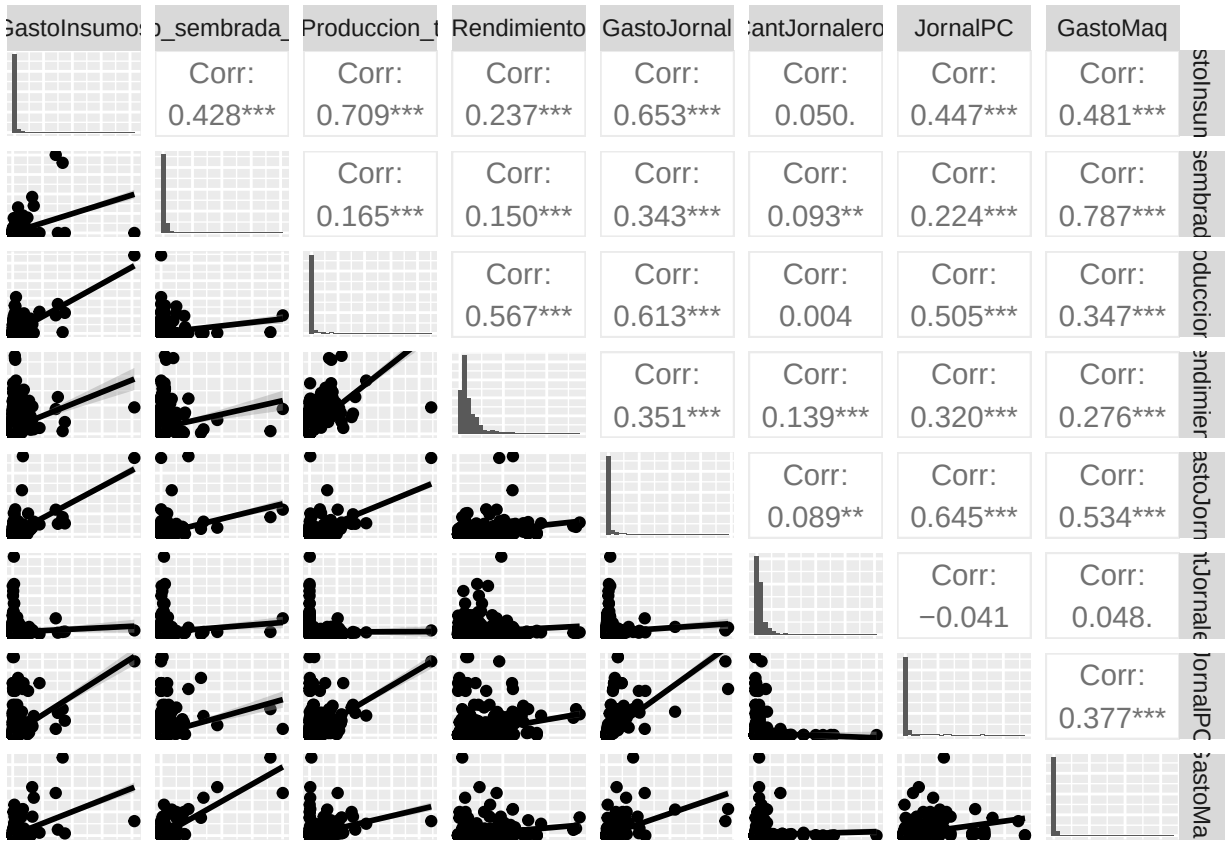
Evaluación por procesos 4

Miguel Roca
Rodrigo Huaman
Gerson Julca
Fabrizzio Arce
Jesus paucar

Contents

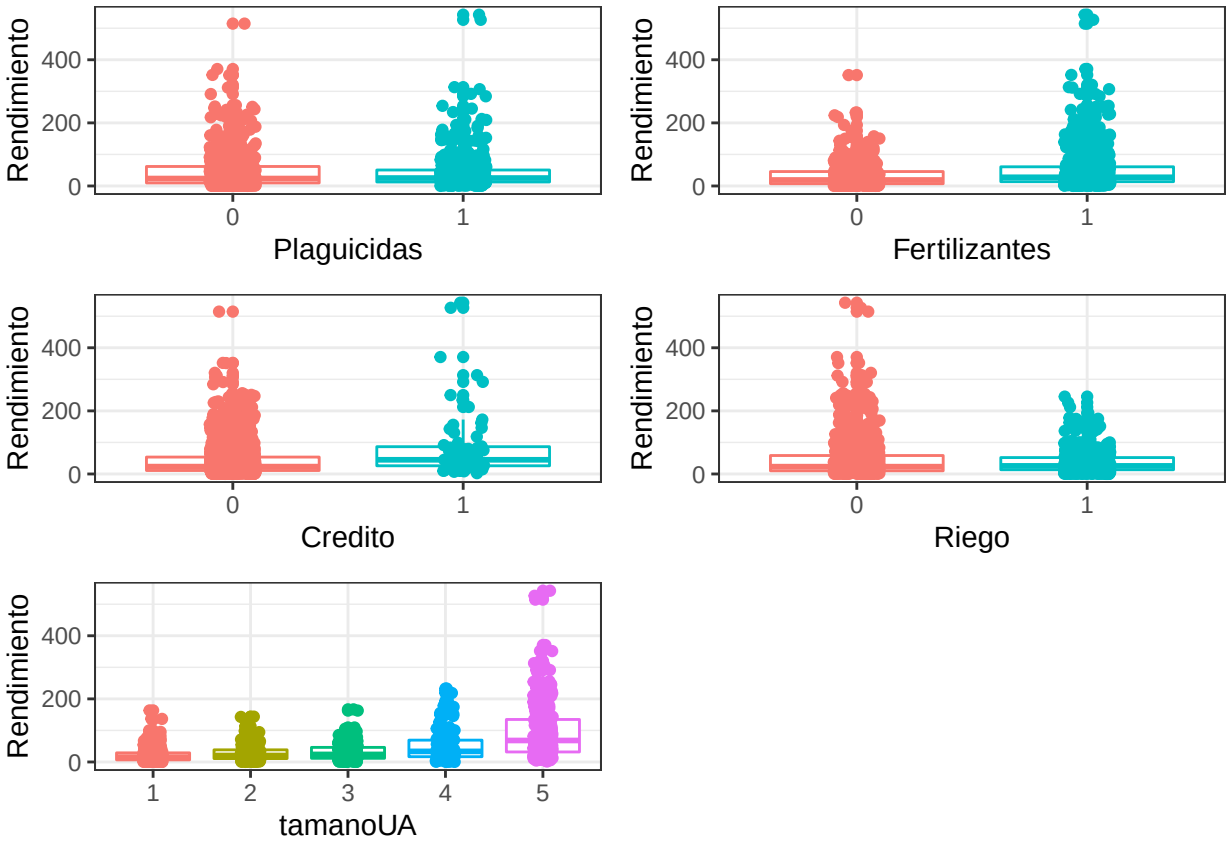
1. Análisis descriptivo de la variable dependiente y sus covariables.

1.1. Analizar la correlación entre cada par de variables cuantitativas



- La variable que tiene una mayor correlación con el rendimiento es la producción (r = 0.567).
- Gasto en maquinaria y Superficie sembrada están medianamente correlacionados (r = 0.787) por lo que posiblemente no sea útil introducir ambos predictores en el modelo.
- Gasto en insumos y Produccción están medianamente correlacionados (r = 0.709) por lo que posiblemente no sea útil introducir ambos predictores en el modelo.
- Gasto por Jornal y la Producción están medianamente correlacionados (r = 0.613) por lo que posiblemente no sea útil introducir ambos predictores en el modelo.
- Las variables Rendimiento y Cantidad de Jornaleros muestran una distribución exponencial, una transformación logarítmica posiblemente haría más normal su distribución.

1.2. Analizar las diferencias del valor promedio entre las categóricas



- La variable Plaguicidas parece no influir de forma significativa en el Rendimeinto
- La variable Fertilizantes parece influir de forma significativa en el Rendimeinto
- La variable Credito parece influir de forma significativa en el Rendimeinto
- La variable Riego no parece influir de forma significativa en el Rendimeinto
- La variable tamanoUA parece influir de forma significativa en el Rendimeinto

2. Análisis sobre su elección del mejor modelo econométrico y Análisis de sus pruebas de hipótesis

Hay diferentes formas para obtener el modelo más adecuado. Obtaremos por emplear el método mixto iniciando el modelo con todas las variables como predictores y realizando la selección de los mejores predictores con la medición Akaike(AIC).

