Trabajo Práctico Nº 2

Regresión Lineal Múltiple

Presentado en la fecha: 09/09/2023

Hecho por: Huarca Brian

Nicolas Benitez

Facundo Rodriguez

Derlis Walter Hodge

Yasmin Salvi

Contents

Resumen

El presente informe detalla los pasos necesarios para generar un modelo de regresión lineal múltiple en el que dado un conjunto de variables permita estimar el valor de una variable dependiente, ademas busca concientizar sobre el daño producido y el que se sigue produciendo al riachuelo, y la forma en el que dicho daño repercute en nuestras vidas.
Para ello se busca replicar mediante un modelo de Regresion Lineal Multiple los distintos factores que tienen incidencia en la calidad de vida de las personas que viven al rededor del Riachuelo.

Sumario

0.1 Sumario

- Preparación del dataset (Valores Null, Outliers).
- Análisis de la relación entre variables(Correlograma).
- Correlación entre posibles predictores.
- Análisis sobre la influencia de las variables categóricas en relación al indice de calidad de vida.
- Elección de variables para el Modelo de Regresión Lineal.
- Intervalos de confianza.
- Ejemplo de Estimación o Predicción.

Objetivo

El informe tiene por objetivo concientizar, entender y comprender mediante un modelo predictivo la calidad de vida de las personas que viven cerca los riachuelo.

Dataset

1.1 Informacion del dataset

Este documento presenta, de forma resumida el grado en que la población, de un territorio específico, logra disponer de recursos socioeconómicos, culturales, de infraestructura, y ambientales para satisfacer una variada gama de necesidades humanas que posibilitensu desarrollo integral e incrementen sus posibilidades para elegir trayectorias vitales significativas en un marco de equidad. La metodología de cálculo para la construcción (y posterior actualización) del índice de referencia, desarrollado por la Unidad de Investigación, Desarrollo, Extensión y Transferencia Gestión Ambiental, perteneciente a la Facultad de Ingeniería de la Universidad Nacional de La Plata (UNLP).

EstructuraLa propuesta de ICV para la Cuenca Hídrica Matanza Riachuelo (CHMR)posee la estructura general que se muestra en la Tabla 1.

ICV para la CHMR	
Dimensión	**Variable/indicador**
VIVIENDA	Hacinamiento Cobertura de gas por red Calidad constructiva de la vivienda Certeza de uso de dominio
SALUD PÚBLICA	Disponibilidad de Centros de Atención Primaria Servicios sanitarios básicos Áreas de disposición de residuos
EDUCACIÓN	Años esperados de educación Años promedio de educación
ENTORNO	Accesibilidad a espacios verdes públicos Presencia de cavas Transporte público Presencia de industrias Riesgo de inundación

Figure 1: Dimensiones y variables/indicadoresque componen el ICV para la CHMR

DOCUMENTACION ICV: Palabras Clave y Atributos a manejar en cada dataset

Dataset ICV:

- sup_ha:Superficie en hectáreas (ha) del radio censal.
- pobl_tot:Población del radio censal según el CNPHV 2010.
- dens_ha:Densidad poblacional bruta del radio censal expresada en Hab/Ha.
- icv:Valor que adquiere el Índice de Calidad de Vida.

Dataset EDUCACION (dim_edu_red)

- a_esp: Años esperados de escolarización que un niño puede recibir con tasas de matriculación por edad se mantuvieran constantes toda su vida.
- a_prom: Años promedio de educación promedio de la población que en teoría está fuera del ciclo lectivo.

Fuente: [Destacar fuente]

Dataset SALUD (dim_salud_red)

- ind_efe_sa: Indicador Efectores de Salud: porcentaje de población que vive a una distancia aceptable de un Centro de Atención Primaria (CAP). Definiendo áreas de influencia de 1000m para cada uno de los CAP.
- ind_ssb:Indicador Servicios Sanitarios Básicos. Está definido como el porcentaje de población que cuenta en sus viviendas con los servicios de agua de red y desagüe cloacal a red pública en forma simultánea.
- ind_rsu:Indicador Disposición de residuos. Población que no vive dentro del área de influencia de basurales (punto de arrojo, microbasural, basural, macrobasural, según definición de ACUMAR).

Fuente: [Destacar Fuente]

Dataset ENTORNO (dim_entorno_red)

- ind_trpu:Indicador Transporte público. Acceso de la población en la CHMR. Mejor situación: vivir a $<$ 300 m del recorrido de una línea de transporte inter/urbano o $<$ 1000 m de una estación de tren.

- ind_inunda:Indicador Riesgo por inundación: Determinación del total de riesgos a las personas causados por una inundación, metodología del Department for Environment Food and Rural Affairs del Reino Unido.

- ind_evp:Indicador Espacios verdes públicos. Categorización según su superficie. Plaza: 10.000 a 40.000 m2; Parque a escala urbana: 40.000 a 200.000 m2; Parque a escala metropolitana: más de 200.000 m2.

- ind_cavas:Indicador Cavas. En función de los riesgos que implican, se tomó como zona de influencia una distancia de 600 m. Mejor Situación: vivir a $>$ de 600 m de una cava.

- ind_indus:Cantidad de población que reside en cercanías a una industria considerada riesgosa, según lo establecido en la Resolución ACUMAR 12/2019 y/o la que en un futuro la reemplace.

Dataset VIVIENDA (dim_vivienda_red)

- ind_hac: indicador de Hacinamiento.

- ind_gr: Indicador de Gas por Red.

- ind_matv: Indicador Calidad de Vivienda.

- ind_dom: Indicador Certeza del Dominio.

Fuente: [Destacar Fuente]

1.2 Valores Null

Verificamos que no existiesen valores Null (desconocidos) en el dataset.

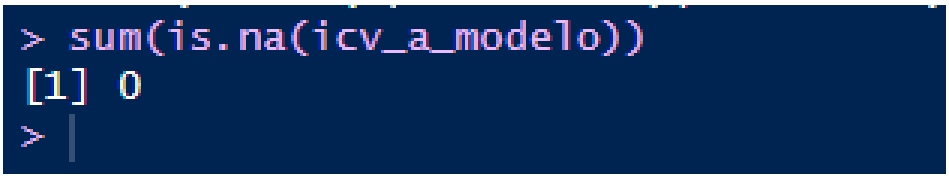


Figure 2: Cantidad de Valores Null

1.3 Manejo de Outliers

Como podemos observar existe una gran cantidad de outliers en los datos. Procedemos a eliminarlos para obtener un modelo más preciso.

