Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores

Unidad Morelia

Tarea

Ejercicios sección 2.6

PRESENTA:

Luis Alberto García Orozco

PROFESOR:

Haydeé Peruyero

GRADO

Licenciatura en Tecnologías para la Información en Ciencias

Asignatura: Estadística Multivariada

A: 22 de septiembre de 2025

Ejercicios:

Ejercicio 1: Para los datos de la Liga Nacional de Fútbol. Realizar tanto con las funciones de R y Python como con las fórmulas que usan matrices.

a) Ajustar un modelo de regresión lineal múltiple que relacione la cantidad de juegos ganados con las yardas por aire del equipo (x_2), el porcentaje de jugadas por tierra (x_7) y las yardas por tierra del contrario (x_8).

```
> lm_coefficients(datos, 'y', c('x2', 'x7', 'x8'))
      Intercept          x2         x7            x8
[1,]  -1.808372  0.00359807  0.1939602  -0.004815494
> m1 = lm('y ~ x2 + x7 + x8', datos)
> m1
Call:
lm(formula = "y ~ x2 + x7 + x8", data = datos)
Coefficients:
(Intercept)           x2            x7            x8
  -1.808372     0.003598      0.193960     -0.004815
```

Parece que el porcentaje de jugadas echas por tierra (x_8) tiene más peso para predecir la cantidad de juegos ganados, lo que suguiere que jugar más por tierra que por aire hace más probable que se gane un juego.

El hecho de que la constante para las yardas por tierra del equipo contrario sea negativa hace sentido si aceptamos lo dicho anteriormente, si el enemígo hace yardaje por tierra, puede deberse a que hicieron más jugadas por tierra, haciendo menos probable que el equipo en cuestión no gane.

El coeficiente de las yardas por aire del equipo (x_2) puede ser significante o no dependíendo de el rango de valores de yardas por aire, es decir, si x_2 tiene valores vajos, entonces casi no aporta, sin embargo, si x_2 tiene valores muy grandes, el coeficiente de x_2 podria mitigar estos valores, ajustando lo necesario.

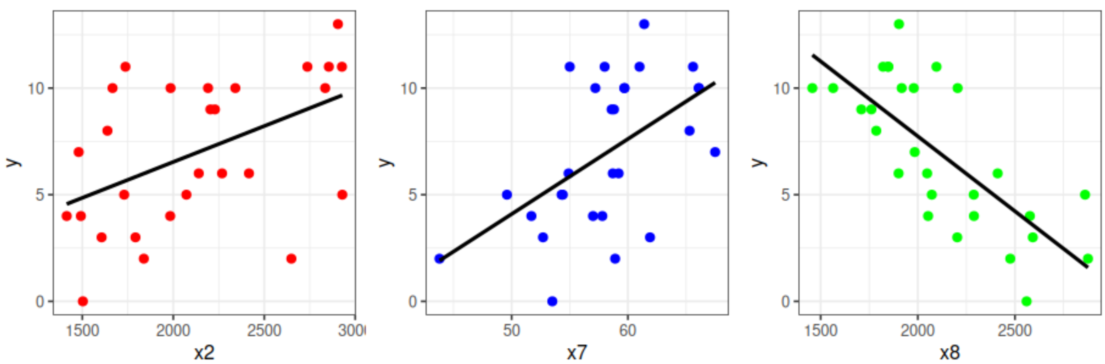

Figura 1: Modelos lineales individuales

b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

```
> anova_table(m1)
  Fuente.de.Variacion Suma de cuadrados Grados de libertad
1           Regresion          257.0943                  3
2          Residuales           69.8700                 24
3               Total          326.9643                 27
  Cuadrados medios      F_0
1         85.69809 29.43687
2          2.91125       NA
3               NA       NA

> F0_test_values(m1)
           F0      p-value       F1
[1,] 29.43687 3.273458e-08 3.008787
```

Como el valor de F_0 es mayor que el valor tabulado de $F_{\alpha;p,n-p-1} = F_{0,05;3,24} = 3{,}008787$, se rechaza H_0. Lo cual implica que la cantidad de juegos ganados depende de las yardas por aire del equipo, el porcentaje de jugadas por tierra y/o las yardas por tierra del contrario.

c) Calcular el estadístico t para probar las hipótesis $H_0 : \beta_2 = 0,\ H_0 : \beta_7 = 0$ y $H_0 : \beta_8 = 0$. ¿Qué conclusiones se pueden sacar acerca del papel de las variables $x_2,\ x_7$ y x_8 en el modelo?

```
> t0_test_values(m1)
         Intercept           x2         x7            x8
t0       -0.228883 5.177090e+00 2.19826168 -3.7710364517
p-value   0.820899 2.655723e-05 0.03781516  0.0009377699
tt        2.063899 2.063899e+00 2.06389856  2.0638985616
```

Los datos que se ven, tanto los de t_0 como los de p-values están ordenados en orden de evaluación para $\beta_0,\ \beta_2,\ \beta_7$ y β_8 respectivamente; y lo que podemos observar de estos datos es que para $\beta_2,\ \beta_7$ y β_8 tanto su p-value asociado como la comparacion de su t_0: $|t_0| > t_{\alpha/2,n-p}$ nos indican rechazar las hipotesis $H_0 : \beta_2 = 0,\ H_0 : \beta_7 = 0$ y $H_0 : \beta_8 = 0$, sin embargo es lo contrario para β_0, nuestro intercepto, que su valor t y su p-value nos indican aceptar $H_0 : \beta_0 = 0$.

