M5_AI4_CANOJORGE

JORGE CANO

2024-10-25

EJERCICIOS

```
##   V1        V2
## 1 x1 0.1207267
```

1. Proponed una especificación que a vuestra intuición sea un buen modelo para explicar la variable y en base a las x que tenemos anteriormente.

Resulta dificil estimar una especificación con la variable obtenidas, únicamente x1 parece representativa en el modelo, por tanto, a pesar de ser sencillo, vamos a proponer un glm (y ~ x1) y en pasos siguiente iremos explorando diferentes posibilidades para ampliar las variables y mejorar la especificad del modelo.

Para quedarnos únicamente con x1 he realizado un análisis de los errores y p-valor, no siendo significativo para el resto de variable. Otra prueba de contraste ha consistido en analizar el AIC, siendo mejor con este modelo planteado.

```
##
## Call:
## glm(formula = formula, family = gaussian, data = df)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.490010   1.535476  21.811  < 2e-16 ***
## x1          -0.047026   0.004985  -9.434 3.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.736062)
##
##     Null deviance: 1139.11  on 29  degrees of freedom
## Residual deviance:  272.61  on 28  degrees of freedom
## AIC: 157.34
```

```
##
## Number of Fisher Scoring iterations: 2
```

```
## [1] 161.5452
```

2. Utilizar la técnica STEPWISE para elegir el modelo de tal forma que minimicemos el BIC.

Si aplicamos la técnica STEPWISE vemos como nos devuelve un modelo simplificado en número de variables (y~ x1 + x4), dónde no podemos considerar significativa x4, pue presenta un p valor elevado (0.273). Hacemos un análisis de los residuos para entender como se comportan.

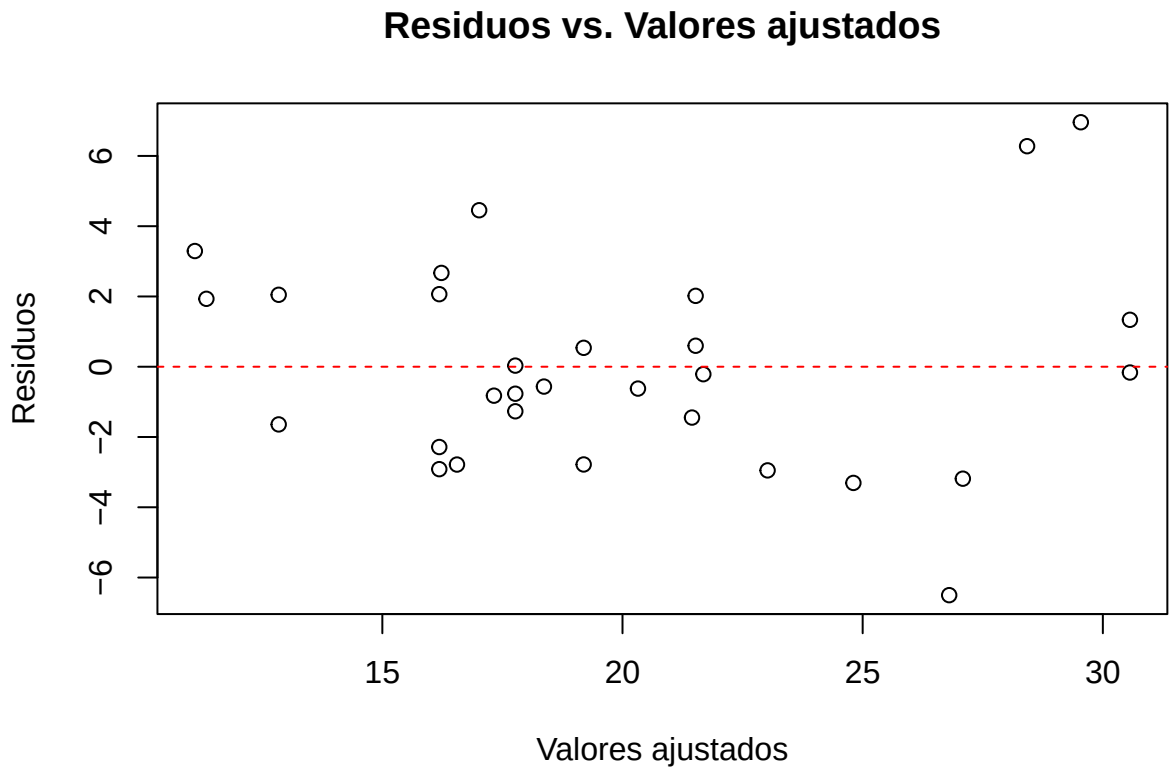
```