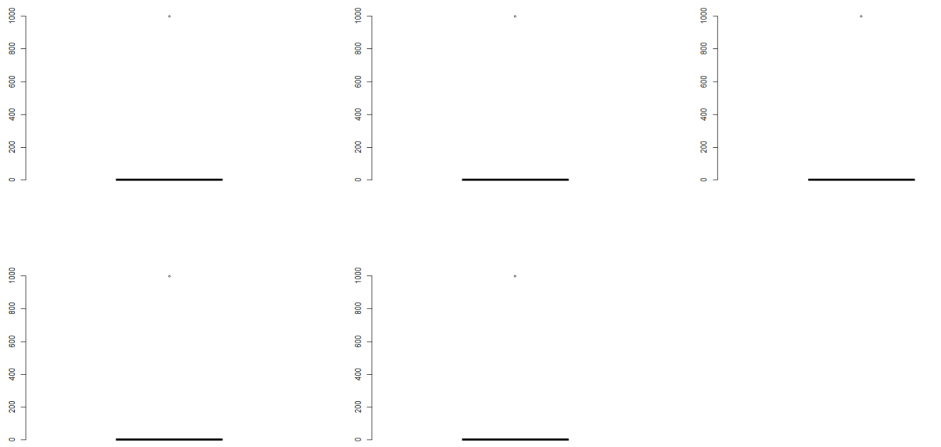


Figure 3: Boxplots con outliers

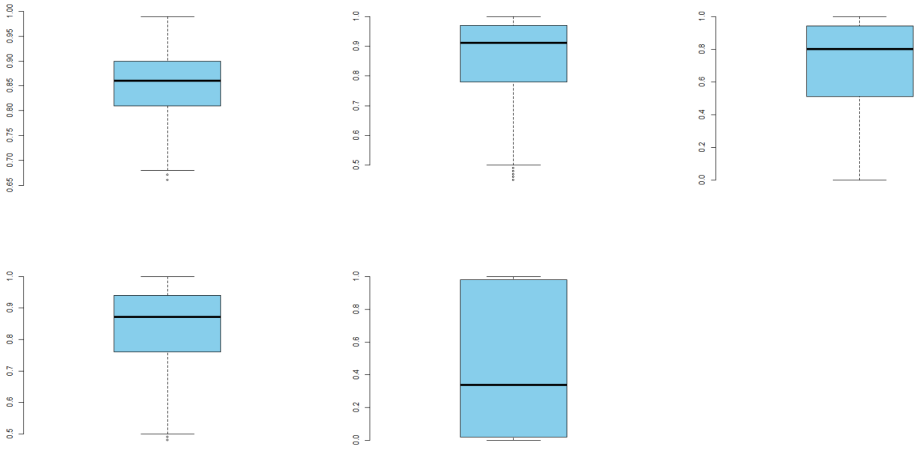


Figure 4: Boxplots sin outliers

1.4 Análisis de la relación entre variables

1.4.1 Summary del dataframe

Se hace un visualizacion rápida de la distribución de las variables. Con esta información, podemos obtener una idea de la media o mediana, la dispersión, rango entre el mínimo y el máximo y si hay valores outliers.

```
     sup_ha          pobl_tot        dens_ha            icv          edu_escolar_ninos
Min.   :  0.33   Min.   : 144.0   Min.   :  10.00   Min.   : 0.2700   Min.   : 8.34
1st Qu.:  6.54   1st Qu.: 805.2   1st Qu.:  68.58   1st Qu.: 0.6100   1st Qu.:12.66
Median : 10.63   Median :1025.5   Median : 101.72   Median : 0.6900   Median :13.34
Mean   : 14.85   Mean   :1144.7   Mean   : 141.50   Mean   : 0.7605   Mean   :13.45
3rd Qu.: 16.52   3rd Qu.:1381.0   3rd Qu.: 149.31   3rd Qu.: 0.8200   3rd Qu.:14.21
Max.   :328.03   Max.   :5895.0   Max.   :8375.76   Max.   :23.1700   Max.   :17.25
edu_escolar_adul salud_ate_prim   salud_serv_sanit  salud_disp_rsu    entor_acceso_transp
Min.   : 5.990   Min.   :0.0000   Min.   :  0.000   Min.   :-0.4707   Min.   :0.0000
1st Qu.: 7.710   1st Qu.:0.6300   1st Qu.:  0.020   1st Qu.: 1.0000   1st Qu.:0.8400
Median : 8.990   Median :1.0000   Median :  0.380   Median : 1.0000   Median :1.0000
Mean   : 9.271   Mean   :0.7696   Mean   :  5.277   Mean   : 0.9548   Mean   :0.8358
3rd Qu.:10.648   3rd Qu.:1.0000   3rd Qu.:  0.980   3rd Qu.: 1.0000   3rd Qu.:1.0000
Max.   :14.470   Max.   :1.0000   Max.   :999.000   Max.   : 1.0000   Max.   :1.0000
 entor_inunda    entor_area_verde  entor_cavas       entor_indus     vivi_hacinamiento
Min.   :0.0000   Min.   :0.0000   Min.   :0.05139   Min.   :0.0000   Min.   :  0.290
1st Qu.:1.0000   1st Qu.:0.9748   1st Qu.:1.00000   1st Qu.:0.2400   1st Qu.:  0.750
Median :1.0000   Median :1.0000   Median :1.00000   Median :0.5000   Median :  0.860
Mean   :0.8433   Mean   :0.8182   Mean   :0.99702   Mean   :0.4856   Mean   :  3.094
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.00000   3rd Qu.:0.6700   3rd Qu.:  0.940
Max.   :1.0000   Max.   :1.0028   Max.   :1.00000   Max.   :1.0000   Max.   :999.000
 vivi_gas_red     vivi_calid_casa    vivi_dominio
Min.   :  0.000   Min.   :  0.110   Min.   :  0.270
1st Qu.:  0.470   1st Qu.:  0.760   1st Qu.:  0.800
Median :  0.780   Median :  0.900   Median :  0.860
Mean   :  2.926   Mean   :  3.105   Mean   :  3.101
3rd Qu.:  0.940   3rd Qu.:  0.970   3rd Qu.:  0.900
Max.   :999.000   Max.   :999.000   Max.   :999.000
>
```

Figure 5: Summary del dataframe

1.4.2 Matriz de correlación

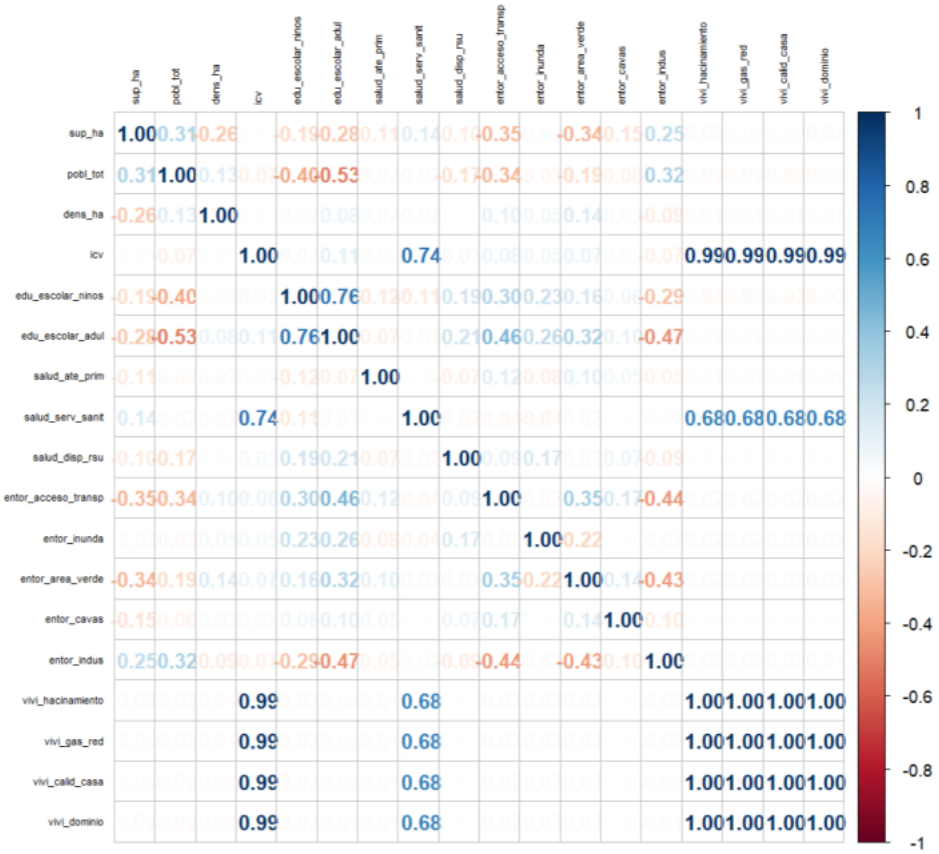


Figure 6: Matriz de correlación entre variables

De la matriz podemos observar que nuestra variable dependiente (icv) se encuentra altamente correlacionada con las variables de la dimension de vivienda. y con la variable salud servicios sanitarios basicos

Al mismo tiempo, estas variables se encuentran altamente correlacionadas entre sí.

Las demas variables no ofrecen una correlación importante. A primera vista podemos identificar alguna variables que no van a portar a la prediccion

1.4.3 Summary del modelo de regresion

```
Call:
lm(formula = icv ~ ., data = icv_a_modelo)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55628 -0.02929  0.00031  0.02912  1.03725

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.615e-02  3.188e-02   2.075  0.03807 *
sup_ha             -3.353e-04  7.398e-05  -4.532 6.04e-06 ***
pobl_tot            8.285e-06  2.553e-06   3.246  0.00118 **
dens_ha             2.637e-05  5.308e-06   4.968 7.08e-07 ***
edu_escolar_ninos  -8.362e-04  1.474e-03  -0.567  0.57047
edu_escolar_adul    4.273e-02  1.164e-03  36.719  < 2e-16 ***
salud_ate_prim      8.860e-02  2.856e-03  31.027  < 2e-16 ***
salud_serv_sanit    2.082e-03  2.165e-05  96.195  < 2e-16 ***
salud_disp_rsu      6.855e-02  7.633e-03   8.980  < 2e-16 ***
entor_acceso_transp 4.315e-02  4.339e-03   9.944  < 2e-16 ***
entor_inunda        5.583e-02  3.627e-03  15.395  < 2e-16 ***
entor_area_verde    4.607e-02  3.609e-03  12.763  < 2e-16 ***
entor_cavas        -2.420e-02  2.567e-02  -0.943  0.34584
entor_indus         6.704e-03  4.198e-03   1.597  0.11032
vivi_hacinamiento  -9.383e-02  1.717e-02  -5.466 4.93e-08 ***
vivi_gas_red        2.520e-02  7.798e-03   3.232  0.00124 **
vivi_calid_casa     1.072e-01  1.610e-02   6.663 3.11e-11 ***
vivi_dominio       -1.931e-02  1.431e-02  -1.349  0.17734
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06174 on 3520 degrees of freedom
Multiple R-squared:  0.9964,    Adjusted R-squared:  0.9964
F-statistic: 5.75e+04 on 17 and 3520 DF,  p-value: < 2.2e-16
```

Figure 7: summary del modelo full

Se aplico la funcion lm() al modelo full. Podemos observar un resumen estadistco de los residuos, los coeficientes de regresion que se van utilizar para armar la recta de regresion.

