Proyecto de Análisis de Datos

Docentes a cargo: Dra. María Gabriela Palacio – Mgtr. Sergio Martín Buzzi

Problema 2

Los datos del archivo base_regresion.xls corresponden a 5396 alojamientos alquilados por medio de la plataforma Airbnb en la Ciudad de Buenos Aires durante el día 25/11/20151. La base citada contiene información sobre el identificador del alojamiento (id), el precio de una noche de alojamiento medido en dólares (precio), el tipo de alojamiento (tipo), la cantidad de valoraciones que ha recibido el alojamiento (valoraciones), el puntaje promedio de un máximo de 5 que recibió el alojamiento (puntaje), la cantidad de personas que puede recibir el alojamiento (personas), la cantidad de dormitorios (dormitorios), la cantidad de baños (baños), la estadía mínima (estadía), la distancia a la estación de trenes más cercana (distancia) y la cantidad de dependencias culturales que se encuentran a menos de 300 metros a la redonda (dependencias).

El objetivo general del problema es ajustar un modelo de regresión para explicar precio en dólares pagado por la estadía por una noche (precio).

a) Realice un análisis descriptivo de las variables utilizando medidas resumen y gráficos.

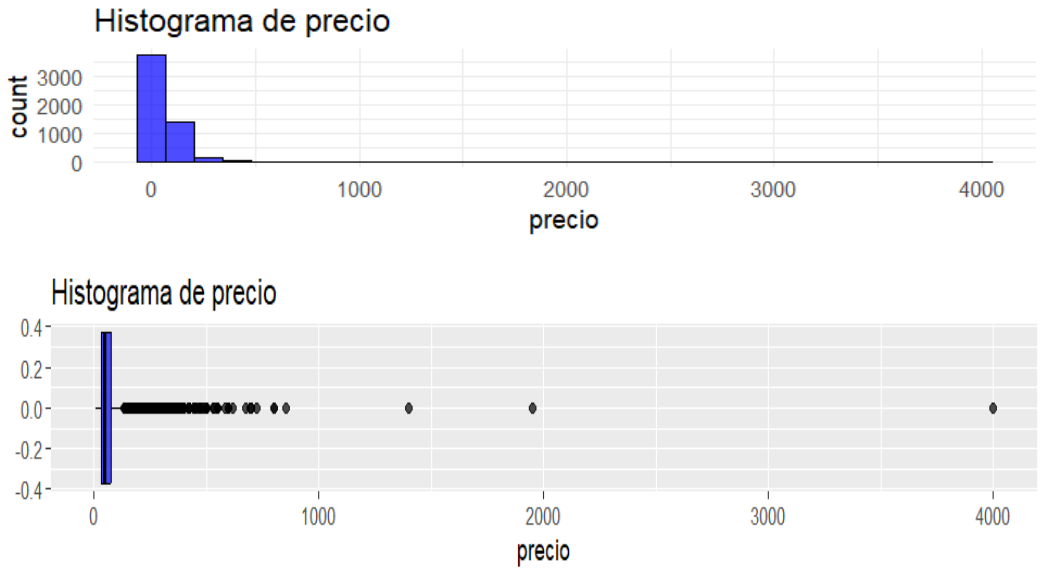
A continuación, se presenta en la tabla 1 las medidas descriptivas de la variable dependiente (precio) e independientes, a considerar en el análisis de regresión.

Tabla 1. Medidas descriptivas

Variable	Media	D.E.	CV	Mín	Máx	Q1	Mediana	Q3	Asimetría
precio	70,43	92,66	131,57	10	4000	34	50	75	17,91
valoraciones	7,81	16,88	216,31	0	180	0	1	7	4,04
puntaje	2,54	2,33	91,99	0	5	0	4	5	-0,13
personas	2,43	1,28	52,55	0	6	2	2	3	0,09
dormitorios	1,23	0,88	71,56	0	10	1	1	1	2,23
baños	1,07	0,8	75,12	0	8	1	1	1	2,16
estadía	3,07	2,73	88,76	0	27	1	3	4	2,74
distancia	3591231600	3028751733	84,34	1142094	9999061008	1259525917	1776957466	6271459342	0,73
dependencias	13,31	14,39	108,1	0	97	4	8	18	2,13

Acá no coincidimos con el Pablo.