2.1. Modelo completo

```
##
## Call:
## lm(formula = Rendimiento ~ Plaguicidas + Fertilizantes + GastoInsumos +
##     Credito + Riego + sup_sembrada_ha + Produccion_t + tamanoUA +
##     GastoJornal + CantJornaleros + JornalPC + GastoMaq, data = df)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -264.538  -20.098   -6.568   12.126  315.465
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.3612241  2.9700838   3.489 0.000503 ***
## Plaguicidas1     0.2808060  2.9699193   0.095 0.924688
## Fertilizantes1   1.7159691  2.9800922   0.576 0.564849
## GastoInsumos    -0.0015787  0.0001191 -13.253  < 2e-16 ***
## Credito1         9.0636454  5.2261546   1.734 0.083120 .
## Riego1          13.7701969  2.6807369   5.137 3.25e-07 ***
## sup_sembrada_ha  0.9208545  0.5562396   1.655 0.098080 .
## Produccion_t     0.0192217  0.0008864  21.685  < 2e-16 ***
## tamanoUA2        6.1367803  3.4475178   1.780 0.075315 .
## tamanoUA3       10.7376455  3.4560254   3.107 0.001934 **
## tamanoUA4       29.1968413  4.2934618   6.800 1.63e-11 ***
## tamanoUA5       47.9131359  4.0708443  11.770  < 2e-16 ***
## GastoJornal      0.0003236  0.0001968   1.644 0.100359
## CantJornaleros   0.4244438  0.0996565   4.259 2.21e-05 ***
## JornalPC        -0.0008472  0.0009941  -0.852 0.394287
## GastoMaq         0.0026587  0.0012749   2.085 0.037235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.55 on 1221 degrees of freedom
## Multiple R-squared:  0.5075, Adjusted R-squared:  0.5014
## F-statistic: 83.88 on 15 and 1221 DF,  p-value: < 2.2e-16
```

- El modelo con todas las variables introducidas como predictores tiene un R^2 muy bajo (0.5014), es capaz de explicar el 50.14% de la variabilidad observada en el rendimiento.