d) Calcular R^2 y R^2_{adj} para este modelo.

```
R2_test(m1)
            R2 R2 ajustada
[1,] 0.7863069   0.7595953
```

R^2 nos indica que el modelo explica la mayor parte de la variabilidad de los datos, pero R^2_{adj} castiga un poco el uso de 4 coeficientes, según los datos de las pruebas t nos podría indicar que hay que eliminar el intercepto ($\beta_0 = 0$).

e) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

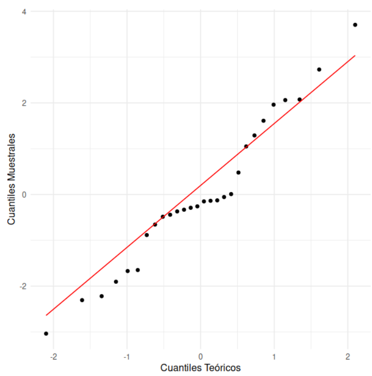

Figura 2: Gráfica de residuales

```
> qq_residuals(m1)

        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.96508, p-value = 0.4566
```

Tanto por lo visto en la gráfica como por lo visto en la prueba de shapiro los residuales tienen una buena correlación, con pequeñas desviaciones al la probabilidad de normalidad con un p-value de 0.4566 mucho mayor a 0.05, por lo tanto se puede afirmar que los residuales presentan normalidad.

f) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

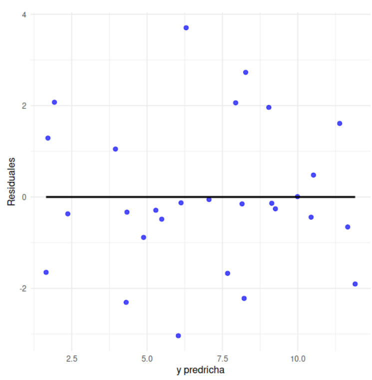

Figura 3: Residuales vs Juegos ganados predichos por el modelo

```
> res_vs_fitt(m1)

        studentized Breusch-Pagan test

data:  modelo
BP = 3.345, df = 3, p-value = 0.3414

`geom_smooth()` using formula = 'y ~ x'
```

En general los residuales parecen una nueve aleatoria de datos, un poco cargada hacia abajo, pero nada fuera de lo esperado, y nuestra prueba de Breush-Pagan nos lo comfirma.

g) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?

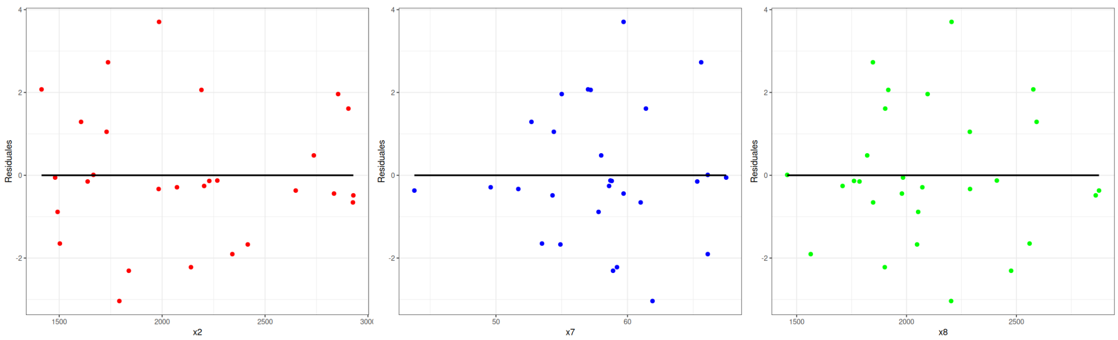

Figura 4: Residuales vs Variables regresoras

Lo mismo sucede aqui que en el inciso anterior.

h) Calcular un intervalo de confianza de 95 % para β_7 y un intervalo de confianza de 95 % para la cantidad media de juegos ganados por un equipo cuando $x_2 = 2300, x_7 = 56$ y $x_8 = 2100$.

```
> intervalos_conf_beta(m1)
            Intercept          x2         x7           x8
izq        -18.114944 0.002163664 0.01185532 -0.007451027
beta_value  -1.808372 0.003598070 0.19396021 -0.004815494
der         14.498200 0.005032477 0.37606510 -0.002179961
long        32.613145 0.002868813 0.36420978  0.005271066
> intervalos_conf_media_y(c(2300, 56, 2100), m1)
        izq      der     long
y0 6.436203 7.996645 1
60442
```

Tenemos que para $\beta_7 = 0{,}19396021$ su intervalo de confianza: $IC:\ [0{,}01185532, 0{,}37606510]$.

Y tenemos que el intervalo de confianza para la media de juegos ganados por un equipo cuando $x_2 = 2300, x_7 = 56$ y $x_8 = 2100$ es de: $[6{,}436203, 7{,}996645]$.

Además de esto de esta tabla podemos interpretar nuevamente la insignigicancia del intercepto al tener un IC bastante amplio que incluso abarca el 0, lo que además de añadirle incertidumbre, le quita significancia.

i) Ajustar un modelo a esos datos, usando solo x_7 y x_8 como regresores y probar la significancia de la regresión.

```
> lm_coefficients(datos, 'y', c('x7', 'x8'))
     Intercept         x7           x8
[1,]  17.94432 0.04837087 -0.006536593
> m2 = lm('y ~ x7 + x8', datos)
> m2

Call:
lm(formula = "y ~ x7 + x8", data = datos)

Coefficients:
(Intercept)          x7          x8
  17.944319    0.048371   -0.006537
> F0_test_values(m2)
           F0    p-value      F1
[1,] 15.13425 4.9349e-05 3.38519
```

Como el valor de F_0 es mayor que el valor tabulado de $F_{\alpha;p,n-p-1} = F_{0{,}05;3{,}24} = 3{,}008787$, se rechaza H_0. Lo cual implica que la cantidad de juegos ganados depende del porcentaje de jugadas por tierra y/o las yardas por tierra del contrario.

