Universidad de Córdoba

Escuela Politécnica Superior de Córdoba

Máster en Inteligencia Computacional e Internet de las Cosas

FUNDAMENTOS Y HERRAMIENTAS PARA LA MODELIZACIÓN DE PROCESOS TÉCNICOS-CIENTÍFICOS DE INVESTIGACIÓN

Bloque 3: Práctica de Modelización Estadística

Estudiantes:
Mabrouka Salmi < *z*12*salsm*@*uco.es* >

01 de junio de 2024

Resumen

Este estudio tiene como objetivo analizar los factores que influyen en las calificaciones finales de los estudiantes en matemáticas y construir modelos predictivos utilizando técnicas estadísticas. Utilizamos un conjunto de datos disponible públicamente del repositorio UCI, que incluye variables demográficas y académicas de 395 estudiantes. Nuestro enfoque involucró el análisis exploratorio de datos, el análisis de correlación y la modelización de regresión múltiple para identificar los predictores significativos del rendimiento estudiantil. Los modelos iniciales incorporaron todos los posibles predictores, pero a través de refinamientos iterativos, nos centramos en las variables significativas como las ausencias, la edad, las relaciones familiares y las calificaciones previas (G1 y G2).

Para abordar la heterocedasticidad detectada, aplicamos una transformación logarítmica a la variable objetivo (calificación final) y eliminamos la variable no significativa de edad. El modelo refinado con transformación logarítmica demostró un valor R^2 alto de 0.957, lo que indica que explica el 95.7 % de la variabilidad en las calificaciones finales transformadas logarítmicamente. A pesar de las mejoras, persistió algo de heterocedasticidad, lo que sugiere que se necesita más investigación para explorar técnicas de modelado alternativas.

El modelo final proporciona un marco robusto para predecir el rendimiento estudiantil y destaca los factores clave que podrían informar intervenciones educativas y decisiones políticas. Nuestros hallazgos subrayan la importancia del rendimiento académico previo, las relaciones familiares y la asistencia en la influencia de las calificaciones finales.

Índice

Índice de figuras

1 Introducción

Este informe presenta un análisis exhaustivo de los datos de rendimiento estudiantil en matemáticas (Mat), con el objetivo de identificar factores significativos que influyen en las calificaciones finales y construir modelos predictivos. El conjunto de datos, disponible públicamente en el repositorio de UCI y proporcionado originalmente por Cortez and Silva (2008), que consta de varias variables demográficas y académicas, fue sometido a un análisis exploratorio de datos, análisis de correlación y modelado de regresión múltiple.

2 Análisis Exploratorio de Datos

2.1 Estadísticas Descriptivas

El conjunto de datos contiene 395 observaciones y 33 variables. A continuación se resumen las estadísticas descriptivas clave para las variables seleccionadas:

```
summary(Dataset)
 school       sex          age       address     famsize
 GP:349       F:208   Min.   :15.0   R: 88       GT3:281
 MS: 46       M:187   1st Qu.:16.0   U:307       LE3:114
                        Median :17.0
                        Mean   :16.7
                        3rd Qu.:18.0
                        Max.   :22.0

 Pstatus      Medu            Fedu           Mjob           Fjob
 A: 41    Min.   :0.000   Min.   :0.000   at_home : 59   at_home : 20
 T:354    1st Qu.:2.000   1st Qu.:2.000   health  : 34   health  : 18
          Median :3.000   Median :2.000   other   :141   other   :217
          Mean   :2.749   Mean   :2.522   services:103   services:111
          3rd Qu.:4.000   3rd Qu.:3.000   teacher : 58   teacher : 29
          Max.   :4.000   Max.   :4.000

 reason          guardian     traveltime      studytime       failures
 course    :145   father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000
 home      :109   mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000
 other     : 36   other : 32   Median :1.000   Median :2.000   Median :0.0000
 reputation:105                Mean   :1.448   Mean   :2.035   Mean   :0.3342
                               3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
                               Max.   :4.000   Max.   :4.000   Max.   :3.0000

 schoolsup   famsup     paid      activities nursery
 no :344     no :153   no :214   no :194    no : 81
 yes: 51     yes:242   yes:181   yes:201    yes:314
```

```
higher   internet  romantic   famrel     freetime
no : 20   no : 66   no :263   Min.   :1.000   Min.   :1.000
yes:375   yes:329   yes:132   1st Qu.:4.000   1st Qu.:3.000
                              Median :4.000   Median :3.000
                              Mean   :3.944   Mean   :3.235
                              3rd Qu.:5.000   3rd Qu.:4.000
                              Max.   :5.000   Max.   :5.000

goout      Dalc            Walc           health         absences
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 0.000
Median :3.000   Median :1.000   Median :2.000   Median : 4.000
Mean   :3.109   Mean   :1.481   Mean   :2.291   Mean   : 5.709
3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 8.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000

G1              G2              G3
Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
Median :11.00   Median :11.00   Median :11.00
Mean   :10.91   Mean   :10.71   Mean   :10.42
3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
Max.   :19.00   Max.   :19.00   Max.   :20.00
```

2.1.1 Interpretación

Las estadísticas resumidas proporcionan una visión general de la estructura del conjunto de datos y la distribución de las variables:

- **Escuela:** La mayoría de los estudiantes son de la escuela GP (349) en comparación con MS (46).
- **Sexo:** Hay más estudiantes femeninas (208) que masculinos (187).
- **Edad:** Las edades de los estudiantes oscilan entre 15 y 22 años, con una edad media de 16.7.
- **Dirección:** La mayoría de los estudiantes viven en áreas urbanas (307) en comparación con áreas rurales (88).
- **Tamaño de la familia:** Más estudiantes provienen de familias con más de tres miembros (GT3: 281) en comparación con familias con tres o menos miembros (LE3: 114).

