

# Asignacion 04 28 2022

Sergio Andrés Hernández

4/28/2022

## Asignación

- Escoger una propia variable (Relación ej. Ca/Mg)
- Resultados (en clase) próximo JUEVES

## Selección de una variable

Modelo de regresión simple (Cuestionarlo)

## Librerías a usar

## Carga de datos

```
datos_xp <- read_excel("C:/Users/57321/Downloads/XPABLO.XLSX")  
#View(datos_xp)
```

## Modelo de regresión simple (Cuestionarlo)

```
names(datos_xp)
```

```
## [1] "id" "Long" "Lat" "z" "MO" "Ca" "Mg" "K" "Na" "CICE"  
## [11] "CE" "Fe" "Cu" "Zn" "cos" "mod1" "mod2" "mod3" "mod4"
```

## Redefiniendo df

```
df <- datos_xp[-c(15,16,17,18,19)]  
names(df)
```

```
## [1] "id" "Long" "Lat" "z" "MO" "Ca" "Mg" "K" "Na" "CICE"  
## [11] "CE" "Fe" "Cu" "Zn"
```

```
names(df)
```

```
## [1] "id" "Long" "Lat" "z" "MO" "Ca" "Mg" "K" "Na" "CICE"  
## [11] "CE" "Fe" "Cu" "Zn"
```

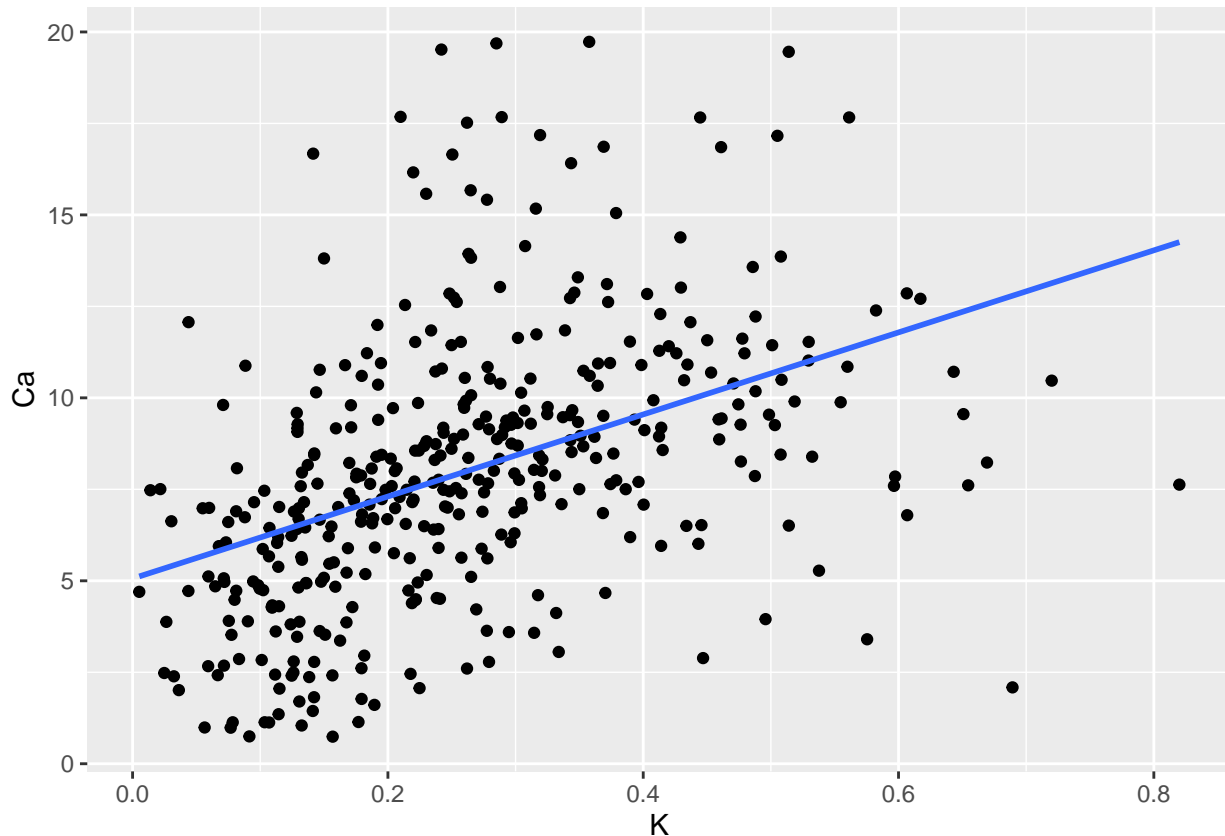
```
model_1 <- lm(Ca ~ K, data = df)  
summary(model_1)
```

```
##
## Call:
## lm(formula = Ca ~ K, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.703  -1.979  -0.182   1.373  11.745
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0613     0.3442  14.705  <2e-16 ***
## K             11.2126     1.1499   9.751  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.327 on 401 degrees of freedom
## Multiple R-squared:  0.1917, Adjusted R-squared:  0.1896
## F-statistic: 95.07 on 1 and 401 DF,  p-value: < 2.2e-16
```

$$Y_{Ca} = 5.0613 + 11.21K$$

```
ggplot(df, aes(y = Ca, x = K)) +
  geom_point()+
  geom_smooth(method='lm', se = F)
```

