Regresión, modelos y métodos - PEC 2

Alejandro Keymer

1/6/2020

Contents

Ejercicio 1 (25 pt.) Mecanismo neural de la nocicepción.

El gusano *Caenorhabditis elegans* de la familia *Rhabditidae* mide aproximadamente 1 mm de longitud y vive enambientes templados. Es un importante modelo de estudioen biología, muy especialmente en genética del desarrollodes de los años 70, en parte debido a su pequeño número de neuronas y su genoma fácilmente manipulable. La nocicepción es la percepción neural de un estímulo real o potencialmente dañino. En *C. elegans* evoca un comportamiento de abstinencia autoconservante. Sin embargo, la estimulación repetida puede derivar en una respuesta de abstinencia o habituación reducida. Los investigadores compararon la respuesta de retirada a estímulos de luz perturbadores en *C. elegans* de tipo salvaje y una línea mutante de *C. elegans* que exhibe una respuesta más lenta de las neuronas sensoriales PVD. La falta de reacción indica habituación. Aquí están los porcentajes de animales probados que exhibieron una reacción de abstinencia a un estímulo nocivo que consiste en un número variable depulsos de luz consecutivos (IndMutant = 1 para la línea mutante y 0 para el tipo salvaje):

(a) En una primera aproximación para comparar los porcentajes de reacción de los gusanos mutantes con los salvajes, nos planteamos hacer un test de comparación de dos muestras independientes y un gráfico adecuado, sin tener en cuenta los pulsos. Para ello debemos observar que la variable respuesta es un porcentaje y el test t de Student habitual no sirve. Entre las posibles soluciones tenemos las siguientes:

1. Aplicar un test de Welch a los datos transformados con la función normalizante $\arcsin(\sqrt{p})$.
2. Hacer un test no paramétrico como el de Mann-Whitney-Wilcoxon.
3. Aplicar un test de permutaciones para comparar las dos muestras.
4. Realizar una regresión logística binomial.
5. Calcular una regresión beta que es especialmente apropiada para proporciones.

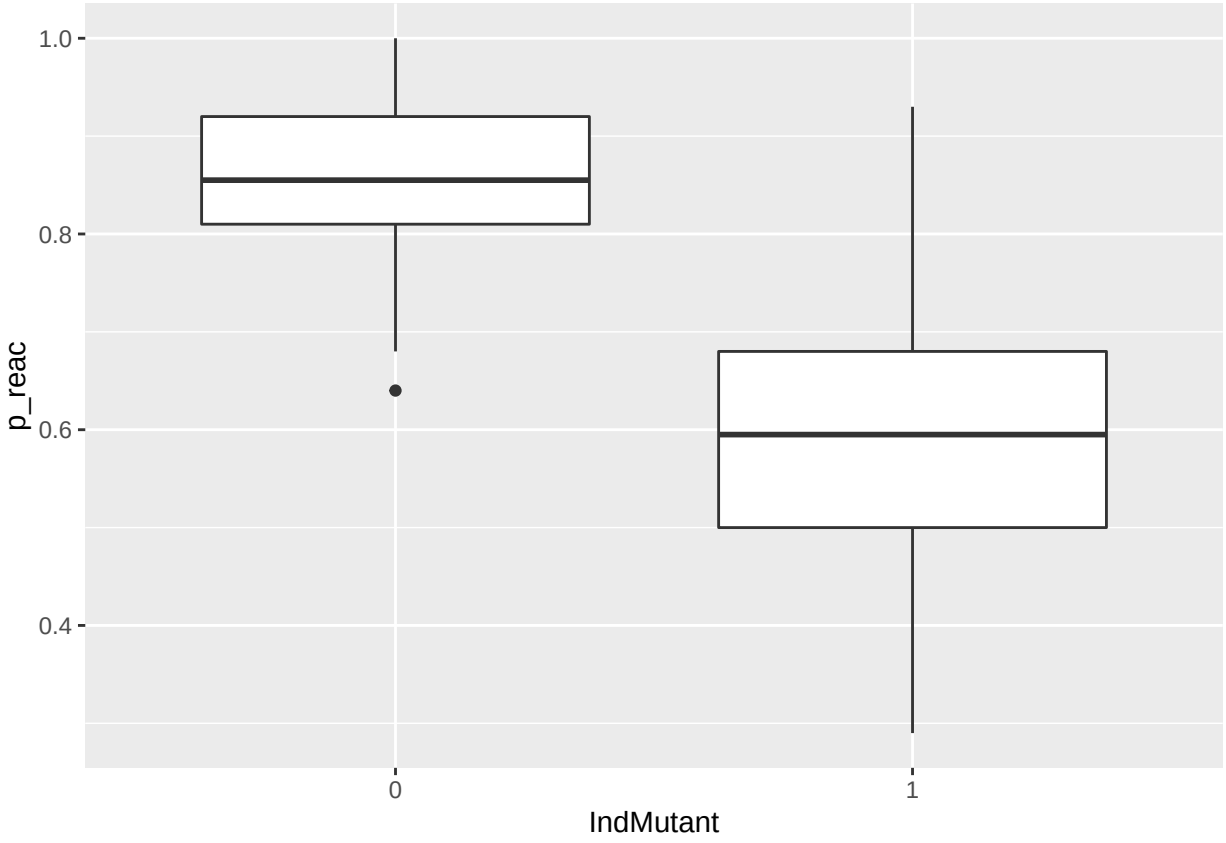
Las dos últimas propuestas son las más acertadas, ya que las otras contemplan comparar medias o medianas de porcentajes. En todo caso, realizar tres de las cinco soluciones con una de ellas entre las dos últimas y comentar su resultado.

```
df <-
    read_csv2("C_elegans.csv") %>%
    mutate(p_reac = Perc_Reacting/100,
           IndMutant = factor(IndMutant))
```

```
## Using ',' as decimal and '.' as grouping mark. Use read_delim() for more control.
```

```
## Parsed with column specification:
## cols(
##   Perc_Reacting = col_double(),
##   Pulses = col_double(),
##   IndMutant = col_double()
## )
```

```r
# boxplot de los datos
ggplot(df, aes(x = IndMutant, y = p_reac)) +
    geom_boxplot()
```