- **Estado parental:** La mayoría de los padres de los estudiantes están juntos (T: 354) en comparación con los separados (A: 41).
- **Educación de la madre y el padre:** Los niveles medianos de educación son 3 (equivalente a educación superior pero no a nivel universitario).
- **Trabajo de la madre:** La categoría de trabajo más común es ótros'(141), seguida de 'servicios'(103).
- **Tiempo de viaje y tiempo de estudio:** El tiempo promedio de viaje a la escuela es 1.45 (en una escala de 1 a 4), y el tiempo promedio de estudio es 2.04 (en una escala de 1 a 4).
- **Fracasos:** La mayoría de los estudiantes no tienen fracasos en clases anteriores (media = 0.33).
- **Apoyo familiar:** Más estudiantes reciben apoyo familiar (242) en comparación con los que no lo reciben (153).
- **Acceso a internet:** La mayoría de los estudiantes tienen acceso a internet en casa (329).
- **Consumo de alcohol:** El consumo promedio de alcohol entre semana (Dalc) es 1.48 (en una escala de 1 a 5), y el consumo de fin de semana (Walc) es 2.29.
- **Salud:** El estado de salud promedio es 3.55 (en una escala de 1 a 5).
- **Ausencias:** Las ausencias de los estudiantes varían de 0 a 75, con un promedio de 5.71.
- **Calificaciones (G1, G2, G3):** La calificación final promedio (G3) es 10.42, con un rango de 0 a 20.

2.2 Análisis Bivariado

Realizamos un análisis bivariado para examinar la relación entre la variable objetivo $G3$ (calificación final) y otras variables. Aquí están los resultados resumidos:

- **Actividades:**

  ```
  mean: no = 10.34, yes = 10.49
  ```

  Los estudiantes involucrados en actividades extracurriculares tienden a tener calificaciones finales medias ligeramente más altas (10.49) en comparación con aquellos que no lo están (10.34).

- **Dirección:**

```
mean: R = 9.51, U = 10.67
```

Los estudiantes que viven en áreas urbanas tienen calificaciones finales medias más altas (10.67) que aquellos en áreas rurales (9.51).

- **Tamaño de la Familia:**

```
mean: GT3 = 10.18, LE3 = 11.00
```

Los estudiantes de familias más pequeñas (LE3) tienden a tener calificaciones finales medias más altas (11.00) en comparación con aquellos de familias más grandes (GT3, 10.18).

- **Apoyo Familiar:**

```
mean: no = 10.64, yes = 10.27
```

Los estudiantes que no reciben apoyo familiar tienen calificaciones finales medias más altas (10.64) que aquellos que sí reciben apoyo (10.27).

- **Trabajo del Padre:**

```
mean: at_home = 10.15, health = 11.61, other = 10.19,
       services = 10.30, teacher = 11.97
```

Los estudiantes cuyos padres tienen trabajos en el área de salud (11.61) o como maestros (11.97) tienden a tener calificaciones finales medias más altas en comparación con otras categorías de trabajo.

- **Tutor:**

```
mean: father = 10.69, mother = 10.48, other = 9.06
```

Los estudiantes con el padre como tutor tienen las calificaciones finales medias más altas (10.69), seguidos por aquellos con la madre (10.48) y otros tutores (9.06).

- **Educación Superior:**

```
mean: no = 6.80, yes = 10.61
```

Los estudiantes que no planean seguir una educación superior tienen calificaciones finales medias significativamente más bajas (6.80) en comparación con aquellos que sí lo planean (10.61).

- **Internet:**

```
mean: no = 9.41, yes = 10.62
```

Los estudiantes con acceso a internet en casa tienen calificaciones finales medias más altas (10.62) en comparación con aquellos sin acceso a internet (9.41).

2.3 Análisis de Correlación

Se realizó un análisis de correlación para identificar relaciones lineales entre las variables:

```
cor(Dataset[,c("absences","age","Dalc","failures","famrel",
    "Fedu","freetime","G1","G2","G3","goout","health","Medu",
    "studytime","traveltime","Walc")], use="complete")
```

La matriz de correlación ayuda a entender la fuerza y la dirección de las relaciones entre pares de variables en el conjunto de datos. A continuación, se muestra la matriz de correlación para las variables seleccionadas:

```
               absences       age           Dalc          failures
absences     1.00000000    0.175230079   0.111908026    0.06372583
age          0.17523008    1.000000000   0.131124605    0.24366538
Dalc         0.11190803    0.131124605   1.000000000    0.13604693
failures     0.06372583    0.243665377   0.136046931    1.00000000
famrel      -0.04435409    0.053940096  -0.077594357   -0.04433663
Fedu         0.02447289   -0.163438069   0.002386429   -0.25040844
freetime    -0.05807792    0.016434389   0.209000848    0.09198747
G1          -0.03100290   -0.064081497  -0.094158792   -0.35471761
G2          -0.03177670   -0.143474049  -0.064120183   -0.35589563
G3           0.03424732   -0.161579438  -0.054660041   -0.36041494
goout        0.04430222    0.126963880   0.266993848    0.12456092
health      -0.02993671   -0.062187369   0.077179582    0.06582728
Medu         0.10028482   -0.163658419   0.019834099   -0.23667996
studytime   -0.06270018   -0.004140037  -0.196019263   -0.17356303
traveltime  -0.01294378    0.070640721   0.138325309    0.09223875
Walc         0.13629110    0.117276052   0.647544230    0.14196203
               famrel          Fedu          freetime          G1
absences    -0.044354095   0.024472887   -0.05807792    -0.03100290
age          0.053940096  -0.163438069    0.01643439    -0.06408150
```