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



## Filtrado Ca > 15

```
df_2 <- df |>
  filter(Ca <= 10)
df_2
```

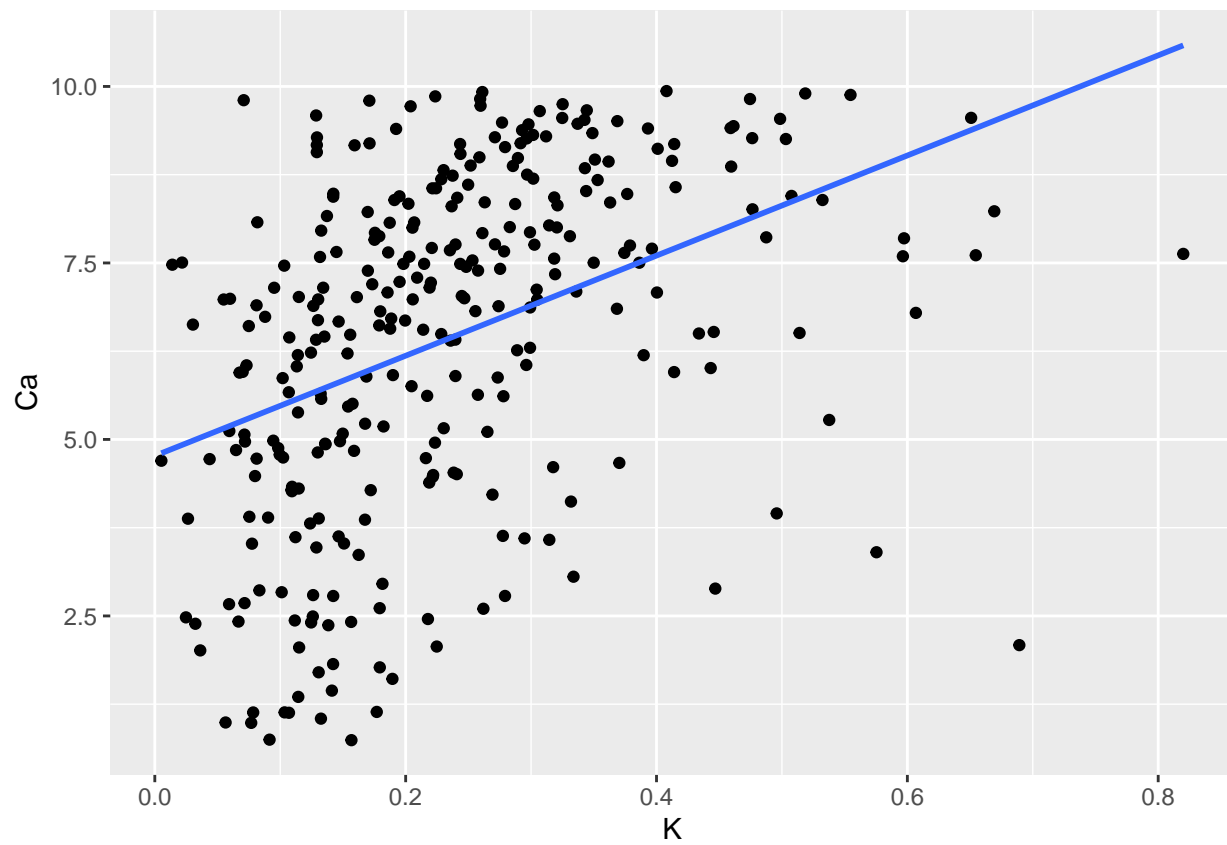
```
## # A tibble: 305 x 14
##       id Long  Lat    z    MO    Ca    Mg    K    Na  CICE    CE    Fe
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1   -72.6  8.08  120  2.09  7.83  1.56  0.175  0.291  9.85  0.130  133.
## 2     2   -72.6  8.08  119  1.65  3.95  0.771  0.496  0.136  5.36  0.126  29.7
## 3     3   -72.6  8.08  111  1.65  5.88  1.23  0.273  0.135  7.52  0.287  237.
## 4     4   -72.6  8.08  114  2.48  5.62  1.13  0.217  0.163  7.13  0.415  331.
## 5     6   -72.6  8.09  109  1.93  7.49  1.56  0.244  0.115  9.41  0.410  258.
## 6    10   -72.6  8.10  115  2.06  7.96  1.78  0.133  0.308 10.2  0.245  446.
## 7    13   -72.6  8.10  111  2.88  9.07  2.14  0.129  0.247 11.6  0.169  225.
## 8    15   -72.6  8.10  114  1.68  4.48  0.789  0.0799 0.197  5.55  0.84  196.
## 9    16   -72.6  8.10  125  2.63  7.48  1.52  0.0140 0.257  9.26  0.213  198.
## 10   17   -72.6  8.11   97  1.33  5.95  1.32  0.0676 0.285  7.62  0.572  12.4
## # ... with 295 more rows, and 2 more variables: Cu <dbl>, Zn <dbl>
```

```
model_2 <- lm(Ca ~ K, data = df_2)
summary(model_2)
```

```
##
## Call:
## lm(formula = Ca ~ K, data = df_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.5685 -1.3974  0.4821  1.6435  4.5350
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7677     0.2484  19.193 < 2e-16 ***
## K              7.0908     0.9136   7.762  1.3e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.189 on 303 degrees of freedom
## Multiple R-squared:  0.1658, Adjusted R-squared:  0.1631
## F-statistic: 60.24 on 1 and 303 DF, p-value: 1.297e-13
```