2.2. Elección del mejor modelo

```
step(object = modelo, direction = "both", trace = 1)
```

```
## Start:  AIC=9176.17
## Rendimiento ~ Plaguicidas + Fertilizantes + GastoInsumos + Credito +
##     Riego + sup_sembrada_ha + Produccion_t + tamanoUA + GastoJornal +
##     CantJornaleros + JornalPC + GastoMaq
##
##                   Df Sum of Sq     RSS    AIC
## - Plaguicidas      1        15 2008033 9174.2
## - Fertilizantes    1       545 2008563 9174.5
## - JornalPC         1      1194 2009212 9174.9
## <none>                         2008018 9176.2
## - GastoJornal      1      4447 2012465 9176.9
## - sup_sembrada_ha  1      4507 2012525 9176.9
## - Credito          1      4946 2012965 9177.2
## - GastoMaq         1      7152 2015171 9178.6
## - CantJornaleros   1     29832 2037850 9192.4
## - Riego            1     43393 2051412 9200.6
## - tamanoUA         4    264838 2272857 9321.4
## - GastoInsumos     1    288859 2296878 9340.4
## - Produccion_t     1    773319 2781337 9577.2
```

```
##
## Step:  AIC=9174.18
## Rendimiento ~ Fertilizantes + GastoInsumos + Credito + Riego +
##     sup_sembrada_ha + Produccion_t + tamanoUA + GastoJornal +
##     CantJornaleros + JornalPC + GastoMaq
##
##                   Df Sum of Sq     RSS    AIC
## - Fertilizantes    1       891 2008923 9172.7
## - JornalPC         1      1195 2009228 9172.9
## <none>                         2008033 9174.2
## - GastoJornal      1      4454 2012486 9174.9
## - sup_sembrada_ha  1      4497 2012530 9174.9
## - Credito          1      4967 2013000 9175.2
## + Plaguicidas      1        15 2008018 9176.2
## - GastoMaq         1      7188 2015221 9176.6
## - CantJornaleros   1     30134 2038167 9190.6
## - Riego            1     45830 2053863 9200.1
## - tamanoUA         4    272162 2280195 9323.4
## - GastoInsumos     1    289637 2297670 9338.9
## - Produccion_t     1    781132 2789164 9578.6
##
## Step:  AIC=9172.73
## Rendimiento ~ GastoInsumos + Credito + Riego + sup_sembrada_ha +
##     Produccion_t + tamanoUA + GastoJornal + CantJornaleros +
##     JornalPC + GastoMaq
##
##                   Df Sum of Sq     RSS    AIC
## - JornalPC         1      1075 2009998 9171.4
## <none>                         2008923 9172.7
## - GastoJornal      1      4319 2013243 9173.4
## - sup_sembrada_ha  1      4638 2013561 9173.6
## - Credito          1      5194 2014118 9173.9
## + Fertilizantes    1       891 2008033 9174.2
## + Plaguicidas      1       360 2008563 9174.5
## - GastoMaq         1      7179 2016103 9175.1
## - CantJornaleros   1     31042 2039965 9189.7
## - Riego            1     51132 2060056 9201.8
## - tamanoUA         4    276955 2285878 9324.5
## - GastoInsumos     1    289856 2298779 9337.4
## - Produccion_t     1    792719 2801642 9582.2
##
## Step:  AIC=9171.39
## Rendimiento ~ GastoInsumos + Credito + Riego + sup_sembrada_ha +
##     Produccion_t + tamanoUA + GastoJornal + CantJornaleros +
##     GastoMaq
##
##                   Df Sum of Sq     RSS    AIC
## <none>                         2009998 9171.4
## - GastoJornal      1      3260 2013258 9171.4
## - sup_sembrada_ha  1      4751 2014750 9172.3
## - Credito          1      4919 2014917 9172.4
## + JornalPC         1      1075 2008923 9172.7
## + Fertilizantes    1       771 2009228 9172.9
## + Plaguicidas      1       323 2009676 9173.2
```

```
## - GastoMaq         1      7026 2017025 9173.7
## - CantJornaleros   1     33262 2043260 9189.7
## - Riego            1     51341 2061339 9200.6
## - tamanoUA         4    279924 2289923 9324.7
## - GastoInsumos     1    288955 2298953 9335.5
## - Produccion_t     1    795080 2805078 9581.7
```

```
##
## Call:
## lm(formula = Rendimiento ~ GastoInsumos + Credito + Riego + sup_sembrada_ha +
##     Produccion_t + tamanoUA + GastoJornal + CantJornaleros +
##     GastoMaq, data = df)
##
## Coefficients:
##     (Intercept)     GastoInsumos         Credito1           Riego1
##      11.0180872       -0.0015751        9.0043342       14.2746278
## sup_sembrada_ha     Produccion_t        tamanoUA2        tamanoUA3
##       0.9443799        0.0192001        6.3161816       10.9992852
##       tamanoUA4        tamanoUA5      GastoJornal   CantJornaleros
##      29.1476468       47.6435950        0.0002478        0.4413926
##        GastoMaq
##       0.0026321
```