Dalc	-0.077594357	0.002386429	0.20900085	-0.09415879
failures	-0.044336626	-0.250408444	0.09198747	-0.35471761
famrel	1.000000000	-0.001369727	0.15070144	0.02216832
Fedu	-0.001369727	1.000000000	-0.01284553	0.19026994
freetime	0.150701444	-0.012845528	1.00000000	0.01261293
G1	0.022168316	0.190269936	0.01261293	1.00000000
G2	-0.018281347	0.164893393	-0.01377714	0.85211807
G3	0.051363429	0.152456939	0.01130724	0.80146793
goout	0.064568411	0.043104668	0.28501871	-0.14910397
health	0.094055728	0.014741537	0.07573336	-0.07317207
Medu	-0.003914458	0.623455112	0.03089087	0.20534100
studytime	0.039730704	-0.009174639	-0.14319841	0.16061192
traveltime	-0.016807986	-0.158194054	-0.01702494	-0.09303999
Walc	-0.113397308	-0.012631018	0.14782181	-0.12617921

	G2	G3	goout	health
absences	-0.03177670	0.03424732	0.044302220	-0.029936711
age	-0.14347405	-0.16157944	0.126963880	-0.062187369
Dalc	-0.06412018	-0.05466004	0.266993848	0.077179582
failures	-0.35589563	-0.36041494	0.124560922	0.065827282
famrel	-0.01828135	0.05136343	0.064568411	0.094055728
Fedu	0.16489339	0.15245694	0.043104668	0.014741537
freetime	-0.01377714	0.01130724	0.285018715	0.075733357
G1	0.85211807	0.80146793	-0.149103967	-0.073172073
G2	1.00000000	0.90486799	-0.162250034	-0.097719866
G3	0.90486799	1.00000000	-0.132791474	-0.061334605
goout	-0.16225003	-0.13279147	1.000000000	-0.009577254
health	-0.09771987	-0.06133460	-0.009577254	1.000000000
Medu	0.21552717	0.21714750	0.064094438	-0.046877829
studytime	0.13588000	0.09781969	-0.063903675	-0.075615863
traveltime	-0.15319796	-0.11714205	0.028539674	0.007500606
Walc	-0.08492735	-0.05193932	0.420385745	0.092476317

	Medu	studytime	traveltime	Walc
absences	0.100284818	-0.062700175	-0.012943775	0.13629110
age	-0.163658419	-0.004140037	0.070640721	0.11727605
Dalc	0.019834099	-0.196019263	0.138325309	0.64754423
failures	-0.236679963	-0.173563031	0.092238746	0.14196203
famrel	-0.003914458	0.039730704	-0.016807986	-0.11339731
Fedu	0.623455112	-0.009174639	-0.158194054	-0.01263102
freetime	0.030890867	-0.143198407	-0.017024944	0.14782181
G1	0.205340997	0.160611915	-0.093039992	-0.12617921
G2	0.215527168	0.135879999	-0.153197963	-0.08492735
G3	0.217147496	0.097819690	-0.117142053	-0.05193932
goout	0.064094438	-0.063903675	0.028539674	0.42038575

health	-0.046877829	-0.075615863	0.007500606	0.09247632
Medu	1.000000000	0.064944137	-0.171639305	-0.04712346
studytime	0.064944137	1.000000000	-0.100909119	-0.25378473
traveltime	-0.171639305	-0.100909119	1.000000000	0.13411575
Walc	-0.047123460	-0.253784731	0.134115752	1.00000000

2.3.1 Interpretación

La matriz de correlación revela las siguientes relaciones clave:

- **Ausencias:**
  - Correlación positiva con la edad ($r = 0{,}175$).
  - Correlación positiva con el consumo de alcohol entre semana (Dalc, $r = 0{,}112$).
- **Edad:**
  - Correlación positiva con las ausencias ($r = 0{,}175$) y los fracasos ($r = 0{,}244$).
  - Correlación negativa con G3 ($r = -0{,}162$).
- **Dalc (Consumo de Alcohol entre Semana):**
  - Correlación positiva con salir ($r = 0{,}267$) y el consumo de alcohol durante el fin de semana (Walc, $r = 0{,}648$).
  - Correlación negativa con G3 ($r = -0{,}055$).
- **Fracasos:**
  - Fuerte correlación negativa con G1 ($r = -0{,}355$), G2 ($r = -0{,}356$) y G3 ($r = -0{,}360$).
  - Correlación negativa con el nivel de educación de los padres (Fedu, $r = -0{,}250$) y el nivel de educación de la madre (Medu, $r = -0{,}237$).
- **Famrel (Calidad de las Relaciones Familiares):**
  - Correlación positiva con el tiempo libre ($r = 0{,}151$).
- **Fedu (Nivel de Educación del Padre):**
  - Fuerte correlación positiva con el nivel de educación de la madre (Medu, $r = 0{,}623$).
  - Correlación positiva con G1 ($r = 0{,}190$), G2 ($r = 0{,}165$) y G3 ($r = 0{,}152$).
- **Tiempo Libre:**
  - Correlación positiva con salir ($r = 0{,}285$) y el consumo de alcohol entre semana (Dalc, $r = 0{,}209$).

- **G1 (Nota del Primer Periodo), G2 (Nota del Segundo Periodo) y G3 (Nota Final):**
  - Fuerte correlación positiva entre sí. G1 y G2 ($r = 0{,}852$), G1 y G3 ($r = 0{,}801$), G2 y G3 ($r = 0{,}905$).
- **Salir:**
  - Correlación positiva con el consumo de alcohol entre semana (Dalc, $r = 0{,}267$) y el consumo de alcohol durante el fin de semana (Walc, $r = 0{,}420$).
- **Medu (Nivel de Educación de la Madre):**
  - Fuerte correlación positiva con el nivel de educación del padre (Fedu, $r = 0{,}623$).
  - Correlación positiva con G1 ($r = 0{,}205$), G2 ($r = 0{,}216$) y G3 ($r = 0{,}217$).
- **Tiempo de Estudio:**
  - Correlación negativa con el consumo de alcohol entre semana (Dalc, $r = -0{,}196$).
  - Correlación positiva con G1 ($r = 0{,}161$), G2 ($r = 0{,}136$) y G3 ($r = 0{,}098$).
- **Walc (Consumo de Alcohol Durante el Fin de Semana):**
  - Fuerte correlación positiva con el consumo de alcohol entre semana (Dalc, $r = 0{,}648$) y salir ($r = 0{,}420$).