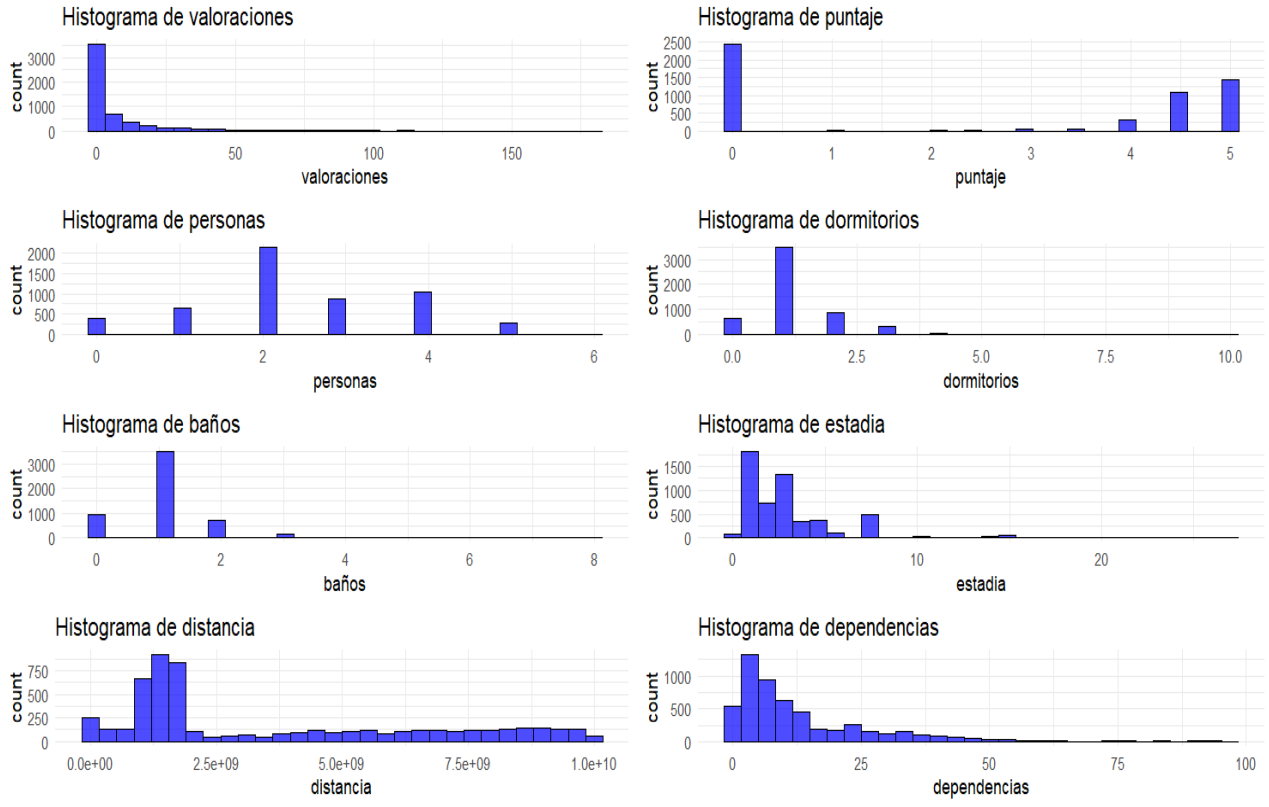
Gráfico 1.

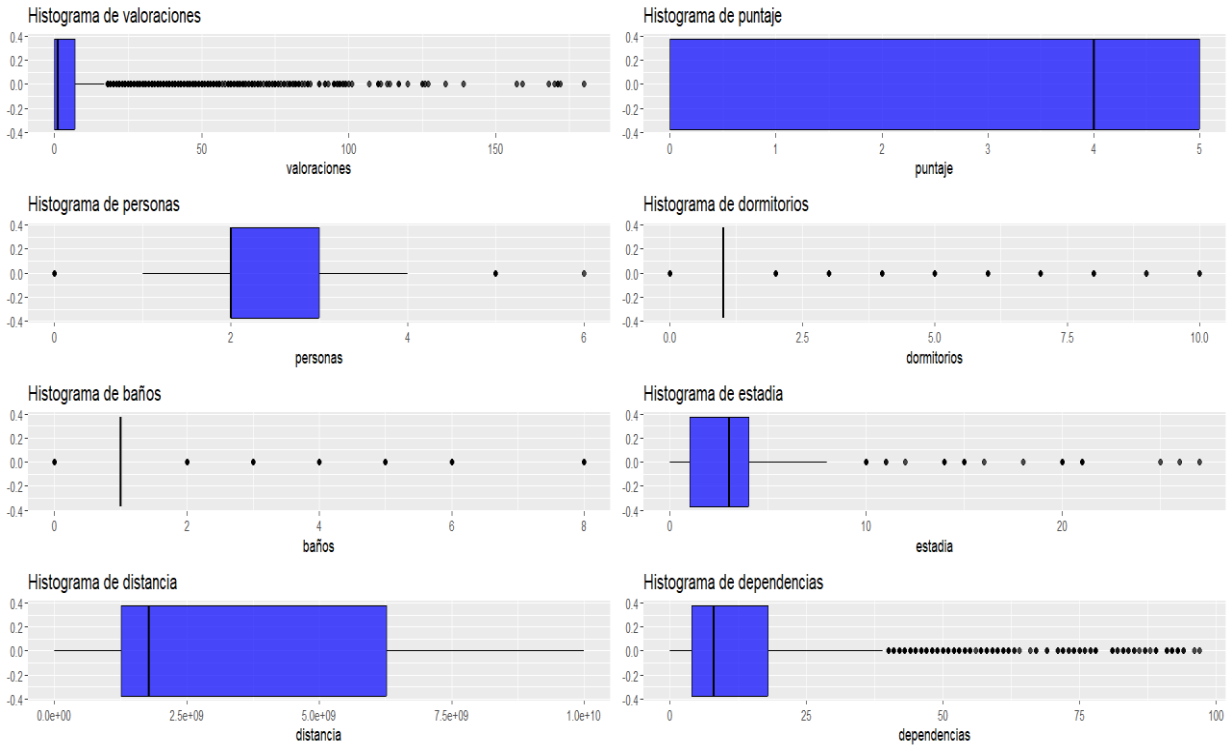


Es de interés observar que la media (70,43) y la mediana (50) de la variable dependiente precio no se encuentran próximas, como así también su coeficiente de variación es del 131% aproximadamente, confirmando a través de la gráfica también una fuerte asimetría derecha.