1.4.4 Recta de regresion

Al realizar el plot de la recta de regresion del modelo full podemos identificar ouliers que posiblemete afecten al modelo de prediccion

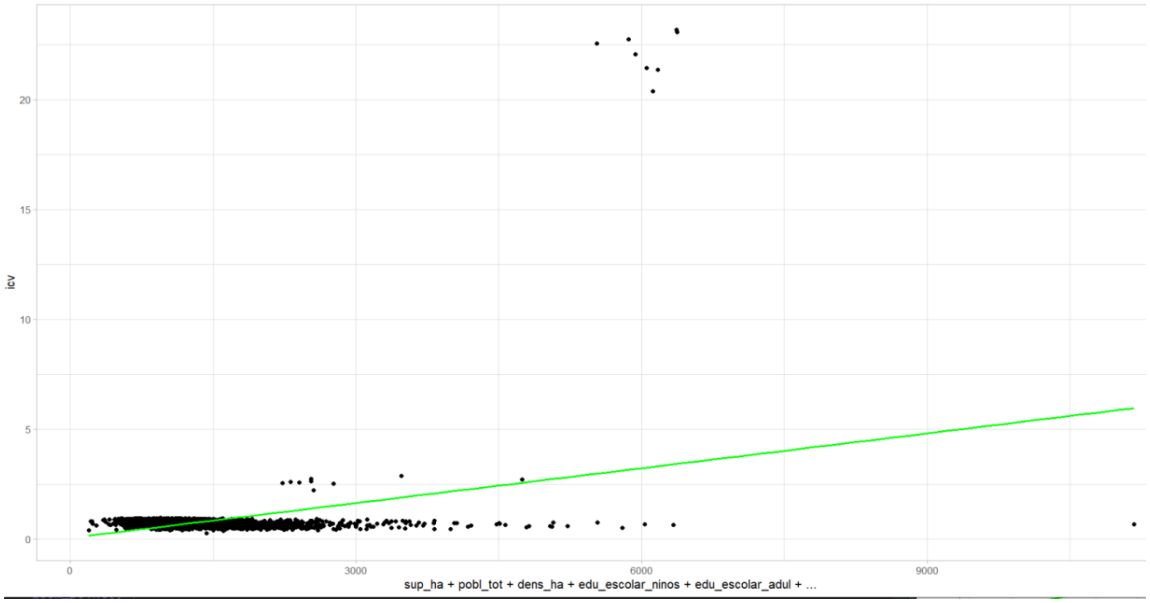


Figure 8: Comparación de transformaciones

1.4.5 Metodo fordward, backward, del modelo full

Se utiliza los metodos backward forward para buscar la mejor seleccion de predictores para el modelo, se puede observar que el mejor modelo con AIC=-19687.67 con todas las variables

```
> icv_mod_full_fordward <- step(icv_mod_full, direction = "forward",trace=T)
Start:  AIC=-19687.67
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_ninos + edu_escolar_adul +
    salud_ate_prim + salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_cavas + entor_indus +
    vivi_hacinamiento + vivi_gas_red + vivi_calid_casa + vivi_dominio
```

Figure 9: Método forward

```
> icv_mod_full_backward <- step(icv_mod_full, direction = "backward",trace=T)
Start:  AIC=-19687.67
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_ninos + edu_escolar_adul +
    salud_ate_prim + salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_cavas + entor_indus +
    vivi_hacinamiento + vivi_gas_red + vivi_calid_casa + vivi_dominio

                     Df Sum of Sq    RSS    AIC
- edu_escolar_ninos   1     0.001 13.418 -19689
- entor_cavas         1     0.003 13.420 -19689
- vivi_dominio        1     0.007 13.424 -19688
<none>                            13.417 -19688
- entor_indus         1     0.010 13.427 -19687
- vivi_gas_red        1     0.040 13.457 -19679
- pobl_tot            1     0.040 13.457 -19679
- sup_ha              1     0.078 13.495 -19669
- dens_ha             1     0.094 13.511 -19665
- vivi_hacinamiento   1     0.114 13.531 -19660
- vivi_calid_casa     1     0.169 13.586 -19645
- salud_disp_rsu      1     0.307 13.724 -19610
- entor_acceso_transp 1     0.377 13.794 -19592
- entor_area_verde    1     0.621 14.038 -19530
- entor_inunda        1     0.903 14.320 -19459
- salud_ate_prim      1     3.669 17.086 -18834
- edu_escolar_adul    1     5.139 18.556 -18542
- salud_serv_sanit    1    35.271 48.687 -15130
```

Figure 10: Método backward

```
Step:  AIC=-19689.35
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim +
    salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_cavas + entor_indus +
    vivi_hacinamiento + vivi_gas_red + vivi_calid_casa + vivi_dominio

                     Df Sum of Sq    RSS    AIC
- entor_cavas         1     0.003 13.421 -19691
- vivi_dominio        1     0.007 13.425 -19690
<none>                            13.418 -19689
- entor_indus         1     0.009 13.427 -19689
- vivi_gas_red        1     0.039 13.457 -19681
- pobl_tot            1     0.040 13.458 -19681
- sup_ha              1     0.078 13.496 -19671
- dens_ha             1     0.097 13.515 -19666
- vivi_hacinamiento   1     0.117 13.535 -19661
- vivi_calid_casa     1     0.174 13.592 -19646
- salud_disp_rsu      1     0.306 13.725 -19611
- entor_acceso_transp 1     0.380 13.798 -19593
- entor_area_verde    1     0.631 14.049 -19529
- entor_inunda        1     0.904 14.322 -19461
- salud_ate_prim      1     3.686 17.104 -18833
- edu_escolar_adul    1     7.508 20.926 -18119
- salud_serv_sanit    1    36.012 49.430 -15078
```

Figure 11: Método backward continuacion

```
Step:  AIC=-19690.45
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim +
    salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_indus + vivi_hacinamiento +
    vivi_gas_red + vivi_calid_casa + vivi_dominio

                     Df Sum of Sq    RSS    AIC
- vivi_dominio        1     0.007 13.428 -19691
<none>                            13.421 -19691
- entor_indus         1     0.009 13.431 -19690
- vivi_gas_red        1     0.038 13.460 -19682
- pobl_tot            1     0.039 13.461 -19682
- sup_ha              1     0.076 13.498 -19672
- dens_ha             1     0.098 13.519 -19667
- vivi_hacinamiento   1     0.115 13.537 -19662
- vivi_calid_casa     1     0.172 13.593 -19648
- salud_disp_rsu      1     0.304 13.725 -19613
- entor_acceso_transp 1     0.377 13.798 -19595
- entor_area_verde    1     0.628 14.049 -19531
- entor_inunda        1     0.904 14.325 -19462
- salud_ate_prim      1     3.682 17.104 -18835
- edu_escolar_adul    1     7.512 20.934 -18120
- salud_serv_sanit    1    36.026 49.447 -15079
```

Figure 12: Método backward continuacion

```
Step:  AIC=-19690.71
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim +
    salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_indus + vivi_hacinamiento +
    vivi_gas_red + vivi_calid_casa

                      Df Sum of Sq    RSS    AIC
<none>                             13.428 -19691
- entor_indus          1     0.008 13.436 -19691
- pobl_tot             1     0.040 13.468 -19682
- vivi_gas_red         1     0.046 13.474 -19681
- sup_ha               1     0.079 13.507 -19672
- dens_ha              1     0.098 13.526 -19667
- vivi_calid_casa      1     0.172 13.600 -19648
- vivi_hacinamiento    1     0.215 13.643 -19637
- salud_disp_rsu       1     0.308 13.736 -19612
- entor_acceso_transp  1     0.410 13.838 -19586
- entor_area_verde     1     0.639 14.067 -19528
- entor_inunda         1     0.917 14.345 -19459
- salud_ate_prim       1     3.676 17.104 -18837
- edu_escolar_adul     1     8.298 21.726 -17990
- salud_serv_sanit     1    36.837 50.265 -15023
> icv_mod_full_both <- step(icv_mod_full, direction = "both",trace=F)
>
```

Figure 13: Método backward continuacion

1.5 Modelo de regresión lineal

1.5.1 Selección del modelo

Procedemos a generar los modelos de forma incremental para hacer comparaciones

```
Start:  AIC=-19687.67
icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_ninos + edu_escolar_adul +
    salud_ate_prim + salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_cavas + entor_indus +
    vivi_hacinamiento + vivi_gas_red + vivi_calid_casa + vivi_dominio
```

Figure 14: Modelos

Como se puede apreciar, en nuestro modelo que contiene a todos nuestros predictores, puede explicar 99.64% de la variabilidad observada en el icv, por lo que continuamos el análisis con este modelo.

