Análise exploratória dos dados

Carregando os dados

```
dados = read_xlsx("Dados_trabalho_1.xlsx")
dados = data.frame(dados)
```

Carregando tabela de referência das variáveis

```
dados_dictionary = read_xlsx("Referência_variaveis.xlsx")
dados_dictionary = data.frame(dados_dictionary)
```

Transformando N/A em zero

Como a base tinha muitos registros vazios, decidimos substituir os N/A por zero para facilitar a modelagem.

```
dados_sem_na <- dados[-1]
dados_sem_na[is.na(dados_sem_na)] <- 0
```

Analisando a qtd de observações distintas das variáveis

Na primeira tentativa de rodar um modelo linear com os dados, recebemos como resultado o erro de que havia variáveis com singularidades, o que não era permitido pelo modelo. Pesquisando, vimos que esse ocorre quando há variáveis explicativas que tem poucos valores únicos, e consequentemente pouca variabilidade. Por esse motivo, decidimos selecionar variáveis a partir de um critério de quantidade de valores únicos: variáveis com mais de 20 valores únicos.

```
dados_temp = data.frame(rep(NA,nrow(dados)))
for (i in (1:ncol(dados))) {
  dados_temp[,i] <- factor(dados[,i])
}

classe_variaveis = NULL
for (i in (1:ncol(dados_temp))) {
  classe_variaveis[i] <- class(dados[,i])
}

niveis_variaveis = NULL
for (i in (1:ncol(dados_temp))) {
  niveis_variaveis[i] <- length(levels(dados_temp[,i]))
}
```

Listando variáveis com mais de 20 níveis

```
variaveis = data.frame(names(dados),classe_variaveis,niveis_variaveis)
names(variaveis) = c("Coluna","Classe","Niveis")
variaveis_explicativas = variaveis[-379,]
ind = variaveis_explicativas$Niveis > 20
variaveis_explicativas = data.frame(variaveis_explicativas,ind)

variaveis_modelo = variaveis_explicativas[ind==TRUE,]
x = paste(variaveis_modelo[-1,1],collapse="'+'")
write.table(x,"output/Teste.txt")

cols_1<-c('AG_AGR_TRAC_NO','AG_LND_AGRI_K2','AG_LND_AGRI_ZS','AG_LND_ARBL_HA','AG_LND_ARBL
_HA_PC','AG_LND_ARBL_ZS','AG_LND_CREL_HA','AG_LND_CROP_ZS','AG_LND_FRST_K2','AG_LND_FRST_Z
S')
cols_2<-c('AG_LND_TRAC_ZS','AG_PRD_CREL_MT','AG_PRD_CROP_XD','AG_PRD_FOOD_XD','AG_PRD_LVSK
_XD','AG_YLD_CREL_KG','BM_TRF_PWKR_CD_DT','BN_TRF_CURR_CD','BX_GRT_EXTA_CD_WD','BX_GRT_TEC
H_CD_WD','BX_KLT_DINV_CD_WD','BX_KLT_DINV_WD_GD_ZS','BX_KLT_DREM_CD_DT','BX_PEF_TOTL_CD_W
D','BX_TRF_PWKR_CD_DT','BX_TRF_PWKR_DT_GD_ZS','CM_MKT_LCAP_CD','CM_MKT_LCAP_GD_ZS','CM_MKT
_LDOM_NO','CM_MKT_TRAD_CD','CM_MKT_TRAD_GD_ZS','CM_MKT_TRNR','DC_DAC_AUTL_CD','DC_DAC_BELL
 CD','DC_DAC_CANL_CD','DC_DAC_CECL_CD','DC_DAC_CHEL_CD','DC_DAC_DEUL_CD','DC_DAC_DNKL_CD',
```

Análise de componentes principais

Selecionando apenas as vaiáveis com mais de vinte observações distintas, ficamos com 721 variáveis explicativas passíveis de entrar no modelo, e o modelo linear passou a rodar. No entanto, no summary do modelo vimos que boa parte das variáveis explicativas ficaram com N/A em parâmetro estimado, Desvio Padrão e Teste t. Por esse motivo, decidimos avançar com uma abordagem mais sofisticada para redução de dimensionalidade. A técnica escolhida foi a Análise de Componentes Principais. Este método agrupa variáveis a partir de sua covariâncias, gerando combinações lineares das variáveis, que se chamam Componentes Principais.

```
acomp <- prcomp(dados_relevantes, scale = TRUE)
summary(acomp)
```

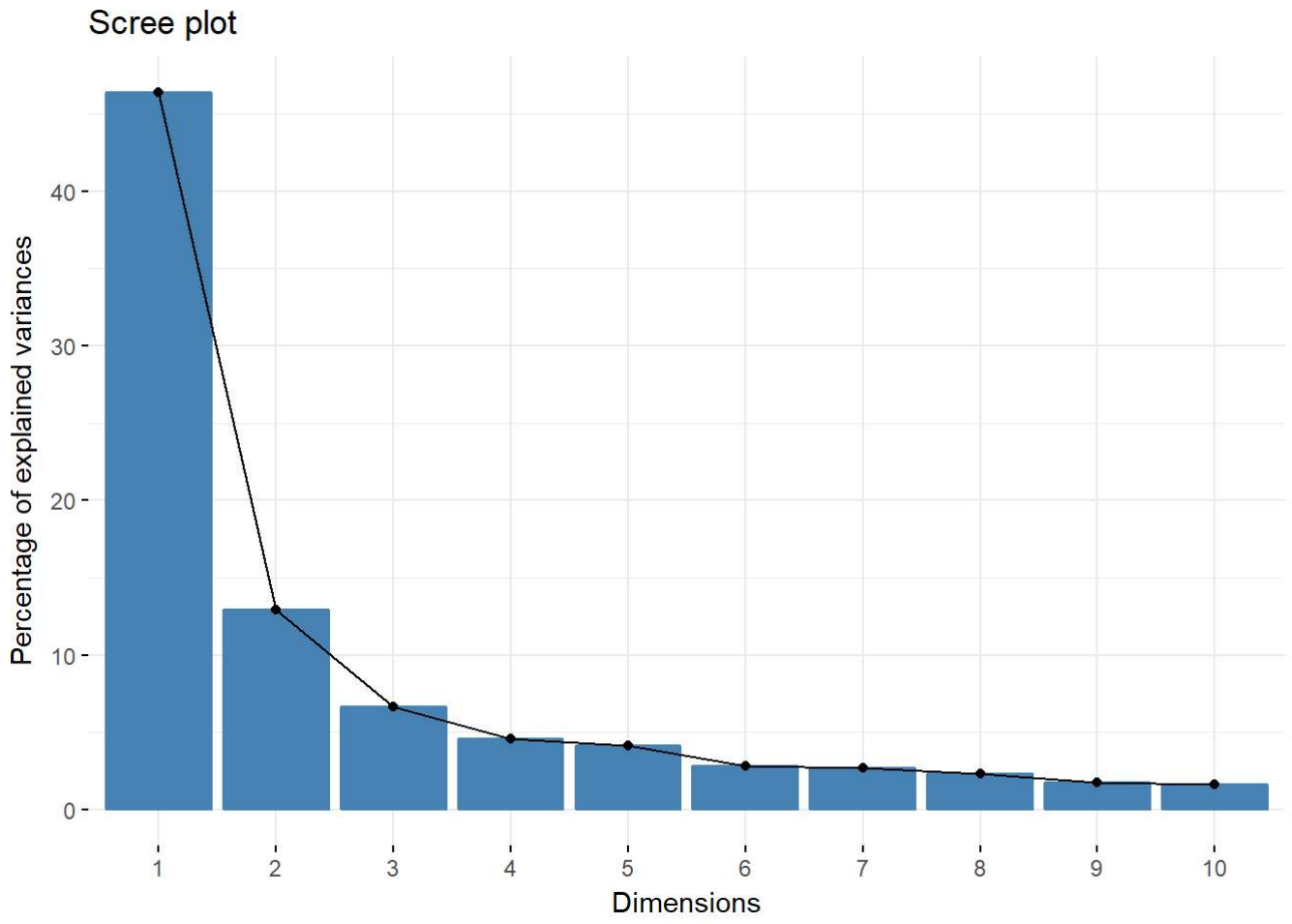
```
## Importance of components:
##                             PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     18.2934 9.6619 6.9194 5.76893 5.46266 4.51541
## Proportion of Variance  0.4642 0.1295 0.0664 0.04616 0.04139 0.02828
## Cumulative Proportion   0.4642 0.5936 0.6600 0.70619 0.74757 0.77585
##                             PC7     PC8     PC9   PC10    PC11    PC12
## Standard deviation     4.42783 4.11473 3.57064 3.4282 3.05433 2.96292
## Proportion of Variance 0.02719 0.02348 0.01768 0.0163 0.01294 0.01218
## Cumulative Proportion  0.80304 0.82653 0.84421 0.8605 0.87345 0.88563
##                            PC13   PC14    PC15    PC16    PC17    PC18
## Standard deviation     2.86419 2.6581 2.55666 2.37356 2.27255 1.99459
## Proportion of Variance 0.01138 0.0098 0.00907 0.00781 0.00716 0.00552
## Cumulative Proportion  0.89700 0.9068 0.91587 0.92368 0.93085 0.93636
##                            PC19    PC20    PC21    PC22   PC23    PC24
## Standard deviation     1.8613 1.83854 1.83080 1.80955 1.6773 1.59068
## Proportion of Variance 0.0048 0.00469 0.00465 0.00454 0.0039 0.00351
## Cumulative Proportion  0.9412 0.94586 0.95051 0.95505 0.9589 0.96246
##                            PC25    PC26    PC27    PC28    PC29    PC30
## Standard deviation     1.54419 1.48666 1.39097 1.33870 1.32894 1.25629
## Proportion of Variance 0.00331 0.00307 0.00268 0.00249 0.00245 0.00219
## Cumulative Proportion  0.96577 0.96883 0.97152 0.97400 0.97645 0.97864
##                            PC31    PC32    PC33    PC34    PC35    PC36
## Standard deviation     1.20797 1.14172 1.10081 1.08267 1.06162 1.01547
## Proportion of Variance 0.00202 0.00181 0.00168 0.00163 0.00156 0.00143
## Cumulative Proportion  0.98066 0.98247 0.98415 0.98578 0.98734 0.98877
##                            PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.97071 0.93637 0.87435 0.85457 0.82041 0.77759
## Proportion of Variance 0.00131 0.00122 0.00106 0.00101 0.00093 0.00084
## Cumulative Proportion  0.99008 0.99129 0.99235 0.99337 0.99430 0.99514
##                            PC43    PC44   PC45    PC46    PC47   PC48
## Standard deviation     0.72870 0.68827 0.6566 0.63080 0.61112 0.5363
## Proportion of Variance 0.00074 0.00066 0.0006 0.00055 0.00052 0.0004
## Cumulative Proportion  0.99588 0.99653 0.9971 0.99768 0.99820 0.9986
##                            PC49    PC50    PC51    PC52    PC53    PC54
## Standard deviation     0.49508 0.47726 0.40621 0.35403 0.31736 0.28398
## Proportion of Variance 0.00034 0.00032 0.00023 0.00017 0.00014 0.00011
## Cumulative Proportion  0.99894 0.99926 0.99948 0.99966 0.99980 0.99991
##                            PC55      PC56
## Standard deviation     0.25476 8.124e-15
## Proportion of Variance 0.00009 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00
```