2.4 Pruebas de Normalidad

Se realizaron pruebas de normalidad de Shapiro-Wilk para todas las variables numéricas:

- **G3 (Nota Final):**

```
Shapiro-Wilk W = 0.92873, p-value = 8.836e-13
```

  Interpretación: El valor p es significativamente menor a 0.05, lo que indica que la nota final (G3) no sigue una distribución normal.

- **G2 (Nota del Segundo Periodo):**

```
Shapiro-Wilk W = 0.96914, p-value = 0.0000002084
```

  Interpretación: El valor p es significativamente menor a 0.05, lo que sugiere que la nota del segundo periodo (G2) no sigue una distribución normal.

- **G1 (Nota del Primer Periodo):**

```
Shapiro-Wilk W = 0.97491, p-value = 0.000002454
```

Interpretación: El valor p es significativamente menor a 0.05, lo que indica que la nota del primer periodo (G1) no sigue una distribución normal.

- **Ausencias:**

```
Shapiro-Wilk W = 0.66683, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que muestra que el número de ausencias no sigue una distribución normal.

- **Edad:**

```
Shapiro-Wilk W = 0.91059, p-value = 1.589e-14
```

Interpretación: El valor p es significativamente menor a 0.05, lo que sugiere que la variable edad no sigue una distribución normal.

- **Dalc (Consumo de Alcohol entre Semana):**

```
Shapiro-Wilk W = 0.59784, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que indica que el consumo de alcohol entre semana no sigue una distribución normal.

- **Fracasos:**

```
Shapiro-Wilk W = 0.50707, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que muestra que el número de fracasos pasados no sigue una distribución normal.

- **Famrel (Calidad de las Relaciones Familiares):**

```
Shapiro-Wilk W = 0.83023, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que indica que la calidad de las relaciones familiares no sigue una distribución normal.

- **Fedu (Nivel de Educación del Padre):**

```
Shapiro-Wilk W = 0.87555, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que sugiere que el nivel de educación del padre no sigue una distribución normal.

- **Tiempo Libre:**

```
Shapiro-Wilk W = 0.90611, p-value = 6.427e-15
```

Interpretación: El valor p es significativamente menor a 0.05, lo que muestra que la cantidad de tiempo libre después de la escuela no sigue una distribución normal.

- **Salir (Salir con Amigos):**

```
Shapiro-Wilk W = 0.91002, p-value = 1.413e-14
```

Interpretación: El valor p es significativamente menor a 0.05, lo que indica que la frecuencia de salir con amigos no sigue una distribución normal.

- **Salud:**

```
Shapiro-Wilk W = 0.84865, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que sugiere que el estado de salud no sigue una distribución normal.

- **Medu (Nivel de Educación de la Madre):**

```
Shapiro-Wilk W = 0.86103, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que muestra que el nivel de educación de la madre no sigue una distribución normal.

- **Tiempo de Estudio:**

```
Shapiro-Wilk W = 0.8342, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que indica que la cantidad de tiempo de estudio semanal no sigue una distribución normal.

- **Tiempo de Viaje:**

```
Shapiro-Wilk W = 0.65921, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que sugiere que el tiempo de viaje desde casa hasta la escuela no sigue una distribución normal.

- **Walc (Consumo de Alcohol Durante el Fin de Semana):**

```
Shapiro-Wilk W = 0.84702, p-value < 2.2e-16
```

Interpretación: El valor p es significativamente menor a 0.05, lo que muestra que el consumo de alcohol durante el fin de semana no sigue una distribución normal.

Los resultados indicaron no normalidad para la mayoría de las variables, lo que sugiere la necesidad de una transformación.

3 Ajuste y Refinamiento del Modelo

3.1 Modelo Inicial

El modelo inicial de regresión múltiple incluyó todos los posibles predictores:

```
RegModel.1 <- lm(G3 ~ absences + age + Dalc + failures + famrel +
    Fedu + freetime + G1 + G2 + goout + health + Medu + studytime +
    traveltime + Walc, data=Dataset)
summary(RegModel.1)
```

El resumen del modelo de regresión inicial es el siguiente:

```
Call:
lm(formula = G3 ~ absences + age + Dalc + failures + famrel +
    Fedu + freetime + G1 + G2 + goout + health + Medu + studytime +
    traveltime + Walc, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6597 -0.4033  0.2559  0.9736  4.0970