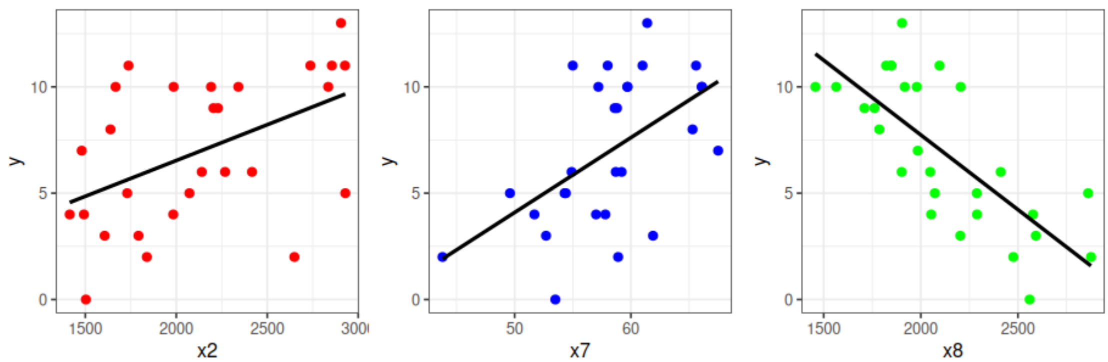

Figura 5: Modelos lineales individuales para el modelo 2

j) Calcular R^2 y R^2_{adj}. Compararlos con los resultados del modelo anterior.

```
> R2_test(m2)
            R2 R2 ajustada
[1,] 0.5476628   0.5114759
> R2_test(m1)
            R2 R2 ajustada
[1,] 0.7863069   0.7595953
```

Clararmente hay una diferencia significativa entre los modelos, lo que nos habla mucho de la dependencia las yardas por aire de un equipo con respecto los juegos que ha ganano, o que la información solo de los juegos por tierra no es suficiente para predecir bien los juegos ganados de un equipo.

k) Calcular un intervalo de confianza de 95 % para β_7. También, un intervalo de confianza de 95 % para la cantidad media de juegos ganados por un equipo cuando $x_7 = 56$ y $x_8 = 2100$. Comparar las longitudes de esos intervalos de confianza con las longitudes de los correspondientes al modelo anterior.

```
> intervalos_conf_beta(m2)
           Intercept          x7           x8
izq        -2.367848 -0.19716429 -0.010156368
beta_value 17.944319  0.04837087 -0.006536593
der        38.256485  0.29390602 -0.002916818
long       40.624334  0.49107031  0.007239550
> intervalos_conf_media_y(c(56, 2100), m2)
        izq      der     long
y0 5.828643 8.023842 2.195198
```

Con respecto al modelo anterior hay una diferencia notable entre los intervalos, pues mientras en el primer modelo los intervalos para β_7 y y_0 eran de aproximadamente $\pm 0{,}18$ y $\pm 0{,}75$ de manera respectiva, para el modelo 2 resultan ser de $\pm 0{,}25$ y 1,1, generando más incertidumbre en las constantes.

l) ¿Qué conclusiones se pueden sacar de este problema, acerca de las consecuencias de omitir un regresor importante de un modelo?

Al omitir un regresor importante en nuestro modelo, este se vuelve bastante más impreciso, lo que termina dando como consecuencia estimaciones menos certeras para nuevos datos obtenidos, por lo que es importante hacer las pruebas de significancia para cada constante regresora de las variables de nuestro modelo antes de eliminarlas.

Ejercicio 2: Véase los datos de rendimiento de gasolina. Realizar el ejercicio en R.

a) Ajustar un modelo de regresión lineal múltiple que relacione el rendimiento de la gasolina en millas por galón (y), la cilindrada del motor (x_1) y la cantidad de gargantas del carburador (x_6).

```
> lm_coefficients(datos, 'y', c('x1','x6'))
     Intercept          x1        x6
[1,]  32.88455 -0.05314767 0.9592231
> m1 = lm('y ~ x1 + x6', datos)
> m1

Call:
lm(formula = "y ~ x1 + x6", data = datos)

Coefficients:
(Intercept)           x1           x6
   32.88455     -0.05315      0.95922
```

Entre los valores de los coeficientes regresores β_1 y β_6 parece que mientras más gargantas tenga un vehículo, más le rinde la gasolina, a diferencia del volumen de la cilindrada que parece afectar negativamente al rendimiento, aunque su constante de regresión pequeña.

b) Formar la tabla de análisis de varianza y probar la significancia de la regresión.

```
> anova_table(m1)
  Fuente.de.Variacion Suma de cuadrados Grados de libertad
1           Regresion          974.3095                  2
2          Residuales          263.2345                 29
3               Total         1237.5441                 31
  Cuadrados medios      F_0
1       487.154770 53.66882
2         9.077053       NA
3               NA       NA
> F0_test_values(m1)
           F0      p-value       F1
[1,] 53.66882 1.789955e-10 3.327654
```