Interpretando os componentes principais

- Artigo Referência (http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/)

Visualização do percentual de variância explicada por cada Componente

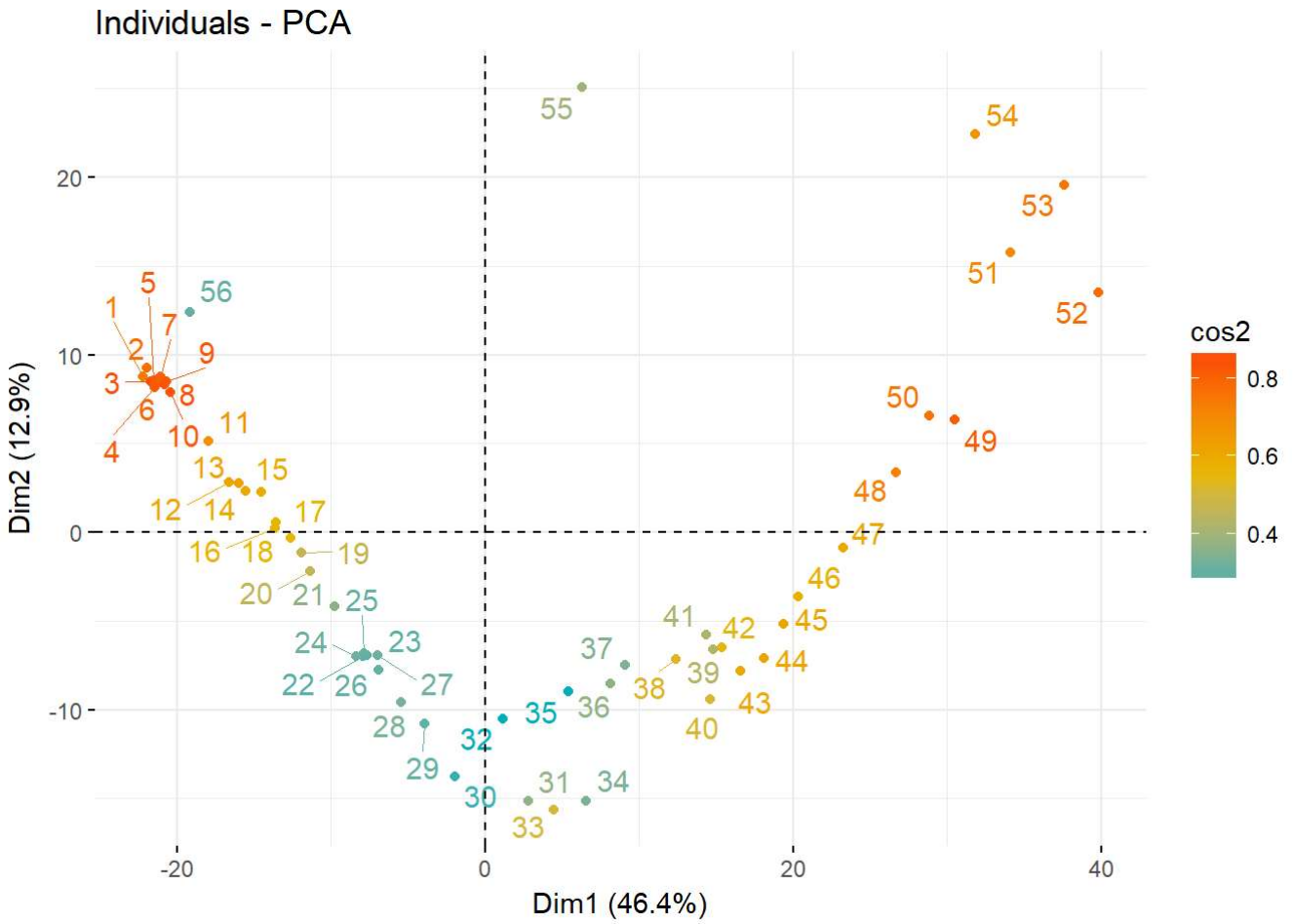
```
fviz_eig(acomp)
```



Visualização do grau de associação de cada registro com os primeiros dois Componentes

Como cada registro é de um ano, o registro 1 representa o ano 1960 e consecutivamente até o resgistro 56, que representa o ano 2005.

```
fviz_pca_ind(acomp,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```



Entendendo a relação das variáveis iniciais com os Componentes