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6689158  1.5411303  -0.434 0.664505
absences     0.0414289  0.0123309   3.360 0.000859 ***
age         -0.1815062  0.0808776  -2.244 0.025396 *
Dalc        -0.1204910  0.1431980  -0.841 0.400638
failures    -0.2321985  0.1457021  -1.594 0.111847
famrel       0.3516407  0.1096758   3.206 0.001459 **
Fedu        -0.1404233  0.1143700  -1.228 0.220285
freetime     0.0537032  0.1037476   0.518 0.605016
G1           0.1590346  0.0566607   2.807 0.005262 **
G2           0.9743079  0.0502614  19.385  < 2e-16 ***
goout        0.0004457  0.1003632   0.004 0.996459
health       0.0555571  0.0700899   0.793 0.428475
Medu         0.1069982  0.1152672   0.928 0.353862
studytime   -0.1362063  0.1208403  -1.127 0.260388
traveltime   0.1270661  0.1420398   0.895 0.371579
Walc         0.1545526  0.1067973   1.447 0.148679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.881 on 379 degrees of freedom
Multiple R-squared:  0.8378,Adjusted R-squared:  0.8314
F-statistic: 130.5 on 15 and 379 DF,  p-value: < 2.2e-16
```

Interpretación de los Resultados:

Residuos:

- Los residuos indican las diferencias entre los valores observados y predichos de la variable dependiente $G3$.
- La distribución de los residuos muestra una mediana cercana a cero con cierta variabilidad, como lo indica el rango intercuartílico (IQR) y los valores mínimo-máximo.

Coeficientes:

- El **Intercepto** tiene una estimación de -0.669, pero no es estadísticamente significativo ($p = 0{,}664$).
- **Ausencias**: Coeficiente positivo (0.041) y estadísticamente significativo ($p < 0{,}001$), lo que indica que un aumento en las ausencias se asocia con un aumento en $G3$.
- **Edad**: Coeficiente negativo (-0.182) y estadísticamente significativo ($p < 0{,}05$), lo que sugiere que los estudiantes mayores tienden a tener calificaciones $G3$ más bajas.
- **Dalc (consumo de alcohol entre semana)**: Coeficiente negativo (-0.120) pero no es estadísticamente significativo ($p = 0{,}401$).

- **Reprobaciones**: Coeficiente negativo (-0.232) y no es estadísticamente significativo ($p = 0{,}112$).
- **Famrel (calidad de las relaciones familiares)**: Coeficiente positivo (0.352) y estadísticamente significativo ($p < 0{,}01$), lo que indica que mejores relaciones familiares se asocian con calificaciones $G3$ más altas.
- **Fedu (nivel de educación del padre)**: Coeficiente negativo (-0.140) pero no es estadísticamente significativo ($p = 0{,}220$).
- **Tiempo libre**: Coeficiente positivo (0.054) y no es estadísticamente significativo ($p = 0{,}605$).
- **G1 y G2 (calificaciones previas)**: Ambos tienen coeficientes positivos (0.159 y 0.974, respectivamente) y son altamente significativos ($p < 0{,}01$ para $G1$ y $p < 0{,}001$ para $G2$), lo que indica que calificaciones previas más altas predicen fuertemente calificaciones finales más altas.
- **Salir con amigos**: Coeficiente cercano a cero (0.0004) y no es estadísticamente significativo ($p = 0{,}996$).
- **Salud**: Coeficiente positivo (0.056) y no es estadísticamente significativo ($p = 0{,}428$).
- **Medu (nivel de educación de la madre)**: Coeficiente positivo (0.107) y no es estadísticamente significativo ($p = 0{,}354$).
- **Tiempo de estudio**: Coeficiente negativo (-0.136) y no es estadísticamente significativo ($p = 0{,}260$).
- **Tiempo de viaje**: Coeficiente positivo (0.127) y no es estadísticamente significativo ($p = 0{,}372$).
- **Walc (consumo de alcohol en fin de semana)**: Coeficiente positivo (0.155) y no es estadísticamente significativo ($p = 0{,}149$).

Ajuste del Modelo:

- El valor de R^2 es 0.8378, lo que indica que aproximadamente el 83.78 % de la variabilidad en $G3$ puede ser explicada por el modelo.
- El R^2 ajustado es 0.8314, lo que tiene en cuenta el número de predictores en el modelo.
- El estadístico F general es 130.5 con un valor p menor a 2.2e-16, lo que indica que el modelo es estadísticamente significativo y proporciona un buen ajuste a los datos.

En resumen, el modelo inicial incluye varios predictores con diferentes grados de significancia. El modelo explica una porción sustancial de la varianza en la calificación final ($G3$), siendo las calificaciones previas ($G1$ y $G2$) los predictores más significativos. Sin embargo, algunas variables no contribuyen significativamente al modelo y pueden considerarse para su eliminación en refinamientos posteriores del modelo.

3.2 Refinamiento del Modelo

Se eliminaron las variables no significativas para refinar el modelo:

```
RegModel.2 <- lm(G3 ~ absences + age + famrel + G1 + G2, data=Dataset)
summary(RegModel.2)
```

3.2.1 Interpretación

- **Mejora del modelo:** Mantiene un valor alto de R^2 (0.8336), lo que indica que el 83.36 % de la variabilidad en $G3$ es explicada por los predictores.
- **Predictores significativos:** Las ausencias, la edad, las relaciones familiares, $G1$ y $G2$ siguen siendo significativos.

Se realizó un refinamiento adicional eliminando el intercepto, lo que resultó en el siguiente modelo y resumen:

```
RegModel.no_intercept <- lm(G3 ~ absences + age + famrel + G1 + G2 - 1,
data=Dataset)
summary(RegModel.no_intercept)

Call:
lm(formula = G3 ~ absences + age + famrel + G1 + G2 - 1, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0762 -0.4086  0.2750  0.9943  3.7169