El mejor modelo resultante del proceso de selección ha sido:

```
##
## Call:
## lm(formula = Rendimiento ~ GastoInsumos + Credito + Riego + sup_sembrada_ha +
##     Produccion_t + tamanoUA + GastoJornal + CantJornaleros +
##     GastoMaq, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -266.136  -20.256   -6.526   11.901  313.382
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.0180872  2.7715626   3.975 7.44e-05 ***
## GastoInsumos    -0.0015751  0.0001187 -13.265  < 2e-16 ***
## Credito1         9.0043342  5.2028008   1.731  0.08376 .
## Riego1          14.2746278  2.5529385   5.591 2.77e-08 ***
## sup_sembrada_ha  0.9443799  0.5551911   1.701  0.08920 .
## Produccion_t     0.0192001  0.0008726  22.004  < 2e-16 ***
## tamanoUA2        6.3161816  3.4353744   1.839  0.06622 .
## tamanoUA3       10.9992852  3.4304456   3.206  0.00138 **
## tamanoUA4       29.1476468  4.2746211   6.819 1.44e-11 ***
## tamanoUA5       47.6435950  3.9390477  12.095  < 2e-16 ***
## GastoJornal      0.0002478  0.0001759   1.409  0.15912
## CantJornaleros   0.4413926  0.0980751   4.501 7.42e-06 ***
## GastoMaq         0.0026321  0.0012724   2.069  0.03880 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.52 on 1224 degrees of freedom
## Multiple R-squared:  0.507,  Adjusted R-squared:  0.5022
```

```
## F-statistic: 104.9 on 12 and 1224 DF,  p-value: < 2.2e-16
```

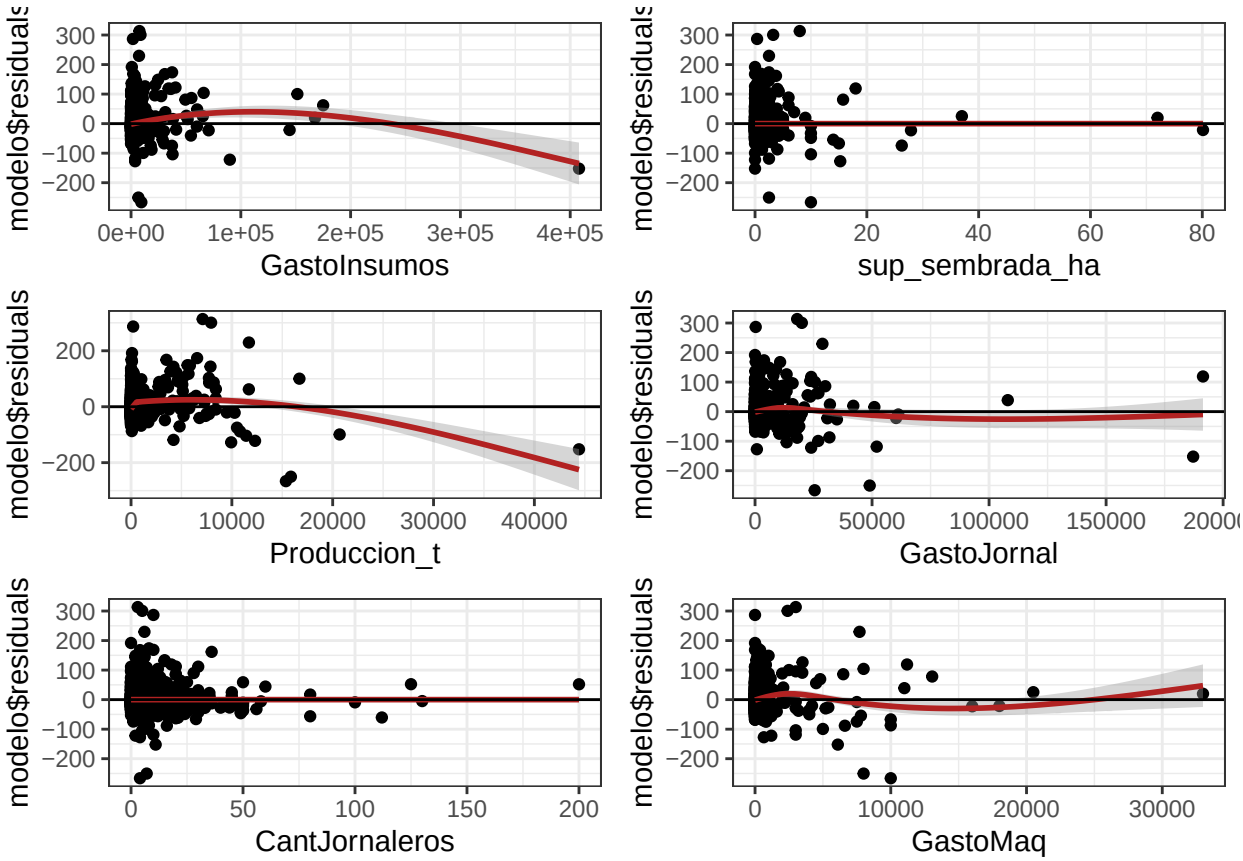
- Con la elección del mejor modelo el R^2 no mejoro significativamente (0.5022), ahora el modelo es capaz de explicar el 50.22% de la variabilidad observada en el rendimiento.
- El p-value del modelo es significativo (2.2e-16) por lo que se puede aceptar que el modelo no es por azar, al menos uno de los coeficientes parciales de regresión es distinto de 0.
- Por otro lado, al igual que en el modelo inicial tenemos coeficientes que no son significativos (Credito1, sup_sembrada_ha, tamanoUA2 y GastoJornal) lo que es un indicativo de que podrían no contribuir al modelo y probablemente sean retiradas.