```
res.var <- get_pca_var(acomp)
res.var$coord          # Coordenadas
```

```
##                                Dim.1         Dim.2         Dim.3
## AG_AGR_TRAC_NO          0.185442343 -0.8969051335 -4.846777e-02
## AG_LND_AGRI_K2          0.623148204 -0.4593621688  2.873067e-01
## AG_LND_AGRI_ZS          0.623148204 -0.4593621688  2.873067e-01
## AG_LND_ARBL_HA          0.826390919 -0.3114716465  2.438793e-01
## AG_LND_ARBL_HA_PC       0.159452939 -0.3448360780  4.389083e-01
## AG_LND_ARBL_ZS          0.826390919 -0.3114716465  2.438793e-01
## AG_LND_CREL_HA          0.338584873 -0.5308387744  5.538255e-01
## AG_LND_CROP_ZS          0.344212207 -0.4612051093  2.437061e-01
## AG_LND_FRST_K2          0.792316204 -0.3144256948 -4.282559e-01
## AG_LND_FRST_ZS          0.851389101 -0.2299074406 -3.638628e-01
## AG_LND_TRAC_ZS          0.011398014 -0.9295257349  5.810846e-02
## AG_PRD_CREL_MT          0.904356230 -0.0395197945  2.052226e-01
## AG_PRD_CROP_XD          0.934429857 -0.0415926038  1.474315e-01
## AG_PRD_FOOD_XD          0.954542022 -0.0276880898  8.818118e-02
## AG_PRD_LVSK_XD          0.959755727 -0.0200354102  3.261228e-02
## AG_YLD_CREL_KG          0.900300270  0.0066001019  1.003699e-01
## BM_TRF_PWKR_CD_DT       0.803920692  0.4266683091  1.977344e-02
## BN_TRF_CURR_CD          0.903438799  0.0424531983 -2.241472e-01
```

```
res.var$contrib        # Contribuições para os Componentes Principais
```

```
##                                  Dim.1        Dim.2        Dim.3
## AG_AGR_TRAC_NO            1.027609e-02 8.617234e-01 4.906522e-03
## AG_LND_AGRI_K2            1.160360e-01 2.260400e-01 1.724087e-01
## AG_LND_AGRI_ZS            1.160360e-01 2.260400e-01 1.724087e-01
## AG_LND_ARBL_HA            2.040708e-01 1.039230e-01 1.242274e-01
## AG_LND_ARBL_HA_PC         7.597574e-03 1.273797e-01 4.023604e-01
## AG_LND_ARBL_ZS            2.040708e-01 1.039230e-01 1.242274e-01
## AG_LND_CREL_HA            3.425666e-02 3.018562e-01 6.406393e-01
## AG_LND_CROP_ZS            3.540482e-02 2.278574e-01 1.240511e-01
## AG_LND_FRST_K2            1.875888e-01 1.059036e-01 3.830667e-01
## AG_LND_FRST_ZS            2.166038e-01 5.662144e-02 2.765305e-01
## AG_LND_TRAC_ZS            3.882114e-05 9.255453e-01 7.052553e-03
## AG_PRD_CREL_MT            2.443931e-01 1.673032e-03 8.796668e-02
## AG_PRD_CROP_XD            2.609176e-01 1.853135e-03 4.539922e-02
## AG_PRD_FOOD_XD            2.722701e-01 8.212225e-04 1.624125e-02
## AG_PRD_LVSK_XD            2.752525e-01 4.300028e-04 2.221417e-03
## AG_YLD_CREL_KG            2.422059e-01 4.666337e-05 2.104139e-02
## BM_TRF_PWKR_CD_DT         1.931240e-01 1.950094e-01 8.166425e-04
## BN_TRF_CURR_CD            2.438975e-01 1.930615e-03 1.049384e-01
```

Entendendo a relação dos registros com os Componentes

```
res.ind <- get_pca_ind(acomp)
res.ind$coord          # Coordenadas
```

```
##         Dim.1       Dim.2       Dim.3       Dim.4       Dim.5        Dim.6
## 1  -22.219361   8.7702975  -7.4493503  0.73410202 -0.82406575   3.03616303
## 2  -21.957175   9.2364288  -6.1442348 -3.24304620  0.10385771   4.59762815
## 3  -21.700927   8.4864877  -6.1914778 -2.81417196  0.71196733   5.01445654
## 4  -21.434700   8.1708500  -5.8622874 -2.11790235  0.26315120   6.06295373
## 5  -21.525276   8.5404754  -5.9652008 -2.63873233  0.41156370   6.18253401
## 6  -21.016176   8.4216023  -5.6303678 -2.65563721  0.21431707   6.21823248
## 7  -21.087628   8.7666519  -5.5089771 -3.34184187  0.80661298   5.99189820
## 8  -20.814338   8.3424318  -5.3905579 -2.95447923  0.59793364   4.74351531
## 9  -20.698567   8.4929509  -4.5671313 -4.41774147  0.89345754   3.40044010
## 10 -20.442444   7.8672560  -4.5510698 -3.23319229  0.40114653   3.32217189
## 11 -17.933854   5.1263197   0.6088323 -5.90256918  1.35729673  -2.98551825
## 12 -16.644359   2.8221745   3.1509743 -7.08585804  1.02777667  -6.33633683
## 13 -15.957860   2.7654768   3.7472683 -6.70588173  1.12610849  -6.67998647
## 14 -15.523944   2.3079589   4.5818789 -6.58870911  0.96902662  -7.03769224
## 15 -14.534692   2.2379623   5.4424148 -5.21633917  0.58847787  -5.82648217
## 16 -13.661440   0.2256432   6.8087956 -2.12119924  0.22937875  -4.20221420
## 17 -13.599737   0.5367634   7.1185338 -3.55965044  0.83261197  -5.96963076
## 18 -12.647580  -0.3390958   7.3467871 -1.69797936  0.25548854  -4.64740168
```

