Exámen final Regresión Avanzada 2023

Marcos Buccellato

2023-07-18

Contents


```
datos <- read.csv2('data_pancreas_resumen.csv',sep=";")
datos[, 'sexo'] <- as.factor(datos[, 'sexo'])
datos[, 'estadio'] <- as.factor(datos[, 'estadio'])
datos[, 'diagnosis'] <- as.factor(datos[, 'diagnosis'])
```

```
library(splitstackshape)
```

```
## Warning: package 'splitstackshape' was built under R version 4.3.1
```

```
set.seed(24564066)
strat_data <- stratified(datos, "diagnosis", 300/nrow(datos))
```

Ejercicio 1

1) Construya un modelo lineal simple para explicar el valor de la creatinina en función de alguna de las restantes variables numéricas y evalúe la bondad del ajuste.

Probamos un modelo lineal muy sencillo

```
model11_simple <- lm(creatinina ~ LYVE1,data=strat_data)
summary(model11_simple)
```

```
##
## Call:
## lm(formula = creatinina ~ LYVE1, data = strat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0558 -0.4192 -0.1578  0.2666  2.8200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.622677   0.049260  12.641  < 2e-16 ***
## LYVE1       0.062810   0.009519   6.598  1.9e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6039 on 298 degrees of freedom
## Multiple R-squared:  0.1275, Adjusted R-squared:  0.1245
## F-statistic: 43.54 on 1 and 298 DF,  p-value: 1.898e-10
```

El p-valor d del modelo general y el test de wald de la variable LYVE1 son menores a 0.05, por lo tanto el modelo es significativo. Sin embargo el R cuadrado indica que esta variable sólo explica el 12.45% de los datos, si bien es significativa, no es muy útil para explicar los datos.

2) Realice un análisis diagnóstico y de puntos influyentes e indique si el modelo es adecuado.

Vamos a testear las condiciones que un modelo lineal debe cumplir:

Diagnóstico

```
shapiro.test(model11_simple$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model11_simple$residuals
## W = 0.9094, p-value = 1.845e-12
```

Rechaza normalidad

```
dwtest(model11_simple,alternative ="two.sided",iterations=1000)
```

```
##
##  Durbin-Watson test
##
## data:  model11_simple
## DW = 1.6372, p-value = 0.001481
```

```
## alternative hypothesis: true autocorrelation is not 0
```

Rechaza independencia

```
bptest(model11_simple)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model11_simple
## BP = 16.1, df = 1, p-value = 6.009e-05
```

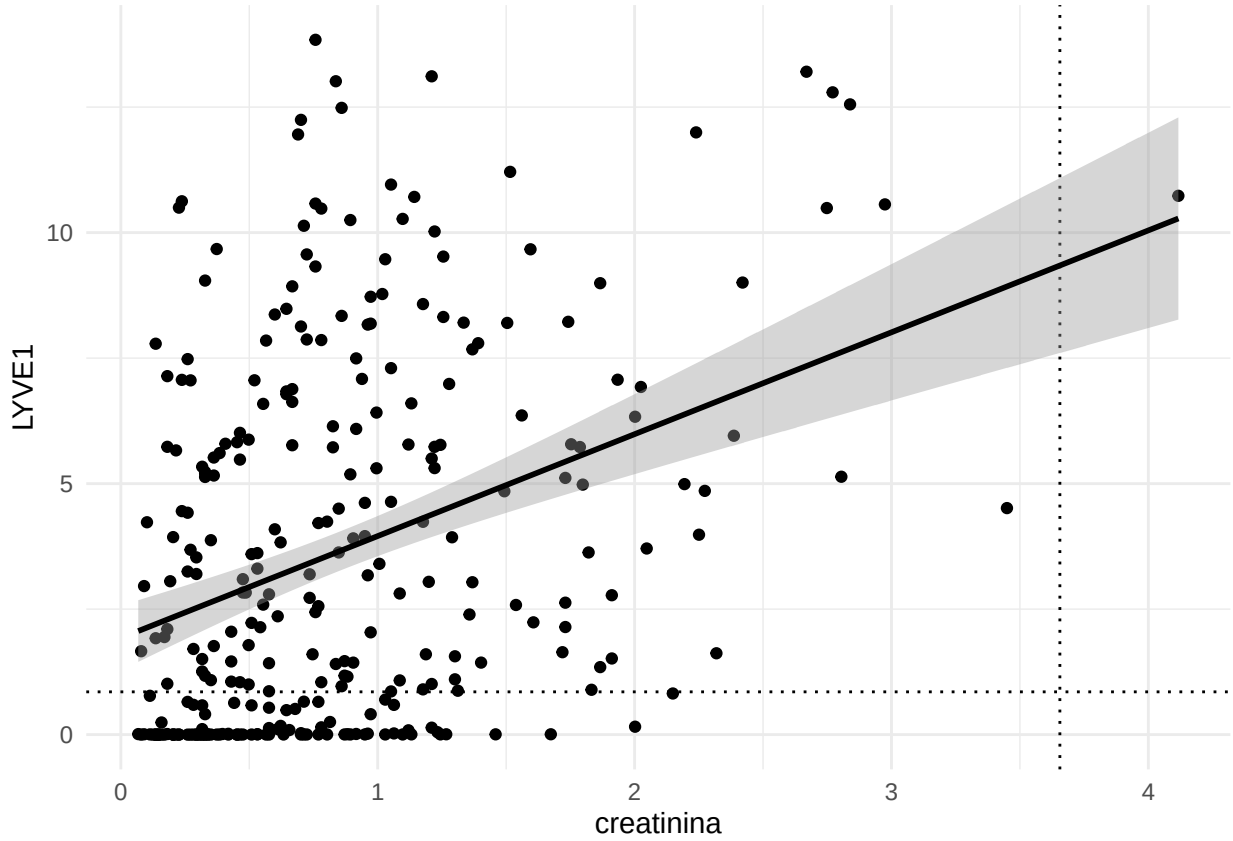
Rechaza homocedasticidad.

El modelo no cumple con ninguno de los supuestos....

** Puntos influyentes **

```
model11_simple <- lm(creatinina ~ LYVE1, data = strat_data)
promedios11 <- colMeans(strat_data[,6:7])
ggplot(strat_data, aes(creatinina,LYVE1)) +
  geom_point() +
  geom_vline(xintercept=promedios11[2],linetype="dotted") +
  geom_hline(yintercept=promedios11[1],linetype="dotted") +
   geom_smooth(method = "lm", se = TRUE, color = "black") +
  theme_minimal()
```

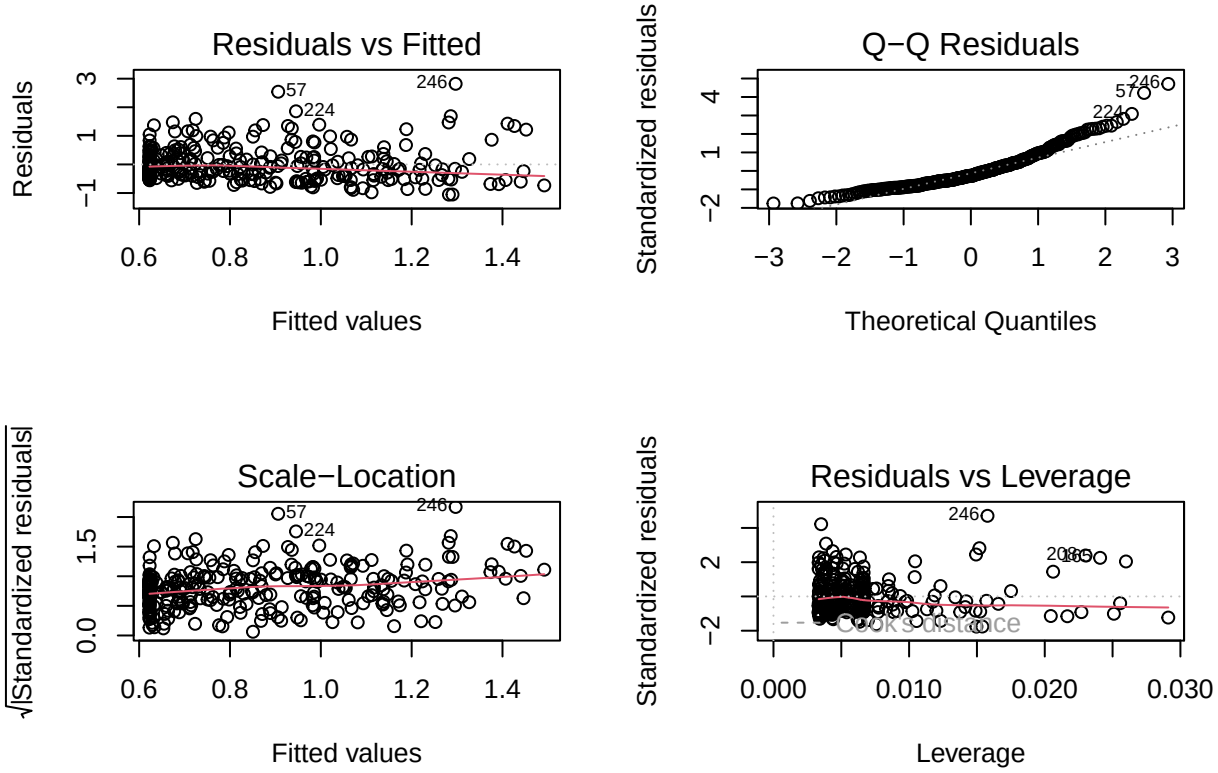
```
## `geom_smooth()` using formula = 'y ~ x'
```