Lo mismo se puede observar con las variables independientes valoraciones y dependencias, cuyas medias (7,81; 13,31) y medianas (1; 8) respectivamente también se encuentran alejadas.

Gráfico 2.



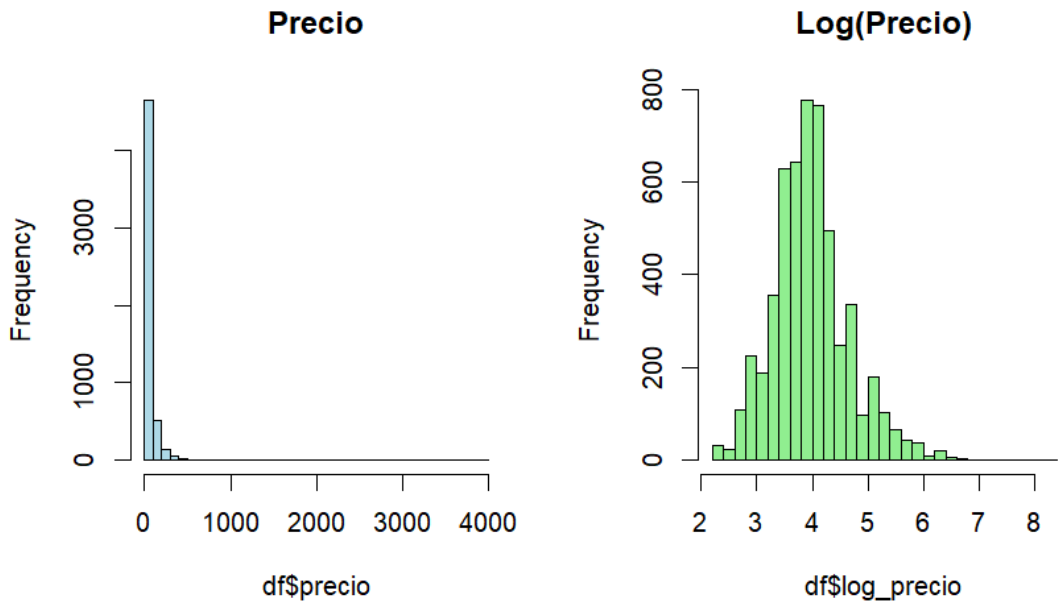


Finalmente, en las variables dormitorios, baños y estadías, el 50% de los alojamientos, cuenta con hasta 1 dormitorio, 1 baño y le reservan hasta 3 días de estadía. Estas 3 variables si bien presentan sus medidas de tendencia central similares, presentan también asimetría derecha, dado que entre los alojamientos analizados, algunos tienen un máximo de 10 dormitorios, 8 baños y 27 días de estadía, tal como se aprecia en el gráfico 2.

b) De acuerdo con lo observado en el análisis descriptivo de la variable precio, ¿considera conveniente transformarla tomando logaritmo natural para especificar el modelo de regresión?

Bajo el supuesto de normalidad de los residuos, como consecuencia, la variable dependiente Y tiene distribución normal. Sin embargo, se puede apreciar una fuerte asimetría derecha de la variable precio, por lo que se considera conveniente la transformación de la misma.

A continuación, se puede observar que la variable log_precio presenta mayor simetría.



c) Estime un modelo de regresión lineal sin considerar las variables tipo, distancia y dependencias.

Tabla 2. Modelo de regresión

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
LN_precio	5396	0,28	0,28	0,34	9486,91	9539,66

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	3,36	0,02	3,31	3,41	134,99	<0,0001		
valoraciones	-4,10E-04	5,20E-04	-1,40E-03	6,10E-04	-0,78	0,433	5,61	1,24
puntaje	-0,02	3,80E-03	-0,03	-0,01	-4,99	<0,0001	29,88	1,23
personas	0,1	0,01	0,09	0,11	16,17	<0,0001	266,43	1,02
dormitorios	0,39	0,01	0,37	0,41	41,53	<0,0001	1729,61	1,07
baños	-0,04	0,01	-0,06	-0,02	-3,58	0,0003	17,84	1,07
estadia	-0,01	2,90E-03	-0,01	-3,00E-03	-2,98	0,0029	13,87	1,02

1) ¿Son significativas todas las variables explicativas incluidas?

   Considerando un nivel de significancia del 5%, la única variable que no resulta significativamente distinta de cero es “valoraciones”.

2) Interprete los coeficientes estimados.
   La interpretación de los coeficientes estimados es la siguiente:

   - Al aumentar en 1 el puntaje promedio el precio tiende a disminuir en promedio un 0,02%.
   - Al aumentar en 1 la cantidad de personas que puede recibir el alojamiento el precio tiende a aumentar en promedio un 0,1%.