regresión logista binomial

```r
mod <-
    glm(IndMutant ~ p_reac, df, family = binomial(link = "logit"))
tidy(mod)
```

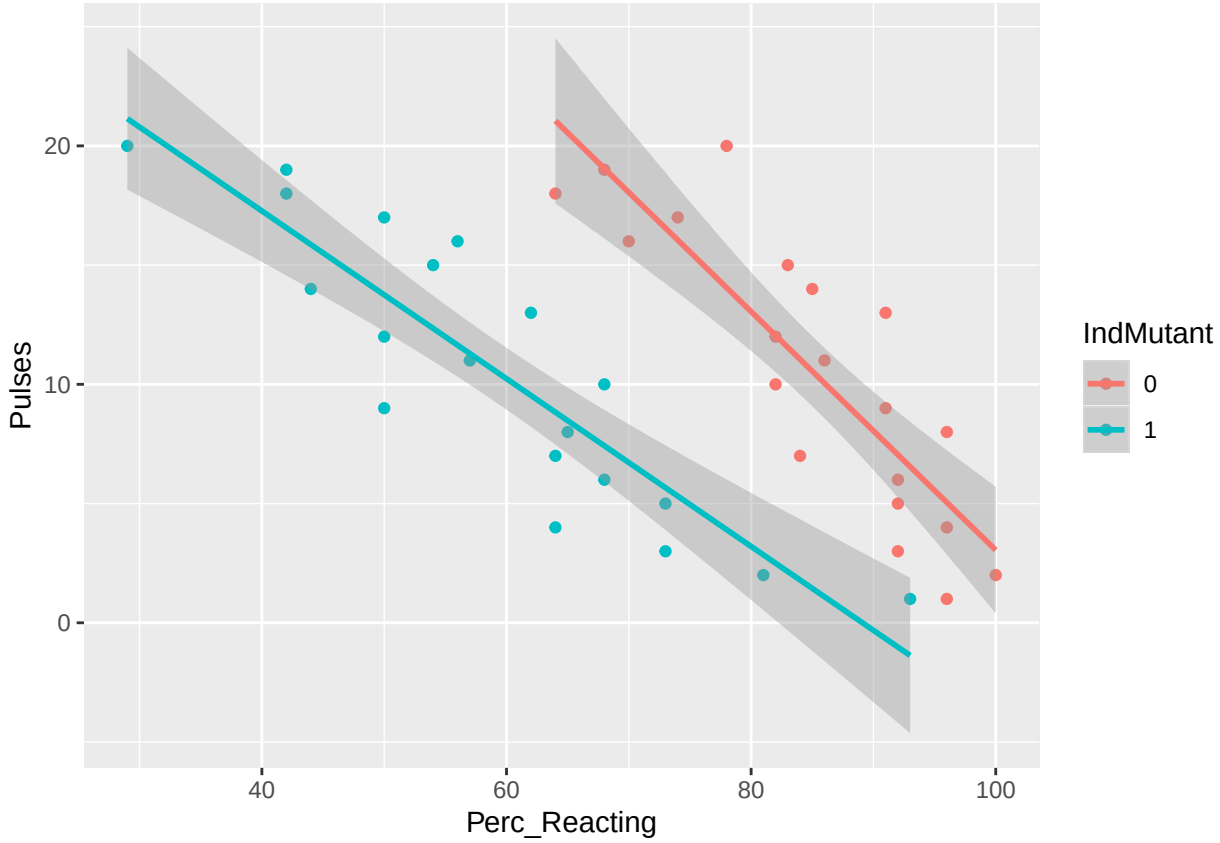
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     11.1      3.34      3.31 0.000924
## 2 p_reac         -15.1      4.48     -3.37 0.000755
```

Test de permutaciones