```
res.ind$contrib        # Contribuições para os Componentes Principais
```

```
##          Dim.1        Dim.2       Dim.3        Dim.4        Dim.5
## 1  2.634419951 1.471348e+00 2.069742867 2.891569e-02 4.063754e-02
## 2  2.572614890 1.631905e+00 1.408040699 5.643221e-01 6.454776e-04
## 3  2.512918684 1.377662e+00 1.429776775 4.249345e-01 3.033359e-02
## 4  2.451639949 1.277089e+00 1.281780953 2.406759e-01 4.143949e-03
## 5  2.472403398 1.395246e+00 1.327179731 3.736039e-01 1.013626e-02
## 6  2.356835364 1.356676e+00 1.182369261 3.784062e-01 2.748636e-03
## 7  2.372888512 1.470125e+00 1.131935134 5.992286e-01 3.893445e-02
## 8  2.311782957 1.331288e+00 1.083794712 4.683631e-01 2.139487e-02
## 9  2.286137935 1.379762e+00 0.777976713 1.047180e+00 4.776959e-02
## 10 2.229910824 1.183950e+00 0.772514439 5.608979e-01 9.629633e-03
## 11 1.716204733 5.026876e-01 0.013825312 1.869403e+00 1.102436e-01
## 12 1.478277663 1.523542e-01 0.370313325 2.694050e+00 6.321224e-02
## 13 1.358848954 1.462941e-01 0.523732084 2.412862e+00 7.588643e-02
## 14 1.285955624 1.018927e-01 0.783009097 2.329279e+00 5.619207e-02
## 15 1.127284517 9.580592e-02 1.104747132 1.459998e+00 2.072354e-02
## 16 0.995897921 9.739377e-04 1.729100326 2.414258e-01 3.148546e-03
## 17 0.986922153 5.511282e-03 1.889995212 6.798850e-01 4.148479e-02
## 18 0.853565232 2.199539e-03 2.013142454 1.546983e-01 3.906128e-03
```

Modelo Linear com a variável Emissões de CO2 explicada pelas componentes principais

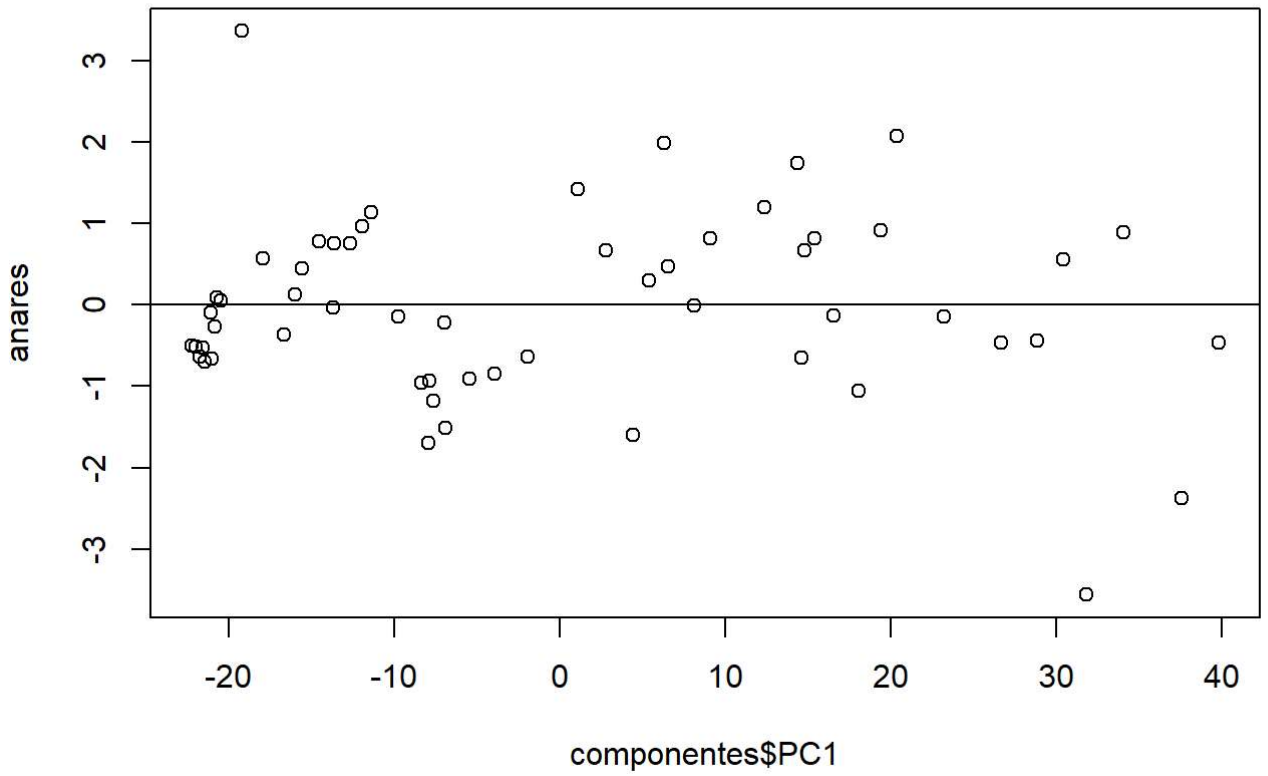
```
componentes <- data.frame(dados_sem_na$EN_ATM_CO2E_KT,acomp$x)
nomes <- NULL
for (i in 1:(ncol(componentes)-1)) {nomes[i] <- paste0("PC",i)}
names(componentes) = c("EN_ATM_CO2E_KT",nomes)


mod=lm(EN_ATM_CO2E_KT ~ PC1+PC2+PC4+PC5+PC7+PC9+PC10+PC13, data = componentes)
summary(mod)
```