Como el valor de F_0 es mayor que el valor tabulado de $F_{\alpha;p,n-p-1} = F_{0,05;3,24} = 3{,}327654$, se rechaza H_0. Lo cual implica que el rendimiento de gasolina depende de la cilindrada y/o la cantidad de gargantas del carburador.

c) Calcular R^2 y R^2_{adj} para este modelo. Compararlas con las R^2 y R^2_{adj} ajustado para el modelo de regresión lineal simple, que relaciona las millas con la cilindrada.

```
> R2_test(m1)
            R2 R2 ajustada
[1,] 0.7872928   0.7726233
> m2 = lm('y ~ x1', datos)
> R2_test(m2)
            R2 R2 ajustada
[1,] 0.7722712   0.7646803
```

Hay una pequeña diferencia notable entre estos dos modelos, ambas (R^2 y R^2_{adj} disminuyen en el segundo modelo.

d) Determinar un intervalo de confianza para β_1

```
> intervalos_conf_beta(m1)
           Intercept          x1         x6
izq         29.74429 -0.06569892 -0.4116474
beta_value  32.88455 -0.05314767  0.9592231
der         36.02481 -0.04059641  2.3300935
long        6.280524  0.02510251  2.7417409
> intervalos_conf_beta(m2)
           Intercept          x1
izq         30.77383 -0.05694883
beta_value  33.72268 -0.04735958
der         36.67152 -0.03777032
long        5.897686  0.01917851
```

Las longitudes de los intervalos son de 0.02510251 en el modelo 1 y 0.01917851 en el modelo 2, lo que indíca que el modelo linear simple tuvo menos incertidumbre que el modelo 1 para el coeficiente de β_1

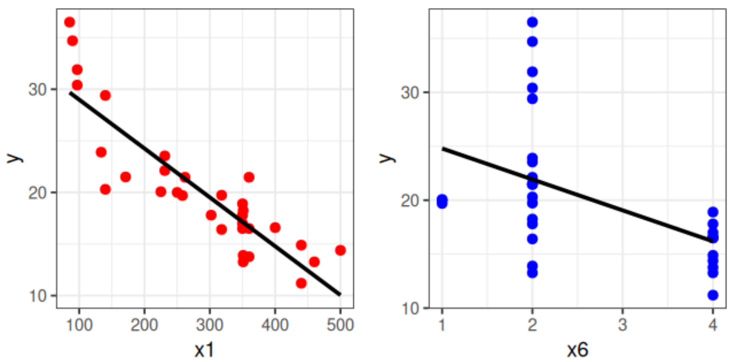

Figura 6: Modelos lineales individuales

Veamos que x_6 por si solo no se ajusta a su modelo linear simple, pues visualmente es más una variable categórica que no parece decir mucho del rendimiento de la gasolina.

e) Determinar un intervalo de confianza de 95 % para el rendimiento promedio de la gasolina, cuando $x_1 = 225$ $pulg^3$ y $x_6 = 2$ gargantas.

```
> intervalos_conf_media_y(c(255, 2),m1)
        izq      der     long
y0 19.99071 22.50997 2.519265
```

f) Determinar un intervalo de predicción de 95 % para una nueva observación de rendimiento de gasolina, cuando $x_1 = 225$ $pulg^3$ y $x_6 = 2$ gargantas.

```
> intervalos_pred_y(c(255, 2),m1)
        izq      der     long
y0 14.96101 27.53967 12.57866
```

g) Considerar el modelo de regresión lineal simple, que relaciona las millas con la cilindrada. Construir un intervalo de confianza de 95 % para el rendimiento promedio de la gasolina y un intervalo de predicción para el rendimiento, cuando $x_1 = 225$ $pulg^3$. Comparar las longitudes de estos intervalos con los intervalos obtenidos en los dos incisos anteriores. ¿Tiene ventajas agregar x_6 al modelo?

```
> intervalos_conf_media_y(c(255),m2)
        izq      der    long
y0 20.50255 22.78941 2.28686
> intervalos_pred_y(c(255),m2)
        izq     der     long
y0 15.28287 28.0091 12.72623
```

Las longitudes en el modelo 1 para el intervalo de confianza y el intervalo de predicción son 2.519265 y 4.804015 millas/galón correspondientemente a diferencia de las del modelo 2 que son 2.286865 y 4.681157 millas/galón respectivamente, donde podemos apreciar que nuevamente en el modelo 2 se presenta menos incertidumbre para para predecir o estimar la media de las millas por galón recorridas.

Apesar de que el modelo con x_6 parecía describir ligeramente mejor los datos según R^2 y R^2_{adj}, el quitarlo resulta mucho mejor, pues al obtener intervalos de confianza más estrechos para la media de y e intervalos de predicción para también más estrechos, el modelo 2 podría considerarse mejor modelo o un modelo más confiable.

h) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

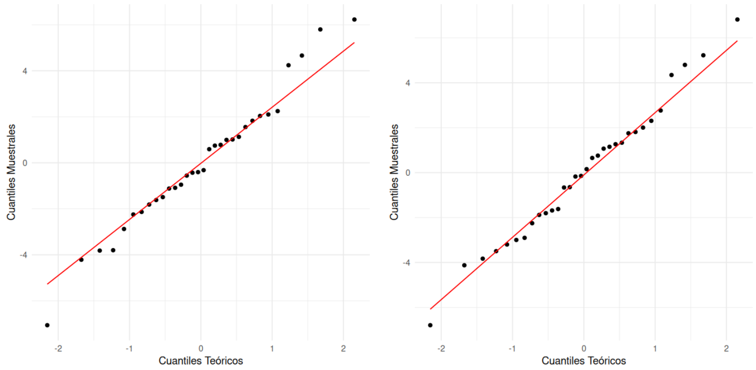

Figura 7: Gráfica de residuales de modelos $m1$ y $m2$

```
> qq_residuals(m1)

        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.98065, p-value = 0.8183

> qq_residuals(m2)

        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.98718, p-value = 0.961
```