1.6 Bondad del modelo

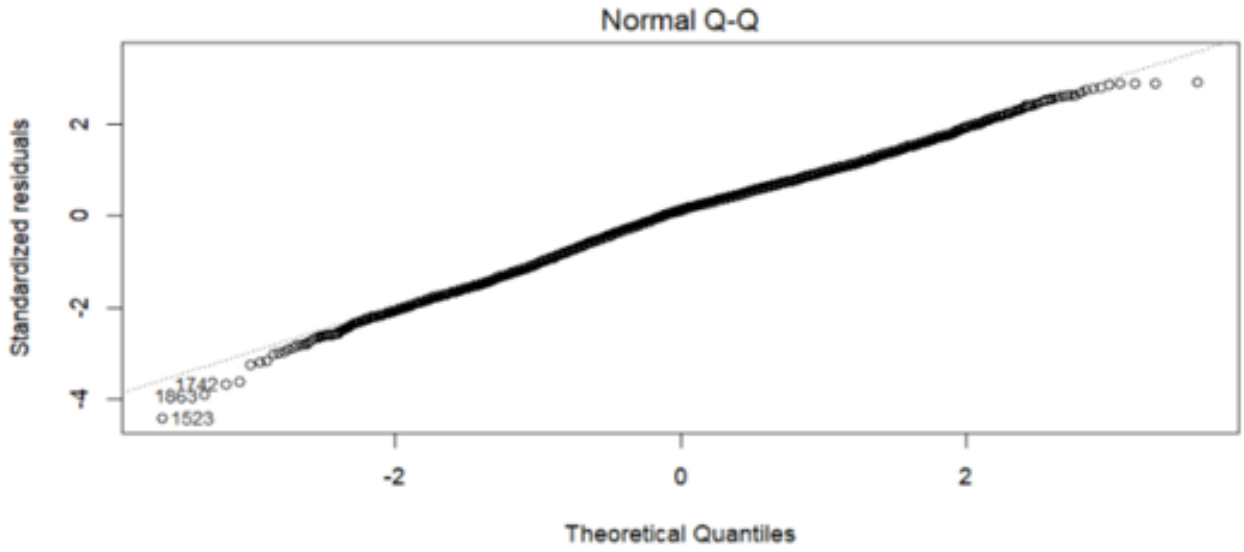
```
Call:
lm(formula = icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul +
    salud_ate_prim + salud_serv_sanit + salud_disp_rsu + entor_acceso_transp +
    entor_inunda + entor_area_verde + entor_indus + vivi_hacinamiento +
    vivi_gas_red + vivi_calid_casa, data = icv_a_modelo)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55500 -0.02937  0.00039  0.02877  1.03622

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.038e-02  1.390e-02   2.186 0.028882 *
sup_ha              -3.345e-04  7.365e-05  -4.542 5.76e-06 ***
pobl_tot             8.266e-06  2.550e-06   3.242 0.001200 **
dens_ha              2.679e-05  5.287e-06   5.067 4.26e-07 ***
edu_escolar_adul     4.272e-02  9.156e-04  46.658  < 2e-16 ***
salud_ate_prim       8.849e-02  2.849e-03  31.055  < 2e-16 ***
salud_serv_sanit     2.079e-03  2.115e-05  98.309  < 2e-16 ***
salud_disp_rsu       6.845e-02  7.614e-03   8.991  < 2e-16 ***
entor_acceso_transp  4.400e-02  4.244e-03  10.370  < 2e-16 ***
entor_inunda         5.613e-02  3.619e-03  15.509  < 2e-16 ***
entor_area_verde     4.637e-02  3.582e-03  12.945  < 2e-16 ***
entor_indus          5.885e-03  4.159e-03   1.415 0.157181
vivi_hacinamiento   -1.063e-01  1.417e-02  -7.506 7.70e-14 ***
vivi_gas_red         2.642e-02  7.617e-03   3.468 0.000531 ***
vivi_calid_casa      9.923e-02  1.476e-02   6.723 2.07e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06174 on 3523 degrees of freedom
Multiple R-squared:  0.9964,    Adjusted R-squared:  0.9964
F-statistic: 6.982e+04 on 14 and 3523 DF,  p-value: < 2.2e-16
```