```r
mod <- coin::independence_test(p_reac ~ IndMutant, df)
```

(b) Dibujar el gráfico de dispersión del porcentaje de reacción en función del número de pulsos de luz con dos rectas que representen las dos sub poblaciones por separado, sin modelo común. Describir la forma de la relación. ¿Qué sugiere sobre el patrón de respuestas de abstinencia en las dos sub-poblaciones de *C. elegans* para un número creciente de pulsos de luz? ¿Cómo encaja su respuestaen el contexto de habituación? Explique por qué estos resultados sugieren una participación de las neuronas sensoriales de PVD en la nocicepción y la habituación. Nota: En este apartado consideramos la variable respuesta el porcentaje de reacción sin ninguna transformación.

```
ggplot(df, aes(Perc_Reacting, Pulses, color = IndMutant)) +
    geom_point() +
    geom_smooth(method = "lm")
```



(c) Proponer un modelo lineal que permita estudiar la asociación de la variable respuesta porcentajede reacción (sin ninguna transformación) con la variable Pulses y el factor IndMutant.

¿Cual es la ecuación del modelo final propuesto?

¿Es un modelo significativo?

¿Qué tanto por ciento de la variabilidad del porcentaje de reacción queda explicado por el modelo?

¿Todos los coeficientes de las variables explicativas del modelo son significativos?

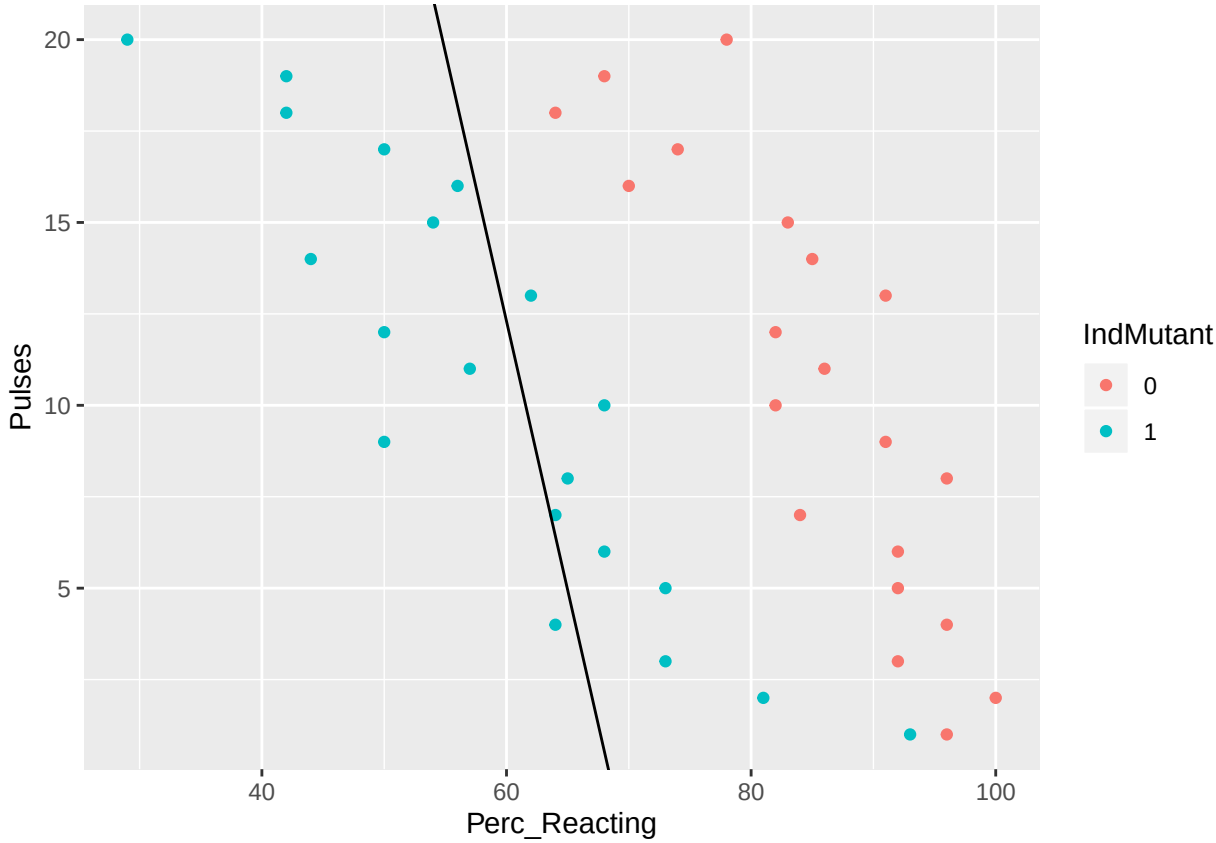
Dibujar el gráfico de dispersión del apartado anterior pero con las rectas resultantes del modelo. Nota: En este apartado no haremos ninguna transformación de la variable respuesta, ni utilizaremos ningún modelo generalizado, ya que confiamos en la validez del modelo balanceado.