En niguno de los dos modelos parece haber un problema de normalidad con los residuales. Solo se puede añadir que los residuales del modelo 2 presentan más normalidad según el test de Shapiro-Wilk, con valores más grandes para el estadístico W y el p-value que los del modelo 1.

i) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

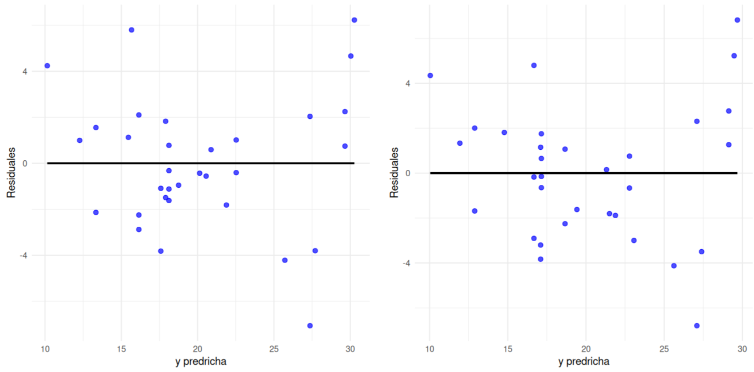

Figura 8: Residuales vs Millas/Galón predichos por los modelos $m1$ y $m2$

```
> res_vs_fitt(m1)

        studentized Breusch-Pagan test

data:  modelo
BP = 3.6593, df = 2, p-value = 0.1605

'geom_smooth()' using formula = 'y ~ x'
> res_vs_fitt(m2)

        studentized Breusch-Pagan test

data:  modelo
BP = 5.3609, df = 1, p-value = 0.02059

'geom_smooth()' using formula = 'y ~ x'
```

En ambos modelos los residuales parecen presentar una nueve aleatoria de datos muy similares, sin embargo el la prueba de Breusch-Pagan nos afirma la existencia de herterocedasticidad en el modelo 2 al presentar un p-value de 0.02059.

j) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?

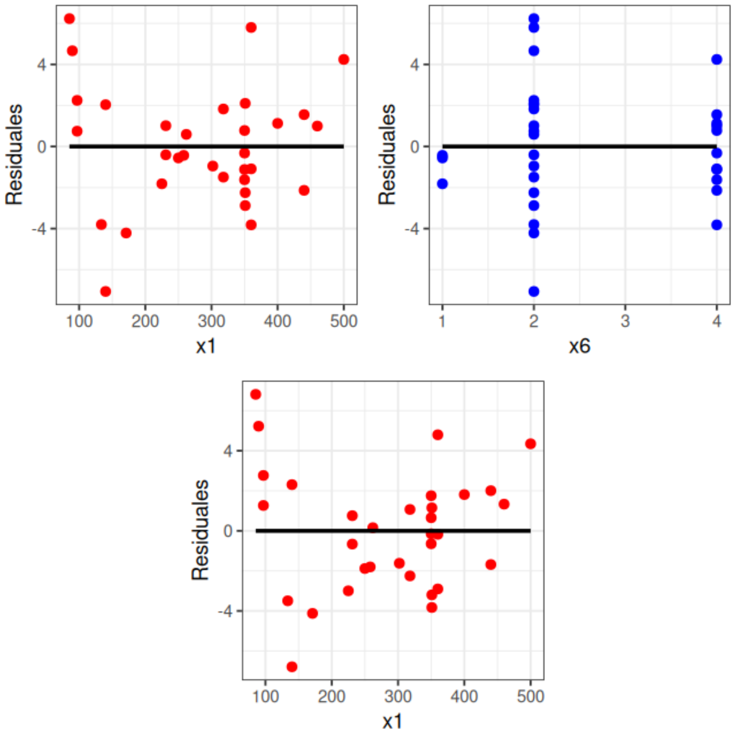

Figura 9: Residuales vs Variables regresoras de los modelos $m1$ y $m2$

Para la variable x_1 en ambos modelos se podría decir con facilidad que sí, sin embargo para x_6 explicarlo es un poco más dificil al tratarse de una variable discreta con tan solo información en tres valores (1, 2, 4), pero considerando que en en $x_6 = 2$ y $x_6 = 4$ los rangos son aplios y distintos a $x_6 = 1$ interpretaré heterocedasticidad, por lo que esta variable no se especificó bien.

Ejercicio 3: Véase los datos sobre precios de viviendas. Realizar el ejercicio en Python.

a) Ajustar un modelo de regresión lineal múltiple que relacione el precio de venta con los nueve regresores.

```
> lm_coefficients(datos, 'y', colnames(datos)[2:10])
     Intercept        x1        x2         x3        x4        x5         x6
[1,]  14.92765 1.924722 7.000534 0.1491779 2.722808 2.006684 -0.4101238
            x7          x8        x9
[1,] -1.403235 -0.03714908 1.559447
> m1 = lm('y ~ .', datos)
> m1
Call:
lm(formula = "y ~ .", data = datos)
Coefficients:
(Intercept)           x1           x2           x3           x4
   14.92765      1.92472      7.00053      0.14918      2.72281
         x5           x6           x7           x8           x9
    2.00668     -0.41012     -1.40324     -0.03715      1.55945
```

De este modelo destaca mucho la cantidad de baños (x_2) pues su constante regresora es la más grande (7.00053) a diferencia de la edad de la casa (x_8) que afecta poco y de manera negativa, lo que hace sentido, pues no todos se fijan en la edad de una casa, pero mientras más vieja es menos deseada.