## Start:  AIC=111.1
## y ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + x1    1    866.50  272.61  70.205
## + x10   1    828.59  310.52  74.111
## + x3    1    822.21  316.89  74.721
## + x2    1    723.26  415.84  82.873
## + x9    1    663.14  475.96  86.924
## + x8    1    645.58  493.53  88.012
## + x11   1    592.62  546.49  91.070
## + x7    1    570.62  568.49  92.253
## + x5    1    458.95  680.15  97.634
## + x6    1    253.57  885.54 105.550
## + x4    1    203.21  935.89 107.209
## <none>              1139.11 111.104
##
## Step:  AIC=70.21
## y ~ x1
##
##        Df Sum of Sq     RSS     AIC
## + x4    1     18.57  254.04  70.089
## <none>               272.61  70.205
## + x6    1     15.32  257.29  70.471
## + x9    1     14.24  258.37  70.596
## + x3    1      9.11  263.50  71.186
## + x10   1      6.35  266.26  71.498
## + x2    1      5.60  267.01  71.583
## + x5    1      4.90  267.71  71.661
## + x7    1      3.36  269.25  71.833
## + x11   1      0.02  272.59  72.203
## + x8    1      0.00  272.61  72.205
## - x1    1    866.50 1139.11 111.104
##
## Step:  AIC=70.09
## y ~ x1 + x4
```

```
##
##        Df Sum of Sq    RSS     AIC
## <none>              254.04  70.089
## - x4    1     18.57 272.61  70.205
## + x9    1      9.30 244.73  70.969
## + x6    1      6.36 247.68  71.328
## + x3    1      6.24 247.80  71.342
## + x10   1      4.44 249.59  71.559
## + x2    1      3.54 250.49  71.667
## + x11   1      1.37 252.66  71.926
## + x5    1      1.28 252.76  71.937
## + x7    1      0.09 253.95  72.078
## + x8    1      0.00 254.04  72.089
## - x1    1    681.85 935.89 107.209
```

```
##
## Call:
## lm(formula = y ~ x1 + x4, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5011 -2.1243 -0.3884  1.9964  6.9582
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.179421  18.787955   0.382    0.705
## x1          -0.044479   0.005225  -8.513 3.98e-09 ***
## x4           3.077228   2.190294   1.405    0.171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.067 on 27 degrees of freedom
## Multiple R-squared:  0.777,  Adjusted R-squared:  0.7605
## F-statistic: 47.03 on 2 and 27 DF,  p-value: 1.594e-09
```

```
## [1] 162.8297
```