- Al aumentar en 1 la cantidad de dormitorios que dispone el alojamiento el precio tiende a aumentar un 0,39% en promedio.
- Al aumentar en 1 la cantidad de baños disponibles en el alojamiento el precio tiende a disminuir un 0,04% en promedio.
- Al aumentar en 1 día la estadía mínima requerida el precio tiende a disminuir en 0,01% en promedio.

3) ¿El modelo tiene buen poder explicativo? ¿Cómo lo mide?

El poder explicativo del modelo es bajo y es del 0,28 según el R2 ajustado.

4) ¿Se cumplen los supuestos en los que se basa el modelo? Justifique adecuadamente.

Normalidad de los errores: se analiza mediante el análisis de los residuos estandarizados utilizando un QQ plot normal y el Test de Kolmogorov-Smirnov.

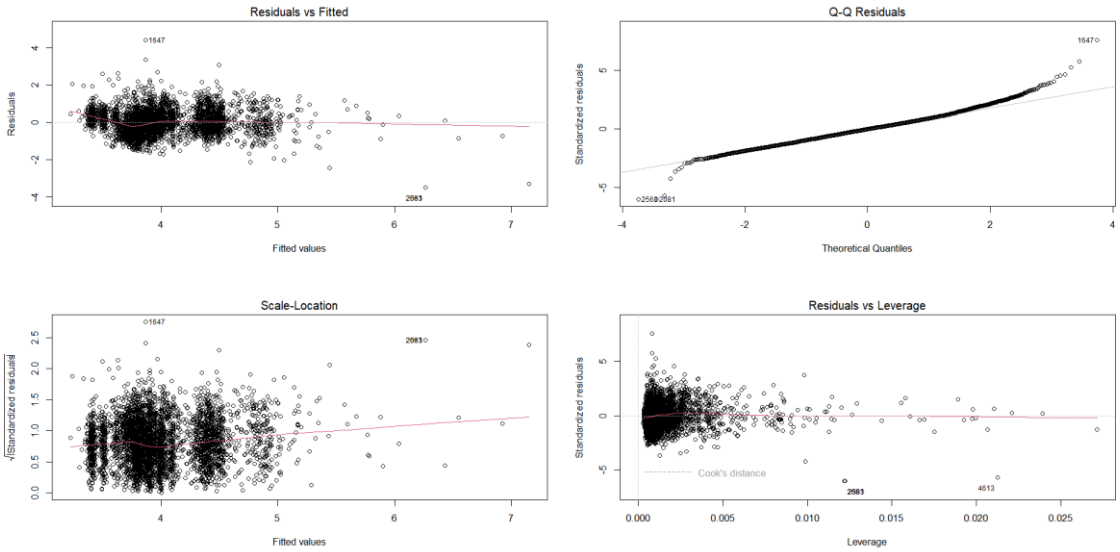
• Hipótesis nula (H0): Los residuos tienen una distribución normal.

• Hipótesis alternativa (Ha): Los residuos no tienen una distribución normal.

El test arroja el siguiente resultado

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuo
D = 0.030564, p-value = 8.37e-05
alternative hypothesis: two-sided
```

Dado que el p-valor es inferior al nivel de significancia, se rechaza la hipótesis nula de normalidad de los residuos. Lo mismo, si se observa el gráfico QQplot, donde se observa que no ajusta bien normalidad en los extremos.



Homocedasticidad (varianza de residuos constantes) de los errores: se analiza mediante el test de Breusch-Pagan, mediante el grafico de dispersión entre residuos estandarizados vs predichos por el modelo. Como así también, a través del análisis de la regresión de los valores absolutos de los residuos con respecto a las regresoras.

Las hipótesis del modelo son:

• Hipótesis nula (H0): La varianza de los residuos es constante.

• Hipótesis alternativa (Ha): La varianza de los residuos no es constante.

El test arroja el siguiente resultado