Me parece sorprendente que la cantidad de recamaras (x_7) y cantidad de habitaciones (x_6) afecten de manera negativa, esperaría un comportamiento distinto.

b) Probar la significancia de la regresión.¿Qué conclusiones se pueden sacar?

```
F0_test_values(m1)
           F0      p-value       F1
[1,] 9.037027 0.0001850299 2.645791
```

El modelo es significativo pues p-value es muy inferior a 0.05 y el estadístico $F_0 > F_1$, por lo que depende de almenos una variable del modelo.

c) Usar pruebas t para evaluar la contribución de cada regresor al modelo.

```
> t0_test_values(m1)
          Intercept         x1        x2        x3        x4        x5
t0       2.52461055 1.86884083 1.6278905 0.3042043 0.6245612 1.4609912
p-value  0.02428304 0.08271059 0.1258361 0.7654469 0.5423043 0.1660965
tt       2.14478669 2.14478669 2.1447867 2.1447867 2.1447867 2.1447867
                 x6         x7         x8        x9
t0       -0.1724264 -0.4132581 -0.5567916 0.8048774
p-value   0.8655702  0.6856776  0.5864610 0.4343472
tt        2.1447867  2.1447867  2.1447867 2.1447867
```

Parece que la única variable significante según esta prueba t, es el intercepto, con un p-value de 0.02428304, seguida talvez del que más se le acerca que es x_1 con un p-value de 0.08271059 cercano a 0.05 y despues le siguen x_2 y x_5 con 0.1258361 y 0.1660965 como sus correspondientes p-values, las demás variables ya son muy poco significantes con p-values superiores a 0.4, rondeando principalmente cerca de 0.6.

d) Calcular R^2 y R^2_{adj} para este modelo.

```
> R2_test(m1)
            R2 R2 ajustada
[1,] 0.8531467    0.758741
```

Contrario a lo esperado por los resultados de las purebas test parece que el modelo se desempeño bien al ajustar y describir los datos, puede deberse a una correlación de los datos o un overfitting.

e) ¿Cuál es la contribución del tamaño del lote y el espacio vital para el modelo, dado que se incluyeron todos los demás regresores?.

```
> t0_test_values(m1)[,'x3']
       t0   p-value        tt
0.3042043 0.7654469 2.1447867
```

Si revisamos su p-value de las pruevas t podemos ver que es 0.76, aparentemente aporta muy poco.

f) En este modelo, ¿la colinealidad es un problema potencial?

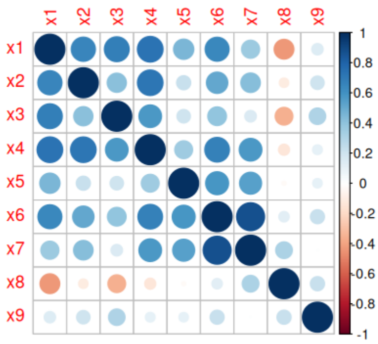

Figura 10: Correlación entre las variables del modelo 1

Parece las variables x_1, x_2, x_3 y x_4 estan muy correlacionadas etre sí así como también pasa con x_4, x_5, x_6 y x_7, por lo que si se podría decir que la multicolinealidad es un problema grave en esta situación.

g) Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber algún problema con la hipótesis de normalidad?

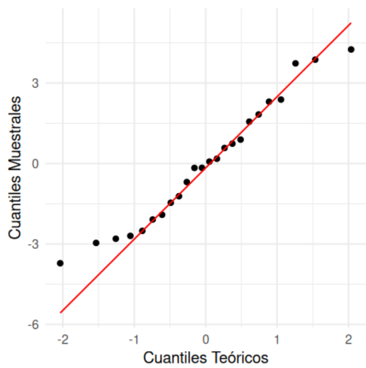

Figura 11: Gráfica de residuales

```
> qq_residuals(m1)

        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.96232, p-value = 0.4868
```

No hay ningún problema con el test de normalidad, los residuales presentan la normalidad esperada.

h) Trazar e interpretar una gráfica de los residuales en función de la respuesta predicha.

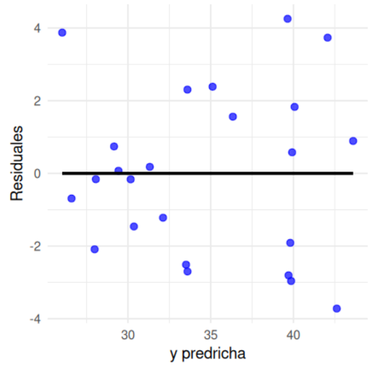

Figura 12: Residuales vs Precios de la casas predichos por el modelo

```
> res_vs_fitt(m1)

        studentized Breusch-Pagan test

data:  modelo
BP = 11.46, df = 9, p-value = 0.2455

`geom_smooth()` using formula = 'y ~ x'
```

Parece que tiene un comportamiento aleatorio normal, que nos confirma la prueba de Breusch-Pagan.

i) Trazar las gráficas de los residuales en función de cada una de las variables regresoras. ¿Implican esas gráficas que se especificó en forma correcta el regresor?.