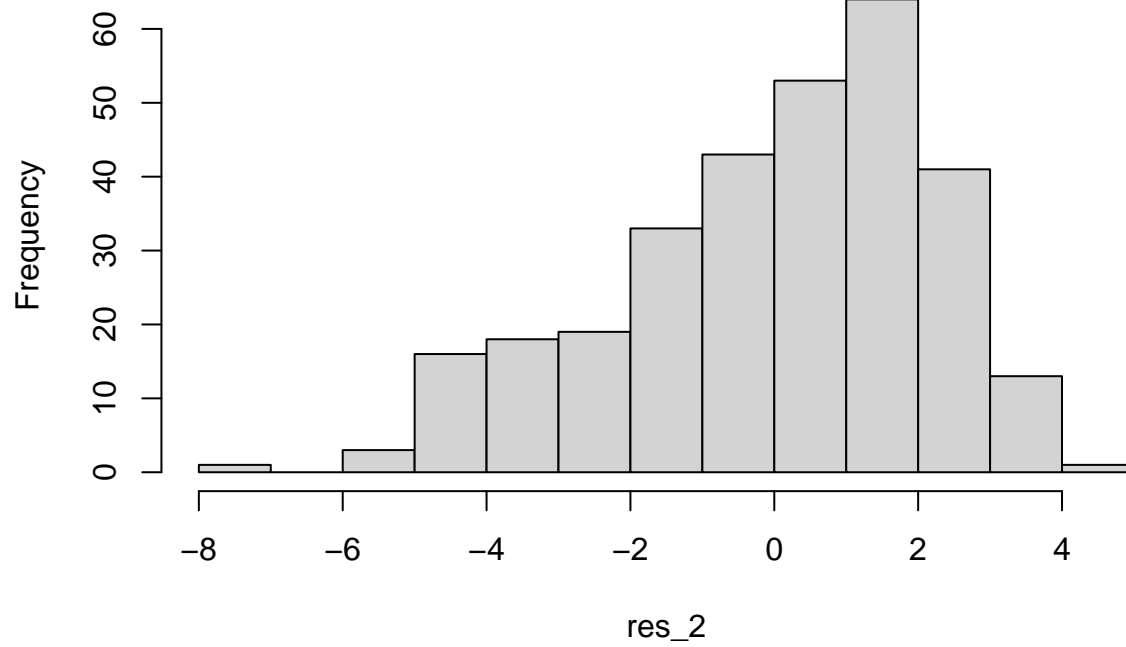
```
ggplot(df_2, aes(y = Ca, x = K)) +
  geom_point()+
  geom_smooth(method='lm', se = F)
```