```
        studentized Breusch-Pagan test

data:  modelo1
BP = 437.61, df = 6, p-value < 2.2e-16
```

Dado que el p-valor es inferior al nivel de significancia, se rechaza la hipótesis nula de homocedasticidad.

Por otro lado, al realizar la regresión entre el valor absoluto de los residuos con las variables regresoras consideradas, todas resultan ser significativas, excepto la variable baños.

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
abs_residuos	5396	0,1	0,1	0,13	4160,36	4213,1

Coeficientes de regresión y estadísticos asociados								
Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	0,55	0,02	0,52	0,58	35,92	<0,0001		
valoraciones	-1,70E-03	3,20E-04	-2,30E-03	-1,10E-03	-5,29	<0,0001	33,03	1,24
puntaje	-0,01	2,30E-03	-0,01	-3,30E-03	-3,39	0,0007	16,5	1,23
personas	-0,06	3,80E-03	-0,07	-0,05	-16,28	<0,0001	269,94	1,02
dormitorios	0,08	0,01	0,07	0,09	14,73	<0,0001	221,94	1,07
baños	-0,01	0,01	-0,02	0,01	-0,82	0,4141	5,67	1,07
estadia	-4,30E-03	1,80E-03	-0,01	-7,90E-04	-2,4	0,0163	10,77	1,02

Por tanto, no cumple con ninguno de los supuestos del modelo de regresión lineal.

d) Mejore la especificación del modelo agregando o quitando variables, sin considerar la variable tipo.

1) ¿Qué criterio utilizó para la selección de variables?

A continuación, se listan algunos tests y herramientas utilizadas para la evaluación de los modelos:

- Prueba de significancia de la regresión: se analiza el estadístico F obtenido en el Análisis de Varianza.
- Contraste de significatividad individual (prueba t).
- Análisis de los residuos parciales: mediante los gráficos de dispersión que nos muestran la relación entre la variable dependiente y las regresoras.
- Multicolinealidad: mediante los coeficientes de correlación, VIF y CP-Mallows.

Acá no coincidimos con el Pablo.

El primer modelo considerado fue con todas las variables regresoras propuestas en el ejercicio, excepto la variable tipo. Del mismo se obtuvo lo siguiente:

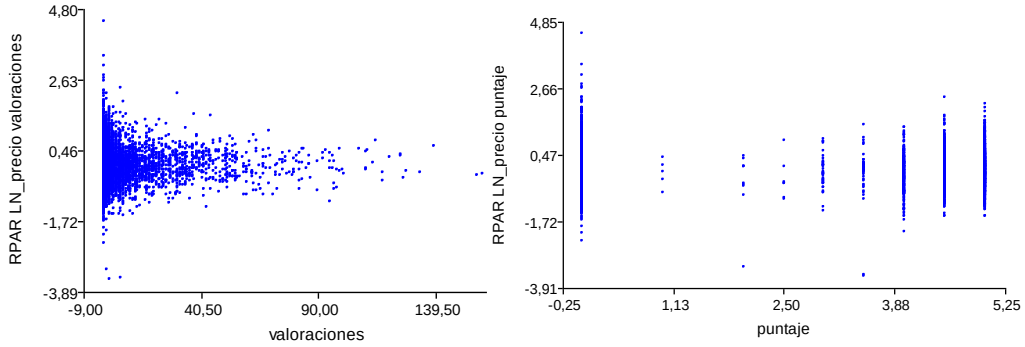
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
LN_precio	5396	0,28	0,28	0,34	9485,82	9551,75

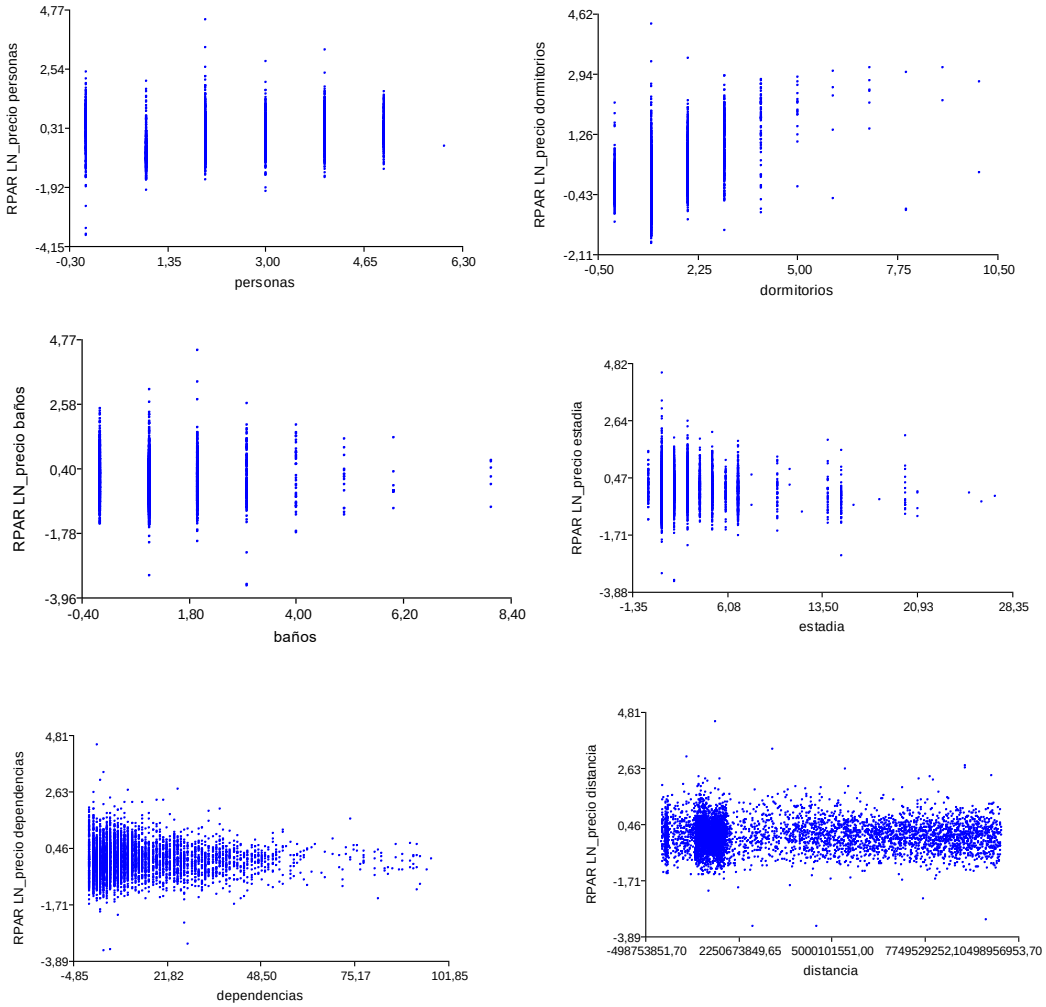
Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	3,39	0,03	3,34	3,44	123,21	<0,0001		
valoraciones	-4,40E-04	5,20E-04	-1,50E-03	5,80E-04	-0,85	0,3943	7,73	1,24
puntaje	-0,02	3,80E-03	-0,03	-0,01	-4,99	<0,0001	31,93	1,23
personas	0,1	0,01	0,09	0,11	16,25	<0,0001	270,97	1,03
dormitorios	0,39	0,01	0,37	0,4	41,4	<0,0001	1720,87	1,07
baños	-0,04	0,01	-0,06	-0,02	-3,53	0,0004	19,47	1,07
estadia	-0,01	2,90E-03	-0,01	-3,10E-03	-3,01	0,0026	16,06	1,02
distancia	-4,50E-12	2,60E-12	-9,60E-12	0	-1,72	0,0862	9,94	1
dependencias	-8,10E-04	5,50E-04	-1,90E-03	2,80E-04	-1,46	0,144	9,14	1,01