```
mod <- lm(Perc_Reacting ~ Pulses * IndMutant, df)
summary(mod)
```

```
##
## Call:
## lm(formula = Perc_Reacting ~ Pulses * IndMutant, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -12.6391  -3.9090  -0.1015   4.6179  12.2857
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       100.5421     2.8911  34.776  < 2e-16 ***
## Pulses             -1.4707     0.2413  -6.094 5.21e-07 ***
## IndMutant1        -17.5684     4.0887  -4.297 0.000126 ***
## Pulses:IndMutant1  -0.7887     0.3413  -2.311 0.026683 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.224 on 36 degrees of freedom
## Multiple R-squared:  0.892,  Adjusted R-squared:  0.883
## F-statistic: 99.09 on 3 and 36 DF,  p-value: < 2.2e-16
```

```
ggplot(df, aes(Perc_Reacting, Pulses, color = IndMutant)) +
    geom_point() +
    geom_abline(intercept = coef(mod)[1], slope = coef(mod)[2])
```




```
coef(mod)
```

```
##       (Intercept)            Pulses        IndMutant1 Pulses:IndMutant1
##       100.5421053        -1.4706767       -17.5684211        -0.7887218
```

(d) Ahora contestaremos las preguntas y dibujaremos el gráfico del apartado anterior si realizamos la transformación normalizante $\arcsin(\sqrt{p})$ sobre la variable respuesta. Hallar el intervalo de confianza al 95 % para la media del porcentaje de mutantes que reaccionan a 10 pulsos de luz en condiciones

experimentales parecidas. Nota: Habrá que deshacer la transformación para dibujar "las rectas".

Ejercicio 2 (50 pt.)

En este ejercicio se estudian diferentes formasde examinar un modelo de regresión múltiple. Los datos de Umetrics (1995) provienen del campo del descubrimiento de fármacos. Se desarrollan nuevos medicamentos a partir de productos químicos que son biológicamente activos. Probar la actividad biológica de un compuesto es un procedimiento costoso, por lo que es útil poder predecir la actividad biológica apartir de mediciones químicas más baratas. De hecho, la química computacional hace posible calcular ciertas mediciones químicas sin siquiera hacer elcompuesto. Estas medidas incluyen el tamaño, la lipofilia y la polaridad en varios lugares de la molécula. Los datos se hallan en el archivo `penta.dat`. Mejor si eliminamos las observaciones con algún valor perdido. Queremos estudiar la relación entre estas mediciones y la actividad del compuesto, representada por el logaritmo de la actividad relativa de bradiquinina `log.RAI`. Debemos tener en cuenta que estos datos contienen muchos predictores en relación con el número de observaciones. Será útil hallar algunos factores predictivos subyacentes que expliquen la mayor parte de la variación en la respuesta. Por lo general, el modelo se ajusta a parte de los datos (el conjunto de "entrenamiento" o "trabajo") y la calidad del ajustese juzga por lo bien que predice la otra parte de los datos (el conjunto de "prueba" o "predicción"). Para este ejercicio tomaremos los siguientes conjuntos:

```
# cargamos los datos
penta <-
    read.table("penta.dat", header = T, na.strings = ".") %>%
    na.omit()