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



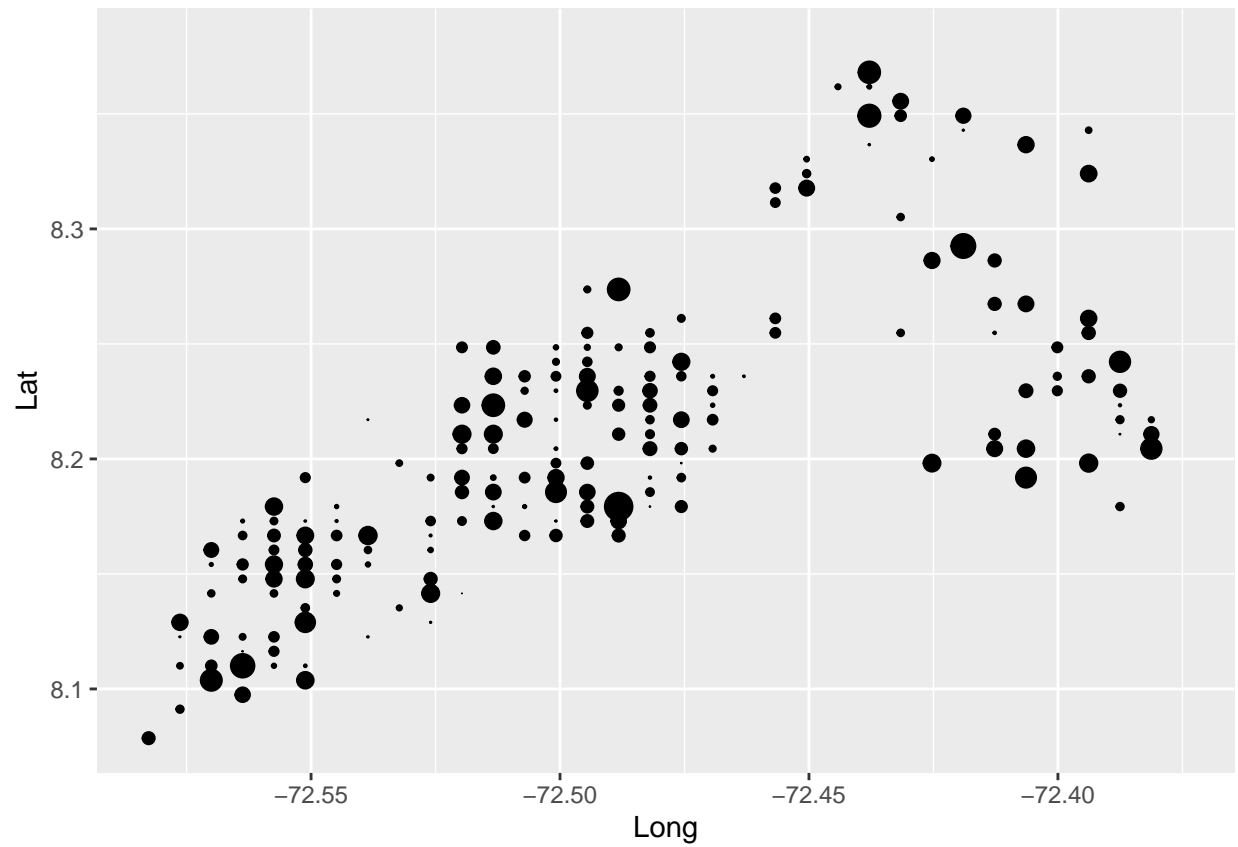
```
res_2 <- model_2$residuals  
hist(res_2)
```