De este plot ya podemos ver que hay muchos puntos muy alejados de la recta, estos son candidatos a ser

puntos influyentes. Es razonable que haya tantos puntos que no estan cercanos a la recta considerando que el r cuadrado es tan bajo

```
par(mfrow=c(2,2))
plot(model11_simple)
```




```
par(mfrow=c(1,1))
```

```
summary(influence.measures(model = model11_simple))
```

```
## Potentially influential observations of
##   lm(formula = creatinina ~ LYVE1, data = strat_data) :
##
##     dfb.1_ dfb.LYVE dffit   cov.r   cook.d hat
## 20   0.06   0.04     0.14    0.98_*  0.01   0.00
## 47   0.04   0.08     0.16    0.98_*  0.01   0.00
## 57   0.14   0.06     0.26_*  0.89_*  0.03   0.00
## 122  0.18  -0.13     0.18    0.98_*  0.02   0.01
## 162  0.09   0.01     0.13    0.98_*  0.01   0.00
## 164 -0.14   0.31     0.34_*  1.00    0.06   0.03_*
## 165 -0.14   0.33     0.36_*  1.00    0.06   0.02_*
## 166  0.06  -0.16    -0.17    1.02    0.01   0.02_*
## 198  0.18  -0.11     0.18    0.97_*  0.02   0.01
## 207 -0.09   0.27     0.30_*  0.98    0.05   0.01
## 208 -0.14   0.34     0.37_*  0.99    0.07   0.02_*
## 210  0.05  -0.13    -0.14    1.02_*  0.01   0.02_*
## 212  0.02  -0.05    -0.06    1.02_*  0.00   0.02
```

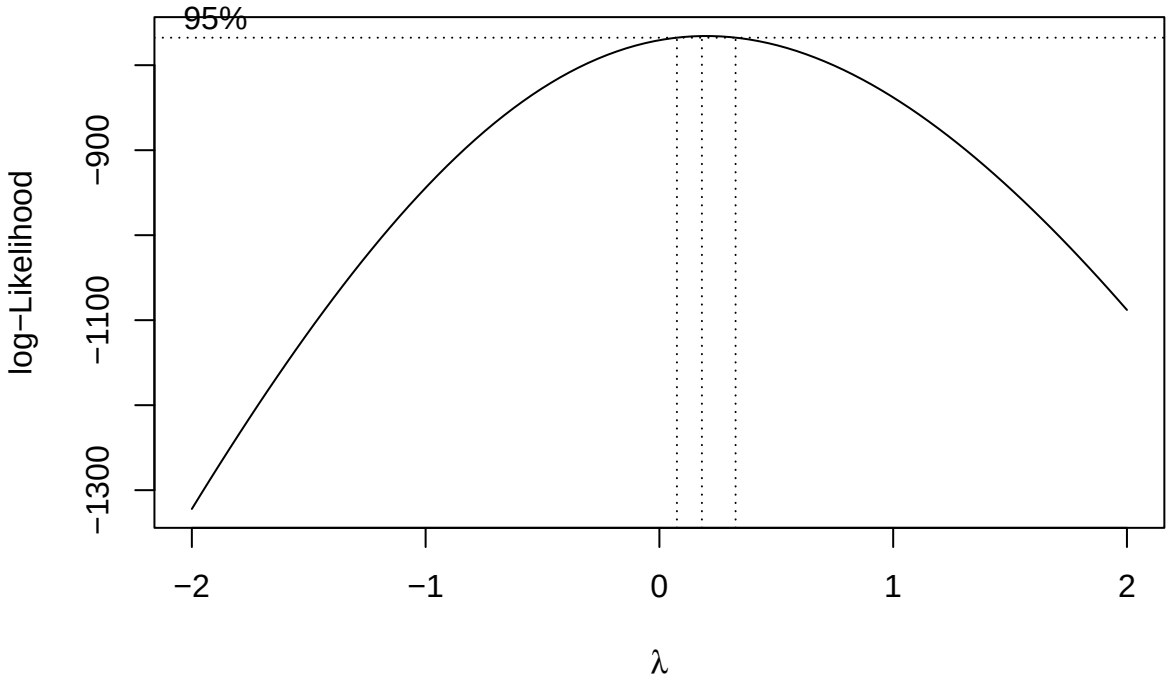
```
## 217  0.01  -0.03    -0.03    1.02_*  0.00   0.01
## 224  0.08   0.07     0.20    0.95_*  0.02   0.00
## 233  0.09  -0.20    -0.21    1.03_*  0.02   0.03_*
## 244 -0.08   0.19     0.21    1.01    0.02   0.02_*
## 246 -0.19   0.55     0.62_*  0.88_*  0.18   0.02
## 249  0.01  -0.03    -0.03    1.02_*  0.00   0.02
## 254 -0.10   0.31     0.35_*  0.97_*  0.06   0.02
## 262  0.06  -0.15    -0.16    1.03_*  0.01   0.03_*
## 271  0.03  -0.06    -0.06    1.03_*  0.00   0.03_*
## 276  0.17  -0.09     0.18    0.96_*  0.02   0.00
## 278 -0.01   0.04     0.04    1.02_*  0.00   0.02
## 290  0.06  -0.15    -0.17    1.02    0.01   0.02_*
## 297  0.00  -0.01    -0.01    1.02_*  0.00   0.01
```

Todos estos puntos son puntos influyentes según alguna de las métricas propiestas, por ejemplo el 188,221,272,287,291,295,203,319,323,326, y 361 los son en base a l inidicador de HAT.

El modelo claramente no es adecuado por más que el p-valor del modelo sea bajo no cumple con ninguno de los supuestos (normalidad de los residuos, independcencia y homocedasticidad), el R cuadrado es muy bajo y, en ese contexto, podemos considerar que hay muchos puntos influyentes.

3. Realice una transformación de la variable respuesta para intentar lograr normalidad en la distribución de los residuos. Indique si el modelo con esta transformación resulta adecuado.

```
box_cox_result <-boxcox(creatinina ~ LYVE1, data = strat_data)
```



```r
best_box_cox <- box_cox_result$x[which.max(box_cox_result$y)]
model11_box_cox <- lm((creatinina)^(best_box_cox) ~ LYVE1, data = strat_data)
summary(model11_box_cox)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox) ~ LYVE1, data = strat_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.289093 -0.083801  0.002202  0.085115  0.310078
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.884276   0.010165  86.994  < 2e-16 ***
## LYVE1       0.012883   0.001964   6.559 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1246 on 298 degrees of freedom
## Multiple R-squared:  0.1262, Adjusted R-squared:  0.1232
## F-statistic: 43.02 on 1 and 298 DF,  p-value: 2.391e-10
```

Realizo un ajuste box y cox y el p-valor del modelo sigue siendo satisfactirio pero el e cuadrado sigue siendo muy bajo.

```r
shapiro.test(model11_box_cox$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model11_box_cox$residuals
## W = 0.99245, p-value = 0.1322
```

NO rechaza normalidad

```r
dwtest(model11_box_cox,alternative ="two.sided",iterations=1000)
```

```
##
##  Durbin-Watson test
##
## data:  model11_box_cox
## DW = 1.7886, p-value = 0.06266
## alternative hypothesis: true autocorrelation is not 0
```

NO Rechaza independencia

```r
bptest(model11_box_cox)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model11_box_cox
## BP = 0.016731, df = 1, p-value = 0.8971
```

No rechaza homocedasticidad.

Ahora el modelo si cumple con todos los supuestos que debe tener. Sin embargo el R cuadrado sigue siendo

muy bajo. No es un buen modelo para explicar los datos.