# funcion para caluclo de RMSE
rmse <- function(x,y) sqrt(mean((x-y)^2))
```

```
set.seed(321)
idx <- sample(1:30, size = 20, replace = F)
ptrain <- penta[idx,]
ptest <- penta[-idx,]
```

Para evaluar el ajuste utilizaremos la raíz cuadrada de la media de los errores al cuadrado (RMSE). Nota: Los mismos datos se hallan en el *data.frame* `Penta` del paquete `mvdalab` de **R**.

(a) Comprobar con los factores de inflación de la varianza que el modelo lineal completo y con todos los datos padece un problema grave de multicolinealidad. Esto justifica que debamos reducir el número de variables predictoras o utilizar algún método alternativo.

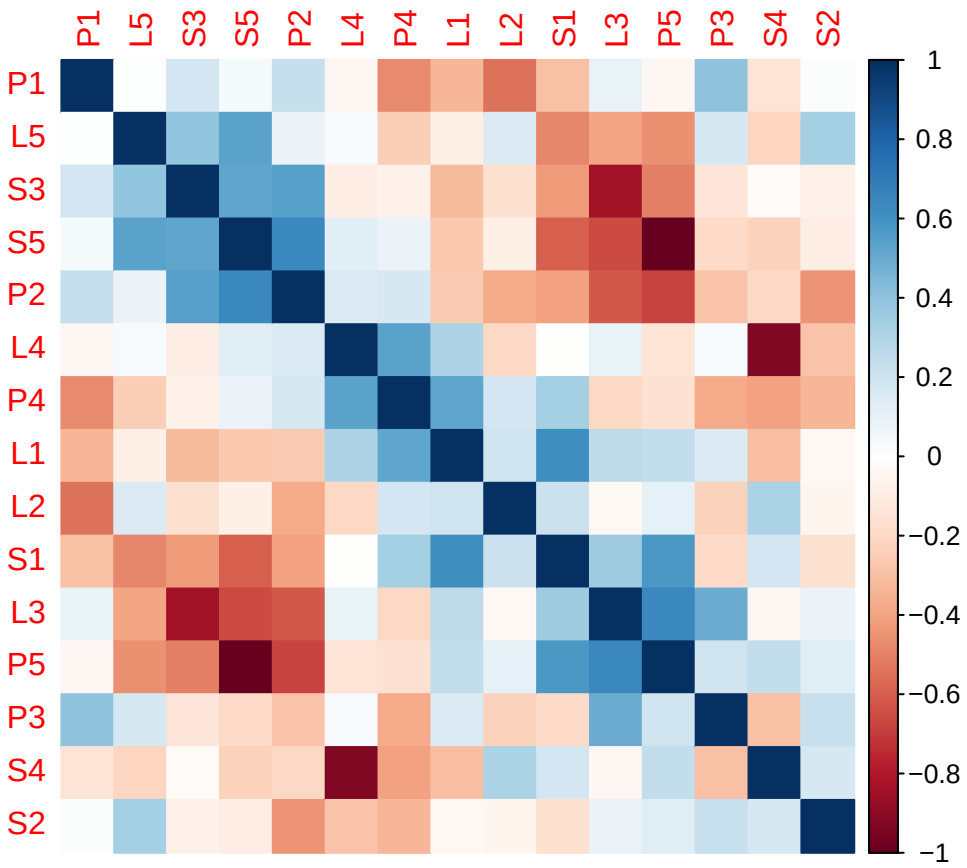
```
mod <- lm(log.RAI ~ ., ptrain[-1])
summary(mod)
```

```
##
## Call:
## lm(formula = log.RAI ~ ., data = ptrain[-1])
##
## Residuals:
##         22         18         13         25         24         16         17
##  4.857e-17 -2.332e-01  4.977e-16  3.021e-01 -1.287e-01  3.014e-01  1.137e-01
##          4         15         11         29         30          9          2
##  1.245e-01 -9.302e-02  4.992e-03 -1.994e-16 -1.087e-16 -2.177e-01 -4.036e-02
##         21         19         26         23          7         14
## -2.177e-01  3.577e-02  4.426e-02  4.842e-16 -2.177e-01  2.216e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  1.06933    1.49165   0.717    0.513
## S1          -0.06191    0.10432  -0.594    0.585
## L1          -0.02261    0.15107  -0.150    0.888
## P1          -0.16102    0.15621  -1.031    0.361
## S2           0.05369    0.06536   0.822    0.457
## L2           0.04448    0.09880   0.450    0.676
## P2          -0.07596    0.15053  -0.505    0.640
## S3          -0.08589    0.09257  -0.928    0.406
## L3           0.15639    0.13825   1.131    0.321
## P3           0.16740    0.32259   0.519    0.631
## S4           0.32417    0.40972   0.791    0.473
## L4           0.35716    0.34837   1.025    0.363
## P4          -0.32232    0.58070  -0.555    0.608
## S5          -1.34432    3.69537  -0.364    0.734
## L5           0.19308    0.44312   0.436    0.686
## P5          -0.99535    2.76249  -0.360    0.737
##
## Residual standard error: 0.3487 on 4 degrees of freedom
## Multiple R-squared:  0.9614, Adjusted R-squared:  0.8165
## F-statistic: 6.637 on 15 and 4 DF,  p-value: 0.04021
```

```r
X <- model.matrix(mod)[,-1]

# hacemos un plot de correlacion
corrplot::corrplot(cor(X), method = "color", order = 'AOE')
```



```r
# Chequeamos las vifs
vifs_x <- vif(X)
kable(rbind(vifs_x, SE = sqrt(vifs_x)), digits = 2)
```

	S1	L1	P1	S2	L2	P2	S3	L3	P3	S4	L4	P4	S5	L5	P5
vifs_x	9.18	10.66	3.83	2.97	3.87	6.42	9.05	22.19	7.88	69.83	55.39	13.07	6374.74	70.37	5912.62
SE	3.03	3.26	1.96	1.72	1.97	2.53	3.01	4.71	2.81	8.36	7.44	3.62	79.84	8.39	76.89

(b) Estudiar el modelo reducido que proporciona el método **stepwise**. Se trata de comparar los RMSE del modelo reducido y del modelo completo para el grupo de entrenamiento y para el grupo de prueba. También de ver si el modelo reducido salva el problemad e multicolinealidad.