**Histogram of res\_2**



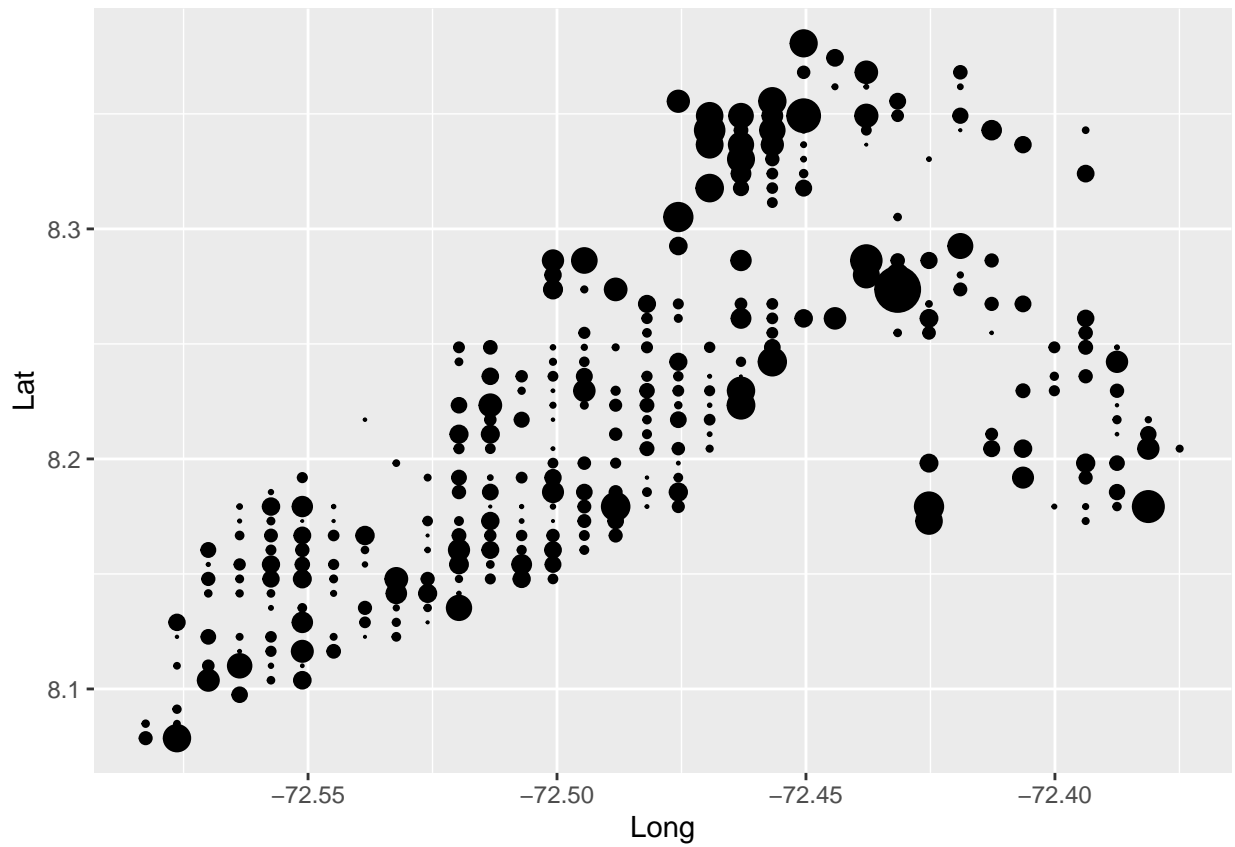
Sin valor absoluto (Con valores negativos)

```
res_2 <- model_2$residuals  
  
ggplot(df_2, aes(Long, Lat))+  
  geom_point(size = res_2)
```

