Modelos Probabilísticos y Análisis Estadístico
Modelos de Regresión Lineal

Mauricio Mejía Castro

11 de mayo de 2021

1. Descripción del conjunto de datos

Debido a la naturaleza del problema, se optó por la exclusión de algunas variables que no aportan a la predicción del precio: `id` y `zipcode`. El *dataframe* de Los siguientes comandos transforman la variable date a tipo fecha e indican a R que la variable `waterfornt` debe ser considerada tratada categórica:

```
king.test$date <- as.Date(substr(king.test$date, 0, 10))
king.vali$date <- as.Date(substr(king.vali$date, 0, 10))
king.test$waterfront <- factor(king.test$waterfront)
king.vali$waterfront <- factor(king.vali$waterfront)
```

Con estos elementos podemos plantear un modelo inicial:

```
lmstart <- lm(price ~ ., data = king.test)
summary(lmstart)
```

Con lo que obtenemos el siguiente resumen:

```
Call:
lm(formula = price ~ ., data = king.test)

Residuals:
Min       1Q   Median       3Q      Max
-1248356   -99179    -9043    75936  4147237

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.858e+07  1.768e+06 -21.827  < 2e-16 ***
date           1.032e+02  1.347e+01   7.662 1.93e-14 ***
bedrooms      -3.183e+04  2.073e+03 -15.358  < 2e-16 ***
bathrooms      3.495e+04  3.598e+03   9.715  < 2e-16 ***
sqft_living    1.486e+02  4.867e+00  30.531  < 2e-16 ***
sqft_lot       1.362e-01  5.332e-02   2.555   0.0106 *
```

```
floors         8.131e+02  3.971e+03   0.205   0.8377
waterfront1    5.909e+05  2.007e+04  29.437  < 2e-16 ***
view           4.883e+04  2.365e+03  20.645  < 2e-16 ***
condition      3.342e+04  2.613e+03  12.787  < 2e-16 ***
grade          9.727e+04  2.393e+03  40.650  < 2e-16 ***
sqft_above     2.845e+01  4.851e+00   5.864 4.60e-09 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.339e+03  8.013e+01 -29.188  < 2e-16 ***
yr_renovated   2.321e+01  4.072e+00   5.700 1.22e-08 ***
lat            5.638e+05  1.163e+04  48.494  < 2e-16 ***
long          -1.143e+05  1.317e+04  -8.679  < 2e-16 ***
sqft_living15  3.144e+01  3.805e+00   8.264  < 2e-16 ***
sqft_lot15    -3.953e-01  8.214e-02  -4.812 1.50e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199800 on 17271 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.697,Adjusted R-squared:  0.6967
F-statistic:  2337 on 17 and 17271 DF,  p-value: < 2.2e-16
```

Inferimos a partir de la prueba de hipótesis (p-value¡2.2e-16) que existe al menos una variable que explica el precio de las casas. Se ve también que la prueba de hipótesis individual para la variable `floors`, cae en la región de rechazo. Por otro lado, la variable `sqft_basement` no genera resultado alguno. Se excluirán estas dos variables y se recalculará nuevamente el modelo.

2. Planteamiento del modelo de regresión

Al ejecutar los siguientes comandos:

```
lmhouses.1 <- lm(price ~ ., data = king.test[, c(-7, -13)])
summary(lmhouses.1)
```

Se obtiene el siguiente resultado

```
lm(formula = price ~ ., data = king.test[, c(-7, -13)])

Residuals:
Min       1Q   Median       3Q      Max
-1248494   -99231    -9034    75933  4145675

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.864e+07  1.741e+06 -22.191  < 2e-16 ***
date           1.031e+02  1.347e+01   7.660 1.96e-14 ***
```

```
bedrooms      -3.185e+04  2.072e+03 -15.369  < 2e-16 ***
bathrooms      3.514e+04  3.477e+03  10.106  < 2e-16 ***
sqft_living    1.483e+02  4.645e+00  31.924  < 2e-16 ***
sqft_lot       1.359e-01  5.330e-02   2.550   0.0108 *
waterfront1    5.910e+05  2.007e+04  29.439  < 2e-16 ***
view           4.885e+04  2.364e+03  20.669  < 2e-16 ***
condition      3.338e+04  2.607e+03  12.803  < 2e-16 ***
grade          9.731e+04  2.385e+03  40.809  < 2e-16 ***
sqft_above     2.889e+01  4.351e+00   6.638 3.26e-11 ***
yr_built      -2.335e+03  7.842e+01 -29.778  < 2e-16 ***
yr_renovated   2.325e+01  4.067e+00   5.717 1.10e-08 ***
lat            5.640e+05  1.156e+04  48.793  < 2e-16 ***
long          -1.146e+05  1.304e+04  -8.788  < 2e-16 ***
sqft_living15  3.134e+01  3.775e+00   8.304  < 2e-16 ***
sqft_lot15    -3.958e-01  8.209e-02  -4.822 1.44e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199800 on 17272 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.697,Adjusted R-squared:  0.6968
F-statistic:  2484 on 16 and 17272 DF,  p-value: < 2.2e-16
```

