la base de datos con la que se trabajó en clase.

El output de la regresión es el siguiente:

REGRESION 1

Dependent Variable: PRECIO
Method: Least Squares

Sample: 1 545
Included observations: 545

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3509.739	3419.390	-1.026423	0.3052
LOTE	3.534066	0.349830	10.10225	0.0000
CUARTOS	1909.324	1046.422	1.824622	0.0686
BANOS	13829.49	1519.347	9.102255	0.0000
PISOS	6492.688	924.6698	7.021628	0.0000
ENTRADA	6722.788	2042.146	3.292022	0.0011
REC	4623.243	1898.206	2.435586	0.0152
SOTANO	5506.627	1585.876	3.472294	0.0006
CALEF	12995.33	3214.103	4.043221	0.0001
AIRE	12622.70	1552.591	8.130086	0.0000
GARAGE	4139.817	841.6709	4.918570	0.0000
NBHD	9469.961	1667.598	5.678803	0.0000

R-squared	0.664891	Mean dependent var	67925.49
Adjusted R-squared	0.657975	S.D. dependent var	26330.72
S.E. of regression	15398.96	Akaike info criterion	22.14376
Sum squared resid	1.26389E+11	Schwarz criterion	22.23846
Log likelihood	-6022.175	Hannan-Quinn criter.	22.18078
F-statistic	96.13890	Durbin-Watson stat	1.697694
Prob(F-statistic)	0.000000		

a) Interprete el coeficiente 4139.8 de la variable GARAGE.

Respuesta: La variable explicativa GARAGE toma los valores 0, 1, 2, 3. El valor predicho de una casa que tiene GARAGE será de CAN $ 4139.82 mayor por cada auto adicional que entre en el garage, *ceteris paribus*. Es decir, si comparamos una casa con garage para un auto versus otra sin garage, manteniendo todas las demás variables constantes, el precio predicho de la casa con garage para un auto será de CAN $ 4139.82 mayor a la que no tiene garage. También se puede decir que *ceteris paribus* una casa con garage para dos autos tendrá un valor superior (precio predicho mayor) en CAN $ 4139.82 en relación a una casa con garage para un solo auto.

b) ¿Para qué sirve el estadístico F (=96.14), cómo está calculado y a qué conclusión puede llegar con el mismo? Indique claramente de qué hipótesis nula y alternativa es este estadístico. Es decir, explique cuál sería el modelo bajo las hipótesis nula y alternativa.

Respuesta: Este estadístico se usa para testear la hipótesis nula de que todas las pendientes son iguales a cero simultáneamente contra la hipótesis alternativa de que al menos una de estas pendientes es distinta de cero. Se calcula como:

$$F = \frac{(RRSS - URSS)/q}{URSS/(n-k)}$$

Donde "RRSS" es la suma de cuadrados residual en el modelo restringido, que coincide solo en este caso con la suma de cuadrados totales de este modelo (ya que el modelo restringido sería PRECIO = constante + error), "URSS" es la suma de cuadrados residual del modelo no restringido, que es 1,26389 * 10^{11}, "q" es el número de restricciones bajo la hipótesis nula (en este caso, q=11), "n" es el número de observaciones (n=545) y "k" es el número de parámetros a estimar en el modelo lineal (constante más pendientes, que en este caso, k=12). Como se mencionó anteriormente, RRSS = TSS solo en este caso en particular, el cual se puede calcular a partir de R^2=1-RSS/TSS=0.664891, de modo que TSS = 1,26389* 10^{11} / (1-0.664891) = 3.7716 * 10^{11}. Este estadístico F tiene una distribución F con q grados de libertad en el numerador y n-k en el denominador. Es decir, 11 grados de libertad en el numerador y 533 en el denominador. Habría que mirar en la tabla cuál es el valor crítico, pero dado que en la salida de la regresión se informa el *p-value* asociado a este estadístico F (*p-value* = 0,0000), sabemos que con una probabilidad de error tipo I del 1% ó mayor vamos a rechazar la hipótesis nula.

c) ¿Puede haber evidencia de multicolinealidad en la Regresión 1? Justifique claramente.