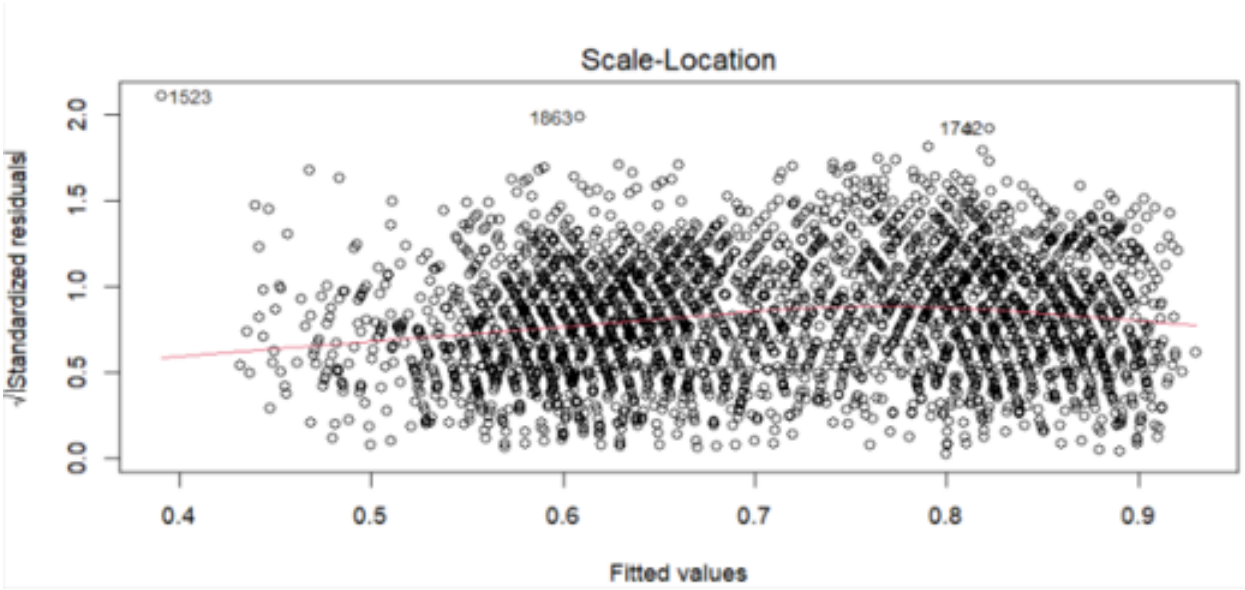
Figure 15: Resumen del modelo

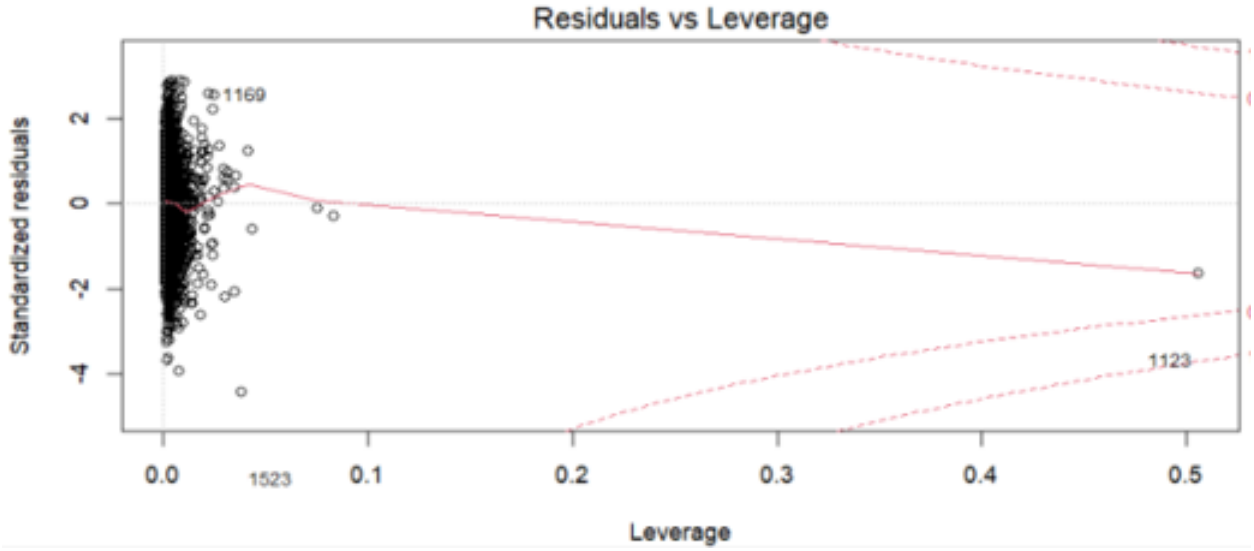
Ademas de un Adjusted R-squared de 0.99, en el resumen podemos ver que el error residual estándar es de 0.06174. La gran mayoria de variable ofrece un valor Pr menor de 0.05, que es nuestro umbral para detectar si se rechaza la hipótesis nula (Que todos los valores β son 0), por lo que rechazamos la hipótesis nula y aceptamos la alternativa en la que al menos un β es distinto de 0, con la excepcion de entor_indus. Podemos decir por lo tanto que nuestro modelo es significante.

1.6.1 Distribución normal de los residuos

Para seguir probando la bondad del modelo podemos visualizar el gráfico de linealidad de los residuos.
Como se ve en la siguiente figura los residuos se acercan a una distribución normal.
Se puede observar la existencia de outliers que pueden estar afectando a nuestro modelo.






Residuals vs Leverage
Standardized residuals
Leverage
1169
1123
1523
0.0
0.1
0.2
0.3
0.4
0.5
2
0
-2
-4

1.6.2 Variabilidad constante de los residuos

Se observa en el gráfico de dispersión que la variabilidad de los residuos del modelo se acerca a una distribución constante, exceptuando ciertas regiones.
A la vez se distinguen los outliers observados previamente.

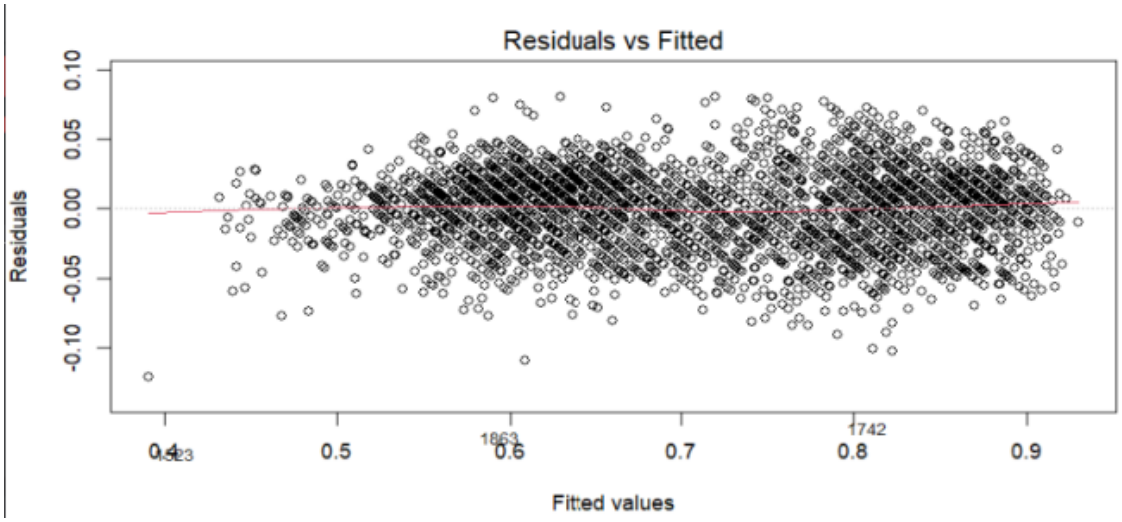


Figure 16: Normalidad

1.7 Test de Shapiro–Wilk

Se usa para contrastar la normalidad de un conjunto de datos.

En los siguientes gráficos se observa nuestro P-Valor es menor a 0.05, rechazamos la hipotesis nula por ende no es normal.

```
        Shapiro-Wilk normality test

data:  icv_mod_ajus_residuos
W = 0.70173, p-value < 2.2e-16
```

Figure 17: Test Shapiro–Wilk

1.7.1 Test de Watson

Observamos la correlacion entre los residuos.

```
        Durbin-Watson test

data:  icv_mod_ajus
DW = 1.6083, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 18: Test de Watson

1.7.2 Fórmula

Quedaría entonces nuestra fórmula de predicción de la siguiente manera.

$$
\begin{aligned}
(\hat{ICV}) &= 1.332\hat{e}-01 + -1.015\hat{e}-04 * sup\_ha \\
&+ (4.108\hat{e}-06) * pobl\_tot + (6.578\hat{e}-06) * dens\_ha \\
&+ (2.011\hat{e}-02) * edu\_escolar\_adul + (8.534\hat{e}-02) * salud\_ate\_prim \\
&+ (1.217\hat{e}-01) * salud\_serv\_sanit + (-0.\hat{2}197) * salud\_disp\_rsu \\
&+ 2.669\hat{e}-02 * entor\_acceso\_transp + 4.038\hat{e}-02 * entor\_inunda \\
&+ 2.226\hat{e}-02 * entor\_area\_verde + 4.364\hat{e}-02 * vivi\_hacinamiento \\
&+ 1.677\hat{e}-02 * vivi\_gas\_red + 6.983\hat{e}-02 * vivi\_calid\_casa
\end{aligned}
$$

Conclusión

A simple vista no pudimos apreciar la bondad del modelo, debimos realizar varios análisis, un summary inicial para observar la distribucon de las variables lo que nos dío un indicio de la posibilidad de que hubieran outliers, la visualizacion de la correlacion nos permitio ver la relacion entre las varibles.
Se hizo un gran uso de los gráficos para entender el dataset y metodos de seleccion de predictores para lograr un mejor modelo de predicción.
Entendimos que el tema de regresión lineal es muy amplio y dependiente de los datos a analizar y que ofrece técnicas potencialmente valiosas. A la vez afianzamos nuestro dominio sobre la herramienta R.

Anexo

```r

# Materia: Explotacion de Datos
# Fecha: 09/09/2023
# Tema: Regresion Lineal Multiple
# Objetivo: Modelo didactico de RLM
library(xtable)
library(corr)
library(readr)
library(dplyr)
library(sqldf)
library(stringi)
library(plotly)
library(tidyverse)
library(TTR)
library(fpp3)
library(scales)
library(xts)
library(plotly)
library(nortest)
library(lmtest)
library (car)
#library(fmsb)
library(broom)

#
    ##########################################################################################

############################## 01. IMPORTACION DE DATOS
    ##################################
#
    ##########################################################################################

rm(icv_corr)
icv <- read.table(file.choose(),
                        sep=',',
                        header=TRUE,
                        fill = TRUE,
                        row.names = NULL,
                        stringsAsFactors = FALSE)

dim_entorno <- read.table(file.choose(),
                  sep=',',
```

```
                  header=TRUE,
                  fill = TRUE,
                  row.names = NULL,
                  stringsAsFactors = FALSE)


dim_edu <- read.table(file.choose(),
                  sep=',',
                  header=TRUE,
                  fill = TRUE,
                  row.names = NULL,
                  stringsAsFactors = FALSE)


dim_salud <- read.table(file.choose(),
                  sep=',',
                  header=TRUE,
                  fill = TRUE,
                  row.names = NULL,
                  stringsAsFactors = FALSE)


dim_vivienda <- read.table(file.choose(),
                  sep=',',
                  header=TRUE,
                  fill = TRUE,
                  row.names = NULL,
                  stringsAsFactors = FALSE)

View(icv)
str(icv)
names(icv)

attach(icv)
attach(dim_vivienda)
attach(dim_salud)
attach(dim_entorno)
attach(dim_edu)