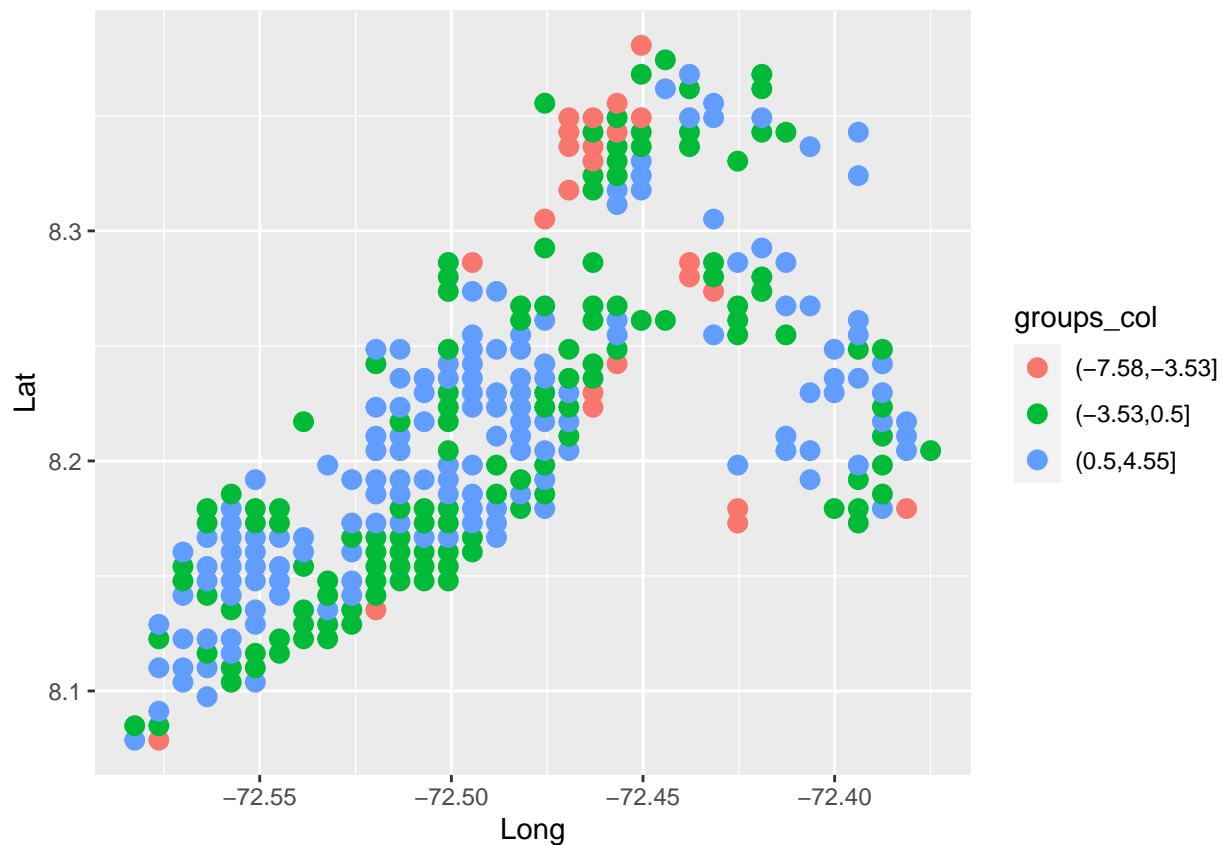


Con corrección

```
ggplot(df_2, aes(Long, Lat))+  
  geom_point(size = abs(res_2))
```



```
groups_col <- cut(res_2, breaks = 3)
#color <-
ggplot(df_2, aes(Long, Lat, color = groups_col))+
  geom_point(size = 3)
```



## Moran Index para residuales

```
matriz_dist <- as.matrix(dist(cbind(x = df_2$Long, y = df_2$Lat)))
```

```
dim(matriz_dist)
```

```
## [1] 305 305
```

```
m_dist_inv <- 1/matriz_dist
```

```
m_dist_inv[is.infinite(m_dist_inv)] <- 0
```

```
diag(m_dist_inv) <- 0
```

```
#m_dist_inv
```

```
Moran.I(res_2, m_dist_inv)
```

```
## $observed
```

```
## [1] 0.09094726
```

```
##
```

```
## $expected
```

```
## [1] -0.003289474
```

```
##
```

```
## $sd
```

```
## [1] 0.005526444
```

```
##
```

```
## $p.value
```

```
## [1] 0
```



## Modelo de regresión multiple

```
model_3 <- lm(Ca ~ K + CICE, data = df)
summary(model_3)

##
## Call:
## lm(formula = Ca ~ K + CICE, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3344  -0.2869   0.2944   0.6983   3.1488
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04521    0.21269  -0.213  0.831764
## K           -2.54957    0.65721  -3.879  0.000122 ***
## CICE         0.79156    0.02121  37.312 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.574 on 400 degrees of freedom
## Multiple R-squared:  0.8196, Adjusted R-squared:  0.8187
## F-statistic: 908.5 on 2 and 400 DF,  p-value: < 2.2e-16
```

$$Y_{Ca} = -0.04521 + -2.54957K + 0.79156z$$

```
res_3 <- model_3$residuals
```

## Moran Index para residuales model 3

```
matriz_dist <- as.matrix(dist(cbind(x = df$Long, y = df$Lat)))
dim(matriz_dist)

## [1] 403 403
m_dist_inv <- 1/matriz_dist
m_dist_inv[is.infinite(m_dist_inv)] <- 0
diag(m_dist_inv) <- 0
#m_dist_inv

Moran.I(res_3, m_dist_inv)

## $observed
## [1] 0.0615839
##
## $expected
## [1] -0.002487562
##
## $sd
## [1] 0.004096863
##
## $p.value
## [1] 0
```

```
model_4 <- lm(Ca ~ K + Long + Lat + I(Long**2) + I(Lat**2), data = df) #datos georreferenciados
summary(model_4)
```

```
##
## Call:
## lm(formula = Ca ~ K + Long + Lat + I(Long^2) + I(Lat^2), data = df)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-11.0367	-1.9749	-0.0537	1.6617	10.1964

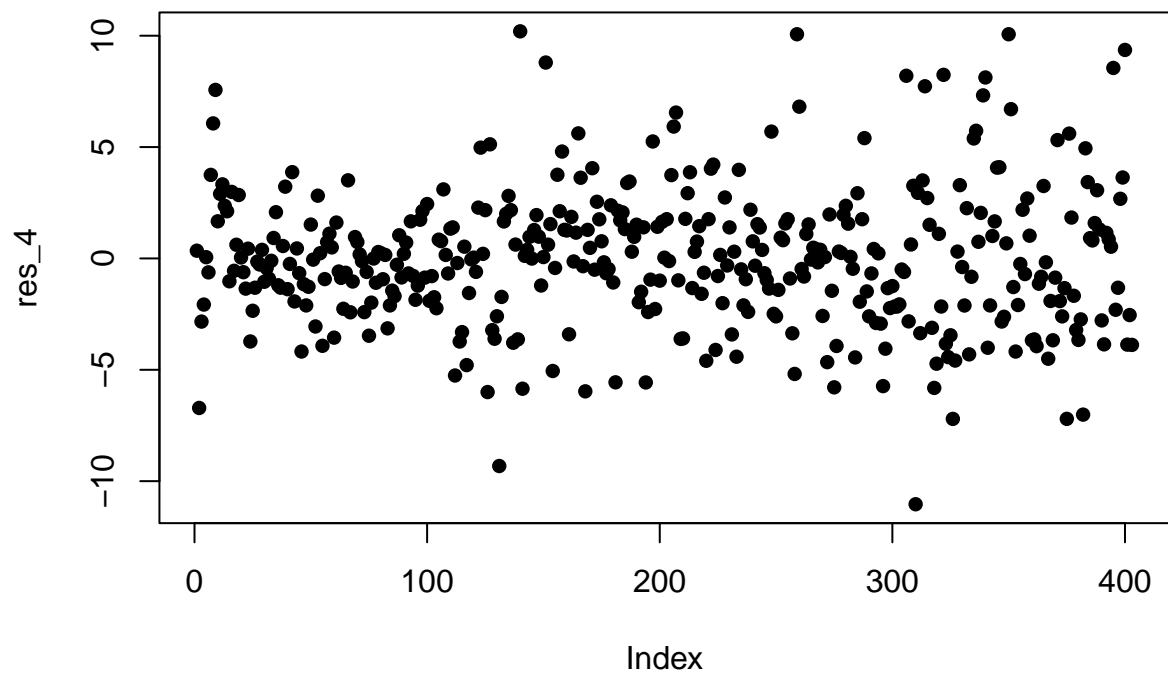
```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.862e+06	3.644e+05	5.110	5.02e-07 ***
K	1.116e+01	1.111e+00	10.049	< 2e-16 ***
Long	5.151e+04	1.009e+04	5.103	5.18e-07 ***
Lat	1.130e+03	5.855e+02	1.931	0.0542 .
I(Long^2)	3.553e+02	6.964e+01	5.101	5.24e-07 ***
I(Lat^2)	-6.861e+01	3.544e+01	-1.936	0.0535 .