Respuesta: para que haya multicolinealidad entre dos o más variables, los *p-values* deberían ser mayores a 0,10. Esto ocurre ya que cuando hay multicolinealidad, los errores estándar de las pendientes (de los beta sombrero) que multiplican a las variables explicativas son colineales entre sí, tendrán valores altos, repercutiendo en un estadístico t bajo y un *p-value* alto. Esto no ocurre en esta regresión, de manera que podemos descartar multicolinealidad. Pero si ese hubiera sido el caso, uno debe eliminar las variables involucradas de a una y fijarse si la(s) que queda(n) pasan a tener *p-values* bajos. Alternativamente uno puede correr un test de hipótesis de que los betas que multiplican a todas esas variables explicativas que parecieran no explicar a la variable dependiente son simultáneamente iguales a cero. Si rechazo esta hipótesis nula, entonces tengo un problema de multicolinealidad.

d) ¿Cuál es el valor esperado de una casa en un lote de 5100 pies cuadrados, con 3 cuartos, 2 baños, construida en dos pisos, con entrada para auto, cuarto de recreación, calefacción central en un vecindario agradable (NBHD=1), sin garage, ni sótano, ni aire acondicionado central? Muestre cómo obtiene ese valor.

Respuesta: armamos una tabla. Haciendo las cuentas, el valor predicho de la casa con esas características es de CAN $ 94697.65.

	Estimador	Variable	mador*Varia
C	-3509.739	1	-3509.739
LOTE	3.534066	5100	18023.7366
CUARTOS	1909.324	3	5727.972
BANOS	13829.49	2	27658.98
PISOS	6492.688	2	12985.376
ENTRADA	6722.788	1	6722.788
REC	4623.243	1	4623.243
SOTANO	5506.627	0	0
CALEF	12995.33	1	12995.33
AIRE	12622.7	0	0
GARAGE	4139.817	0	0
NBHD	9469.961	1	9469.961
Suma			94697.65

La columna de la derecha es la multiplicación de las dos columnas centrales.

REGRESION 2

A continuación, se presenta otra salida. Esta segunda regresión utiliza las mismas observaciones que la regresión 1. En esta regresión se elimina la variable BANOS y se incorporan las variables DB2 y DB3.

Dependent Variable: PRECIO
Method: Least Squares

Sample: 1 545
Included observations: 545

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10308.95	3464.573	2.975534	0.0031
LOTE	3.526753	0.350519	10.06152	0.0000
CUARTOS	1902.874	1047.336	1.816871	0.0698
PISOS	6527.316	928.9431	7.026604	0.0000
ENTRADA	6747.292	2044.525	3.300177	0.0010
REC	4623.607	1899.665	2.433906	0.0153
SOTANO	5521.098	1587.458	3.477948	0.0005
CALEF	13020.30	3217.108	4.047207	0.0001
AIRE	12696.44	1563.394	8.121072	0.0000
GARAGE	4141.753	842.3301	4.917019	0.0000
NBHD	9496.899	1670.077	5.686503	0.0000
DB2	13481.18	1726.361	7.809018	0.0000
DB3	29392.54	5079.824	5.786133	0.0000

R-squared	0.665005	Mean dependent var	67925.49
Adjusted R-squared	0.657449	S.D. dependent var	26330.72
S.E. of regression	15410.80	Akaike info criterion	22.14709
Sum squared resid	1.26346E+11	Schwarz criterion	22.24968
Log likelihood	-6022.082	Hannan-Quinn criter.	22.18719
F-statistic	88.00712	Durbin-Watson stat	1.700277
Prob(F-statistic)	0.000000		

e) ¿Cómo se interpreta el coeficiente 29392.54 correspondiente a la variable DB3? Indique claramente.