4. Sin considerar la variable estadío, ajuste un modelo multivariado robusto para explicar el valor de la creatinina y estime el error absoluto medio cometido.

```
model11_robusto <- rlm(creatinina ~ edad + sexo + diagnosis  + LYVE1 + REG1B + TFF1,psi=psi.huber,data=s
summary(model11_robusto)
```

```
##
## Call: rlm(formula = creatinina ~ edad + sexo + diagnosis + LYVE1 +
##     REG1B + TFF1, data = strat_data, psi = psi.huber)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.89135 -0.30349 -0.06646  0.33568  2.14292
##
## Coefficients:
##                 Value   Std. Error t value
## (Intercept)      0.7253  0.1837     3.9487
## edad            -0.0082  0.0025    -3.2268
## sexoM            0.1302  0.0599     2.1723
## diagnosisnormal  0.3448  0.0782     4.4065
## LYVE1            0.0515  0.0117     4.4180
## REG1B            0.0001  0.0002     0.6112
## TFF1             0.0002  0.0000     5.6419
##
## Residual standard error: 0.4621 on 293 degrees of freedom
```

```
mean(abs(strat_data$creatinina - predict(model11_robusto, newdata = strat_data)))
```

```
## [1] 0.4039833
```

```
library(Metrics)
```

```
## Warning: package 'Metrics' was built under R version 4.3.1
```

```
MAE_robust11 <- mae(strat_data$creatinina,predict(model11_robusto))
MAE_robust11
```

```
## [1] 0.4039833
```

Genero un modelo robusto y calculo el MAE con dos métodos. No se me pide interpretar nada

##5. Sin considerar la variable estadío, utilice un método de selección de variables para proponer un nuevo modelo multivariado que explique el valor de la creatinina. Estudie el cumplimiento de los supuestos y haga una transformación en caso de ser necesario. Analice los coeficientes del modelo final.

```
model11_multi <- lm(creatinina ~ edad + sexo + diagnosis  + LYVE1 + REG1B + TFF1,data=strat_data)
model11_forward <- ols_step_forward_aic(model11_multi)
summary(model11_forward$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75799 -0.36386 -0.09897  0.27759  2.09191
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.378e-01  2.033e-01   4.614 5.91e-06 ***
## TFF1             1.841e-04  3.266e-05   5.637 4.05e-08 ***
## edad            -1.075e-02  2.817e-03  -3.817 0.000165 ***
## LYVE1            6.305e-02  1.257e-02   5.014 9.23e-07 ***
## diagnosisnormal  3.059e-01  8.658e-02   3.533 0.000477 ***
## sexoM            1.301e-01  6.554e-02   1.985 0.048034 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5461 on 294 degrees of freedom
## Multiple R-squared:  0.2962, Adjusted R-squared:  0.2842
## F-statistic: 24.74 on 5 and 294 DF,  p-value: < 2.2e-16
```

El modelo seleccionado como el mejor por el método step forward es el que usa las variables: TFF1, edad, LYVE1, sexo y diagnosis.

Vemos que el p-valor de significación del modelo es menor a 0.05 por ,o cual el modelo es significativbo y también vemos que los test de wald de todas las variables son menores a 0.05 por lo tanto son significativas. Sin embargo el modelo tiene un R cuadrado ajustado muy bajo, de 28,4%.

```
model11_multi_final <- lm(creatinina ~ edad + sexo + diagnosis  + LYVE1 + TFF1,data=strat_data)
```

** Diagnóstico **

```
shapiro.test(model11_multi_final$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model11_multi_final$residuals
## W = 0.94819, p-value = 8.657e-09
```

Rechaza normalidad

```
dwtest(model11_multi_final,alternative ="two.sided",iterations=1000)
```

```
##
##  Durbin-Watson test
##
## data:  model11_multi_final
## DW = 1.8275, p-value = 0.1207
## alternative hypothesis: true autocorrelation is not 0
```

NO Rechaza independencia

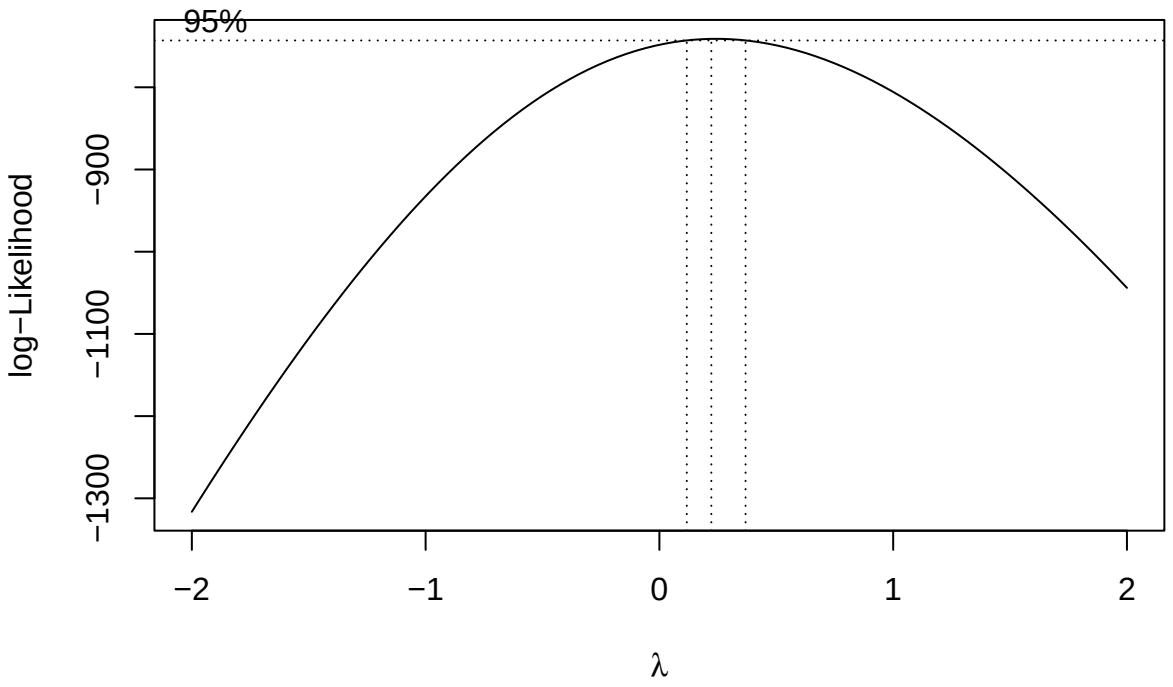
```
bptest(model11_multi_final)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model11_multi_final
## BP = 30.664, df = 5, p-value = 1.091e-05
```

Rechaza homocedasticidad

No cumple con el supuesto de homocedasticidad ni normalidad de los residuos. Pruebo con box y cox

```r
box_cox_result_final <-boxcox(creatinina ~ edad + sexo + diagnosis  + LYVE1 + TFF1, data = strat_data)
```




```r
best_box_cox_final <- box_cox_result_final$x[which.max(box_cox_result_final$y)]
model11_box_cox_final <- lm((creatinina)^(best_box_cox_final) ~ edad + sexo + diagnosis  + LYVE1 + TFF1
summary(model11_box_cox_final)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox_final) ~ edad + sexo +
##     diagnosis + LYVE1 + TFF1, data = strat_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35507 -0.08789 -0.00288  0.09731  0.37520
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.022e-01  5.164e-02  17.470  < 2e-16 ***
## edad            -2.235e-03  7.157e-04  -3.123 0.001968 **
## sexoM            4.560e-02  1.665e-02   2.739 0.006538 **
## diagnosisnormal  8.682e-02  2.200e-02   3.947 9.90e-05 ***
## LYVE1            1.862e-02  3.194e-03   5.830 1.45e-08 ***
## TFF1             3.036e-05  8.297e-06   3.660 0.000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1387 on 294 degrees of freedom
## Multiple R-squared:  0.2591, Adjusted R-squared:  0.2465
## F-statistic: 20.57 on 5 and 294 DF,  p-value: < 2.2e-16
```

El modelo sigue siendo significativo asi como todas sus variables, el R cuadrado ajustado sigue siendo bajo.

** Diagnóstico **

```
shapiro.test(model11_box_cox_final$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model11_box_cox_final$residuals
## W = 0.99707, p-value = 0.8664
```

NO Rechaza normalidad

```
dwtest(model11_box_cox_final,alternative ="two.sided",iterations=1000)
```

```
##
##  Durbin-Watson test
##
## data:  model11_box_cox_final
## DW = 1.9759, p-value = 0.793
## alternative hypothesis: true autocorrelation is not 0
```

NO Rechaza independencia

```
bptest(model11_box_cox_final)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model11_box_cox_final
## BP = 7.4923, df = 5, p-value = 0.1865
```

No rechaza homocedasticidad. Es decir que si aplico una transformación de box y cox cumplo con los supuestos del modelo.

El modelo nos indica que la cantidad de creatinina se ve afectada de forma importante para pacientes de sexo masculino con un diagnóstico "normal" y que disminuye con la edad. También se ve positivamente afectado por los valores de LYVE1 y TFF1.

6) Estime los errores de predicción de los 4 modelos previos y compárelos. Cuál elegiría?