#
    ##############################################################################################

########################################## 02. ETL
    ########################################
#
    ##############################################################################################

#
    ------------------------------------------------------------------------------------------

#2.1 REDUCCION DE DIMENSIONALIDAD DE LOS DATOS:

#ICV
View(icv)
str(icv)
```

```r
rm(icv_red)
icv_red = subset(icv, select = c(sup_ha, pobl_tot, dens_ha, icv))
View(icv_red)
str(icv_red)

#SALUD
View(dim_salud)
str(dim_salud)
rm(dim_salud_red)
dim_salud_red = subset(dim_salud, select = c(ind_efe_sa, ind_ssb, ind_rsu))
View(dim_salud_red)
str(dim_salud_red)

#EDUCACION
View(dim_edu)
str(dim_edu)
rm(dim_edu_red)
dim_edu_red = subset(dim_edu, select = c(a_esp, a_prom))
View(dim_edu_red)
str(dim_edu_red)

#VIVIENDA
View(dim_vivienda)
str(dim_vivienda)
rm(dim_vivienda_red)
dim_vivienda_red = subset(dim_vivienda, select = c(ind_hac, ind_gr, ind_matv, ind_dom))
View(dim_vivienda_red)
str(dim_vivienda_red)

#ENTORNO
View(dim_entorno)
str(dim_entorno)
rm(dim_entorno_red)
dim_entorno_red = subset(dim_entorno, select = c(ind_trpu, ind_inunda, ind_evp, ind_cavas
    , ind_indus))
View(dim_entorno_red)
str(dim_entorno_red)

rm(icv)
rm(dim_edu)
rm(dim_entorno)
rm(dim_salud)
rm(dim_vivienda)

#
    ----------------------------------------------------------------------------------------

#2.2 UNIFICAMOS LOS DATASETS
dim_entorno %>% distinct(ind_inunda)

rm(icv_a_modelo)
str(icv_red)
icv_a_modelo <- cbind(icv_red, dim_edu_red, dim_salud_red, dim_entorno_red, dim_vivienda_
    red)
View(icv_a_modelo)
```

```
str(icv_a_modelo)
#
    ----------------------------------------------------------------------------------------------------

#2.3 RENOMBRAMOS LAS COLUMNAS
icv_a_modelo
View(icv_a_modelo)
rm(normalizacion_nombres)
normalizacion_nombres<-c("sup_ha","pobl_tot","dens_ha","icv","edu_escolar_ninos","edu_
    escolar_adul","salud_ate_prim",
                         "salud_serv_sanit","salud_disp_rsu","entor_acceso_transp","entor
    _inunda","entor_area_verde","entor_cavas",
                         "entor_indus","vivi_hacinamiento","vivi_gas_red","vivi_calid_
    casa","vivi_dominio")
names(icv_a_modelo)<-normalizacion_nombres
#
    ----------------------------------------------------------------------------------------------------

#2.4 LIMPIAMOS LOS DATOS:
# CUANTOS  NA TENGO POR COLUMNA?
View(summarise_all(icv_a_modelo, funs(sum(is.na(.)))))
names(icv_a_modelo)
#CASTEAMOS LOS DATOS, YA QUE A SU VEZ ASIGNAMOS LOS VACIOS O NULOS COMO NAs
icv_a_modelo$sup_ha <- suppressWarnings(as.numeric(icv_a_modelo$sup_ha))
icv_a_modelo$pobl_tot <- suppressWarnings(as.numeric(icv_a_modelo$pobl_tot))
icv_a_modelo$dens_ha <- suppressWarnings(as.numeric(icv_a_modelo$dens_ha))
icv_a_modelo$icv <- suppressWarnings(as.numeric(icv_a_modelo$icv))
icv_a_modelo$edu_escolar_ninos <- suppressWarnings(as.numeric(icv_a_modelo$edu_escolar_
    ninos))
icv_a_modelo$edu_escolar_adul <- suppressWarnings(as.numeric(icv_a_modelo$edu_escolar_
    adul))
icv_a_modelo$salud_ate_prim <- suppressWarnings(as.numeric(icv_a_modelo$salud_ate_prim))
icv_a_modelo$salud_serv_sanit <- suppressWarnings(as.numeric(icv_a_modelo$salud_serv_
    sanit))
icv_a_modelo$salud_disp_rsu <- suppressWarnings(as.numeric(icv_a_modelo$salud_disp_rsu))
icv_a_modelo$entor_acceso_transp <- suppressWarnings(as.numeric(icv_a_modelo$entor_acceso
    _transp))
icv_a_modelo$entor_inunda <- suppressWarnings(as.numeric(icv_a_modelo$entor_inunda))
icv_a_modelo$entor_area_verde <- suppressWarnings(as.numeric(icv_a_modelo$entor_area_
    verde))
icv_a_modelo$entor_cavas <- suppressWarnings(as.numeric(icv_a_modelo$entor_cavas))
icv_a_modelo$entor_indus <- suppressWarnings(as.numeric(icv_a_modelo$entor_indus))
icv_a_modelo$vivi_hacinamiento <- suppressWarnings(as.numeric(icv_a_modelo$vivi_
    hacinamiento))
icv_a_modelo$vivi_gas_red <- suppressWarnings(as.numeric(icv_a_modelo$vivi_gas_red))
icv_a_modelo$vivi_calid_casa <- suppressWarnings(as.numeric(icv_a_modelo$vivi_calid_casa)
    )
icv_a_modelo$vivi_dominio <- suppressWarnings(as.numeric(icv_a_modelo$vivi_dominio))

View(summarise_all(icv_a_modelo, funs(sum(is.na(.)))))
#
    ############################################################################################

######################### 03. ANALISIS EXPLOTATORIO DE DATOS
    ##############################
```

```r
#
    ##############################################################################################

#QUE TIPOS DE ZONAS TENEMOS?
attach(icv_a_modelo)

#3.1 analizamos un poco los datos: descriptivos
str(icv_a_modelo)

#La funcion summary nos permite entender de manera descriptiva los datos:
#Podemos sacar conclusiones sobre: NAs, Outliers, Concenstracion de datos en Quartiles
summary(icv_a_modelo)
summary(icv_a_modelo$salud_serv_sanit)
#
    ----------------------------------------------------------------------------------------------

#3.2 BOXPLOT:
icv_a_modelo_salud_serv_sanit <- plot_ly(y = ~icv_a_modelo$salud_serv_sanit, type = "box"
    )
icv_a_modelo_salud_serv_sanit

icv_a_modelo_vivi_hacinamiento <- plot_ly(y = ~icv_a_modelo$vivi_hacinamiento, type = "
    box")
icv_a_modelo_vivi_hacinamiento

icv_a_modelo_vivi_gas_red <- plot_ly(y = ~icv_a_modelo$vivi_gas_red, type = "box")
icv_a_modelo_vivi_gas_red

icv_a_modelo_vivi_calid_casa <- plot_ly(y = ~icv_a_modelo$vivi_calid_casa, type = "box")
icv_a_modelo_vivi_calid_casa

icv_a_modelo_vivi_dominio <- plot_ly(y = ~icv_a_modelo$vivi_dominio, type = "box")
icv_a_modelo_vivi_dominio

rm(icv_outliers)
str(icv_a_modelo)
icv_outliers <- subset(icv_a_modelo, select = c(salud_serv_sanit, vivi_hacinamiento, vivi
    _gas_red, vivi_calid_casa, vivi_dominio))


#AGRUPAMOS TODOS LOS CAMPOS CON POSIBLES OUTLIERS PARA LATEX
xtable(summary(icv_a_modelo))
xtable(summary(icv_outliers))
#
    ----------------------------------------------------------------------------------------------

#3.3 Vemos correlaciones y graficamos:
#El siguiente codigo muestra el indice de correlacion existente en nuestris atributos
#El mismo nos indica en una matriz la fuerza de correlacion positiva o negativa que se
    tiene.
#Tambien se utiliza dicha matriz para generar o ajustar un modelo de Regresion.

rm(icv_mod_full_corr)
icv_mod_full_corr <- cor(icv_a_modelo) #requiere corrplot. Las variables deben ser
    num ricas.
```

```r
icv_mod_full_corr

#Visualizacion con indice de correlacion para cada atributo
corrplot(icv_mod_full_corr, method="number",tl.col="black",tl.cex=0.5)

corrplot(icv_mod_full_corr, type = "upper",
         method = "square",
         addCoef.col = "black",
         tl.col = "black", tl.srt = 10)