```
##
## Call:
## lm(formula = EN_ATM_CO2E_KT ~ PC1 + PC2 + PC4 + PC5 + PC7 + PC9 +
##     PC10 + PC13, data = componentes)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -51973 -14548  -2545  16413  62924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 193514.9     3248.9  59.564  < 2e-16 ***
## PC1           4575.1      179.2  25.530  < 2e-16 ***
## PC2          -5647.7      339.3 -16.645  < 2e-16 ***
## PC4          -6400.1      568.3 -11.263 6.02e-15 ***
## PC5          -3735.5      600.1  -6.225 1.23e-07 ***
## PC7          -4491.6      740.4  -6.067 2.13e-07 ***
## PC9          -3960.9      918.1  -4.314 8.17e-05 ***
## PC10         -7658.9      956.3  -8.009 2.47e-10 ***
## PC13         -8183.1     1144.6  -7.149 4.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24310 on 47 degrees of freedom
## Multiple R-squared:  0.9642, Adjusted R-squared:  0.9581
## F-statistic: 158.1 on 8 and 47 DF,  p-value: < 2.2e-16
```

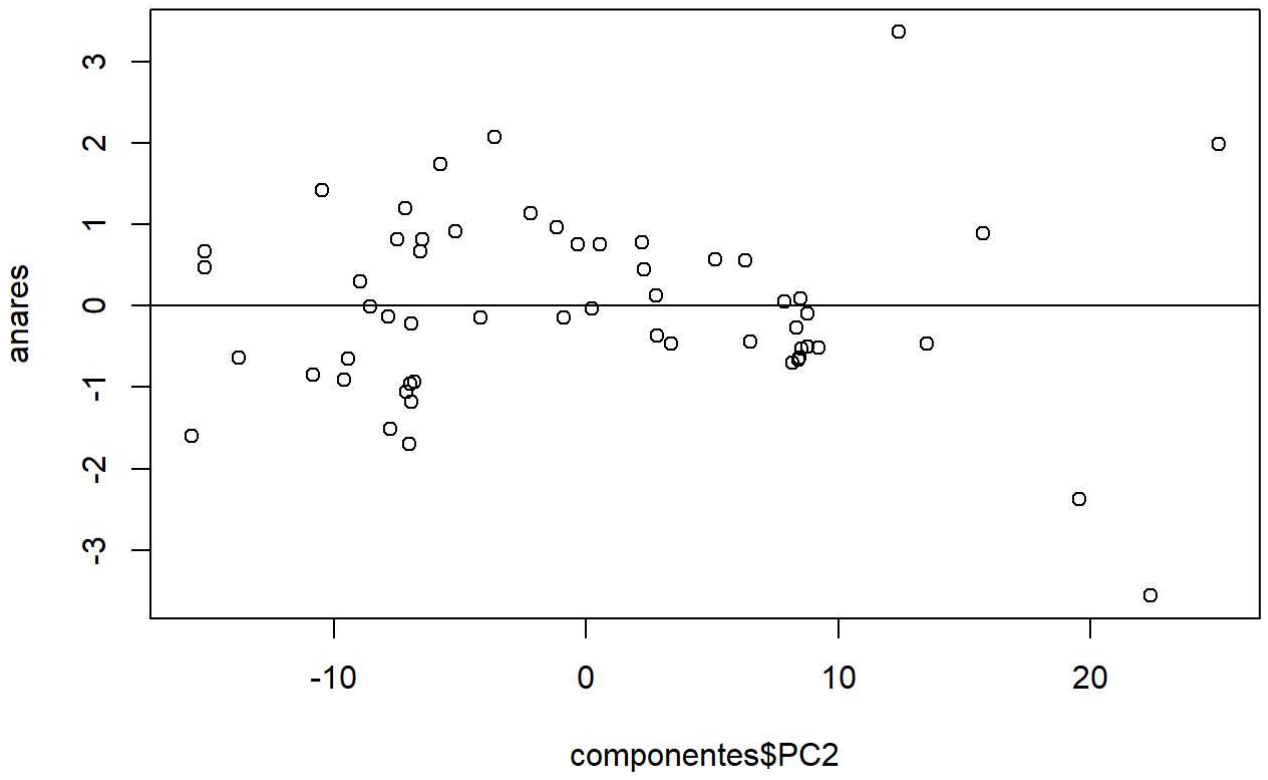
Análise de resíduos

Podemos ver que os resíduos se encontram aleatoriamente distribuidos em relação à variável resposta e também às variáveis explicativas

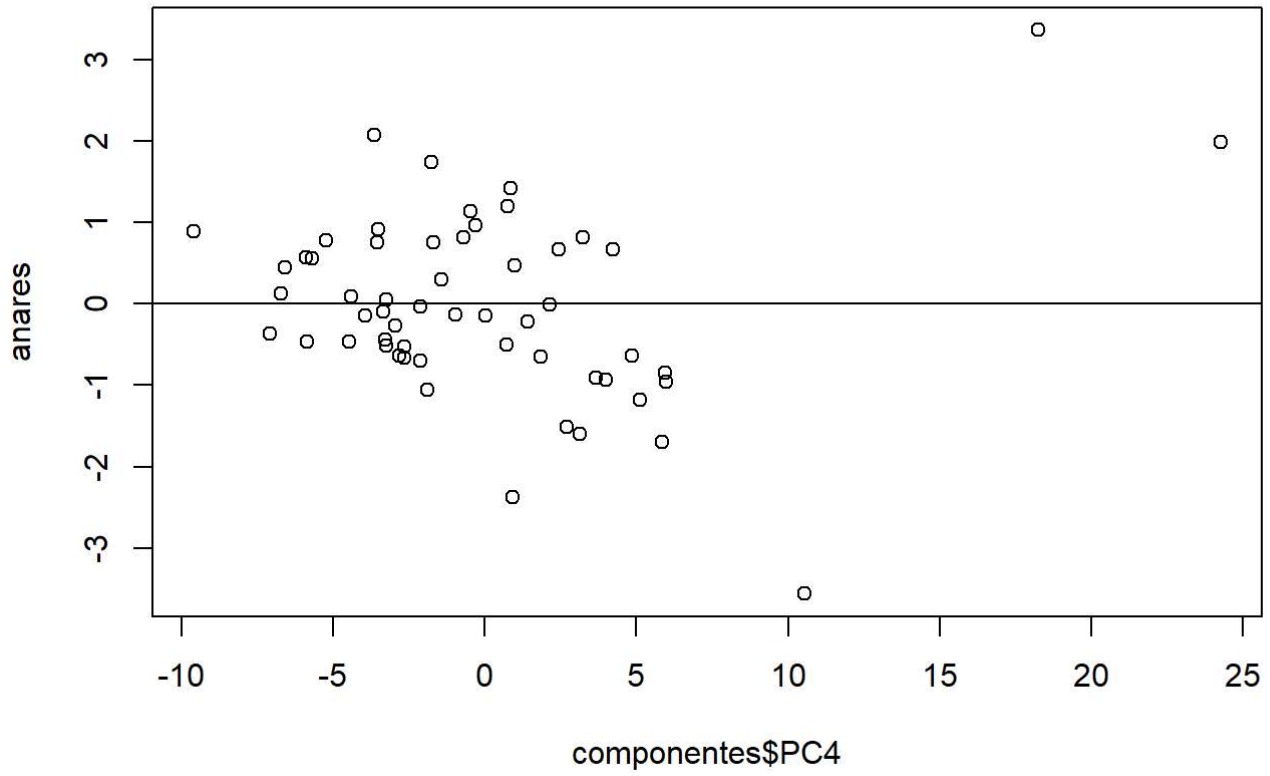
```
anares<-rstandard(mod)
par(mfrow=c(1,1))
plot(componentes$PC1, anares);abline(0,0)
```



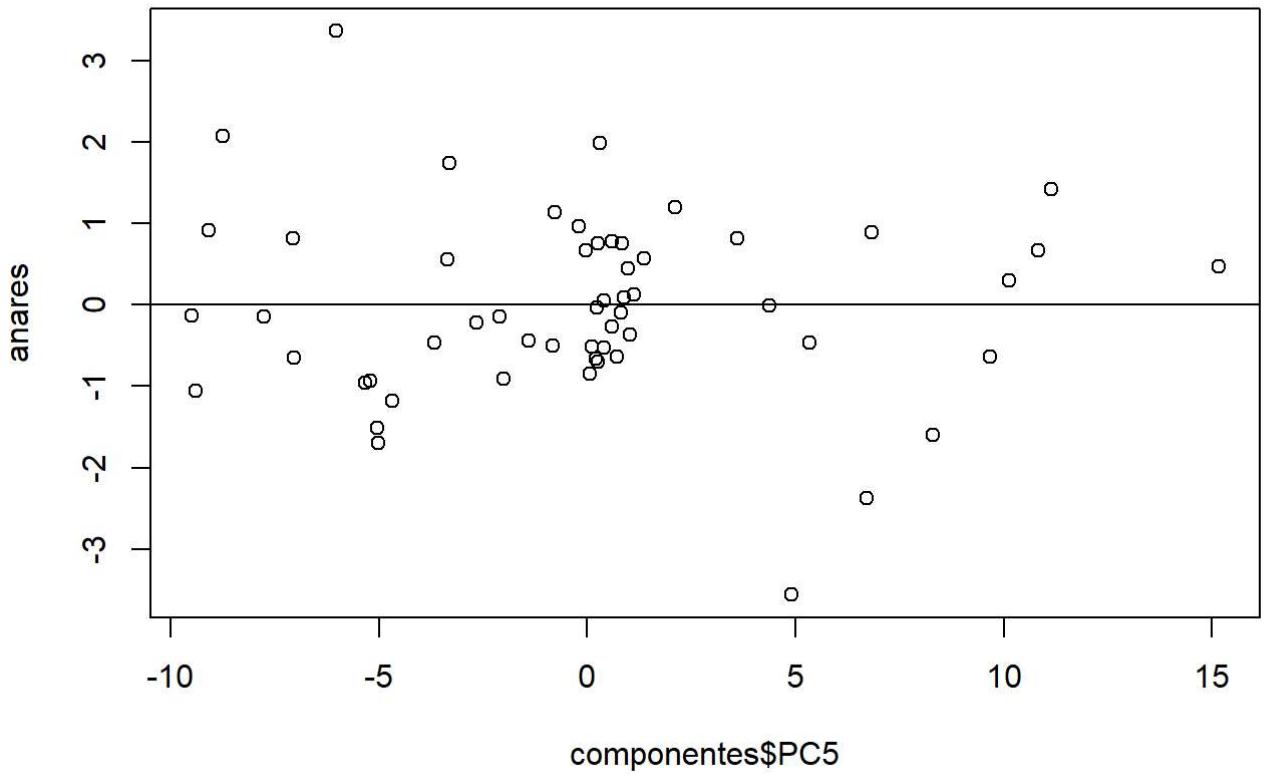