Teniendo en cuenta los criterios mencionados, dado que el p-valor del Test F obtenido del Análisis de la Varianza es menor a 0,05, se rechaza la hipótesis nula de que todos los parámetros son iguales a cero y se considera que al menos uno es significativamente distinto de cero.

En cuanto a la significatividad individual de cada una de las variables regresoras, se obtiene que “valoraciones”, “dependencias” y “distancia” no son significativamente distintas de cero.

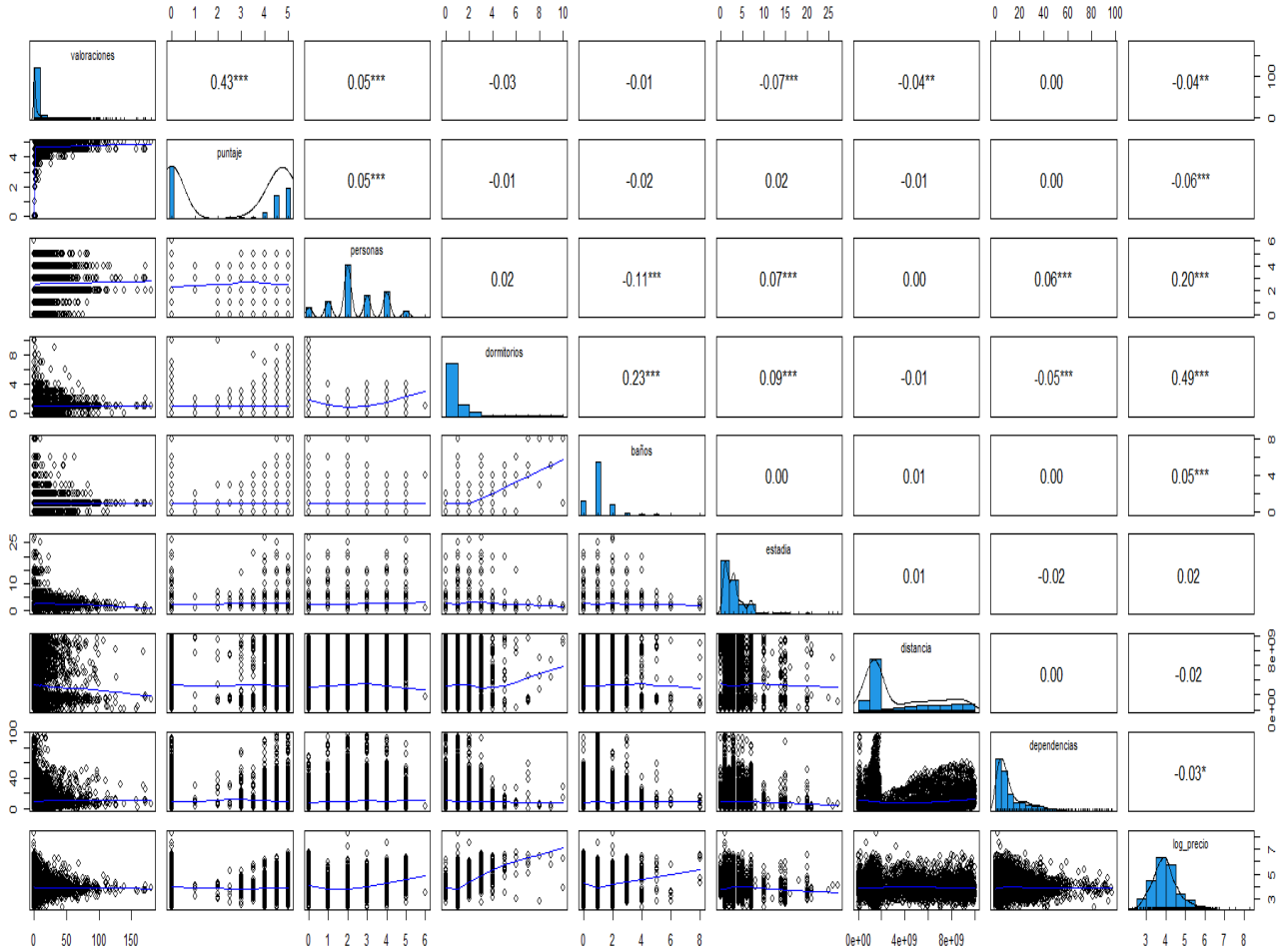
En cuanto a la interpretación gráfica de los residuos parciales de la regresión, no se observa relación lineal clara entre algunas predictoras y la variable de respuesta, principalmente con: valoraciones, puntaje, estadía y dependencia.





Se observa que la que mayor contribución tiene a la variable regresada precio es “dormitorio”, que a su vez es la que mayor coeficiente de correlación de Pearson presenta.

En cuanto a multicolinealidad, es raro porque con VIF se ven todo bonito pero con CP-Mallows se ve muy mal puntaje, personas, dormitorios, baños y estadía, es bastante lógico entre personas dormitorio y baños, pero al observar los coeficientes de correlación, no hay correlaciones fuertes.



Teniendo en cuenta todo esto y realizando

Se probaron los tres métodos de selección de modelo:

• En el método backward, forward y stepwise, las variables que fueron significativas fueron: puntaje, personas, dormitorios, baños, estadía.

• Mientras que en el método a través del cual se prioriza la minimización del AIC, los modelos que resultaron con menor AIC fueron:

1. Primer modelo: puntaje, personas, dormitorios, baños, estadía, distancia y dependencias.

2. Segundo modelo: puntaje, personas, dormitorios, baños, estadía y distancia.

Sin embargo, dado que la variable dependencia y distancia resultaron no ser significativa se considera el modelo que considera las cinco variables (puntaje, personas, dormitorios, baños, estadía) y que resulta ser el cuarto con menor AIC.

Pero además, se incluyó el cuadrado de las variables personas, puntaje y estadía. Por lo que el modelo final resulta:

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
LN_precio	5396	0,31	0,3	0,33	9299,34	9365,27

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	3,68	0,04	3,6	3,76	88,24	<0,0001		
dormitorios	0,31	0,01	0,28	0,33	25,94	<0,0001	679,93	1,76
POT_personas	0,06	0,01	0,05	0,07	11,14	<0,0001	131,18	19,48
personas	-0,19	0,03	-0,24	-0,13	-7,02	<0,0001	56,26	18,73
puntaje	-0,22	0,03	-0,28	-0,17	-7,69	<0,0001	66,07	76,54
POT_puntaje	0,04	0,01	0,03	0,05	6,95	<0,0001	55,26	76,47
baños	-0,04	0,01	-0,06	-0,02	-3,91	0,0001	22,27	1,08
estadia	0,02	0,01	0,01	0,04	3,78	0,0002	21,31	5,1
POT_estadia	-2,30E-03	4,20E-04	-3,10E-03	-1,50E-03	-5,44	<0,0001	36,62	5,07

2) Verifique el cumplimiento de los supuestos.

Test de normalidad de los residuos

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuo
D = 0.028238, p-value = 0.0003663
alternative hypothesis: two-sided
```