Grafico residuos vs. valores ajustados



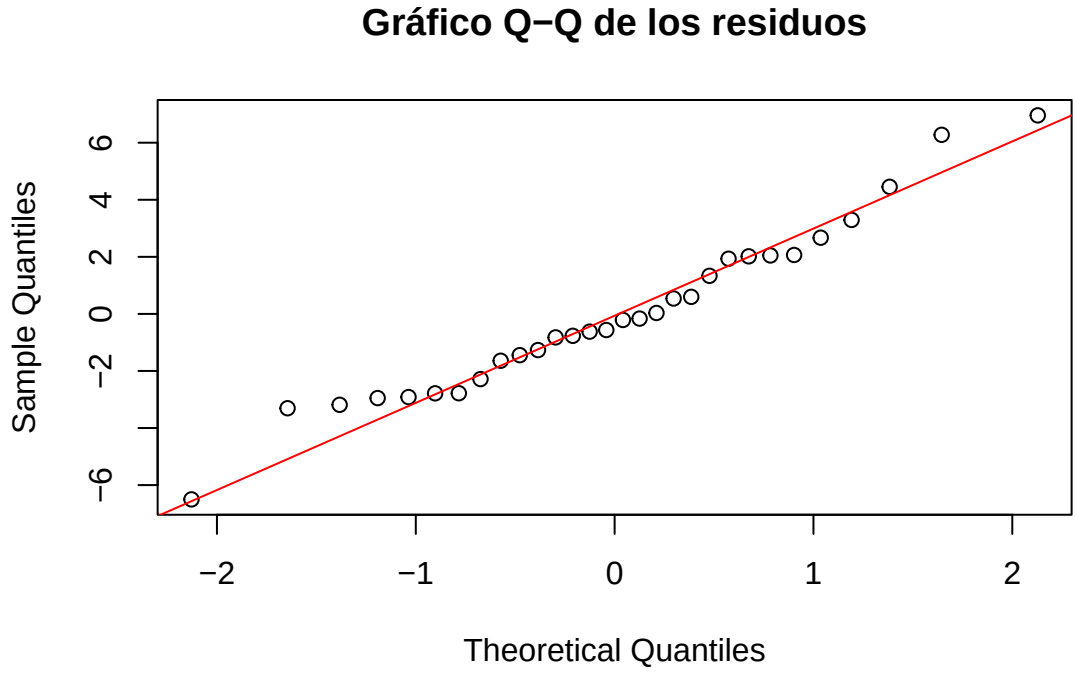
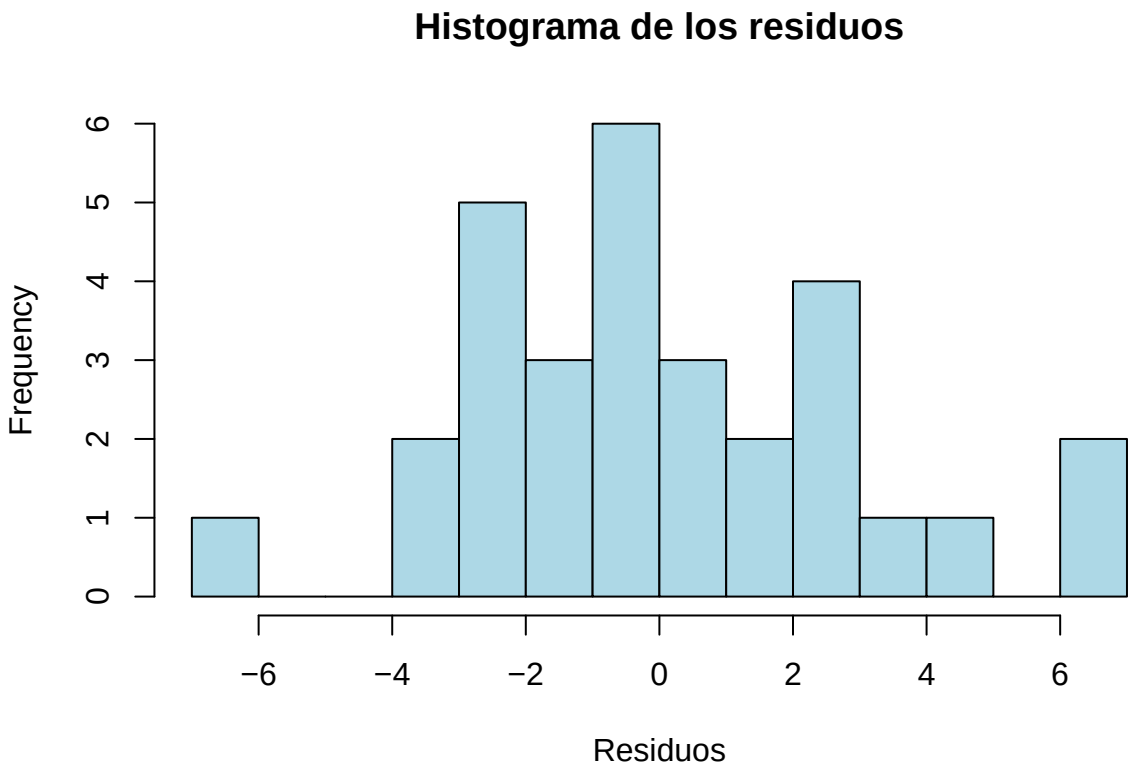


Gráfico Q-Q de los residuos

Histograma de los residuos



No se rechaza la hipótesis nula de homocedasticidad (es decir, la varianza de los errores parece ser constante), pero el valor p cercano a 0.05 sugiere que podrías estar atento a posibles signos de heterocedasticidad en el modelo.

```
##
##  studentized Breusch-Pagan test
##
## data:  stepwise_model
## BP = 6.7029, df = 2, p-value = 0.03503
```

3. Programad vuestro propio STEPWISE (Backward o Forward) para decidir cuál sería el mejor modelo minimizando la siguiente función:

El mejor modelo correspondría a –> y ~ x1

```
## Start:  AIC=157.61
## y ~ x1 + x6
##
```

```
##        Df Deviance    AIC
## - x6    1   272.61 157.34
## <none>      257.29 157.61
## - x1    1   885.54 192.69
##
## Step:  AIC=157.34
## y ~ x1
##
##        Df Deviance    AIC
## <none>      272.61 157.34
## + x6    1   257.29 157.61
## - x1    1  1139.11 198.24
```

```
##
## Call:
## glm(formula = y ~ x1, family = gaussian, data = df)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.490010   1.535476  21.811  < 2e-16 ***
## x1          -0.047026   0.004985  -9.434 3.43e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 9.736062)
##
##     Null deviance: 1139.11  on 29  degrees of freedom
## Residual deviance:  272.61  on 28  degrees of freedom
## AIC: 157.34
##
## Number of Fisher Scoring iterations: 2
```