```r
mod_step <- step(lm(log.RAI ~ ., ptrain[-1]), trace = 0)

anova(mod, mod_step)
```

```
## Analysis of Variance Table
##
## Model 1: log.RAI ~ S1 + L1 + P1 + S2 + L2 + P2 + S3 + L3 + P3 + S4 + L4 +
##     P4 + S5 + L5 + P5
## Model 2: log.RAI ~ S1 + P1 + S2 + L2 + S3 + L3 + L4 + P4
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1      4 0.48628
## 2     11 0.59796 -7  -0.11169 0.1312 0.9892
```

```r
ypred <- predict(mod, newdata = ptest)

rmse(ypred, ptest$log.RAI)
```

```
## [1] 0.582237
```

```r
ypred_step <- predict(mod_step, newdata = ptest)

rmse(ypred_step, ptest$log.RAI)
```

```
## [1] 0.5183494
```

```r
# Chequeamos las vifs
X <- model.matrix(mod_step)[,-1]
vifs_x <- vif(X)
kable(rbind(vifs_x, SE = sqrt(vifs_x)), digits = 2)
```

	S1	P1	S2	L2	S3	L3	L4	P4
vifs_x	1.76	1.97	1.24	1.61	5.12	6.08	2.19	3.42
SE	1.32	1.40	1.12	1.27	2.26	2.47	1.48	1.85

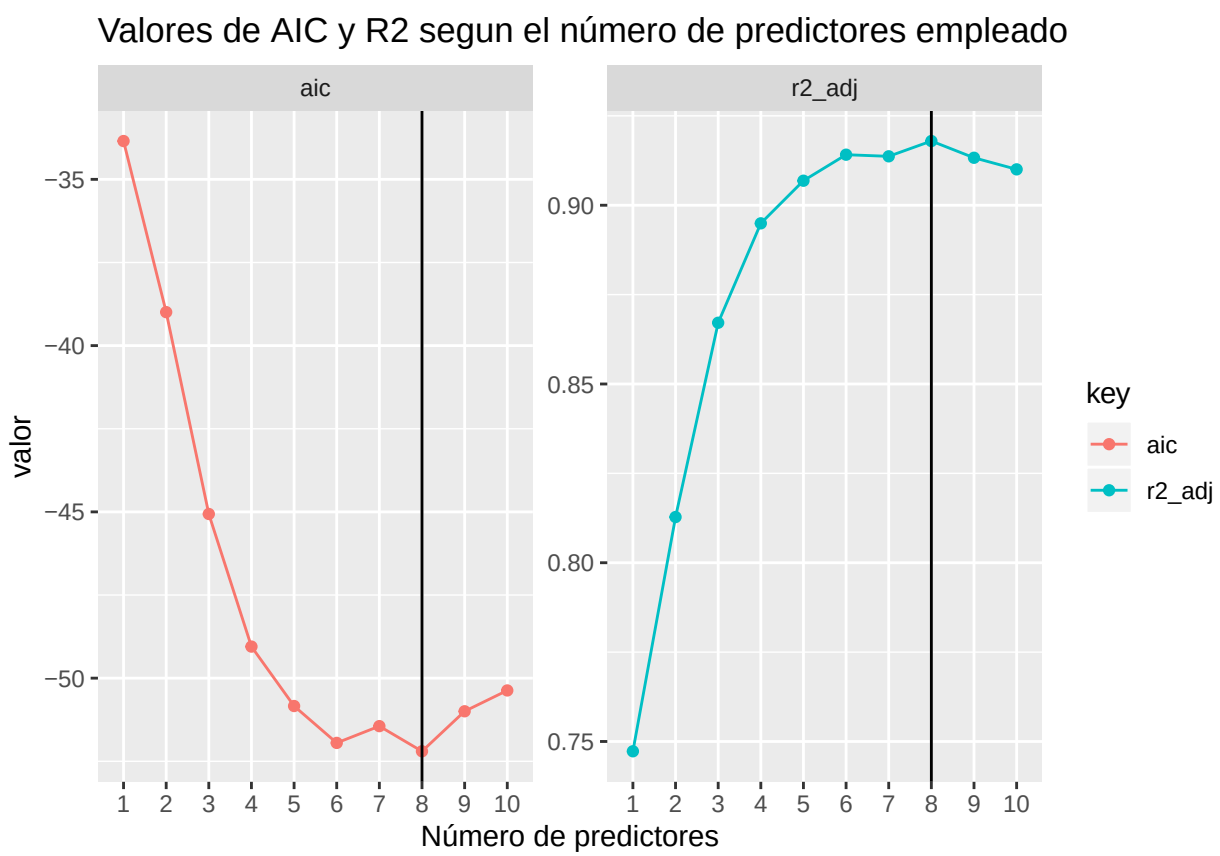
(c) Hallar el mejor modelo por el criterio del AIC y por el R2 ajustado. ¿Coinciden? ¿Es el mismo modelo que selecciona el stepwise?