```
plot(componentes$PC2, anares);abline(0,0)
```



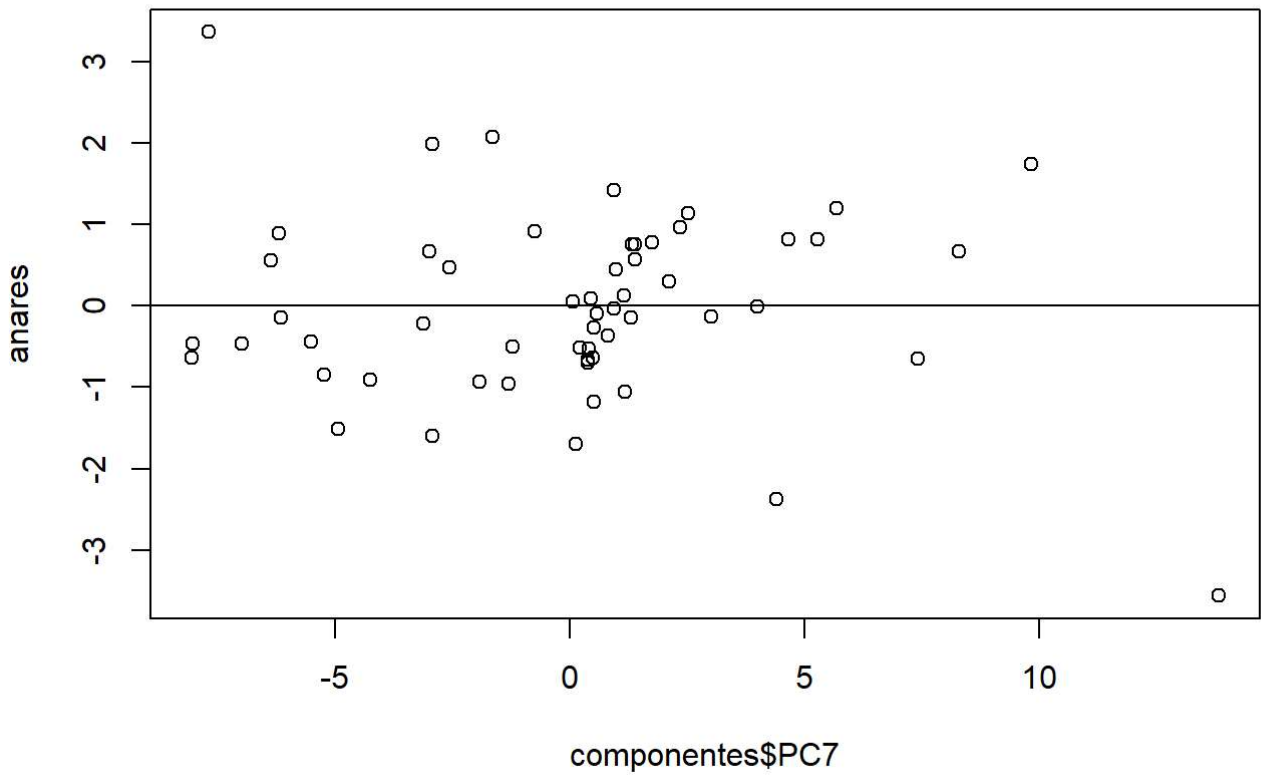
```
plot(componentes$PC4, anares);abline(0,0)
```



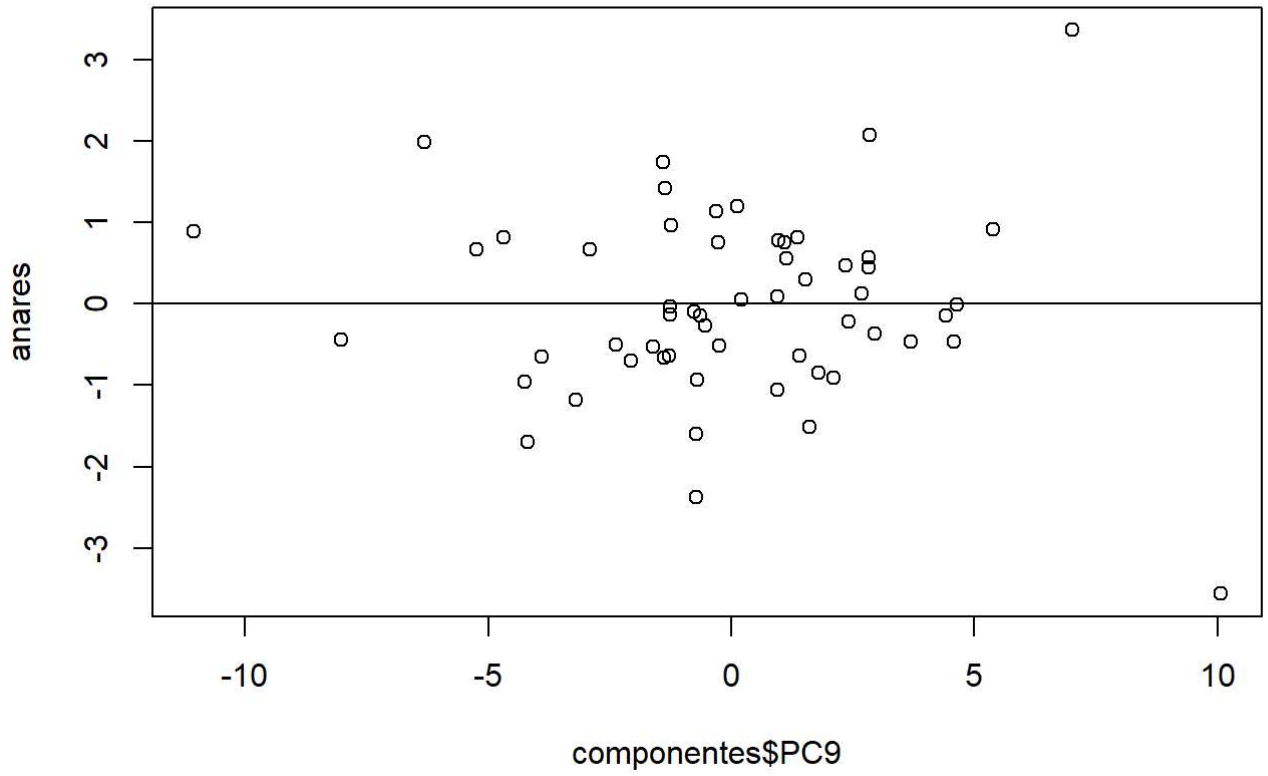

```
plot(componentes$PC5, anares);abline(0,0)
```



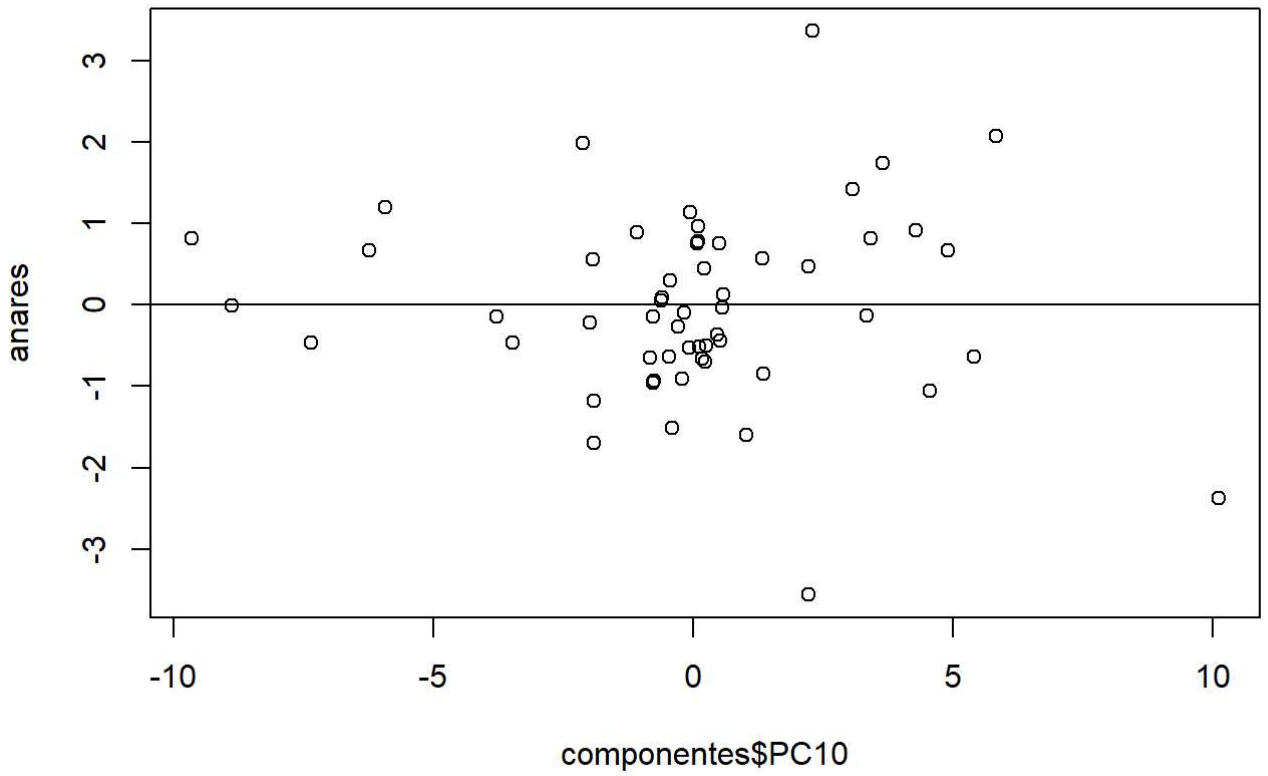