$$\begin{aligned} Rendimiento = 11.01 - 0.0015GastoInsumos + 9.004Credito1 + 14.27Riego1 + \\ 0.94supSembradaHa + 0.01Produccion + 6.31tamanoUA2 + 10.99tamanoUA3 + \\ 29.14tamanoUA4 + 47.64tamanoUA5 + 0.0002GastoJornal + 0.44CantJornaleros + \\ 0.002GastoMaq \end{aligned}$$

- Por cada hectarea sembrada, el rendimeinto aumenta en promedio 0.94 unidades, manteniéndose constantes el resto de predictores.
- El rendimiento de las Unidades agropecurias de tamaño 5 son en promedio 47.64 veces más eficientes que el resto de unidades agropecuarias.

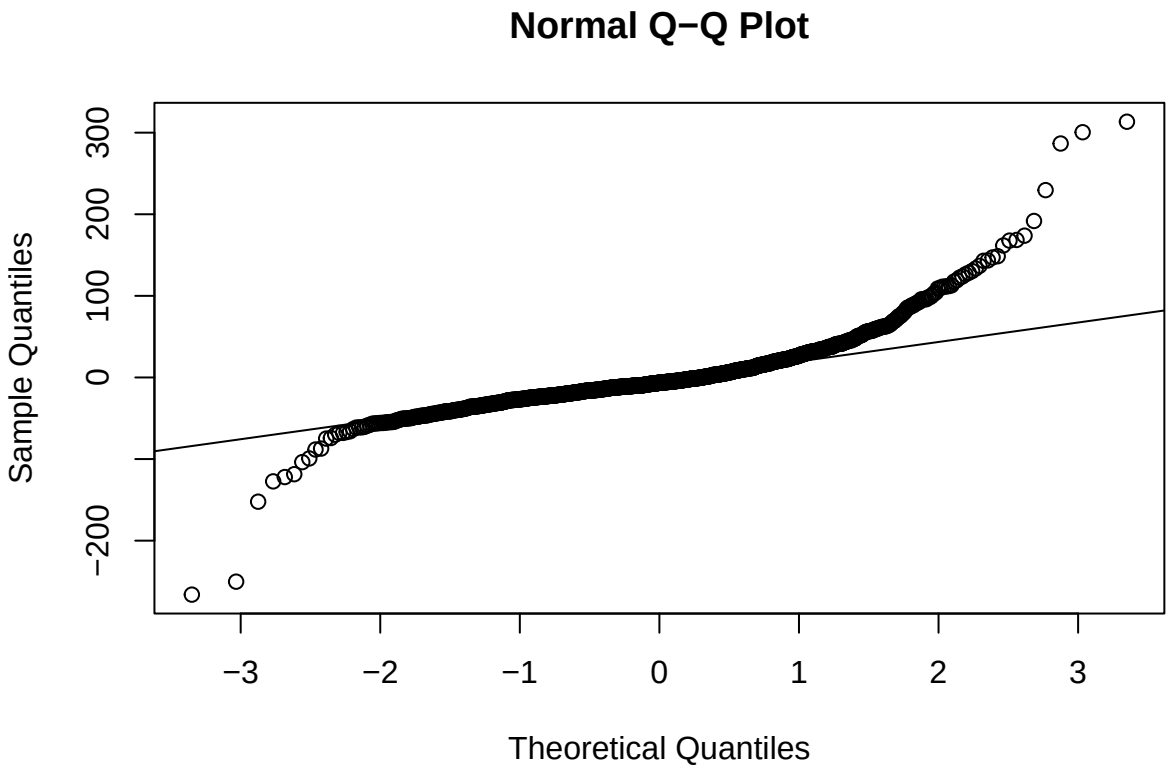
3. Análisis de los supuestos: Multicolinealidad, Heterocedasticidad, Normalidad de los residuos, valores influyentes.

3.1. Relación lineal entre los predictores numéricos y la variable respuesta



No se aprecia una clara linealidad de las variables continuas con los residuos, esto puede ser por que se observan posibles datos atípicos los cuales pueden presentar influencia.