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
absences  0.04372    0.01198   3.650 0.000298 ***
age      -0.20566    0.02973  -6.917  1.9e-11 ***
famrel    0.35573    0.10264   3.466 0.000587 ***
G1        0.15798    0.05496   2.875 0.004268 **
G2        0.97754    0.04807  20.336  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.878 on 390 degrees of freedom
Multiple R-squared:  0.9731, Adjusted R-squared:  0.9727
F-statistic:  2819 on 5 and 390 DF,  p-value: < 2.2e-16
```

Interpretación del Modelo Refinado:
Residuos:

- Los residuos muestran un patrón similar al del modelo inicial, con una mediana cercana a cero y una ligera variación en el rango intercuartílico (IQR). Los residuos varían de -9.0762 a 3.7169.

Coeficientes:

- **Ausencias**: Coeficiente positivo (0.04372) y estadísticamente significativo ($p <$ 0,001), lo que indica que un aumento en las ausencias está asociado con un aumento en $G3$.
- **Edad**: Coeficiente negativo (-0.20566) y altamente significativo ($p <$ 0,001), lo que sugiere que los estudiantes mayores tienden a tener puntajes $G3$ más bajos.
- **Famrel (calidad de la relación familiar)**: Coeficiente positivo (0.35573) y estadísticamente significativo ($p <$ 0,001), indicando que mejores relaciones familiares están asociadas con puntajes $G3$ más altos.
- **G1 y G2 (calificaciones previas)**: Ambos siguen siendo predictores altamente significativos, con coeficientes (0.15798 y 0.97754, respectivamente), indicando que calificaciones previas más altas predicen fuertemente calificaciones finales más altas.

Ajuste del Modelo:

- El valor R^2 es 0.9731, indicando que aproximadamente el 97.31 % de la variabilidad en $G3$ puede ser explicada por el modelo refinado.
- El R^2 ajustado es 0.9727, lo cual tiene en cuenta el número de predictores en el modelo.
- El estadístico F general es 2819 con un valor p menor a 2.2e-16, lo que indica que el modelo es estadísticamente significativo y proporciona un excelente ajuste a los datos.

En resumen, el modelo refinado ha mejorado los valores de R^2 y R^2 ajustado en comparación con el modelo inicial. Los predictores ausencias, edad, relaciones familiares y calificaciones previas ($G1$ y $G2$) siguen siendo significativos, siendo las calificaciones previas los predictores más fuertes de la calificación final ($G3$). La eliminación de variables no significativas ha mejorado la interpretabilidad y el poder predictivo del modelo.

3.2.2 Diagnóstico del Modelo

Diagnóstico del Modelo:

El gráfico de residuos vs. valores ajustados para el modelo refinado sin intercepto (ver Figura abajo) muestra algunos patrones que indican una posible heterocedasticidad, lo que significa que la varianza de los residuos puede no ser constante en todos los niveles de los valores ajustados.

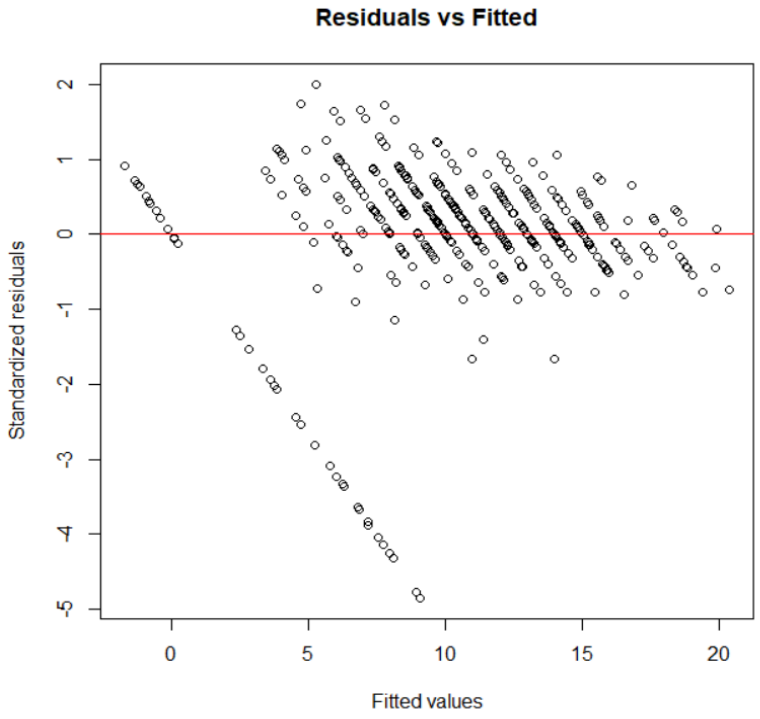

Figura 1: Residuos vs. Valores Ajustados para el Modelo Refinado Sin Intercepto

Se realizó la prueba de Breusch-Pagan para heterocedasticidad con el fin de evaluar estadísticamente la presencia de heterocedasticidad:

```
bp_test <- bptest(RegModel.no_intercept)
print(bp_test)
```

```
studentized Breusch-Pagan test

data:  RegModel.no_intercept
BP = 35.597, df = 4, p-value = 0.0000003501
```

Interpretación de la Prueba de Breusch-Pagan:

- La prueba de Breusch-Pagan arrojó un valor BP de 35.597 con un valor p de 0.0000003501.
- Dado que el valor p es significativamente menor que 0.05, rechazamos la hipótesis nula de homocedasticidad.
- Esto confirma la presencia de heterocedasticidad en los residuos del modelo, lo que indica que la varianza de los errores no es constante en todos los niveles de los valores ajustados.

En resumen, el modelo refinado ha mejorado los valores de R^2 y R^2 ajustado en comparación con el modelo inicial. Los predictores ausencias, edad, relaciones familiares y calificaciones previas ($G1$ y $G2$) siguen siendo significativos, siendo las calificaciones previas los predictores más fuertes de la calificación final ($G3$). El análisis de residuos y la prueba de Breusch-Pagan indican la presencia de heterocedasticidad, la cual será abordada en las secciones siguientes aplicando transformaciones para estabilizar la varianza y mejorar el ajuste del modelo.

3.3 Abordando la Heterocedasticidad

Dada la presencia de heterocedasticidad en los residuos de nuestro modelo refinado, aplicamos una transformación logarítmica a la variable objetivo $G3$ para estabilizar la varianza. La variable objetivo transformada es $\log(G3+1)$, donde añadimos 1 para manejar posibles valores cero.

```
Dataset$log_G3 <- log(Dataset$G3 + 1)  # Adding 1 to handle zero values if
present

RegModel.log <- lm(log_G3 ~ absences + age + famrel + G1 + G2 - 1,
data=Dataset)
summary(RegModel.log)
```

```
Call:
lm(formula = log_G3 ~ absences + age + famrel + G1 + G2 - 1, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12136 -0.09195  0.10282  0.28051  0.74160