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.131 on 397 degrees of freedom
## Multiple R-squared:  0.2911, Adjusted R-squared:  0.2822
## F-statistic: 32.61 on 5 and 397 DF,  p-value: < 2.2e-16
```

```
res_4 <- model_4$residuals
shapiro.test(res_4)
```

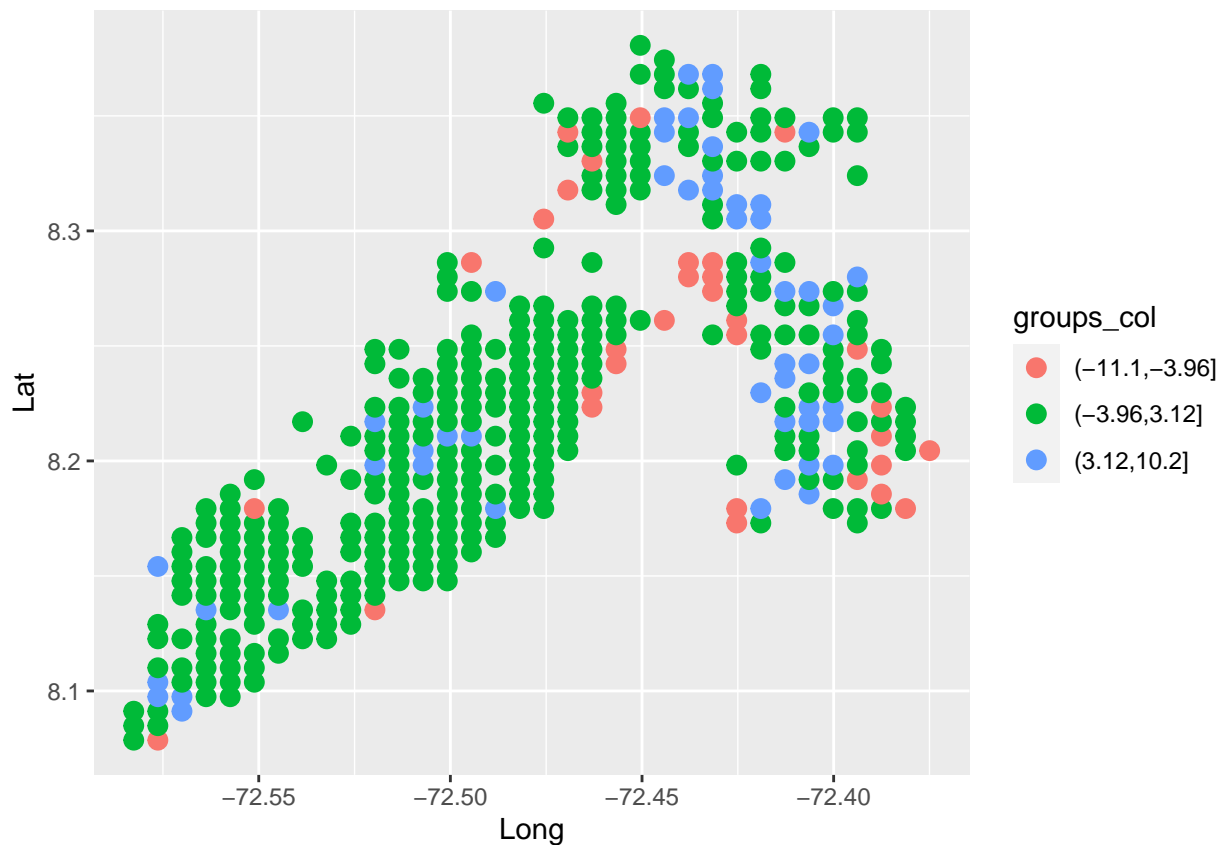
```
##
## Shapiro-Wilk normality test
##
## data:  res_4
## W = 0.98119, p-value = 4.236e-05
plot(res_4, pch = 16)
```