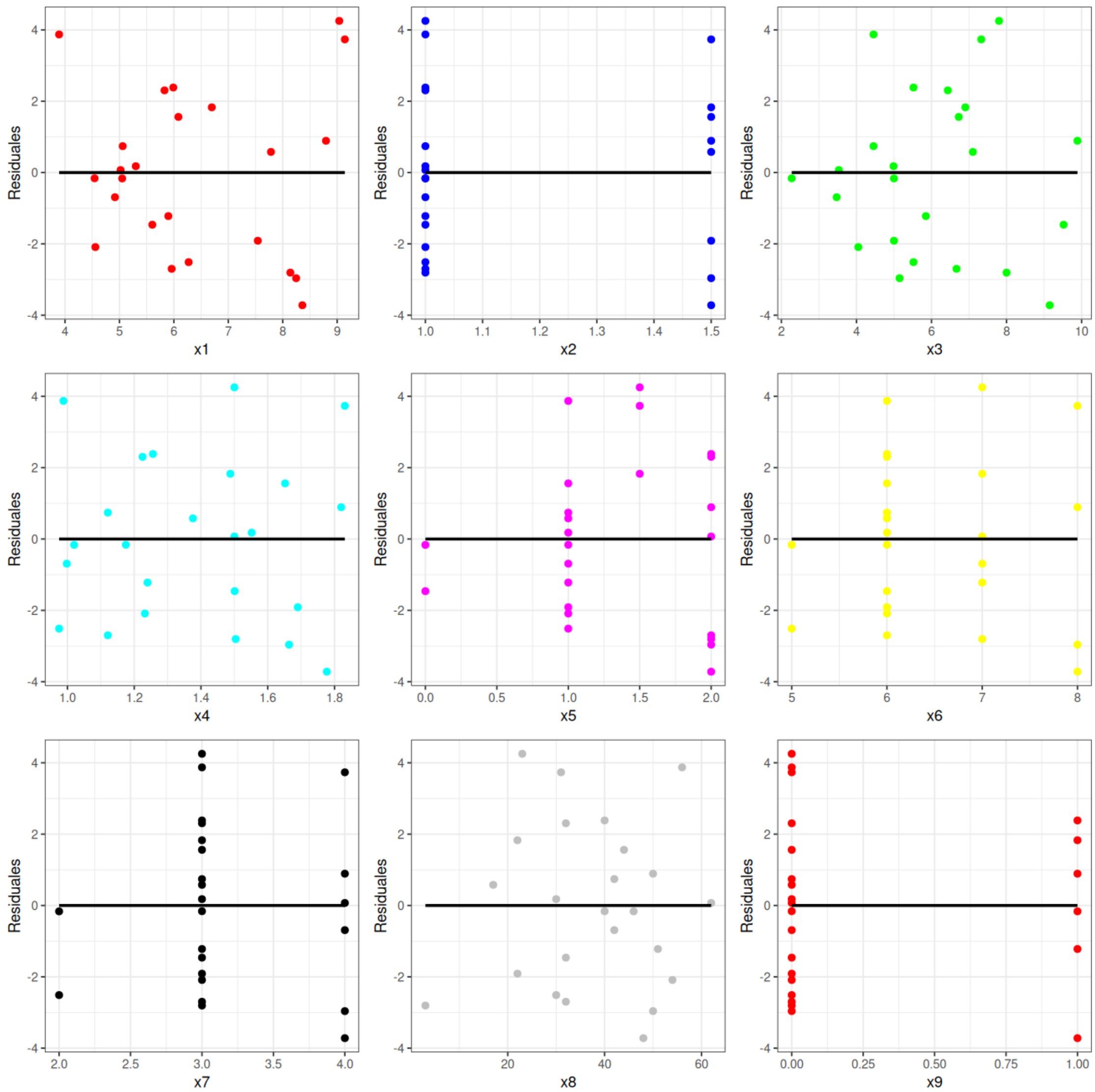

Figura 13: Residuales vs variables del modelo 1

Algunas de las variables parecen tener una distribución aleatoria, pero no todas, como por ejemplo es el caso de x_5, x_7, y x_8 que probablemente presentan heterocedasticidad.

Ejercicio 4: Explica lo siguiente.

a) ¿Qué supuestos del modelo de regresión lineal múltiple deben verificarse?

   I . Multicolinealidad: hay que checar que las variables no se coorelacionen mucho entre si para evitar redundancias.

   II . Normalidad de los residuales: hay que verificar que los residuales sigan una distribución normal, de lo contrario probablemente haya q hacer una transformación en las variables predictoras para seguir considerando un modelo multilineal.

   III . Homocedasticidad: se busca que cuando se grafíquen los residuales contra los valores predichos se forme una nube de aleatoria de puntos pues si no se comporta de esta manera implicaría que el modelo esta más ajustado en alguna parte de los datos pero se comporta distinto en otras.

   IV . No autocorrelacion: se confirma que los residuales no se correlacionen entre si, es decir que dado un error calculado el siguiente no dependa de la posición de este anterior o en dado caso de sus anteriores.

b) ¿Cómo se interpretan los intervalos de confianza? Si construimos un intervalo de confianza del 95 % para un coeficiente β_j, ¿cuál sería la lectura correcta o interpretación correcta sobre este intervalo?

c) Describe los métodos de selección de variables y sus ventajas y desventajas:

   I . Selección hacia adelante (forward)

   Consiste en inciar el modelo que mejor ajusta tu variable predictora por sí sola, y después agregar la siguiente variable que en conjunto con las anteriores ajusten mejor el modelo, y así de manera suscesiva.