```r
rs <-
    regsubsets(log.RAI ~ ., ptrain[-1], nvmax = 10) %>%
    summary()
```

```r
n <- dim(ptrain)[1]
p <- length(rs$rss)


rs_crit <-
    rs$which %>%
    as_tibble() %>%
    rowid_to_column() %>%
    mutate(
        rowid = factor(rowid),
        rss = rs$rss,
        r2_adj = rs$adjr2,
        aic = n * log(rss/n) + (2:(p+1)) * 2)

# Creamos el plot de AIC y R2
rs_crit %>%
    select(rowid, r2_adj, aic) %>%
    gather(key, value, -rowid) %>%
    ggplot(aes(rowid, value, color = key, group = key)) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept = 8) +
    facet_wrap(~key, scales = "free") +
    labs(title = "Valores de AIC y R2 segun el número de predictores empleado",
         x = "Número de predictores", y = "valor")
```



```r
# modelo elegido
vars_aic <- rs$obj$xnames[rs$which[8,]]
vars_step <- colnames(model.matrix(mod_step))


kable(rbind(vars_aic, vars_step))
```

vars_aic	(Intercept)	S1	P1	S2	L2	S3	L3	L4	P4
vars_step	(Intercept)	S1	P1	S2	L2	S3	L3	L4	P4

En este caso los criterios de AIC y de R2 ajustado, coinciden en que el mejor modelo es el que utiliza 8 parámetros (+ el intercepto). El modelo elegido por le criterio AIC / R2, resulta igual al modelo elegido mediante el procedimiento *stepwise*

(d) Estudiar una regresión por componentes principales con estos datos. ¿Qué tal funciona con 8 componentes para el grupo de entrenamiento? ¿Cual es el número de componentes que se recomienda si hacemos una validación cruzada? Con ese número mínimo de componentes, ¿cual es el RMSE para el conjunto de prueba?

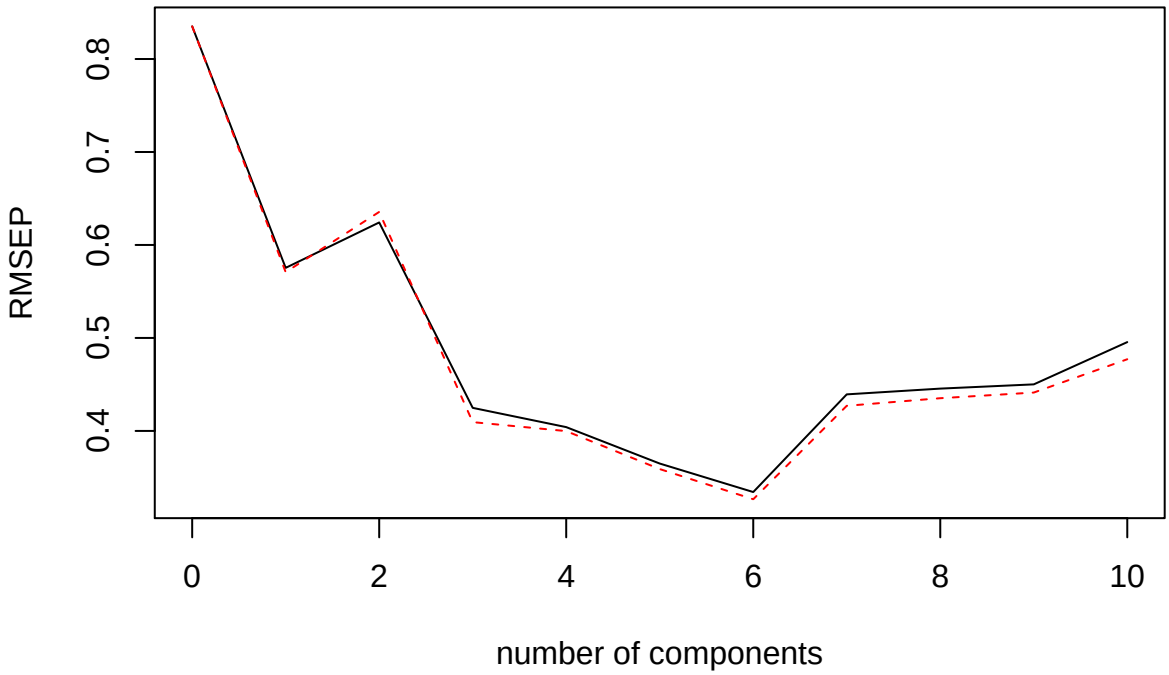
```r
pca_1 <-
    ptrain %>%
    select(S1:P5) %>%
    FactoMineR::PCA(graph = F)