```
plot(componentes$PC7, anares);abline(0,0)
```



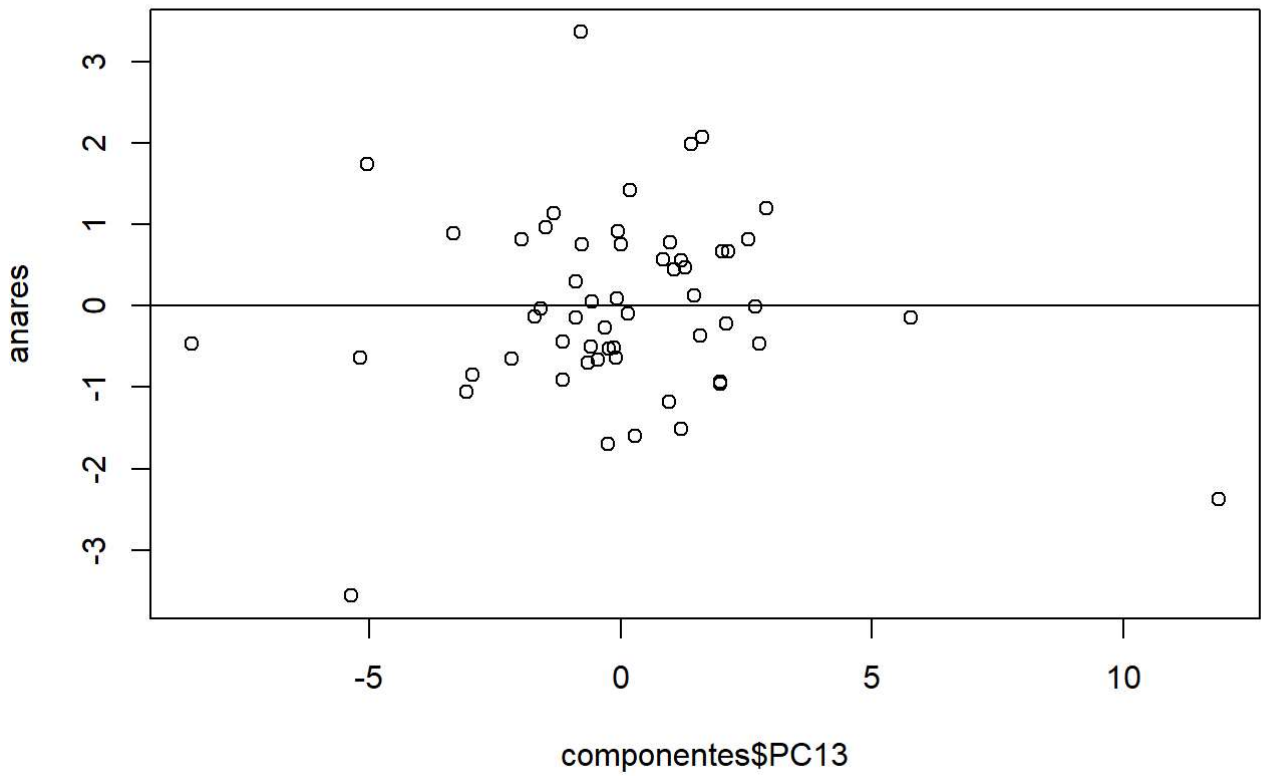
```
plot(componentes$PC9, anares);abline(0,0)
```




```
plot(componentes$PC10, anares);abline(0,0)
```

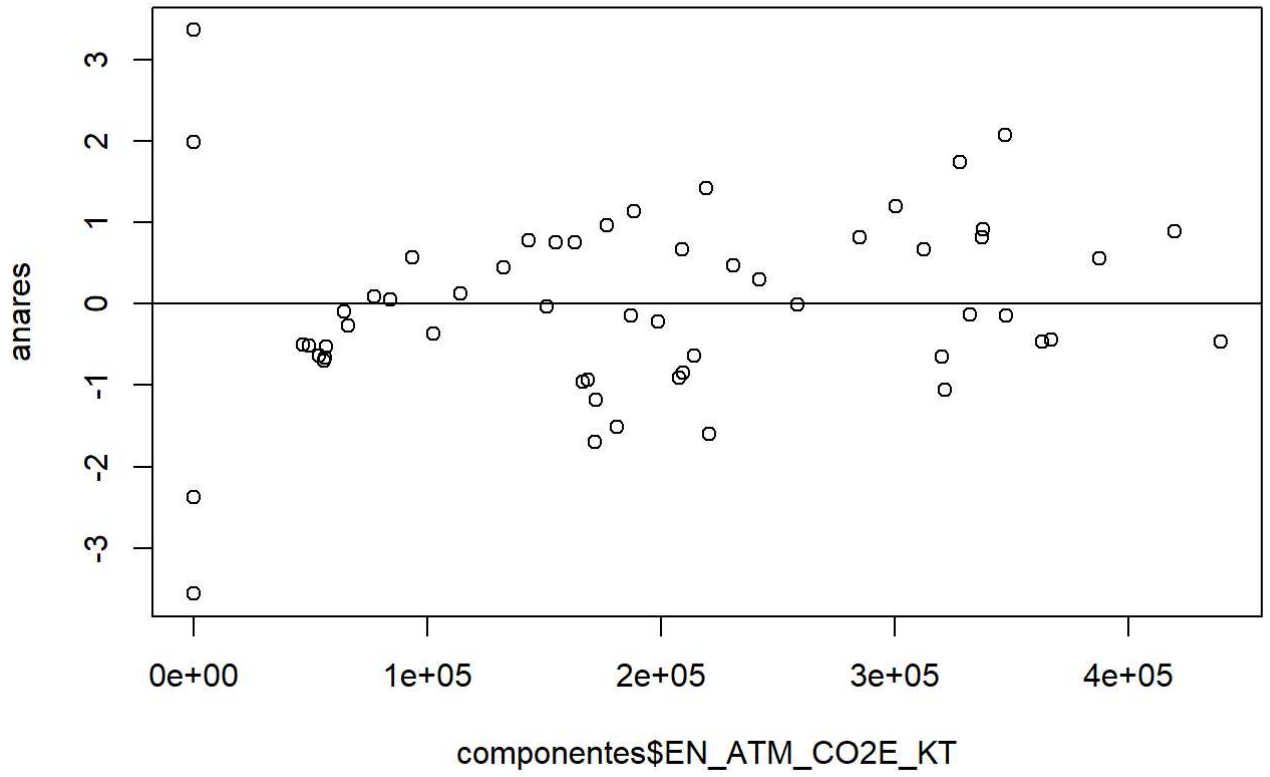