Podemos usar cualquier indicador de error para la comparación, ya veniamos con MAE, podemos

```
MAE_simple11 <- mae(strat_data$creatinina,predict(model11_simple))
MAE_simplebc11 <- mae(strat_data$creatinina,predict(model11_box_cox))
MAE_robust11 <- mae(strat_data$creatinina,predict(model11_robusto))
MAE_multibc11 <- mae(strat_data$creatinina,predict(model11_box_cox_final))
MAE_simple11
```

```
## [1] 0.4614605
```

```
MAE_simplebc11
```

```
## [1] 0.4933007
```

```
MAE_robust11
```

```
## [1] 0.4039833
```

```
MAE_multibc11
```

```
## [1] 0.4739662
```

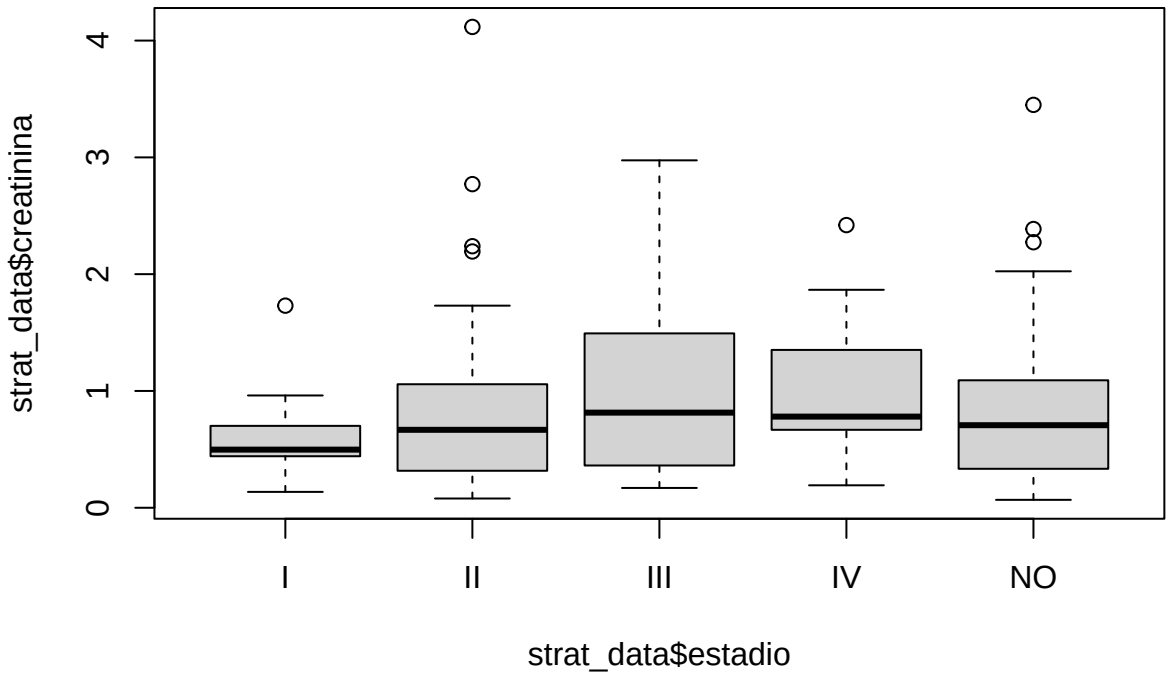
Comparando los MAE de los cuatro modelos: simple, simple con box_cox, robusto, y multivariado con box y cox, el que tiene menor MAE es el modelo robusto. Este sería el que elegiria.

7. Le parece adecuado un modelo GAMLSS en este caso? Justifique.

Ejercicio 2

Estudie analítica y gráficamente si: ## 1. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío.

```
plot(strat_data$creatinina~strat_data$estadio)
```



Si miramos los boxplots de los gráficos podemos ver que las medianas de cada estadío son similares y las categorias parecen estar alineadas en los mismos valores. Visualmente no parece haber mayores diferencias. Veamos analíticamente:

```
AOV_estadio<- aov(strat_data$creatinina~strat_data$estadio)
summary(AOV_estadio)
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## strat_data$estadio   4   3.38  0.8459   2.059 0.0862 .
```

```
## Residuals          295 121.18  0.4108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que el p-valor del test ANOVA da 0.0862 que es mayor a 0.05. Esto indica que no se rechaza la hipótesis nula de que las medias son iguales, sin embargo, HAY que verificar los supuesto de normalidad y homogeneidad de las varianzas primero para cconcluir algo. De cumplirse los mismos, podremos decir que no hay diferencia significativa entre las varianzas.

Analicemos la igualdad de las varianzas primero con el test de levene:

```
leveneTest(strat_data$creatinina~strat_data$estadio)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4   1.995 0.0953 .
##       295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

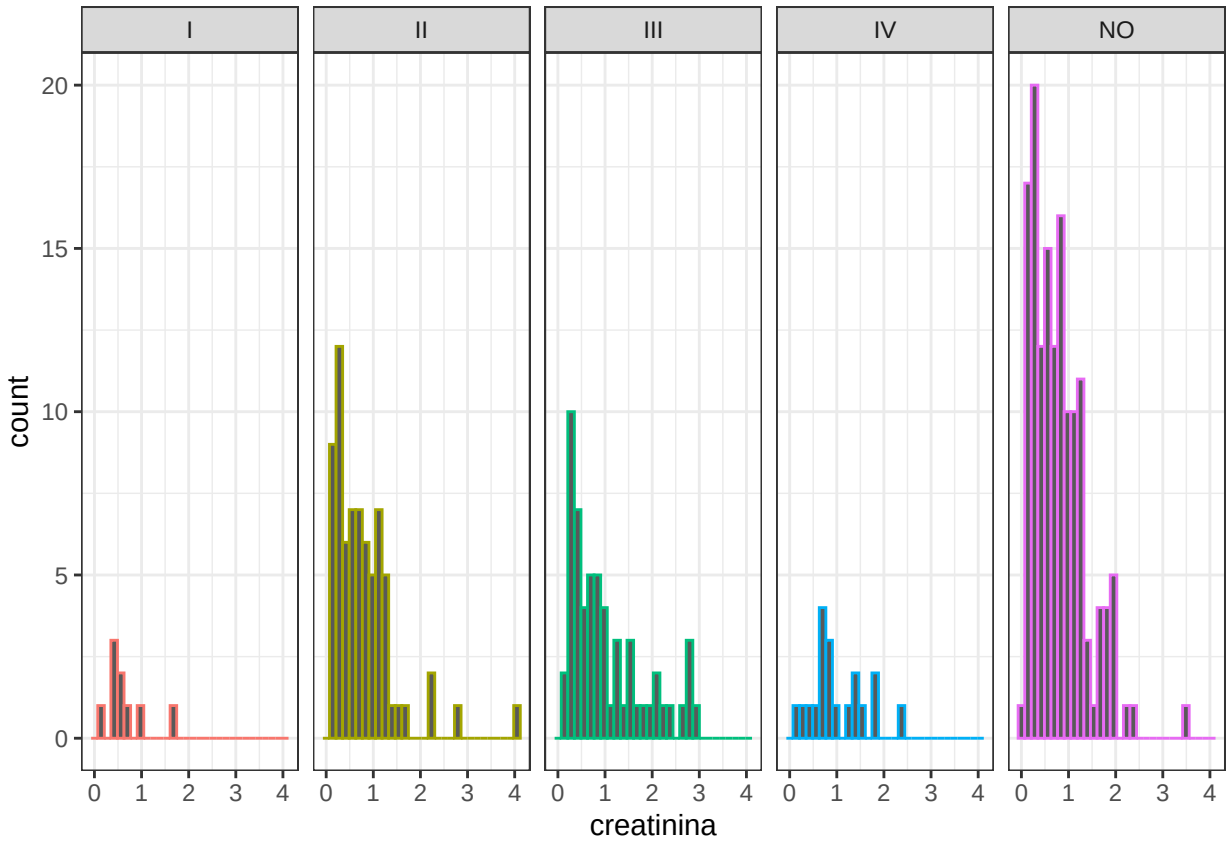
```
shapiro.test(residuals(AOV_estadio))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(AOV_estadio)
## W = 0.88756, p-value = 4.476e-14
```