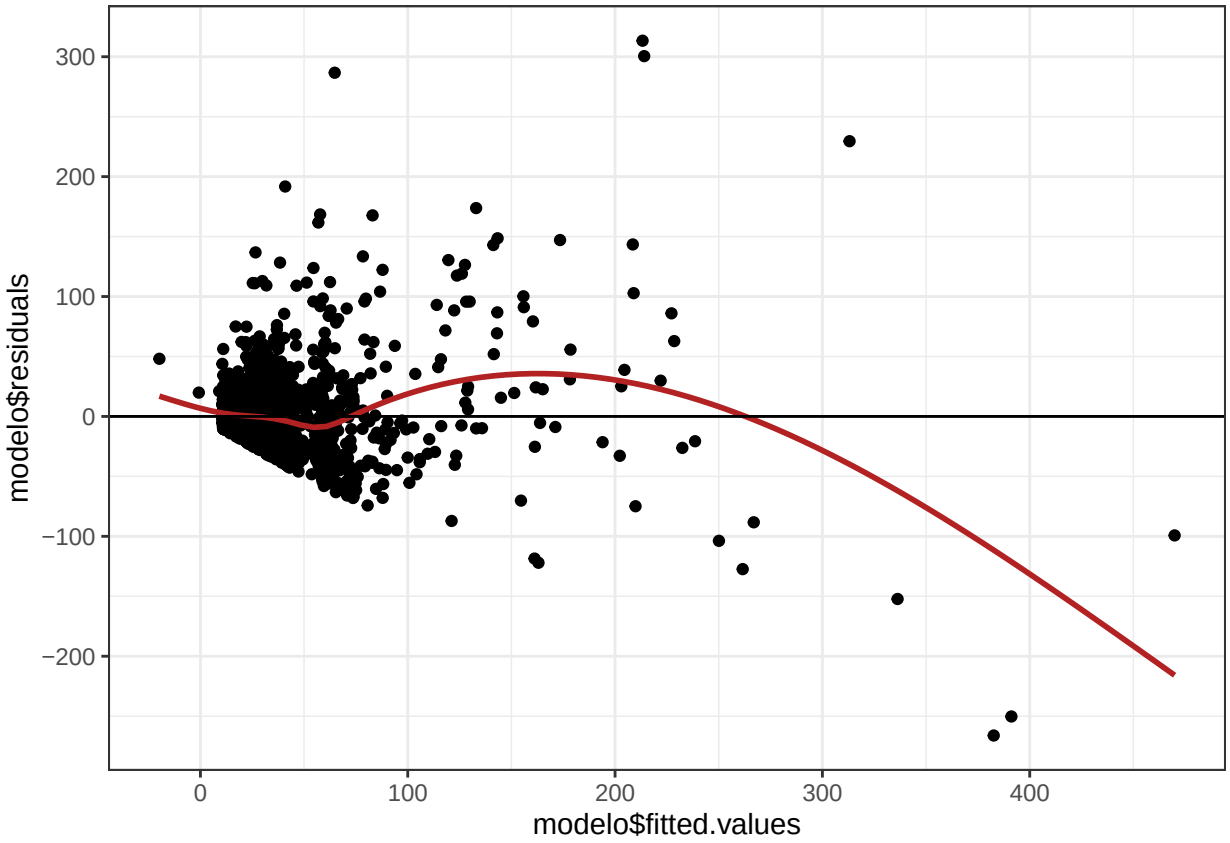
3.2. Distribución normal de los residuos




```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.82728, p-value < 2.2e-16
```

La condición de normalidad no se satisface, posiblemente debido a un dato atípico.

3.3. Variabilidad constante de los residuos (homocedasticidad)

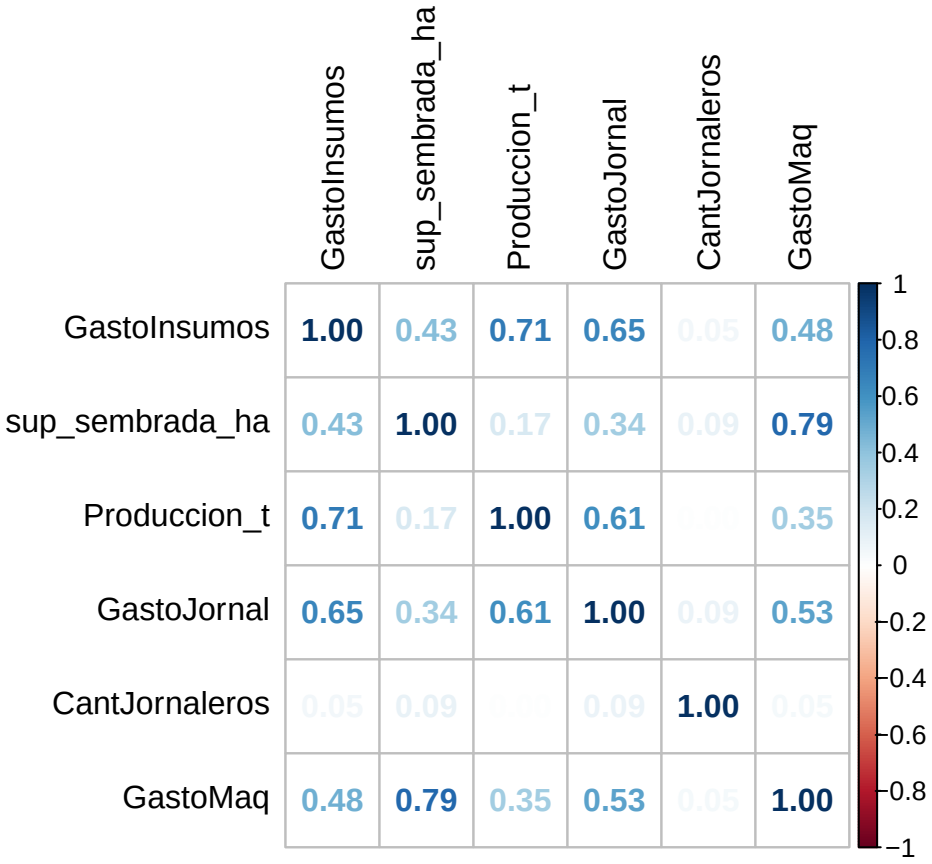



```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 330.68, df = 12, p-value < 2.2e-16
```