4. Probad a variar el 0.05 para elegir un modelo según vuestra visión.

Tras varias iteracciones, ajustamos el k=0.001 (prácticamente nula penalización), obteniendo un modelo (y~x1+x6+x4+x9+x8+x10+x5). En mi opinión, la mejora en el modelo no justifica el incremente de variable que hemos introducido. Por tanto, una vez iterado con el valor de K, podemos defender que un valor igual a 0.05 es óptimo para encontrar un equilibrio.

```
##                              V1        V2
## 9 x1+x4+x9+x8+x10+x5+x7+x3+x2 0.06898312
```

si aplicamos este modelo propuesto con penalización k=0.01 en un stepwise vemos como la mejor especificación corresponde a y ~ x8 + x10 + x5, con un menor AIC y unos p-valor significativos.

```
## Start:  AIC=162.34
## y ~ x1 + x6 + x4 + x9 + x8 + x10 + x5
##
##        Df Deviance    AIC
## - x6    1   216.07 160.37
## - x1    1   216.28 160.40
## - x4    1   217.91 160.62
## - x9    1   220.81 161.02
## - x5    1   228.56 162.05
## <none>      215.88 162.34
## - x10   1   232.15 162.52
## - x8    1   240.76 163.62
##
## Step:  AIC=160.37
## y ~ x1 + x4 + x9 + x8 + x10 + x5
##
##        Df Deviance    AIC
## - x1    1   217.40 158.55
## - x4    1   217.96 158.63
## - x9    1   220.85 159.03
## <none>      216.07 160.37
## - x5    1   231.34 160.42
## - x10   1   232.59 160.58
## - x8    1   241.00 161.64
## + x6    1   215.88 162.34
##
## Step:  AIC=158.55
## y ~ x4 + x9 + x8 + x10 + x5
##
##        Df Deviance    AIC
## - x4    1   218.73 156.74
## - x9    1   221.09 157.06
## <none>      217.40 158.55
## + x1    1   216.07 160.37
## + x6    1   216.28 160.40
## - x5    1   249.28 160.66
## - x8    1   269.17 162.96
## - x10   1   349.34 170.78
##
## Step:  AIC=156.74
## y ~ x9 + x8 + x10 + x5
##
##        Df Deviance    AIC
## - x9    1   223.82 155.43
## <none>      218.73 156.74
## + x4    1   217.40 158.55
## + x1    1   217.96 158.63
## + x6    1   218.18 158.66
## - x5    1   259.13 159.82
```

```
## - x8    1   276.13 161.73
## - x10   1   353.83 169.16
##
## Step:  AIC=155.43
## y ~ x8 + x10 + x5
##
##        Df Deviance    AIC
## <none>      223.82 155.43
## + x9    1   218.73 156.74
## + x4    1   221.09 157.06
## + x6    1   223.69 157.41
## + x1    1   223.82 157.43
## - x5    1   260.14 157.94
## - x8    1   276.78 159.80
## - x10   1   418.66 172.21

##
## Call:
## glm(formula = y ~ x8 + x10 + x5, family = gaussian, data = df)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.590404  11.771925   0.390   0.6998
## x8           0.217814   0.087817   2.480   0.0199 *
## x10         -0.009485   0.001994  -4.757 6.38e-05 ***
## x5           2.597240   1.264562   2.054   0.0502 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.608618)
##
##     Null deviance: 1139.11  on 29  degrees of freedom
## Residual deviance:  223.82  on 26  degrees of freedom
## AIC: 155.43
##
## Number of Fisher Scoring iterations: 2
```

5. En función de los modelos anteriores, ¿cuál de ellos en el caso de que difieran recomendaríais?

Analaizadas diferentes iteraciones con penalizaciones, sus AIC y la significancia de las variables en el modelo, mi propuesta sería simplicar el modelo con (y~x1), pero esto podría llevar al riesgo de simplificar excesivamente el modelo y omitir información importante, lo que puede llevar a sesgos, falta de robustez y problemas en la interpretación.

Por eso, y a pesar de obtener un peor BIC/AIC me quedaría con el modelo y ~ x8 + x10 + x5 limitando la penalización de inclusión de variable a una valor k muy bajo (0.001), reduciendo el sesgo y tratando de evitar de esta forma las posibles limitaciones de simplicar un modelo a una variable.