Sin embargo, no se cumple la normalidad de los residuos. Por lo cual un supuesto falla. Probemos con Kruskal Wallis.

```
ggplot(data = strat_data, mapping = aes(x = creatinina, colour =estadio )) +
  geom_histogram() + theme_bw() + facet_grid(. ~ estadio) +
  theme(legend.position = "none")#
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Verificamos primero que las distribuciones sean similares. Graficamente validamos que los son.

Aplicamos Kruskal-Wallis:

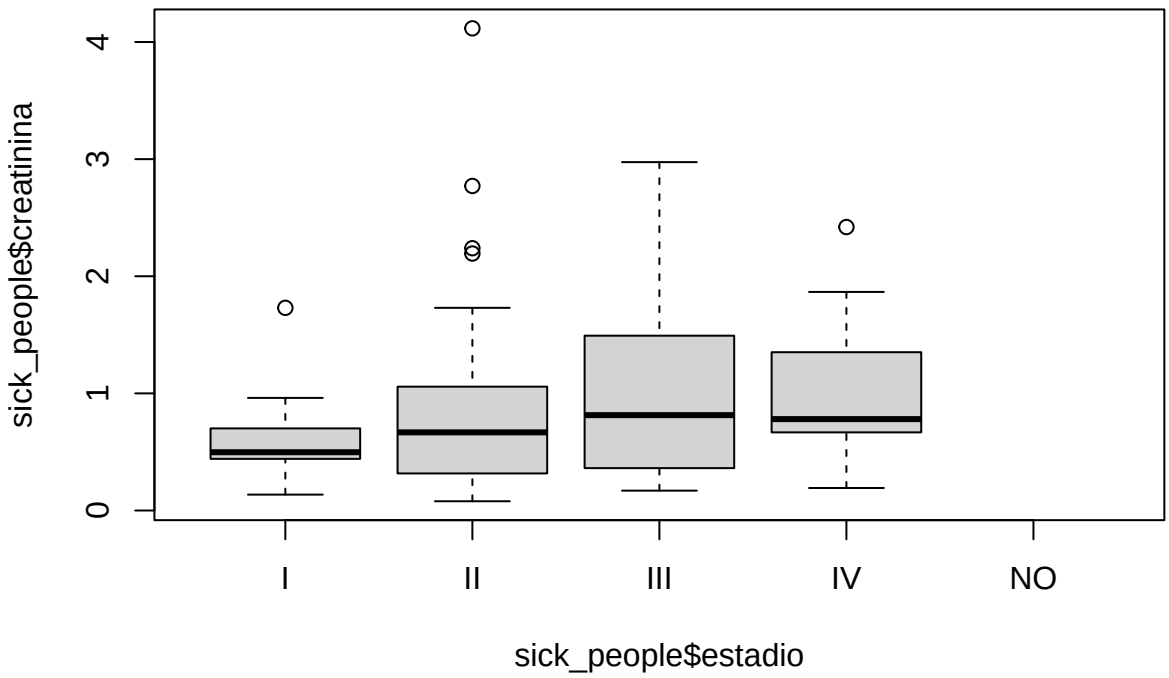
```
kruskal.test(creatinina ~ estadio, data = strat_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinina by estadio
## Kruskal-Wallis chi-squared = 6.6105, df = 4, p-value = 0.158
```

Segun el resultado del test, dado que el p-valor es mayor a 0.05 no puedo conlcuir que haya diferencias en las medias de los diferentes estadíos.

2. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto de la variable estadío considerando sólo la base de pacientes enfermos.

```
sick_people <- strat_data[strat_data$diagnosis!="normal"]
plot(sick_people$creatinina~sick_people$estadio)
```



Filtamos los datos para aquellos que tienen diagnosticado un tumor maligno y volvemos hacer el analisis gráfico. Pareciera haber diferencias.

```
AOV_estadio_sick<- aov(sick_people$creatinina~sick_people$estadio)
summary(AOV_estadio_sick)
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## sick_people$estadio   3   2.75  0.9171    1.87  0.137
## Residuals           152  74.55  0.4905
```

Vemos que el p-valor del test ANOVA da 0.137 que es mayor a 0.05. Esto indica que no se rechaza la hipótesis nula de que las medias son iguales, sin embargo, nuevamente HAY que verificar los supuesto de normalidad y homogeneidad de las varianzas primero para cconcluir algo. De cumplirse los mismos, podremos decir que no hay diferencia significativa entre las varianzas.

```
leveneTest(sick_people$creatinina~sick_people$estadio)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  1.7228 0.1647
##       152
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

```
shapiro.test(residuals(AOV_estadio_sick))
```

```
##
##  Shapiro-Wilk normality test
```

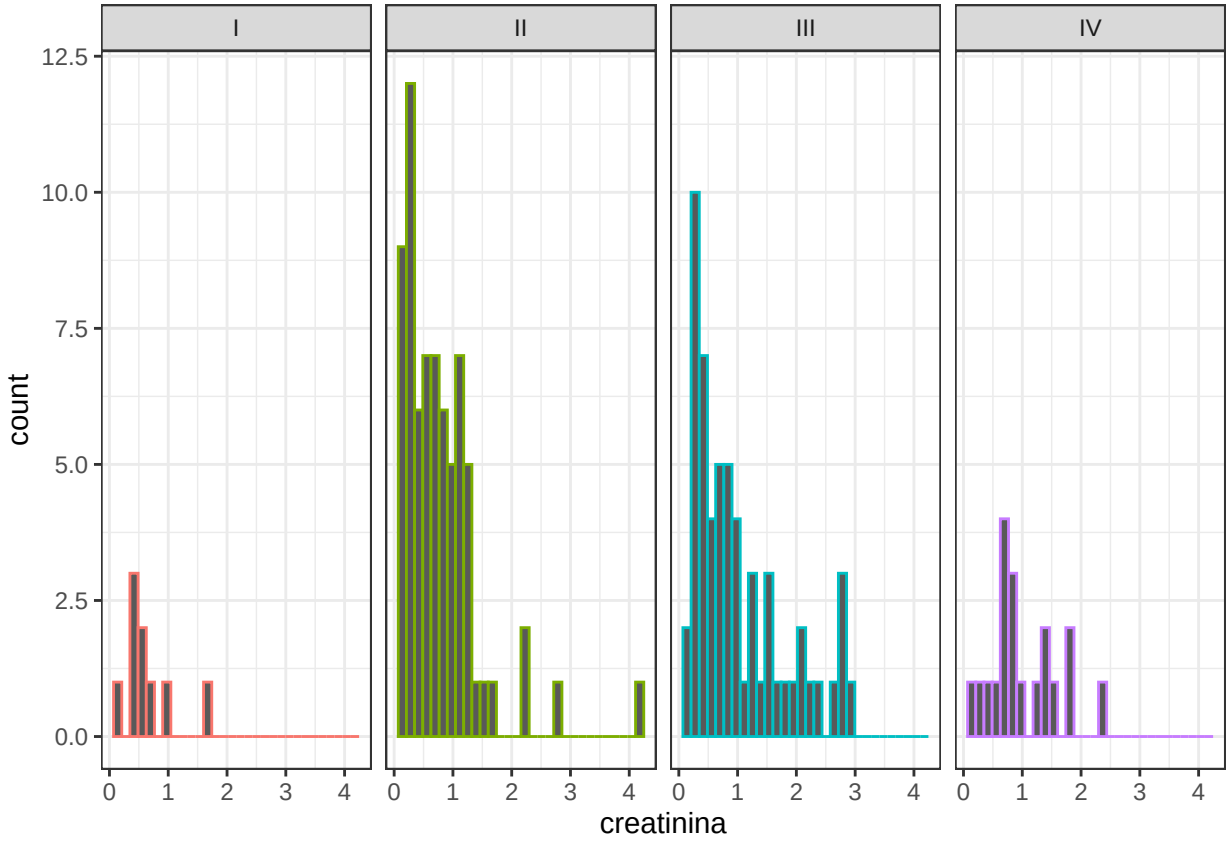
```
##
## data:  residuals(AOV_estadio_sick)
## W = 0.87075, p-value = 2.271e-10
```

Nuevamente se rechaza normalidad de los residuos

Por lo cual un supuesto falla. Probemos con Kruskal Wallis.

```
ggplot(data = sick_people, mapping = aes(x = creatinina, colour =estadio )) +
  geom_histogram() + theme_bw() + facet_grid(. ~ estadio) +
  theme(legend.position = "none")#
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Al igual que el caso anterior se cumple que las distribuciones son similares

```
kruskal.test(creatinina ~ estadio, data = sick_people)
```

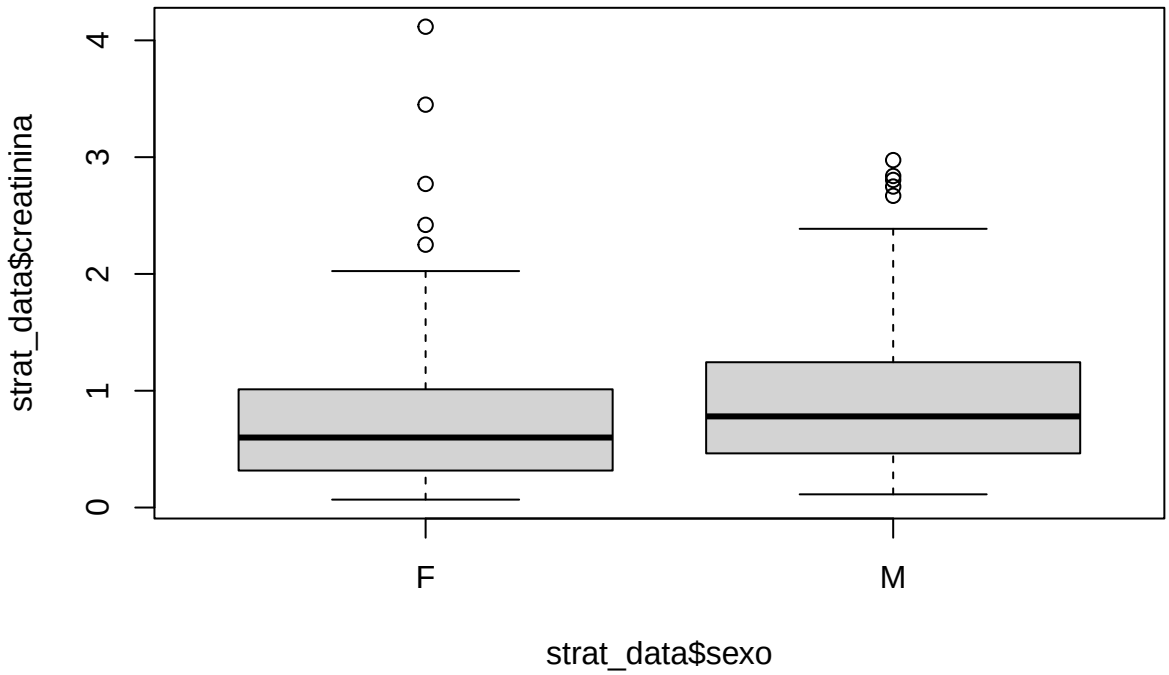
```
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinina by estadio
## Kruskal-Wallis chi-squared = 5.9177, df = 3, p-value = 0.1157
```

Y nuevamente no se rechaza que tengan la misma media.

3. existen diferencias estadísticamente significativas en las medias de los valores de creatinina respecto del sexo.

Bueno... esto más que una pregunta es una afirmación. Veamos:

```r
plot(strat_data$creatinina~strat_data$sexo)
```



MMmmm... dudoso... vamos a los números

```r
AOV_sexo<- aov(strat_data$creatinina~strat_data$sexo)
summary(AOV_sexo)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## strat_data$sexo   1   2.57  2.5691   6.276 0.0128 *
## Residuals       298 121.99  0.4094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vemos que el p-valor del test ANOVA da 0.0128 que es memor a 0.05. Esto indica que se rechaza la hipótesis nula de que las medias son iguales, sin embargo, HAY que verificar los supuesto de normalidad y homogeneidad de las varianzas primero para cconcluir algo. De cumplirse los mismos, podremos decir que no hay diferencia significativa entre las varianzas.

Analicemos la igualdad de las varianzas primero con el test de levene:

```r
leveneTest(strat_data$creatinina~strat_data$sexo)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.6256 0.4296
##       298
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

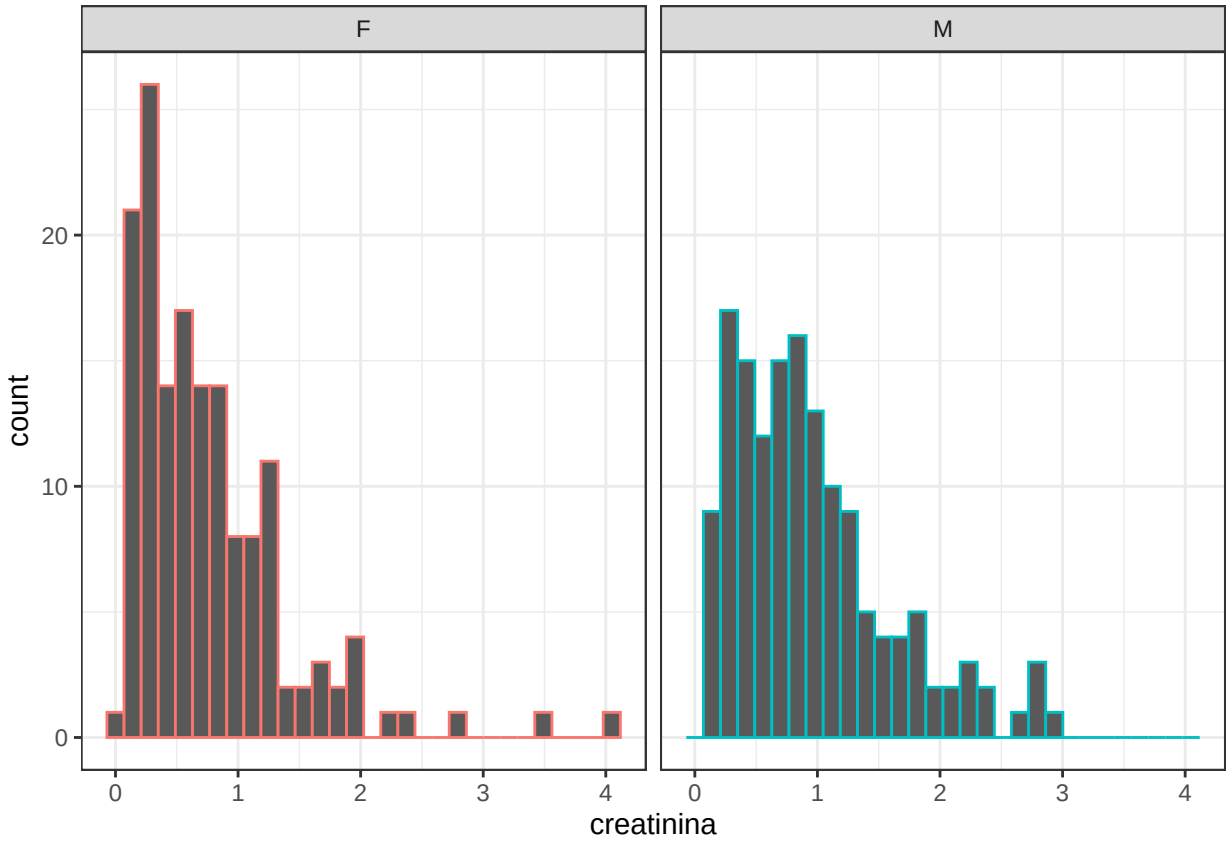
```r
shapiro.test(residuals(AOV_sexo))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(AOV_sexo)
## W = 0.87536, p-value = 6.944e-15
```

uff. nuevamente se rechanzan la normalidad de los residuos....

```r
ggplot(data = strat_data, mapping = aes(x = creatinina, colour =sexo )) +
  geom_histogram() + theme_bw() + facet_grid(. ~ sexo) +
  theme(legend.position = "none")#
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



las distribuciones son similares.

```r
kruskal.test(creatinina ~ sexo, data = strat_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  creatinina by sexo
## Kruskal-Wallis chi-squared = 9.2865, df = 1, p-value = 0.002308
```

En este caso el test de kruskal-wallis si nos indica que las medias pueden ser diferentes.

4. la interacción entre estadío y sexo es significativa cuando se considera la base completa.

```
model24_inter <- lm(creatinina ~  estadio*sexo,data=strat_data)
summary(model24_inter)
```

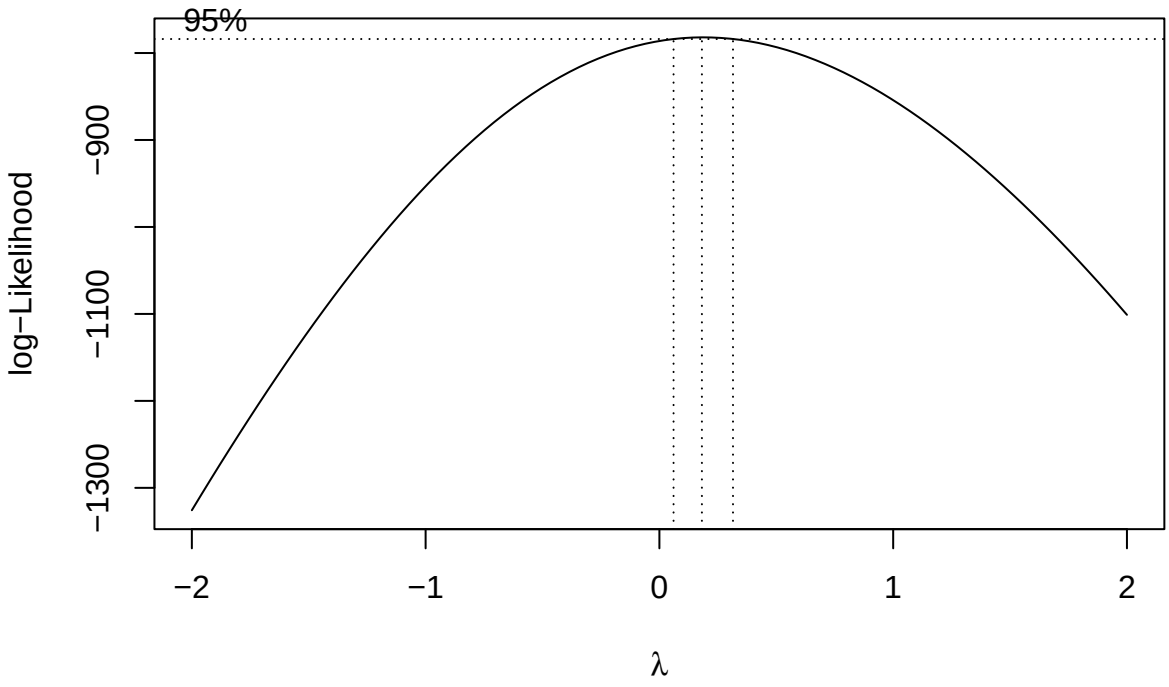
```
##
## Call:
## lm(formula = creatinina ~ estadio * sexo, data = strat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8810 -0.4374 -0.1275  0.2873  3.3073
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.3544     0.3670   0.966   0.3350
## estadioII           0.4551     0.3833   1.187   0.2360
## estadioIII          0.5343     0.3949   1.353   0.1770
## estadioIV           0.8501     0.4494   1.892   0.0596 .
## estadioNO           0.3465     0.3730   0.929   0.3537
## sexoM               0.4467     0.4494   0.994   0.3211
## estadioII:sexoM    -0.4917     0.4742  -1.037   0.3007
## estadioIII:sexoM   -0.2283     0.4836  -0.472   0.6373
## estadioIV:sexoM    -0.7273     0.5481  -1.327   0.1856
## estadioNO:sexoM    -0.1651     0.4627  -0.357   0.7214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6356 on 290 degrees of freedom
## Multiple R-squared:  0.05943,    Adjusted R-squared:  0.03024
## F-statistic: 2.036 on 9 and 290 DF,  p-value: 0.0354
```

Si bien el p-valor de generla del modelo está por debajo de 0.05, las variables y las interacciones, ninguna resulta significativa si miramos los test de wald de cada una de ellas.