A pesar del bajo coeficiente de determinación, el modelo parece exhibir ciertas propiedades deseables. Sin embargo, al graficar los valores predichos contra los residuales, no veremos un comportamiento de homocedasticidad.

Figura 1: Heterocedasticiad del modelo price ~

Para solucionar esto, utilizaremos la transformación Box-Cox. Con los siguientes comandos, hallaremos la grafica y calcularemos el lambda máximo (-0.02) al que elevaremos la variable predictora:

```
trhouses <- boxcox(price ~ ., data = king.train[, c(-1, -7, -8, -13)])
max.lambda <- trhouses$x[which.max(trhouses$y)]
lmhouses.tr <- lm(price^max.lambda ~ .,
  data = king.train[, c(-1, -7, -8, -13)])
summary(lmhouses.tr)
```

Figura 2: Gráfico Box Cox

Esto nos arroja el siguiente resultado

```
Call:
```

```
lm(formula = price^max.lambda ~ ., data = king.train[, c(-1,
-7, -8, -13)])

Residuals:
Min          1Q      Median          3Q         Max
-0.0186987 -0.0025094 -0.0000394  0.0025309  0.0219730

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.882e+00  3.430e-02  54.884  < 2e-16 ***
bedrooms       1.820e-04  4.112e-05   4.425 9.72e-06 ***
bathrooms     -1.280e-03  6.910e-05 -18.520  < 2e-16 ***
sqft_living   -1.937e-06  9.231e-08 -20.988  < 2e-16 ***
sqft_lot      -8.165e-09  1.059e-09  -7.710 1.33e-14 ***
view          -1.144e-03  4.386e-05 -26.072  < 2e-16 ***
condition     -1.029e-03  5.174e-05 -19.881  < 2e-16 ***
grade         -2.506e-03  4.737e-05 -52.896  < 2e-16 ***
sqft_above    -3.880e-07  8.643e-08  -4.489 7.19e-06 ***
yr_built       4.471e-05  1.558e-06  28.690  < 2e-16 ***
yr_renovated  -7.185e-07  8.067e-08  -8.906  < 2e-16 ***
lat           -2.150e-02  2.296e-04 -93.606  < 2e-16 ***
long           1.204e-03  2.592e-04   4.644 3.45e-06 ***
sqft_living15 -1.478e-06  7.498e-08 -19.706  < 2e-16 ***
sqft_lot15     6.279e-09  1.631e-09   3.849 0.000119 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.00397 on 17274 degrees of freedom
(1 observation deleted due to missingness)
Multiple R-squared:  0.763,Adjusted R-squared:  0.7628
F-statistic:  3973 on 14 and 17274 DF,  p-value: < 2.2e-16
```

En este punto, no solo vemos un importante descenso en el Error Residual Estándar, también vemos una leve mejora en el R^2 y el R^2 ajustado. Veremos también que queda corregida la heterocedasticidad del modelo anterior.

Figura 3: Valores residuales contra valores predichos después de la transformación de Box Cox

3. Diagnóstico del modelo

A continuación validaremos los supuestos que el modelo debe cumplir. Veremos los valores atípicos, los valores con alto *leverage* y los valores influyentes.

Además, nos aseguraremos que la estructura del modelo se atenga a un comportamiento lineal.

3.1. Normalidad de los valores residuales

Podemos evaluar visualmente la normalidad de los valores residuales a través del Q-Q Plot. Este gráfico valida el primer supuesto acerca de la normalidad de los valores residuales.

Figura 4: Gráfico Q-Q Plot

3.2. Valores con alto *leverage*

Para identificar los valores con alto *leverage* de manera visual, utilizamos el siguiente comando:

```
infhouses <- influence(lmhouses.tr)
halfnorm(infhouses$hat, labs = row.names(lmhouses.tr))
```

Figura 5: Gráfico para la distancia de Cook

3.3. Valores influyentes

Con la distancia de Cook, podemos identificar también que en el modelo existen mútliples valores influyentes:

Figura 6: Gráfico para la distancia de Cook

Finalmente, la capacidad predictiva del modelo a través de la Raíz del Cuadrado Medio del Error RMSE = 664255.