#Visualizacion con nivel de correlacion para cada atribut
corrplot(icv_mod_full_corr, method="color",tl.col="black",tl.cex=0.5)

#Visualizacion con grafico de dispersion por cada atributo
plot(icv_a_modelo)
#
    ----------------------------------------------------------------------------------------------

################################################################################
###################### 04 APLICACION DE MODELOS DE RLM #########################
################################################################################
#4.1 MODELO FULL DE RLM
str(icv_a_modelo)
names(icv_a_modelo)

icv_mod_full <- lm(icv ~ ., data = icv_a_modelo)
class(icv_mod_full)
str(icv_mod_full)
summary(icv_mod_full)

#SACAR CONCLUSION NUESTRA SOBRE EL:
#P-VALUE Y LOS CONTRASTES DE HIPOTESIS:
#R CUADRADO Y R CUADRADO AJUSTADO
#F ESTADISTICO

#ANALISIS DE SUMMARY
#1: FORMULA DE LA RECTA: ICV ~ RESTO
#2: Residuos: Es la diferencia entre el valor observado(dato real) y el predicho (datos
    de la recta de regresion)
#3: Los coeficientes determinar el incremento o decremento del valor de mi Y (MPG) por
    cada unidad de aumento de Y (MPG)
#4: El Pr(t) nos indica el nivel de significancia que tiene mi variable dependiente con
    mis variables independientes.
#5: P-Value: Nos indica el contraste de hipotesis bajo la distribucion normal: Si el p-
    value es menor a 0.05 entonces se rechaza la Ho (Hipotesis nula)
#5.1: Ho: Si el p-value para nuestra variable es menor a 0.05, entonces rechazamos la Ho
#5.2: Ho: Si el p-value es mayor que 0.05 entonces significa que no nuestro dato no es
    significante

#Ho: Digo que mi valor no es significativo. VOY A DESAPROBAR LA MATERIA
#H1: Digo que mi valor es significativo. VOY A APROBAR LA MATERIA

#Contraste teniendo en cuenta el indice convencional: 0.05
#Para rechazar el valor tiene que ser < 0.05

```

```r
#Concluyo:
#Ho: NOTA tiene un p-value de 0.00002,  Puedo  aceptar la Ho?:
#
    ------------------------------------------------------------------------------------------------

#4.2 ANALISIS DE RESIDUOS SOBRE EL MODELO FULL
names(icv_a_modelo)
#GRAFICO DE RESIDUOS Y RECTA
ggplot(icv_mod_full, aes(x=sup_ha + pobl_tot + dens_ha + edu_escolar_ninos + edu_escolar_
    adul +
                           salud_ate_prim + salud_serv_sanit + salud_disp_rsu + entor_
    acceso_transp +
                           entor_inunda + entor_area_verde + entor_cavas + entor_indus +
    vivi_hacinamiento +
                           vivi_gas_red + vivi_calid_casa + vivi_dominio, y=icv))+
  geom_point() +
  geom_smooth(method='lm',se=FALSE, col='green') +
  theme_light()

#EXPLICAR LOS RESIDUOS EN EL MODELO FULL
##############################################################################
################ 05 AJUSTE Y APLICACION DEL NUEVO MODELO ######################
##############################################################################
#
    ------------------------------------------------------------------------------------------------

#5.1 M todos autom ticos para selecci n de variables: "backward" / "forward" / "both"
#EXPLICAR CADA METODO
icv_mod_full_fordward <- step(icv_mod_full, direction = "forward",trace=T)
icv_mod_full_backward <- step(icv_mod_full, direction = "backward",trace=T)
icv_mod_full_both <- step(icv_mod_full, direction = "both",trace=F)
summary(icv_mod_full_fordward)
summary(icv_mod_full_backward)
summary(icv_mod_full_both)

#PLOT DEL MODELO AJUSTADO
plot(icv_mod_full)
plot(icv_mod_full_backward)
par(mfrow=c(2,2))
plot(icv_mod_full_backward, scale = "adjr2", main = "R^2 ajustado")

#
    ------------------------------------------------------------------------------------------------

#5.2 ANALISIS DE SUPUESTOS DE RESIDUOS: Analizamos supuestos estudiando residuos

#ANALISIS SOBRE EL MODELO BACKWARD:
icv_mod_full_backward_residuos = residuals(icv_mod_full_backward)

boxplot(icv_mod_full_backward_residuos, col = "blue",horizontal=TRUE,ylim = c(-2,2),main=
    "Box-plot de residuos")
plot_icv_mod_full_backward_residuos <- plot_ly(y = ~icv_mod_full_backward_residuos, type
    = "box")
plot_icv_mod_full_backward_residuos
```

```r
#
    ----------------------------------------------------------------------------------------------------

#5.3 CONTRUCCION DEL NUEVO MODELO CON BASE AL ANALISIS PREVIO
#5.3.1 DETECCION DE OUTLIERS CON EL DATASET ORIGINAL

#TABLA CON TODOS LOS CAMPOS CON POSIBLES OUTLIERS A TRATAR (ELIMINAR)
icv_outliers

icv_outliers_vivi_dominio_plotly <- plot_ly(y = ~icv_outliers$vivi_dominio, type = "box")
icv_outliers_vivi_dominio_plotly

rm(icv_outliers_vivi_dominio)
icv_outliers_vivi_dominio<-boxplot(icv_outliers$vivi_dominio, col="skyblue", frame.plot=F
    )
icv_outliers_vivi_dominio$out

icv_outliers_vivi_calid_casa<-boxplot(icv_outliers$vivi_calid_casa, col="skyblue", frame.
    plot=F)
icv_outliers_vivi_calid_casa$out

icv_outliers_vivi_gas_red<-boxplot(icv_outliers$vivi_gas_red, col="skyblue", frame.plot=F
    )
icv_outliers_vivi_gas_red$out

icv_outliers_vivi_hacinamiento<-boxplot(icv_outliers$vivi_hacinamiento, col="skyblue",
    frame.plot=F)
icv_outliers_vivi_hacinamiento$out

icv_outliers_salud_serv_sanit<-boxplot(icv_outliers$salud_serv_sanit, col="skyblue",
    frame.plot=F)
icv_outliers_salud_serv_sanit$out

par(mfrow=c(2, 3))
icv_outliers_vivi_dominio<-boxplot(icv_outliers$vivi_dominio, col="skyblue", frame.plot=F
    )
icv_outliers_vivi_calid_casa<-boxplot(icv_outliers$vivi_calid_casa, col="skyblue", frame.
    plot=F)
icv_outliers_vivi_gas_red<-boxplot(icv_outliers$vivi_gas_red, col="skyblue", frame.plot=F
    )
icv_outliers_vivi_hacinamiento<-boxplot(icv_outliers$vivi_hacinamiento, col="skyblue",
    frame.plot=F)
icv_outliers_salud_serv_sanit<-boxplot(icv_outliers$salud_serv_sanit, col="skyblue",
    frame.plot=F)


#5.3.2 ELIMINO LOS OUTLIERS
rm(icv_a_modelo_v2)
icv_a_modelo_v2 <- icv_a_modelo
View(icv_a_modelo_v2)
View(icv_outliers)
icv_a_modelo_v2 <- icv_a_modelo_v2[!(icv_a_modelo_v2$vivi_dominio %in% icv_outliers_vivi_
    dominio$out),]
icv_a_modelo_v2 <- icv_a_modelo_v2[!(icv_a_modelo_v2$vivi_calid_casa %in% icv_outliers_
    vivi_calid_casa$out),]
```

```r
icv_a_modelo_v2 <- icv_a_modelo_v2[!(icv_a_modelo_v2$vivi_gas_red %in% icv_outliers_vivi_gas_red$out),]
icv_a_modelo_v2 <- icv_a_modelo_v2[!(icv_a_modelo_v2$vivi_hacinamiento %in% icv_outliers_vivi_hacinamiento$out),]
icv_a_modelo_v2 <- icv_a_modelo_v2[!(icv_a_modelo_v2$salud_serv_sanit %in% icv_outliers_salud_serv_sanit$out),]