5. se satisfacen los supuestos del modelo en 1, 2 y 3. En caso negativo intente una transformación adecuada sobre la variable respuesta en cada modelo y revise nuevamente los supuestos.

Ok, en ningún caso satisfizo todos los supuestos. Podemos probar aplicar box y cox en los tres casos. Obtengamos los lambda

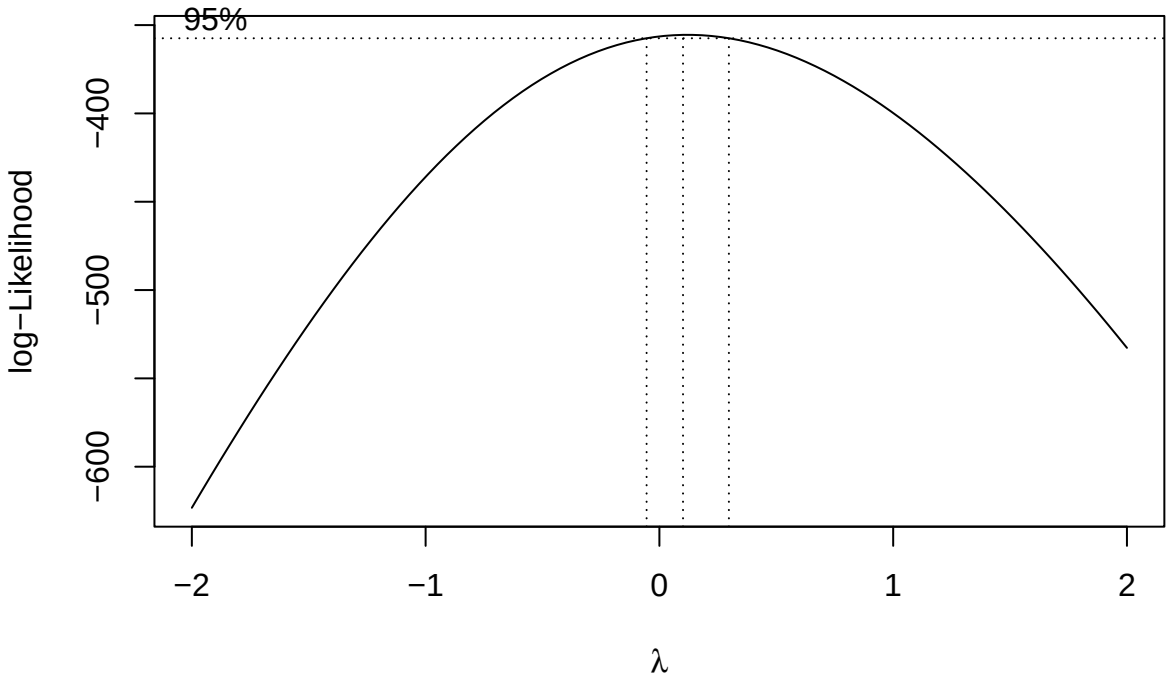
```
box_cox_1 <-boxcox(creatinina ~ estadio, data = strat_data)
```




```
best_box_cox_1 <- box_cox_1$x[which.max(box_cox_1$y)]
model25_1 <- lm((creatinina)^(best_box_cox_1) ~ estadio, data = strat_data)
summary(model25_1)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox_1) ~ estadio, data = strat_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31045 -0.10250  0.00635  0.09533  0.37957
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.89755    0.04408  20.363   <2e-16 ***
## estadioII    0.01629    0.04679   0.348    0.728
## estadioIII   0.06514    0.04743   1.373    0.171
## estadioIV    0.08029    0.05351   1.501    0.135
## estadioNO    0.02605    0.04543   0.573    0.567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1322 on 295 degrees of freedom
## Multiple R-squared:  0.02601,    Adjusted R-squared:  0.0128
## F-statistic: 1.969 on 4 and 295 DF,  p-value: 0.09917
```

```r
box_cox_2 <-boxcox(creatinina ~ estadio, data = sick_people)
```

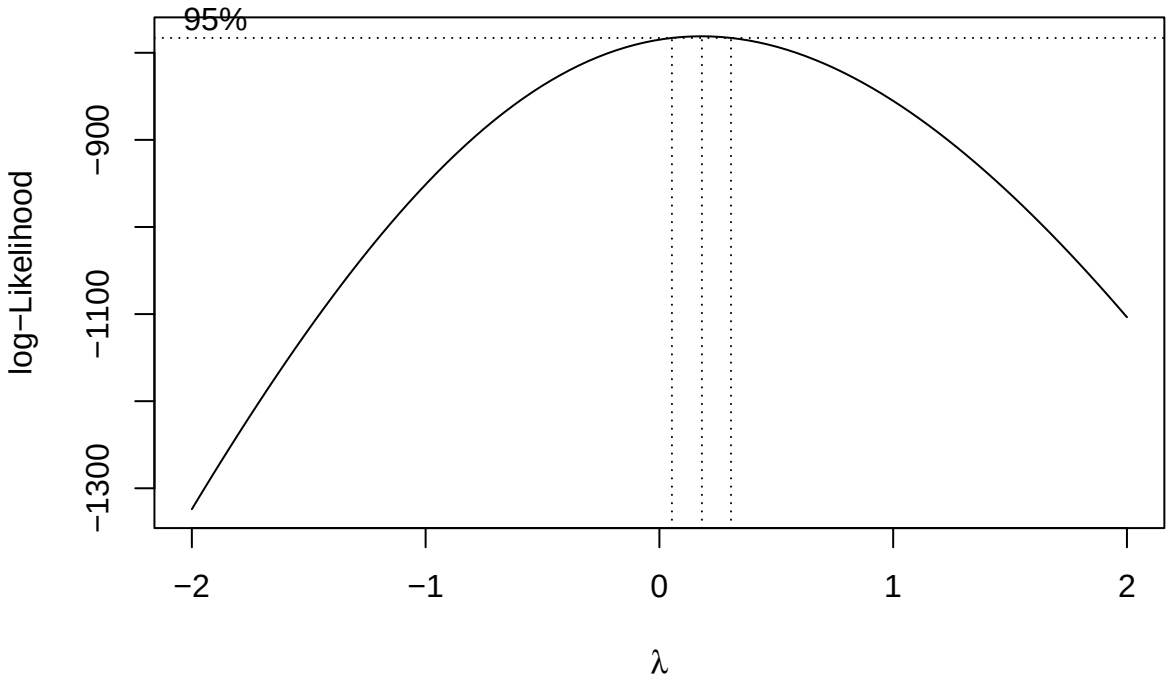



```r
best_box_cox_2 <- box_cox_2$x[which.max(box_cox_2$y)]
model25_2 <- lm((creatinina)^(best_box_cox_2) ~ estadio, data = strat_data)
summary(model25_2)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox_2) ~ estadio, data = strat_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.192356 -0.058208  0.005729  0.055941  0.205106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.940078   0.025466  36.914   <2e-16 ***
## estadioII   0.008474   0.027032   0.313    0.754
## estadioIII  0.036501   0.027403   1.332    0.184
## estadioIV   0.046086   0.030915   1.491    0.137
## estadioNO   0.014324   0.026250   0.546    0.586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0764 on 295 degrees of freedom
## Multiple R-squared:  0.02582,    Adjusted R-squared:  0.01261
```

```
## F-statistic: 1.955 on 4 and 295 DF,  p-value: 0.1014
```

```
box_cox_3 <-boxcox(creatinina ~ sexo, data = strat_data)
```




```
best_box_cox_3 <- box_cox_3$x[which.max(box_cox_3$y)]
model25_3 <- lm((creatinina)^(best_box_cox_3) ~ sexo, data = strat_data)
summary(model25_3)
```

```
##
## Call:
## lm(formula = (creatinina)^(best_box_cox_3) ~ sexo, data = strat_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29476 -0.09138  0.00184  0.08741  0.38551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.90791    0.01064   85.34  < 2e-16 ***
## sexoM        0.04756    0.01515    3.14  0.00186 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1312 on 298 degrees of freedom
## Multiple R-squared:  0.03203,    Adjusted R-squared:  0.02878
## F-statistic:  9.86 on 1 and 298 DF,  p-value: 0.001859
```

** Modelo 1 **

```
leveneTest(model25_1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  1.1081 0.3528
##       295
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

```
shapiro.test(residuals(model25_1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model25_1)
## W = 0.99236, p-value = 0.1264
```

No se rechaza normalidad. Un lujo!

** Modelo 2 **

```
leveneTest(model25_2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  1.0433 0.3851
##       295
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

```
shapiro.test(residuals(model25_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model25_2)
## W = 0.99047, p-value = 0.04824
```

No se rechaza normalidad porque tomamos 0.01 (estuvo cerca...) ** Modelo 3 **

```
leveneTest(model25_3)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  1.1844 0.2773
##       298
```

No se rechaza la hipótesis nula que afirma que las varianzas son iguales, por lo tanto hay homogeneidad.