Se rechaza hipótesis nula de normalidad

Test de homocedasticidad

```
        studentized Breusch-Pagan test

data:  modelo6
BP = 378.37, df = 8, p-value < 2.2e-16
```

Se rechaza la hipótesis de varianza homogénea de los residuos

3) Compare el poder explicativo del nuevo modelo con el estimado en el inciso c. Justifique adecuadamente.

El poder explicativo del modelo aumenta de 0,28 a 0,30.

4) Interprete los coeficientes del nuevo modelo

La interpretación de los coeficientes estimados es la siguiente:

- Al aumentar en 1 el puntaje, el precio tiende a disminuir en promedio un 0,22%.
- Al aumentar en 1 la cantidad de personas que puede recibir el alojamiento el precio tiende a disminuir en promedio un 0,19%.
- Al aumentar en 1 la cantidad de dormitorios que dispone el alojamiento el precio tiende a aumentar un 0,31% en promedio.
- Al aumentar en 1 la cantidad de baños disponibles en el alojamiento el precio tiende a disminuir un 0,04% en promedio.

- Al aumentar en 1 día la estadía mínima requerida el precio tiende a aumentar en 0,02% en promedio.
- Ver interpretación de las potencias

e) Suponga que se desea estudiar el efecto que tienen los distintos tipos de alojamientos (tipo) en el precio, ¿Cómo podemos incorporar dicho análisis en el modelo de regresión?

1) Estime el nuevo modelo.