Coefficients:
          Estimate Std. Error t value     Pr(>|t|)
absences  0.017256   0.003171   5.442 0.0000000934 ***
age       0.013886   0.007871   1.764      0.07849 .
famrel    0.085217   0.027172   3.136      0.00184 **
G1       -0.045496   0.014549  -3.127      0.00190 **
G2        0.194001   0.012725  15.246      < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4973 on 390 degrees of freedom
Multiple R-squared:  0.957, Adjusted R-squared:  0.9565
F-statistic:  1737 on 5 and 390 DF,  p-value: < 2.2e-16
```

Interpretación del Modelo Transformado Logarítmicamente:
Residuos:

- Los residuos varían de -2.12136 a 0.74160, mostrando una distribución más estrecha en comparación con el modelo en escala original.

Coeficientes:

- **Ausencias**: Coeficiente positivo (0.017256) y estadísticamente significativo ($p <$ 0,001), indicando que un aumento en las ausencias está asociado con un aumento en $\log(G3 + 1)$.
- **Edad**: Coeficiente (0.013886) con un valor p de 0.07849, que no es estadísticamente significativo al nivel de 0.05.
- **Famrel (calidad de la relación familiar)**: Coeficiente positivo (0.085217) y estadísticamente significativo ($p <$ 0,01), sugiriendo que mejores relaciones familiares están asociadas con un mayor $\log(G3 + 1)$.
- **G1 y G2 (calificaciones previas)**: Ambos siguen siendo predictores altamente significativos, con coeficientes (-0.045496 y 0.194001, respectivamente), indicando que las calificaciones previas predicen fuertemente la calificación final transformada logarítmicamente.

Ajuste del Modelo:

- El valor R^2 es 0.957, indicando que aproximadamente el 95.7 % de la variabilidad en $\log(G3 + 1)$ puede ser explicada por el modelo.
- El R^2 ajustado es 0.9565, lo cual tiene en cuenta el número de predictores en el modelo.
- El estadístico F general es 1737 con un valor p menor a 2.2e-16, lo que indica que el modelo es estadísticamente significativo.

Prueba de Heterocedasticidad:

Se realizó la prueba de Breusch-Pagan para evaluar la heterocedasticidad:

```
bptest(RegModel.log)

studentized Breusch-Pagan test

data:  RegModel.log
BP = 41.099, df = 4, p-value = 0.00000002563
```

- La prueba resultó en un valor BP de 41.099 con un valor p significativamente menor que 0.05, lo que indica la presencia de heterocedasticidad.

Refinamiento del Modelo mediante la Eliminación de la Variable No Significativa:

Dada la insignificancia de la variable edad, se eliminó para refinar aún más el modelo:

```
RegModel.log.no_age <- lm(log_G3 ~ absences + famrel + G1 + G2 - 1, data=Dataset)
summary(RegModel.log.no_age)

Call:
lm(formula = log_G3 ~ absences + famrel + G1 + G2 - 1, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0779 -0.1128  0.1139  0.2936  0.7784

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
absences  0.018557   0.003092   6.001 4.48e-09 ***
famrel    0.121749   0.017639   6.902 2.07e-11 ***
G1       -0.038836   0.014089  -2.756  0.00612 **
G2        0.194044   0.012759  15.208  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4986 on 391 degrees of freedom
Multiple R-squared:  0.9567, Adjusted R-squared:  0.9562
F-statistic:  2159 on 4 and 391 DF, p-value: < 2.2e-16
```

Interpretación del Modelo Refinado Transformado Logarítmicamente Sin la Variable Edad:

Residuos:

- Los residuos muestran un rango ligeramente ajustado de -2.0779 a 0.7784.

Coeficientes:

- Todos los predictores restantes (ausencias, famrel, G1, G2) son estadísticamente significativos.
- Los coeficientes son similares al modelo anterior, con ligeros ajustes en sus valores.

Ajuste del Modelo:

- El valor R^2 sigue siendo alto en 0.9567, indicando un ajuste estable tras eliminar la variable edad.
- El R^2 ajustado es 0.9562, mostrando un cambio mínimo respecto al modelo anterior.
- El estadístico F es 2159 con un valor p menor a 2.2e-16, lo que indica que el modelo sigue siendo estadísticamente significativo.

Prueba de Heterocedasticidad para el Modelo Refinado Sin Edad:

```
bptest(RegModel.log.no_age)