Analicemos la normalidad:

```
shapiro.test(residuals(model25_3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model25_3)
## W = 0.99365, p-value = 0.2391
```

No se rechaza normalidad.

Los modelos transformados cumplen con los supuestos.

6. Obtenga conclusiones acerca de dónde se observan las diferencias si las hubiere.

Como vimos en el punto 3 observamos diferencias entre las medias de cratinina entre los sexos.

Ejercicio 3

1. Ajuste un modelo logístico para predecir el diagnóstico de cáncer de páncreas en función de las variables en la base que considere razonables.

```
modelo_logistico <- glm(diagnosis ~ sexo+edad+LYVE1+TFF1+REG1B,
                        data=strat_data, family = "binomial")
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = diagnosis ~ sexo + edad + LYVE1 + TFF1 + REG1B,
##     family = "binomial", data = strat_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.0639593  1.1608470   6.085 1.16e-09 ***
## sexoM       -0.9805969  0.3615675  -2.712  0.00669 **
## edad        -0.0783195  0.0170427  -4.595 4.32e-06 ***
## LYVE1       -0.4264511  0.0860025  -4.959 7.10e-07 ***
## TFF1        -0.0008561  0.0005141  -1.665  0.09589 .
## REG1B       -0.0024739  0.0021257  -1.164  0.24452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.41  on 299  degrees of freedom
## Residual deviance: 217.65  on 294  degrees of freedom
## AIC: 229.65
##
## Number of Fisher Scoring iterations: 7
```

Con este modelo vemos que TFF1 y REG1B no resultan significativas, elimino 1

```
modelo_logistico <- glm(diagnosis ~ sexo+edad+LYVE1+REG1B,
                        data=strat_data, family = "binomial")
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = diagnosis ~ sexo + edad + LYVE1 + REG1B, family = "binomial",
##     data = strat_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.923973   1.140037   6.073 1.25e-09 ***
```

```
## sexoM       -0.900482   0.354048  -2.543   0.0110 *
## edad        -0.076783   0.016797  -4.571 4.85e-06 ***
## LYVE1       -0.497562   0.076316  -6.520 7.04e-11 ***
## REG1B       -0.004159   0.001913  -2.175   0.0296 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.41  on 299  degrees of freedom
## Residual deviance: 220.86  on 295  degrees of freedom
## AIC: 230.86
##
## Number of Fisher Scoring iterations: 6
```

REG1B me sigue dando no significativa. Ahora pruebo con la otra:

```
modelo_logistico <- glm(diagnosis ~ sexo+edad+LYVE1+TFF1,
                        data=strat_data, family = "binomial")
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = diagnosis ~ sexo + edad + LYVE1 + TFF1, family = "binomial",
##     data = strat_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.1130419  1.1544477   6.161 7.21e-10 ***
## sexoM       -1.0940435  0.3511780  -3.115  0.00184 **
## edad        -0.0793182  0.0168603  -4.704 2.55e-06 ***
## LYVE1       -0.4365239  0.0848524  -5.145 2.68e-07 ***
## TFF1        -0.0011278  0.0004613  -2.445  0.01448 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 415.41  on 299  degrees of freedom
## Residual deviance: 219.20  on 295  degrees of freedom
## AIC: 229.2
##
## Number of Fisher Scoring iterations: 6
```

Ahora me dan todas significativas. Me quedo con este modelo.

2. Evalúe la calidad de ajuste del modelo con al menos dos criterios distintos.

```
predicciones <- predict(object = modelo_logistico, newdata = strat_data, type = "response")
curva_roc <- pROC::roc(response = strat_data$diagnosis, predictor = predicciones)
```

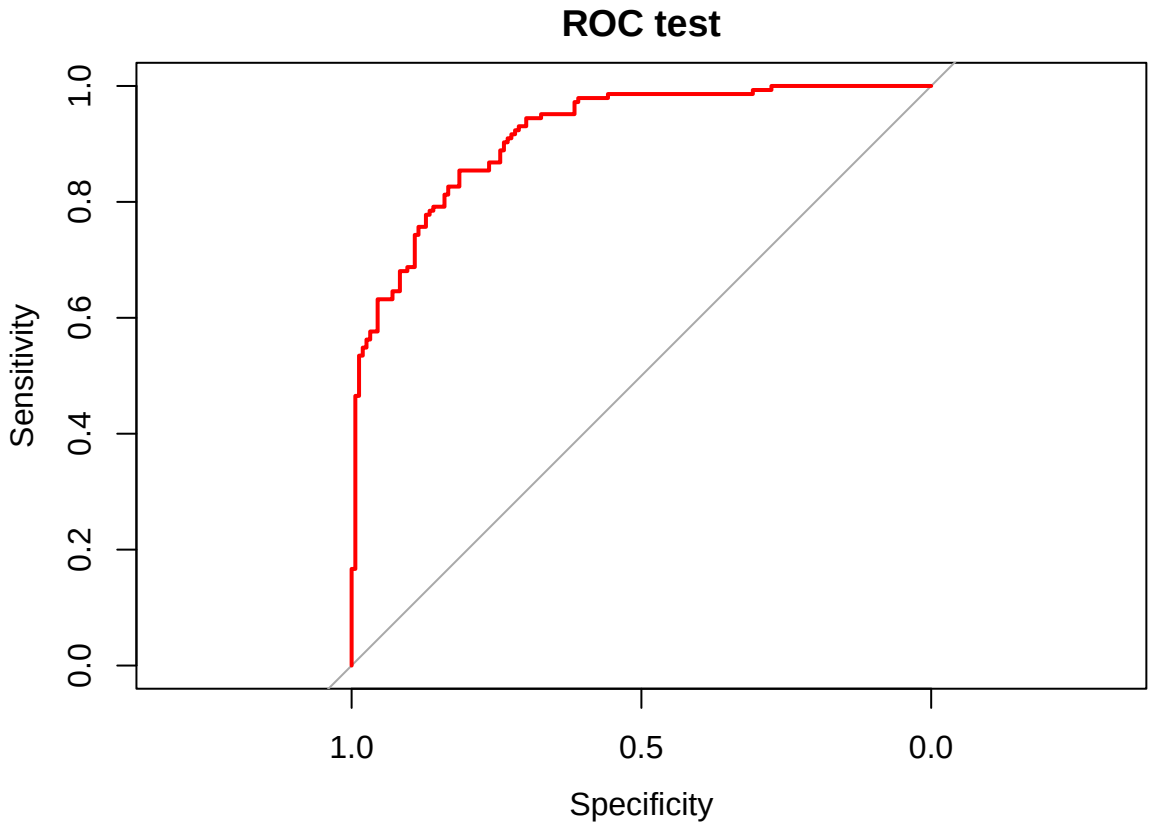
```
## Setting levels: control = maligno, case = normal
```

```
## Setting direction: controls < cases
```

```
curva_roc
```

```
##
## Call:
## roc.default(response = strat_data$diagnosis, predictor = predicciones)
##
## Data: predicciones in 156 controls (strat_data$diagnosis maligno) < 144 cases (strat_data$diagnosis r
## Area under the curve: 0.9159
```

```
plot(curva_roc,col="red",lwd=2,main="ROC test")
```



Vemos que el valor de Área Bajo la Curva ROC es superior a 0,5 y cercano a 1, lo que indica una buena calidad de ajuste

3. Interprete los coeficientes del modelo elegido.

```
coef_sexoM <- modelo_logistico$coefficients["sexoM"]
coef_edad <- modelo_logistico$coefficients["edad"]
coef_LYVE1 <- modelo_logistico$coefficients["LYVE1"]
coef_TFF1 <- modelo_logistico$coefficients["TFF1"]
coef_sexoM
```

```
##     sexoM
## -1.094043
```

```
coef_edad
```

```
##       edad
```

```
## -0.07931825
```

```
coef_LYVE1
```

```
##      LYVE1
## -0.4365239
```

```
coef_TFF1
```

```
##        TFF1
## -0.00112783
```

```
table(strat_data$diagnosis)
```

```
##
## maligno  normal
##     156     144
```

Todos los coeficientes son negativos, el modelo logistico predice entre maligno (0) y normal (1). Entendemos que en ese orden, ser de sexo masculino, tener una edad alta y altos valores de TFF1 y LYVE1 incrementan el riesgo de tener diagnóstico maligno.