Respuesta: el valor predicho de una casa con 3 baños será de CAN $ 29392.54 mayor al de una casa con los mismos atributos (*ceteris paribus*) excepto que con un solo baño. Esto es así porque cuando DB3=1, la casa tiene exactamente 3 baños y cuando DB2 y DB3 son iguales a cero, la casa tiene exactamente un solo baño. Consecuentemente, la interpretación del beta sombrero que multiplica a DB3 es el diferencial del precio predicho de una casa con 3 baños contra una casa con 1 baño, mientras todas las otras variables están fijas en algún valor determinado.

f) Obtenga un intervalo de confianza del 95% para el beta que multiplica a la variable lote a partir de la información de la segunda regresión (REGRESION 2). Puede usar la distribución normal como aproximación. También puede responder usando algún software, en cuyo caso deberá indicar claramente qué hizo.

Respuesta: el intervalo de confianza del 95% se calcula a partir de

$$P\left(-1{,}96 < \frac{\hat{\beta} - \beta}{se(\hat{\beta})} < 1{,}96\right) = 0{,}95$$

Que nos da: $\hat{\beta} \pm 1.96 * se(\hat{\beta})$, de modo que el intervalo de confianza del 95% para el beta que premultiplica a LOTE es $3.536 \pm 1.96 * 0.35$, o sea, $\beta \in (2.85\,, 4.22)$.

Si el intervalo de confianza lo calculan usando algún software, los valores que obtendrán serán los exactos, ya que no estará usando aproximación normal ni redondeos. El valor exacto que obtendrán es $\beta \in (2.838\,,\ 4.215)$.

g) Suponga que le interesa testear con una probabilidad de error tipo I de 5% si el beta que premultiplica a DB3 es el doble al beta que premultiplica a DB2. Es decir, desea testear si $\beta_{DB3} = 2\beta_{DB2}$. ¿Puede realizar este test de hipótesis? ¿Tiene alguna intuición testear particularmente esta hipótesis, $\beta_{DB3} = 2\beta_{DB2}$? ¿Qué significa intuitivamente esta hipótesis nula, es decir, qué es lo que se pretende testear con este hipótesis?

Respuesta: se puede responder usando los RSS de la regresión 2 (que sería el URSS) y usando el RSS de la regresión 1 (que sería el RRSS). El RSS de la regresión 1 es el RSS del modelo restringido ya que BANOS = 1 + DB2 + 2*DB3, de modo que imponer la hipótesis nula, $\beta_{DB3} = 2\beta_{DB2}$, quedaría un modelo que al estimarlo nos va a dar el mismo RSS que el de la regresión 1. El valor del estadístico F (notar que q=1, n=545, k=13) se obtiene haciendo el cálculo del mismo. El estadístico F tiene entonces 1 grado de libertad en el numerador y 532 grados de libertad en el denominador y se calcula usando la fórmula

$$F = \frac{(RRSS - URSS)/q}{URSS/(n-k)}$$

El estadístico F da un valor de 0.1825 y el *p-value* asociado es 0.67 (este valor se obtiene si usan un software), y consecuentemente no rechazamos la hipótesis nula.

Stata: si ustedes quisieran correr esto mismo en Stata (si bien es recomendable hacer la cuenta a mano al menos una vez para entender exactamente cómo funciona el test F), deberían abrir la base de datos de Ejemplo Casa, eliminar la observación con 4 baños escribiendo en la ventana de comandos:

```
drop if banos==4
```

Luego, generan las *dummies* para dos y tres baños desde la ventana de comandos:

```
gen DB2=0
gen DB3=0
replace DB2=1 if bath==2
replace DB3=1 if bath==3
```

A continuación, corren la regresión con DB2 y DB3, sin la variable “baños”. Deberían obtener exactamente los mismos valores que en la regresión 2 de arriba. Después de haber corrido esta regresión, escriben en la ventana de comandos:

```
test (DB3=2*DB2)
```

Stata les va a reportar el estadístico F asociado a esta hipótesis nula y el p-value asociado a este estadístico.