   II . Selección hacia atrás (backward)

   Consiste en iniciar con el modelo completo (el que considere todas las variables posibles) y a partir de ahi ir eliminando variables de tal forma que bajo algún critero se obtenga un modelo mejor ajustado.

   III . Selección por pasos (stepwise) y/o mejor subconjunto (best subset)

   Las técnicas de selección paso a paso consisten en instanciar un modelo a partir del cual se agregan o se eliminan variables hasta encontrar el mejor modelo

d) Explica cómo se utilizan para elegir el modelo final.

Se puede usar una combinacion de estos metódos para reducir los modelos a considerar, también nos sirven para comparar que variables son indispensables y cuales no tanto.

Ejercicio 5: Para los datos del ejercicio 1 de la liga de Fútbol. Realizar el ejercicio en R y Python.

a) Usar el algoritmo de selección hacia adelante para seleccionar un modelo de regresión.

```
> fs = forward_stepwise(m1)
Variables Selected:
=> x8
=> x2
=> x7
=> x9
>fs
                              Stepwise Summary
----------------------------------------------------------------------
Step    Variable        AIC        SBC      SBIC      R2       Adj. R2
----------------------------------------------------------------------
 0      Base Model    152.275    154.939     NA     0.00000    0.00000
 1      x8            132.245    136.242     NA     0.54468    0.52717
 2      x2            118.201    123.530     NA     0.74328    0.72274
 3      x7            115.065    121.726     NA     0.78631    0.75960
 4      x9            115.044    123.037     NA     0.80119    0.76661
----------------------------------------------------------------------
Final Model Output
------------------
                         Model Summary
---------------------------------------------------------------
R                        0.895       RMSE                1.524
R-Squared                0.801       MSE                 2.322
Adj. R-Squared           0.767       Coef. Var          24.140
Pred R-Squared           0.732       AIC               115.044
MAE                      1.107       SBC               123.037
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria
                               ANOVA
-------------------------------------------------------------------
               Sum of
              Squares       DF     Mean Square      F         Sig.
-------------------------------------------------------------------
Regression    261.960        4          65.490    23.172    0.0000
Residual       65.004       23           2.826
Total         326.964       27
-------------------------------------------------------------------
```

b) Usar el algoritmo de selección hacia atrás para seleccionar un modelo de regresión.

```
>bs = backward_stepwise(m1)
Variables Removed:
=> x5
=> x1
=> x6
=> x3
=> x4
>bs
                              Stepwise Summary
--------------------------------------------------------------------------
Step    Variable         AIC         SBC        SBIC        R2        Adj. R2
--------------------------------------------------------------------------
 0      Full Model     122.937     137.591         NA     0.81560     0.72340
 1      x5             120.937     134.259     49.554     0.81560     0.73795
 2      x1             119.193     131.182     45.812     0.81391     0.74877
 3      x6             117.508     128.166     42.492     0.81180     0.75802
 4      x3             116.283     125.608     39.946     0.80652     0.76254
 5      x4             115.044     123.037     37.662     0.80119     0.76661
--------------------------------------------------------------------------
Final Model Output
------------------
                          Model Summary
----------------------------------------------------------------
R                         0.895       RMSE                 1.524
R-Squared                 0.801       MSE                  2.322
Adj. R-Squared            0.767       Coef. Var           24.140
Pred R-Squared            0.732       AIC                115.044
MAE                       1.107       SBC                123.037
----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria
                                ANOVA
--------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F          Sig.
--------------------------------------------------------------------
Regression    261.960         4         65.490    23.172     0.0000
Residual       65.004        23          2.826
Total         326.964        27
--------------------------------------------------------------------
```

c) Usar el algoritmo de regresión por pasos para seleccionar un modelo de regresión.

```
> as = all_models_step(m1)
Variables Selected:
=> x8
=> x2
=> x7
=> x9
> as
                          Stepwise Summary
----------------------------------------------------------------------
Step    Variable        AIC        SBC       SBIC       R2       Adj. R2
----------------------------------------------------------------------
 0      Base Model    152.275    154.939       NA    0.00000    0.00000
 1      x8 (+)        132.245    136.242       NA    0.54468    0.52717
 2      x2 (+)        118.201    123.530       NA    0.74328    0.72274
 3      x7 (+)        115.065    121.726       NA    0.78631    0.75960
 4      x9 (+)        115.044    123.037       NA    0.80119    0.76661
----------------------------------------------------------------------
Final Model Output
------------------
                        Model Summary
---------------------------------------------------------------
R                       0.895       RMSE                 1.524
R-Squared               0.801       MSE                  2.322
Adj. R-Squared          0.767       Coef. Var           24.140
Pred R-Squared          0.732       AIC                115.044
MAE                     1.107       SBC                123.037
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria
                              ANOVA
-------------------------------------------------------------------
               Sum of
              Squares        DF    Mean Square      F         Sig.
-------------------------------------------------------------------
Regression    261.960         4         65.490    23.172     0.0000
Residual       65.004        23          2.826
Total         326.964        27
-------------------------------------------------------------------
```

d) Comenta los modelos finales en cada uno de los casos anteriores. ¿Cuál tiene más sentido? ¿Cuál modelo usarían?

En los tres métodos se eligirieron las variables $x_2,\ x_7,\ x_8$, y x_9, entonces se puede asumir que es muy claro cual es el mejor modelo o hay una diferencia muy grande entre las variables útiles y las que no lo son, al menos para este caso. Tal vez si se dectecta mayor cantidad de variables utiles que inutiles el mejor método de selección sería el backward stepwise, de lo contrario el forward o both stepwise debido a la cantidad de operaciones.