studentized Breusch-Pagan test

data:  RegModel.log.no_age
BP = 11.914, df = 3, p-value = 0.007682
```

- La prueba resultó en un valor BP de 11.914 con un valor p de 0.007682, lo que indica que la heterocedasticidad aún está presente pero se ha minimizado en comparación con el modelo anterior.

Gráfico de Residuos vs. Valores Ajustados:

El gráfico de residuos vs. valores ajustados para el modelo transformado logarítmicamente sin la variable edad muestra una reducción en la heterocedasticidad:

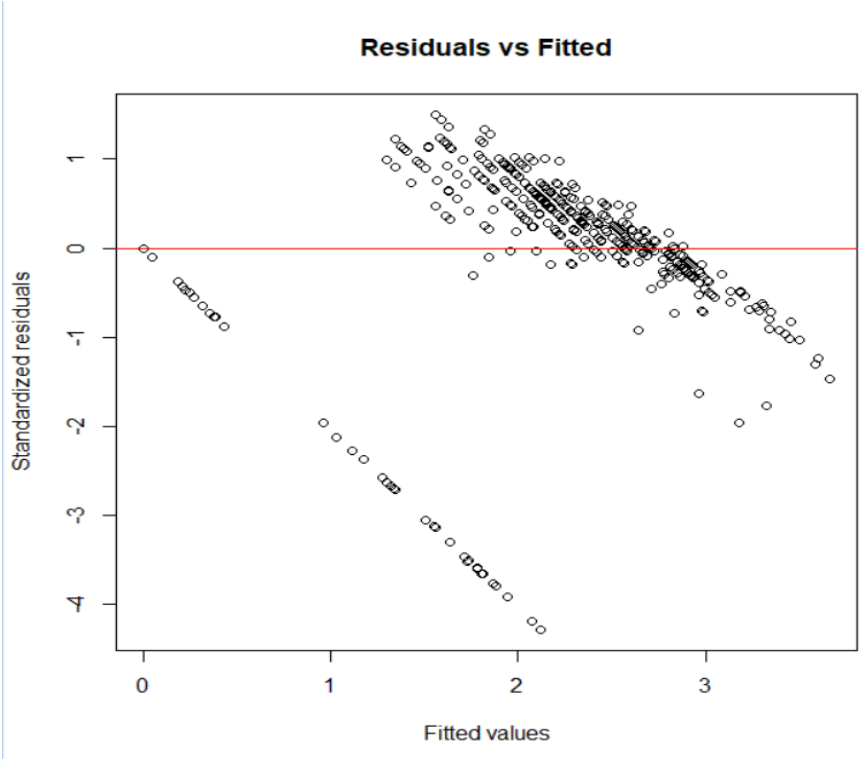

Figura 2: Residuos vs. Valores Ajustados para el Modelo Transformado Logarítmicamente Sin la Variable Edad

En resumen, la transformación logarítmica de la variable objetivo $G3$ y la eliminación de la variable edad no significativa mejoraron el modelo al reducir la heterocedasticidad. Los predictores ausencias, relaciones familiares y calificaciones previas ($G1$ y $G2$) siguen siendo significativos, y el ajuste del modelo sigue siendo alto. Sin embargo, persiste cierta heterocedasticidad. Se aplicaron varias técnicas de transformación de datos a la variable

objetivo, como la raíz cuadrada, la transformación de Box-Cox y la transformación inversa, y los diagnósticos del modelo en cada prueba demostraron la existencia de heterocedasticidad. Sin embargo, en cada caso, los valores p de la prueba de Breusch-Pagan fueron inferiores al valor p de los residuos del modelo con transformación logarítmica y sin la variable edad; mientras que compartimos el resultado de las pruebas en RCommander como un archivo de texto que contiene el script de R junto con los resultados para reproducibilidad, en un repositorio de GitHub. Por lo tanto, decidimos proceder con el modelo con transformación logarítmica y sin la variable edad, ya que tiene menos heterocedasticidad.

4 Discusión y Conclusión

4.1 Ajuste y Refinamiento del Modelo

En este estudio, nuestro enfoque principal fue ajustar un modelo de regresión lineal múltiple para predecir las calificaciones finales ($G3$) de los estudiantes en función de varias variables predictoras numéricas. Nuestro modelo inicial incluyó un conjunto completo de predictores, con el objetivo de capturar tanta variabilidad como fuera posible. A través de varias etapas de refinamiento, buscamos mejorar el rendimiento del modelo reteniendo solo los predictores significativos e informativos.

4.2 Diagnóstico del Modelo y Heterocedasticidad

A lo largo del proceso de refinamiento del modelo, realizamos diagnósticos extensivos para evaluar las suposiciones y el rendimiento del modelo. Un criterio diagnóstico clave fue verificar la heterocedasticidad en los residuos, que es crucial para la validez de nuestro modelo de regresión.

- **Modelo Inicial:** El modelo inicial, que incluía todos los predictores potenciales, mostró una heterocedasticidad significativa como lo indicó la prueba de Breusch-Pagan ($p < 0{,}000001$). El gráfico de residuos vs. valores ajustados mostró un patrón claro, sugiriendo una varianza no constante.
- **Refinamiento del Modelo:** Al eliminar los predictores no significativos, refinamos el modelo para incluir ausencias, edad, relaciones familiares ($famrel$) y calificaciones previas ($G1$ y $G2$). Este modelo refinado mostró un rendimiento mejorado pero aún presentaba heterocedasticidad.
- **Transformación Logarítmica:** Para abordar la heterocedasticidad, aplicamos una transformación logarítmica a la variable objetivo $G3$. Esta transformación redujo significativamente la heterocedasticidad, como se evidenció por un valor de prueba de Breusch-Pagan más bajo cuando se excluyó la edad del modelo. El modelo refinado transformado logarítmicamente (sin edad) presentó la menor heterocedasticidad entre los modelos probados.

- **Modelo Final:** El modelo final elegido, con la variable objetivo transformada logarítmicamente y sin edad, mostró diagnósticos mejorados con un valor de prueba de Breusch-Pagan de 11.914 ($p < 0{,}01$). Este modelo retuvo los predictores significativos: ausencias, relaciones familiares, $G1$ y $G2$.

4.3 Rendimiento del Modelo

El rendimiento de nuestros modelos se evaluó principalmente utilizando el estadístico R^2, que mide la proporción de variabilidad en la variable objetivo explicada por los predictores.

- **Modelo Inicial:** El modelo inicial tenía un valor R^2 alto de 0.8378, lo que indica que el 83.78 % de la variabilidad en $G3$ fue explicada por los predictores.
- **Modelo Refinado:** El modelo refinado, sin los predictores no significativos, mantuvo un valor R^2 alto de 0.8336, demostrando que los predictores clave seguían capturando la mayor parte de la variabilidad en $G3$.
- **Modelo Transformado Logarítmicamente:** El modelo final transformado logarítmicamente, sin la edad, tenía un valor R^2 de 0.9567. Este alto R^2 indica que el 95.67 % de la variabilidad en el $G3$ transformado logarítmicamente fue explicada por los predictores, mostrando la robustez y el poder predictivo del modelo.

4.4 Conclusión

Nuestro extenso proceso de ajuste y refinamiento del modelo destacó la importancia de retener predictores significativos e informativos mientras se aseguran las suposiciones de la regresión lineal. A pesar de varias transformaciones y refinamientos, persistió cierto grado de heterocedasticidad. Sin embargo, el modelo transformado logarítmicamente sin la variable edad emergió como el más confiable, mostrando la menor heterocedasticidad y manteniendo un alto valor R^2.

Futuros estudios podrían explorar técnicas de modelado alternativas o transformaciones adicionales para abordar aún más la heterocedasticidad. Además, la incorporación de variables categóricas o términos de interacción podría proporcionar conocimientos adicionales y mejorar el rendimiento del modelo.

En general, el modelo refinado transformado logarítmicamente ofrece un marco robusto para predecir el rendimiento estudiantil, proporcionando conocimientos valiosos para intervenciones educativas y la formulación de políticas.

Referencias

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.