kable(t(pca_1$eig)[,1:8], digits = 2)
```

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6	comp 7	comp 8
eigenvalue	4.64	2.92	2.41	1.70	0.84	0.74	0.60	0.39
percentage of variance	30.96	19.47	16.06	11.35	5.59	4.92	3.97	2.59
cumulative percentage of variance	30.96	50.43	66.49	77.84	83.44	88.36	92.33	94.92

```r
mod_pca <- pcr(log.RAI ~ ., data = ptrain[-1], validation = "CV", ncomp = 10)

plot(RMSEP(mod_pca))
```




```
ypred <- predict(mod_pca, ptest, ncomp=8)

rmse(ypred, ptest$log.RAI)
```

```
## [1] 0.5548939
```

(e) Estudiar una regresión PLS. ¿Cuantas componentes selecciona la validación cruzada?¿Cual es la variabilidad de las variables predictoras explicada por 3 componentes?Con el paqueteplsdepotse pueden estudiar las correlaciones entre las variables y las compo-nentes y realizar el gráfico llamadocírculo de correlaciones. ¿Cual es la variable predictora mejorcorrelacionada con la primera componente?Nota: Recordemos que ya hemos fijado la semilla aleatoria al principio:set.seed(321).(f) Estudiar también el método de Ridge Regression con la funciónlm.ridge(). Además de los valoresRMSE para el grupo de entrenamiento y de prueba, debemos acompañar el análisis con un gráficode los coeficientes.Para hallar el lambda óptimo podemos empezar por un límite superior bajo y aumentar sucesiva-mente ese límite hasta que el valor óptimo no sea el límite superior.Aunque siempre es mejor estandarizar los datos, en este ejercicio no lo haremos para no compli-car más los cálculos del RMSE del conjunto de prueba. Internamente, lo hace la propia funciónlm.ridge()1.Nota1: Si el resultado del ajuste de las predicciones a los datos es decepcionante podríamos optarpor utilizar la funcióncv.glmnet(…,alpha=0)del paqueteglmnet. Esta función no necesita que1Se puede comprobar si ejecutamos la instrucciónMASS::lm.ridge(sin paréntesis) en la consola deR.Página 4 de 6 le indiquemos una secuencia de valores de lambda, aunque sí requiere que le pasemos los datos comoobjetomatrix. Además podemos utilizar la validaciónleave one outsimplemente haciendo que elnúmero de carpetas (folds) sea exactamente el número de observaciones.Nota2: Aunque no hace falta, si se utilizan las funcinescv.glmnet()yglmnet()para el cálculodel modelo Ridge Regression, se puede comprobar que los resultados difieren notablemente de losque tenemos con la funciónlm.ridge(). Una sencilla explicación se puede leer enhttps://stats.stackexchange.com/questions/74206/ridge-regression-results-different-in-using-lm-ridge-and-glmnet(g) Finalmente estudiaremos el método LASSO con la funciónlars()del paquete del mismo nombre.Nota: Tal vez el resultado es demasiado restrictivo y hay que aumentar el número de variablesseleccionadas, lo que equivale a aumentar ligeramente el valor de lambda.Ejercicio 3 (25 pt.)Es posible utilizar la regresión logística para estudiar la discriminación entre dos grupos

dado un conjuntode variables explicativas. Para más detalles podéis consultar el apartado 8.10 del libro de Manly[1].(a) Realizar un análisis discriminante con ayuda de la regresión logística con los datos de los 49 gorrioneshembra después de una tormenta.Reproducir con todo detalle la tabla 8.6 de la página 153 del libro de Manly.(b) Contrastar con un test 2la significación de la regresión y explicar su resultado.(c) Realizar un gráfico como el de la figura 8.2 de la página 156 del libro de Manly pero con los datosde este ejercicio y valorar el resultado.(d) Calcular un intervalo de confianza para el parámetro de la variablex4=length humerusy para suodds ratio.(e) Calcular la tabla de confusión de la clasificación obtenida con la regresión logística.Página 5 de 6 Referencias[1] Bryan F.J. Manly and Jorge A. Navarro Alberto (2017),Multivariate Statistical Methods: A Primer,Fourth Edition. Chapman and Hall/CRC. Springer-Verlag.[2] SAS/STAT(R) 9.22 User's Guide, "The PLS Procedure".[3]https://v8doc.sas.com/sashtml/stat/chap51/sect19.htm%5B4%5D Umetrics, Inc. (1995), Multivariate Analysis (3-day course), Winchester, MA.