```
plot(componentes$PC13, anares);abline(0,0)
```

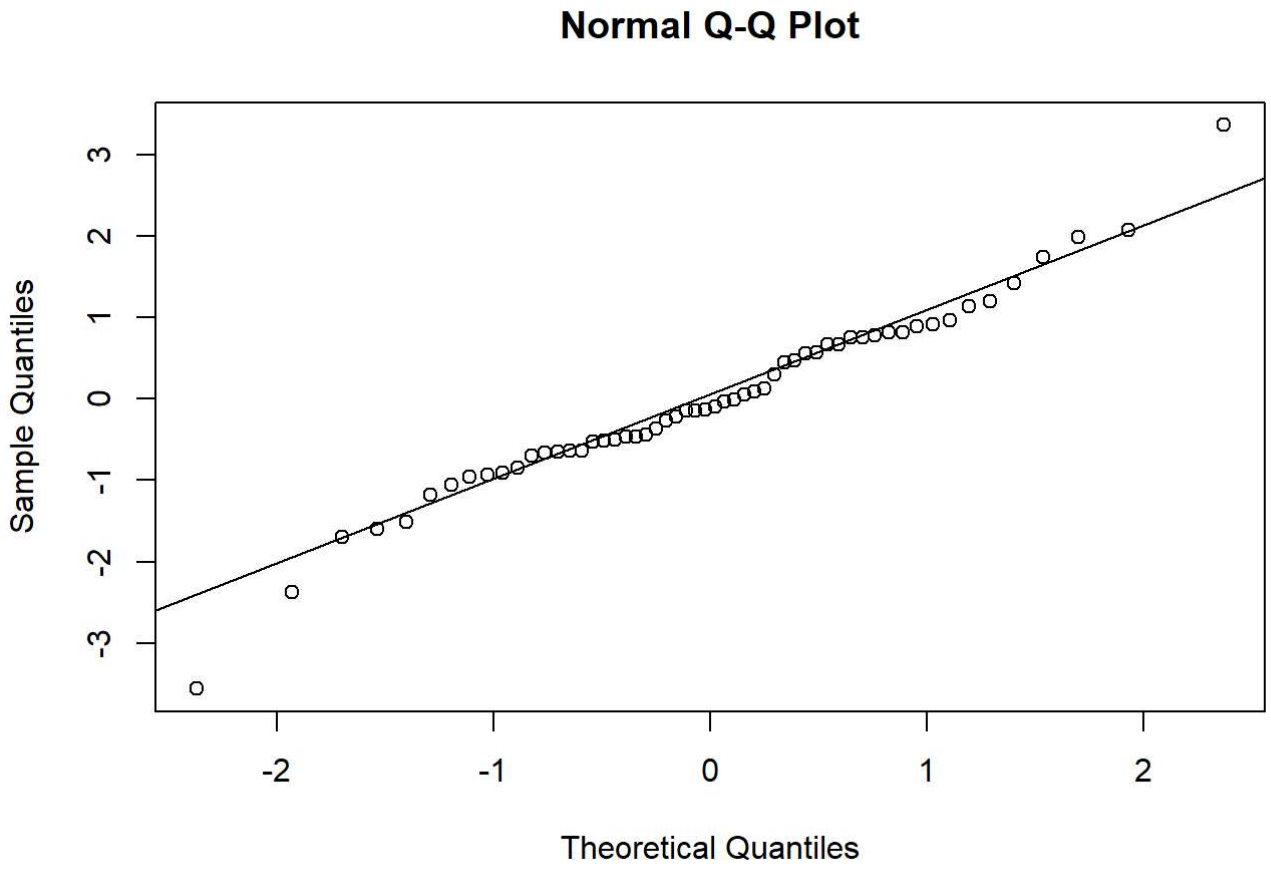


Também é possível ver que os resíduos atendem ao pressuposto de normalidade

```
plot(componentes$EN_ATM_CO2E_KT, anares)
abline(0,0)
```




```
qqnorm(anares)
qqline(anares)
```



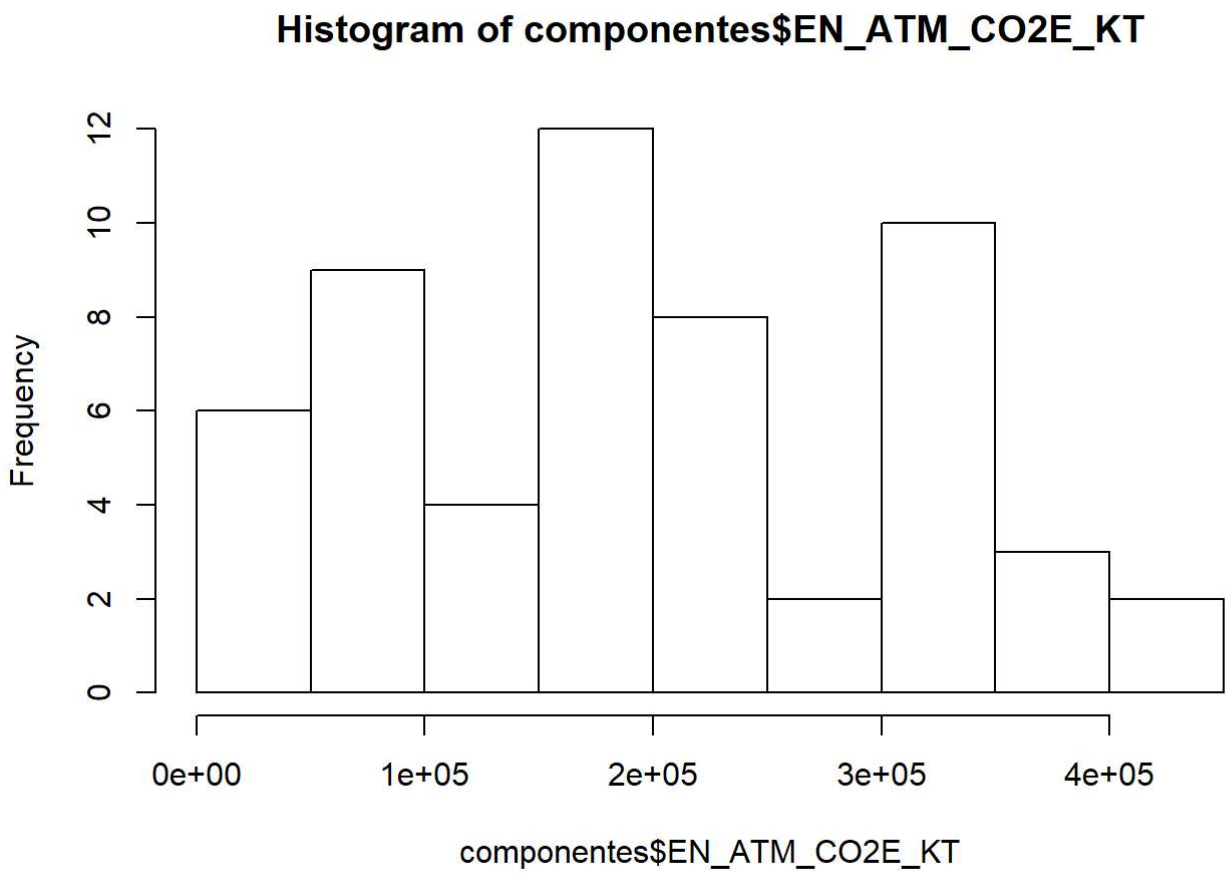
Medida ACI de qualidade do ajuste

```
AIC(mod)
```

```
## [1] 1300.168
```

Testando hipóteses de normalidade da variável resposta Emissões de CO2

```
hist(componentes$EN_ATM_CO2E_KT)
```




```
ks.test(x=componentes$EN_ATM_CO2E_KT, y=pnorm, alternative = c("two.sided", "less", "greater"
),exact = NULL) #Rejeitada hipótese de normalidade
```

```
## Warning in ks.test(x = componentes$EN_ATM_CO2E_KT, y = pnorm, alternative =
## c("two.sided", : ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  componentes$EN_ATM_CO2E_KT
## D = 0.92857, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ad.test(componentes$EN_ATM_CO2E_KT)
```

```
##
##  Anderson-Darling normality test
##
## data:  componentes$EN_ATM_CO2E_KT
## A = 0.61194, p-value = 0.1062
```

```
shapiro.test(componentes$EN_ATM_CO2E_KT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  componentes$EN_ATM_CO2E_KT
## W = 0.96199, p-value = 0.07498
```

A normalidade foi rejeitada no teste Kolmogorov-Smirnov, mas não nos testes Anderson-Darling e Shapiro.

Testando hipóteses de outras distribuições da variável resposta Emissões de CO2

```
ks.test(x=componentes$EN_ATM_CO2E_KT, y=pchisq, df=47, alternative = c("two.sided", "less",
"greater"),exact = NULL)
```

```
## Warning in ks.test(x = componentes$EN_ATM_CO2E_KT, y = pchisq, df = 47, :
## ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  componentes$EN_ATM_CO2E_KT
## D = 0.92857, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(x=componentes$EN_ATM_CO2E_KT, y=ppois, lambda=0.5, alternative = c("two.sided", "les
s", "greater"),exact = NULL)
```

```
## Warning in ks.test(x = componentes$EN_ATM_CO2E_KT, y = ppois, lambda =
## 0.5, : ties should not be present for the Kolmogorov-Smirnov test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  componentes$EN_ATM_CO2E_KT
## D = 0.92857, p-value < 2.2e-16
## alternative hypothesis: two-sided
```