Se evidencia la falta de homocedasticidad.

3.4. No Multicolinealidad

Matriz de correlación entre predictores.



Se puede apreciar una moderada correlación entre algunas variables.

3.5. Análisis de Inflación de Varianza (VIF):

```
##                      GVIF Df GVIF^(1/(2*Df))
## GastoInsumos     2.844297  1        1.686504
## Credito          1.088577  1        1.043349
## Riego            1.221018  1        1.104997
## sup_sembrada_ha 3.115154  1        1.764980
## Produccion_t     2.699327  1        1.642963
## tamanoUA         1.561665  4        1.057301
## GastoJornal      2.225523  1        1.491819
## CantJornaleros   1.076964  1        1.037769
## GastoMaq         3.469732  1        1.862722
```

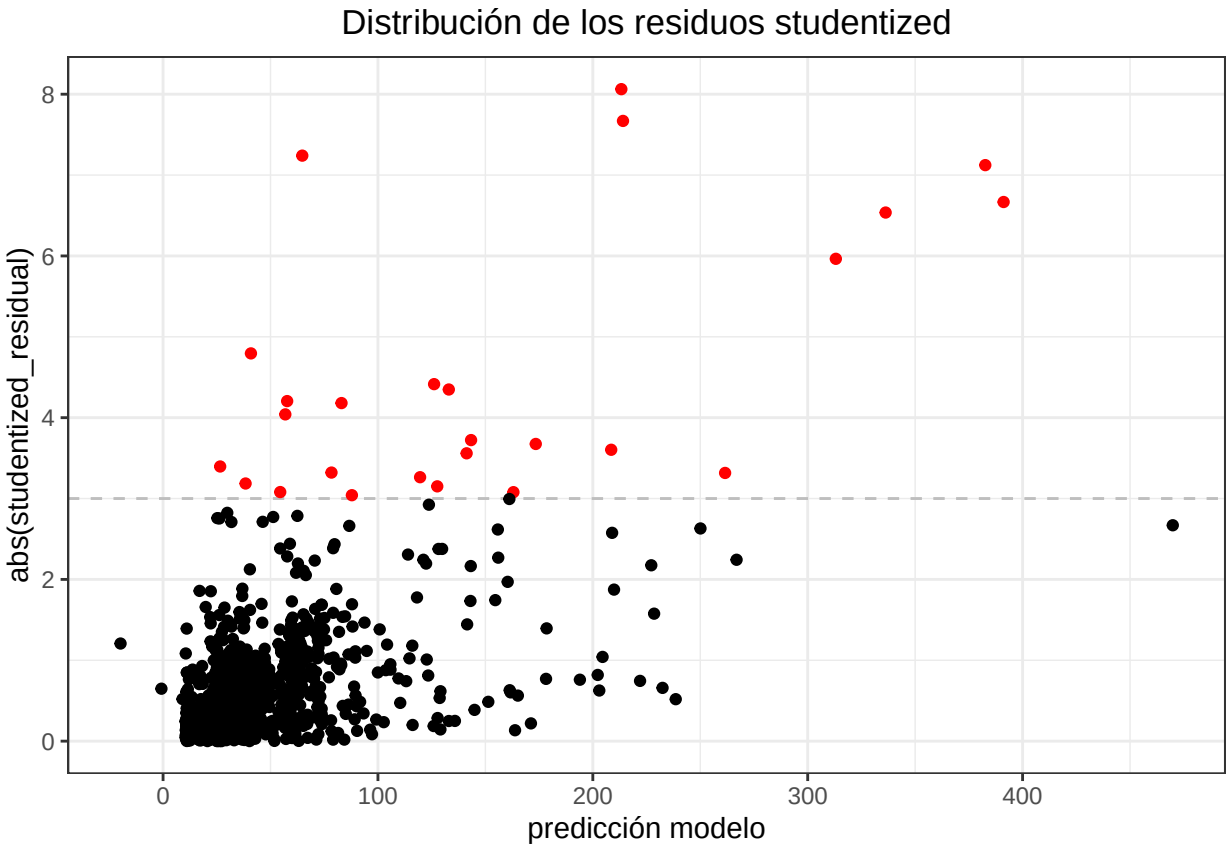
Hay predictores que muestran una correlación lineal muy alta e inflación de varianza, esto se puede deber a valores atípicos o datos influyentes.