```
Moran.I(res_4, m_dist_inv)
```

```
## $observed
## [1] 0.04129557
##
## $expected
## [1] -0.002487562
##
## $sd
## [1] 0.004255854
##
## $p.value
## [1] 0
```

```
groups_col <- cut(res_4, breaks = 3)
ggplot(df, aes(Long, Lat, color = groups_col))+
  geom_point(size = 3)
```



```
model_5 <- lm(Ca ~ K + I(Long**2) + I(Lat**2) + I(K**2) + Long + Lat , data = df) #datos georreferenciados
summary(model_5)
```

```
##
## Call:
## lm(formula = Ca ~ K + I(Long^2) + I(Lat^2) + I(K^2) + Long +
##     Lat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.2475 -1.7323 -0.2072  1.5916  9.4489
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.489e+06  3.565e+05   4.177 3.63e-05 ***
## K             3.138e+01  3.670e+00   8.550 2.71e-16 ***
## I(Long^2)     2.839e+02  6.812e+01   4.167 3.79e-05 ***
## I(Lat^2)     -4.345e+01  3.436e+01  -1.264   0.207
## I(K^2)       -3.163e+01  5.493e+00  -5.758 1.71e-08 ***
## Long         4.116e+04  9.872e+03   4.169 3.76e-05 ***
## Lat           7.125e+02  5.678e+02   1.255   0.210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.011 on 396 degrees of freedom
## Multiple R-squared:  0.3459, Adjusted R-squared:  0.336
## F-statistic: 34.9 on 6 and 396 DF,  p-value: < 2.2e-16
```

```
res_5 <- model_5$residuals
Moran.I(res_5, m_dist_inv)
```

```
## $observed
## [1] 0.04076852
##
## $expected
## [1] -0.002487562
##
## $sd
## [1] 0.004257141
##
## $p.value
## [1] 0
```

## Modelos de regresión espacial

```
xy = as.matrix(df[,c(2,3)])
```

```
contnb <- dnearneigh(coordinates(xy),0,380000,longlat = F)
dlist <- nbdistis(contnb, xy)
dlist <- lapply(dlist, function(x) 1/x)           #inverse distance
Wve <- nb2listw(contnb,glist=dlist,style = "W")    #W matriz-standardized
```

## Modelo autoregresivo puro

```
model_auto <- spautolm(Ca ~ 1,data = df,listw=Wve)
summary(model_auto)
```

```
##
## Call: spautolm(formula = Ca ~ 1, data = df, listw = Wve)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.44226 -2.27552 -0.18596  1.87723 11.05241
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.0623     4.5096   1.566   0.1173
##
## Lambda: 0.96208 LR test value: 51.922 p-value: 5.7765e-13
## Numerical Hessian standard error of lambda: 0.037647
##
## Log likelihood: -1072.116
## ML residual variance (sigma squared): 11.788, (sigma: 3.4333)
## Number of observations: 403
## Number of parameters estimated: 3
## AIC: 2150.2
```

$$Y_{Ca} = \alpha_0 + \lambda W Y_{Ca} + u u = \rho W u + \epsilon$$

Si  $\rho = 0$ ,  $\mathbf{u} = \epsilon$

$$Y_{Ca} = \alpha_0 + \lambda W Y_{Ca} + \epsilon$$

```
res_6 <- model_auto$fit$residuals
```

```
Moran.I(res_6, m_dist_inv)
```

```
## $observed
## [1] 0.05123807
##
## $expected
## [1] -0.002487562
##
## $sd
## [1] 0.004258707
##
## $p.value
## [1] 0
```