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
LN_precio	5396	0,46	0,45	0,26	7992,59	8071,71

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	3,16	0,05	3,06	3,27	57,81	<0,0001		
puntaje	-0,24	0,03	-0,29	-0,19	-9,3	<0,0001	95,5	76,63
personas	-0,28	0,02	-0,33	-0,24	-12,04	<0,0001	154,08	18,98
dormitorios	0,24	0,01	0,22	0,26	23,1	<0,0001	542,71	1,81
baños	-0,02	0,01	-0,04	-0,01	-2,71	0,0067	16,36	1,08
estadia	-0,02	0,01	-0,03	-0,01	-3,73	0,0002	22,93	5,33
POT_puntaje	0,04	0,01	0,03	0,05	7,81	<0,0001	69,95	76,53
POT_personas	0,06	4,60E-03	0,05	0,07	13,6	<0,0001	193,9	19,59
POT_estadia	2,80E-05	3,80E-04	-7,10E-04	7,70E-04	0,07	0,9414	9,01	5,21
Entire home/apt	1,12	0,05	1,03	1,21	24,28	<0,0001	598,62	9,62
Private room	0,54	0,05	0,45	0,63	11,68	<0,0001	145,42	9,25

2) Interprete los coeficientes estimados.

La interpretación de los coeficientes estimados es la siguiente:

- Al aumentar en 1 el puntaje, el precio tiende a disminuir en promedio un 0,24%.
- Al aumentar en 1 la cantidad de personas que puede recibir el alojamiento el precio tiende a disminuir en promedio un 0,28%.
- Al aumentar en 1 la cantidad de dormitorios que dispone el alojamiento el precio tiende a aumentar un 0,24% en promedio.
- Al aumentar en 1 la cantidad de baños disponibles en el alojamiento el precio tiende a disminuir un 0,02% en promedio.
- Al aumentar en 1 día la estadía mínima requerida el precio tiende a dismimuir en 0,02% en promedio.
- Ver interpretación de las potencias
- El precio tiende a aumentar en promedio un 1,12% cuando se alquila el alojamiento completo en relación a un alojamiento con habitaciones compartidas.
- El precio tiende a aumentar en promedio un 0,54% cuando se alquila un alojamiento con habitaciones privadas, en relación a un alojamiento con habitaciones compartidas.

3) Evalúe el poder explicativo del modelo que seleccionó. Compare con los modelos anteriores.

El poder explicativo del modelo pasó de 0,3 a 0,45 según R2 ajustado.

4) ¿Se verifican los supuestos del modelo? Justifique adecuadamente.

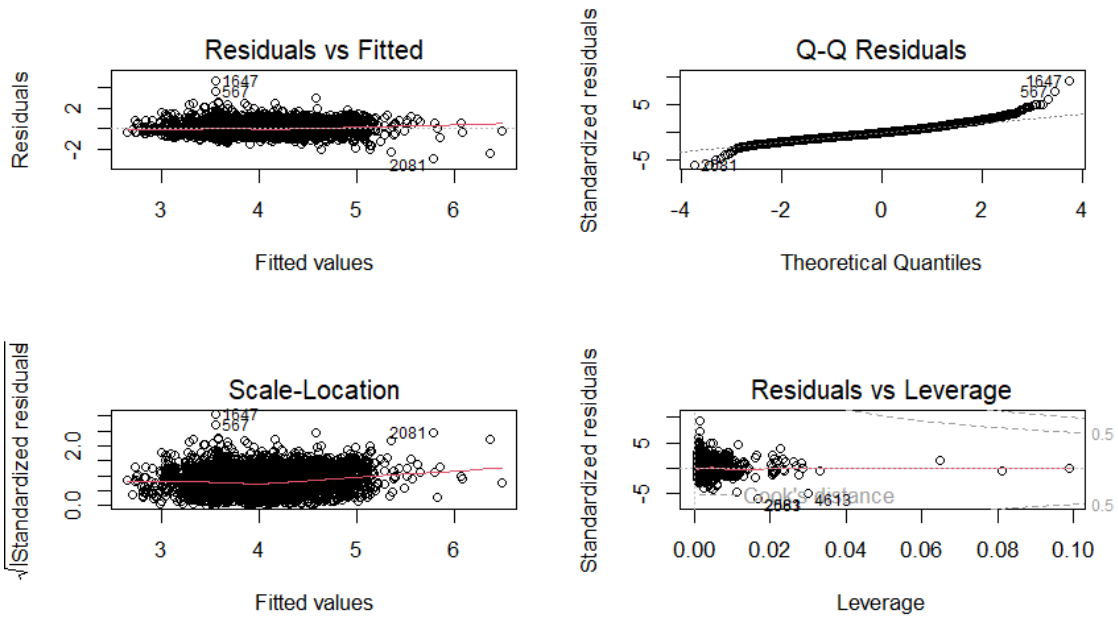
Test de normalidad

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuo
D = 0.053375, p-value = 8.883e-14
alternative hypothesis: two-sided
```

Test de homocedasticidad

```
studentized Breusch-Pagan test

data:  modelo7
BP = 288.79, df = 10, p-value < 2.2e-16
```

No cumple con los supuestos de normalidad y homocedasticidad



5) ¿Existen valores atípicos y/o valores influyentes? En caso afirmativo explique cómo los detecta y que decisión toma respecto a ellos para continuar con el análisis estadístico.

HACER