3.6 Autocorrelación:

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1619735      1.675698       0
##  Alternative hypothesis: rho != 0
```

No hay evidencia de autocorrelación

4.Identificación de posibles valores atípicos o influyentes



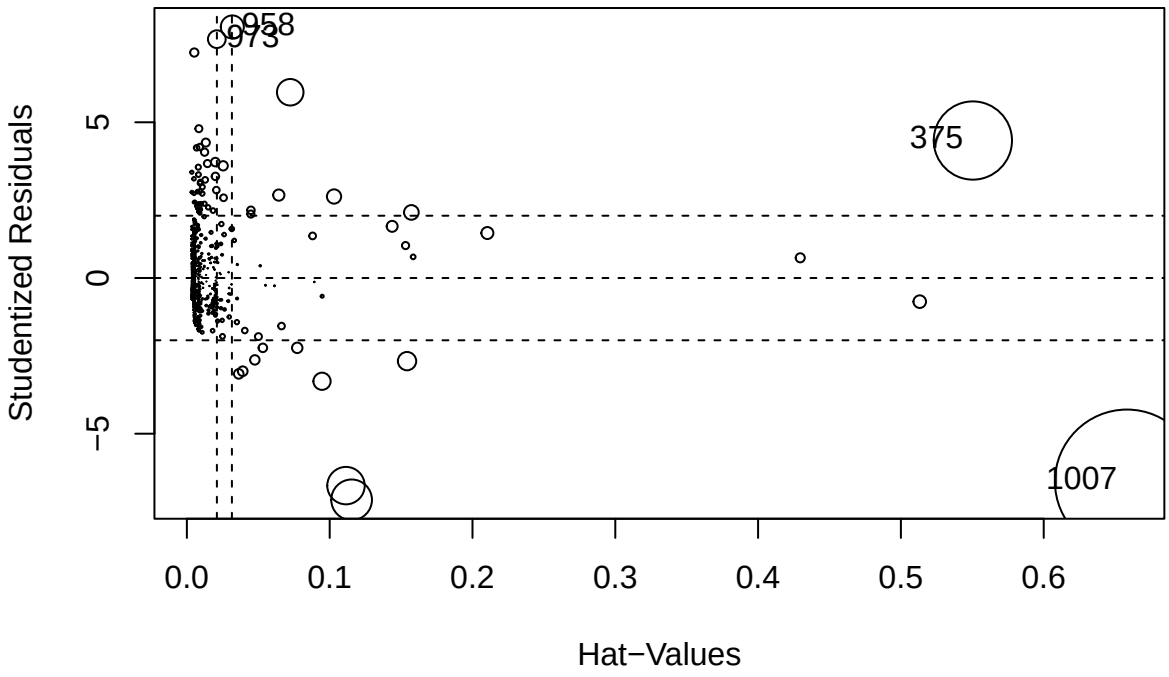

```
##  [1]   34  375  376  534  898  901  904  906  907  924  948  951  954  955  958
## [16]  963  964  973  989 1006 1007 1044 1045 1068 1138 1188
```

Se identifica varias observaciones atípicas.

```
##         cov.r   cook.d       hat
## 375  1.830313 1.807779 0.5503907
## 1007 1.892696 6.122975 0.6582908
```

Según la Distancia Cook (cook.d): Se consideran influyentes valores superiores a 1. Por lo cual, se identifican datos influyentes.

La visualización gráfica de las influencias se obtiene del siguiente modo:




```
##         StudRes        Hat      CookD
## 375   4.414503 0.55039070 1.80777911
## 958   8.062249 0.03182060 0.15616623
## 973   7.669533 0.02096748 0.09253318
## 1007 -6.536596 0.65829084 6.12297503
```

Los análisis muestran varias observaciones influyentes (posición 375 y 1007) que exceden los límites de preocupación para los valores de Leverages o Distancia Cook. Se recomienda realizar el análisis quitando estos datos atípicos.

Análisis de estabilidad de parámetros.

```
##
##  Chow test
##
## data:  df$Rendimiento ~ df$GastoInsumos + df$Credito + df$Riego + df$sup_sembrada_ha +     df$Produc
## F = 5.4605, p-value = 1.082e-09
```

Se puede concluir que no existe estabilidad paramétrica entre el Rendimiento y sus covariables.

5.Conclusión

- El modelo es capaz de explicar tan solo 50.22% de la variabilidad observada en el rendimiento, por lo que no es muy bueno para predecir el rendimiento.

- El test F muestra que es significativo (p-value: 2.2e-16).
- Se encontraron variables que no son significativas como (Credito1, sup_sembrada_ha, tamanoUA2 y GastoJornal).
- No se satisfacen los supuestos de la regresión.
- Se encontro dos observaciones (posición 375 y 1007) podrían estar influyendo de forma notable en el modelo.