#CHEQUEO
rm(icv_a_modelo_v2_vivi_hacinamiento_out)
icv_a_modelo_v2_vivi_dominio_out <- plot_ly(y = ~icv_a_modelo_v2$vivi_dominio, type = "box")
icv_a_modelo_v2_vivi_dominio_out

prueba_salud_serv_sanit_vivi_calid_casa_out <- plot_ly(y = ~icv_a_modelo_v2$vivi_calid_casa, type = "box")
prueba_salud_serv_sanit_vivi_calid_casa_out

icv_a_modelo_v2_vivi_gas_red_out <- plot_ly(y = ~icv_a_modelo_v2$vivi_gas_red, type = "box")
icv_a_modelo_v2_vivi_gas_red_out

icv_a_modelo_v2_vivi_hacinamiento_out <- plot_ly(y = ~icv_a_modelo_v2$vivi_hacinamiento, type = "box")
icv_a_modelo_v2_vivi_hacinamiento_out

icv_a_modelo_v2_salud_serv_sanit_out <- plot_ly(y = ~icv_a_modelo_v2$salud_serv_sanit, type = "box")
icv_a_modelo_v2_salud_serv_sanit_out

View(icv_a_modelo_v2)

par(mfrow=c(2, 3))
icv_a_modelo_v2_vivi_dominio_out<-boxplot(icv_a_modelo_v2$vivi_dominio, col="skyblue", frame.plot=F)
prueba_salud_serv_sanit_vivi_calid_casa_out<-boxplot(icv_a_modelo_v2$vivi_calid_casa, col="skyblue", frame.plot=F)
icv_a_modelo_v2_vivi_gas_red_out<-boxplot(icv_a_modelo_v2$vivi_gas_red, col="skyblue", frame.plot=F)
icv_a_modelo_v2_vivi_hacinamiento_out<-boxplot(icv_a_modelo_v2$vivi_hacinamiento, col="skyblue", frame.plot=F)
icv_a_modelo_v2_salud_serv_sanit_out<-boxplot(icv_a_modelo_v2$salud_serv_sanit, col="skyblue", frame.plot=F)

# CUANTOS  NA TENGO POR COLUMNA?
View(summarise_all(icv_a_modelo_v2, funs(sum(is.na(.)))))
names(prueba)

#AGRUPAMOS TODOS LOS CAMPOS CON POSIBLES OUTLIERS PARA LATEX
xtable(summary(icv_a_modelo_v2))
#
    ----------------------------------------------------------------------------------------------------

#5.4 CORRELACIONES DEL NUEVO DATASET ************************************************
    *****

```

```
rm(icv_a_modelo_v2_corr)
icv_a_modelo_v2_corr <- cor(icv_a_modelo_v2) #requiere corrplot. Las variables deben ser num ricas.
icv_a_modelo_v2_corr

#Visualizacion con indice de correlacion para cada atributo
corrplot(icv_a_modelo_v2_corr, method="number",tl.col="black",tl.cex=0.5)

corrplot(icv_a_modelo_v2_corr, type = "upper",
         method = "square",
         addCoef.col = "black",
         tl.col = "black", tl.srt = 0.8)

#Visualizacion con nivel de correlacion para cada atribut
corrplot(icv_a_modelo_v2_corr, method="color",tl.col="black",tl.cex=0.8)

#Visualizacion con grafico de dispersion por cada atributo
#plot(icv_a_modelo)
#----------------------------------------------------------------------------------------------------

#5.5 MODELO AJUSTADO DE RLM
summary(icv_mod_full) #r-cuadrado: 0.9964 r-cuadrado ajustado: 0.9964
summary(icv_mod_full_backward) #r-cuadrado: 0.9964 r-cuadrado ajustado: 0.9964
str(icv_a_modelo_v2)
names(icv_a_modelo_v2)

rm(icv_mod_ajus)
icv_mod_ajus <- lm(icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim +
                   salud_serv_sanit + salud_disp_rsu + entor_acceso_transp + entor_inunda +
                   entor_area_verde  + vivi_hacinamiento +
                   vivi_gas_red + vivi_calid_casa, data = icv_a_modelo_v2)
class(icv_mod_ajus)
str(icv_mod_ajus)
summary(icv_mod_ajus)

#PRINCIPIO DE PARSIMONIA: MENOS VARIABLES EXPLICAN MEJOR EL MODELO
#----------------------------------------------------------------------------------------------------

#4.2 ANALISIS DE RESIDUOS SOBRE EL MODELO FULL
#names(icv_a_modelo_v2)
#GRAFICO DE RESIDUOS Y RECTA
#ggplot(icv_mod_ajus, aes(x=sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim +
#                           salud_serv_sanit + salud_disp_rsu + entor_acceso_transp + entor_inunda +
#                           entor_area_verde  + vivi_hacinamiento + vivi_gas_red + vivi_calid_casa, y=icv))+
#  geom_point() +
#  geom_smooth(method='lm',se=FALSE, col='green') +
#  theme_light()

```

```r
#EXPLICAR LOS RESIDUOS EN EL MODELO FULL


#
    ----------------------------------------------------------------------------------------------

#5.6 ANALISIS DE SUPUESTOS DE RESIDUOS: Analizamos supuestos estudiando residuos

#ANALISIS SOBRE EL MODELO AJUSTADO:
#residuos alrededor de 0: ver en el boxplot
icv_mod_ajus_residuos = residuals(icv_mod_ajus)

boxplot(icv_mod_ajus_residuos, col = "blue",horizontal=TRUE,ylim = c(-0.2,0.2),main="Box-
    plot de residuos")
plot_icv_mod_ajus_residuos <- plot_ly(y = ~icv_mod_ajus_residuos, type = "box")
plot_icv_mod_ajus_residuos
#
    ----------------------------------------------------------------------------------------------

#5.7 homogeneidad de varianza: ver en residuos vs predichos
#normalidad de residuos: ver en QQplot
par(mfrow=c(2, 2))
plot(icv_mod_ajus)
#
    ----------------------------------------------------------------------------------------------

#5.8 Test de normalidad de Shapiro-Wilk (muestras chicas)
# Test de normalidad de Shapiro-Wilk (muestras chicas)
#SI EL P-VALUE ES MAYOR A 0.10
#H0: La variable presenta una distribuci n normal
#H1: La variable presenta una distribuci n no normal

#Sig(p valor) > alfa: No rechazar H0 (normal).
#Sig(p valor) < alfa: Rechazar H0 (no normal)

#Donde alfa representa la significancia, que en este ejemplo hipot tico es igual al 5%
    (0,05).
icv_mod_ajus_residuos.test <- shapiro.test(icv_mod_ajus_residuos)
print(icv_mod_ajus_residuos.test)
#
    ----------------------------------------------------------------------------------------------

#5.9 Test de normalidad de Kolmogorov-Smirnov (muestras grandes)
lillie.test(icv_mod_ajus_residuos)
#
    ----------------------------------------------------------------------------------------------

# 6.0 Test de Durbin-Watson test para determinar correlacion entre los Residuos
# errores independientes (no correlacionados, es equivalente si hay normalidad)
# DW debe ser cercano a 2 y el p-value cercano a 1 para demostrarlo
# un pvalue muy pequenio indica que no hay independencia (est n correlacionados)
# La hipotesis nula es que no hay autocorrelacion.

dwtest(icv_mod_ajus)
```

```
#
    ------------------------------------------------------------------------------------------

# 6.1 Analizamos multicolinealidad e influyentes:
# VIF: Si es mayor a 10, entonces hay correlacion entra variables
# VIF = 1: NO HAY MULTICOLINEALIDAD
# VIF > 1: MULTICOLINEALIDAD ACEPTABLE
# VIF > 5: MULTICOLINEALIDAD ALTA

vif(icv_mod_ajus)
#
    ------------------------------------------------------------------------------------------

# 6.2 DISTANCIA DE COOK: INFLUYENTES: Se calcula para determinar cuan influyente es cada
    obs
# en las estimaciones del modelo
cooks=cooks.distance(icv_mod_ajus)
plot(cooks.distance(icv_mod_ajus))
#
    ------------------------------------------------------------------------------------------

# 6.3 COMPARAMOS MODELO MEDIANTE ANOVA
# anova(icv_mod_full, icv_a_modelo) #compara si las sumas de cuadrados son signif
    diferentes

#
    ------------------------------------------------------------------------------------------

## PREDICCION de nuevos datos
icv_predecir<-data.frame(disp=c(130,152,305),
                         potencia=c(51,49,200),
                         peso=c(2444,3100,4800),
                         aceleracion=c(11,16,14))

#valores a predecir
predict(mod_full,nuevo)

#
    ------------------------------------------------------------------------------------------

# RECTA DE REGRESION:

#icv ~ sup_ha + pobl_tot + dens_ha + edu_escolar_adul + salud_ate_prim + salud_serv_sanit
     + salud_disp_rsu + entor_acceso_transp + entor_inunda +  entor_area_verde  + vivi_
    hacinamiento + vivi_gas_red + vivi_calid_casa
```