```
In [4]: library(readxl)
library(dplyr)
library(ggplot2)
library(ISLR)
library(dplyr)
```

```
Warning message:
"package 'dplyr' was built under R version 3.6.3"
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
"package 'ggplot2' was built under R version 3.6.3"Warning message:
"package 'ISLR' was built under R version 3.6.3"
```

Trabajo final

Cristóbal Collao

Modelo lineales

Fecha: 12-07-2021

Considere los datos anexos entregados en la tabla de Excel "Datos para evaluación final". Estos corresponden a la evaluación de 32 individuos en una asignatura en la cual se les calificaba inicialmente mediante un examen de conocimientos previos, y luego, después de un tiempo de clases, se tomaron una serie de evaluaciones parciales en donde se registraron las calificaciones promedios obtenidas por cada individuo. Por otra parte, un grupo de ellos siguió una metodología de enseñanza (digamos, 1), y el otro, otra distinta (digamos, 0). La evaluación final está categorizada como 1, si el individuo obtuvo dentro del rango de calificación máxima, y 0 en otro caso. \ El objetivo es evaluar la incidencia de las calificaciones registradas y el método de enseñanza en la calificación final. Para ello, realice lo siguiente: \

1)Plantee un modelo de regresión logística, estime sus parámetros vía estimación de máxima verosimilitud, describiendo el algoritmo numérico a utilizar para aproximar estos, especificando las funciones involucradas en el proceso iterativo.

Primero partirimos leyendo los datos y guardándolos en la variable data y cambiandole los nombres a las variables para manejarlas más facilmente

```
In [5]: data<-read_excel("Datos para evaluación final.xlsx")
data<- as.data.frame(data)
data
```

A data.frame: 32 × 5

Observaciones	Calificaciones de conocimientos previos	Promedio de calificaciones iniciales	Método de enseñanza	Calificación final
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	20	2.66	0	0
2	22	2.89	0	0
3	24	3.28	0	0
4	12	2.92	0	0
5	21	4.00	0	1
6	17	2.86	0	0
7	17	2.76	0	0
8	21	2.87	0	0
9	25	3.03	0	0
10	29	3.92	0	1
11	20	2.63	0	0
12	23	3.32	0	0
13	23	3.57	0	0
14	25	3.26	0	1
15	26	3.53	0	0
16	19	2.74	0	0
17	25	2.75	0	0
18	19	2.83	0	0
19	23	3.12	1	0
20	25	3.16	1	1
21	22	2.06	1	0
22	28	3.62	1	1
23	14	2.89	1	0
24	26	3.51	1	0
25	24	3.54	1	1
26	27	2.83	1	1
27	17	3.39	1	1
28	24	2.67	1	0
29	21	3.65	1	1
30	23	4.00	1	1
31	21	3.10	1	0
32	19	2.39	1	1

```
In [6]: names(data)<- c("observaciones","calificacion_previos","promedio_inicial","metodo_enseñanza","calificac
ion_final")
data
```

A data.frame: 32 × 5

observaciones	calificacion_previos	promedio_inicial	metodo_enseñanza	calificacion_final
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	20	2.66	0	0
2	22	2.89	0	0
3	24	3.28	0	0
4	12	2.92	0	0
5	21	4.00	0	1
6	17	2.86	0	0
7	17	2.76	0	0
8	21	2.87	0	0
9	25	3.03	0	0
10	29	3.92	0	1
11	20	2.63	0	0
12	23	3.32	0	0
13	23	3.57	0	0
14	25	3.26	0	1
15	26	3.53	0	0
16	19	2.74	0	0
17	25	2.75	0	0
18	19	2.83	0	0
19	23	3.12	1	0
20	25	3.16	1	1
21	22	2.06	1	0
22	28	3.62	1	1
23	14	2.89	1	0
24	26	3.51	1	0
25	24	3.54	1	1
26	27	2.83	1	1
27	17	3.39	1	1
28	24	2.67	1	0
29	21	3.65	1	1
30	23	4.00	1	1
31	21	3.10	1	0
32	19	2.39	1	1

Lo siguiente será crear el modelo con la función glm, en este caso se toman las 3 variables explicativas y la variable de respuesta será calificacion_final, además se imprime un resumen del modelo

```
In [8]: modelo <- glm(calificacion_final ~ calificacion_previos +
                   promedio_inicial +
                   metodo_enseñanza , data = data, family=binomial)
summary(modelo)
```

```
Call:
glm(formula = calificacion_final ~ calificacion_previos + promedio_inicial +
    metodo_enseñanza, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9551  -0.6453  -0.2570   0.5888   2.0966

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -13.02135    4.93127  -2.641  0.00828 **
calificacion_previos   0.09516    0.14155   0.672  0.50143
promedio_inicial       2.82611    1.26293   2.238  0.02524 *
metodo_enseñanza       2.37869    1.06456   2.234  0.02545 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.183  on 31  degrees of freedom
Residual deviance: 25.779  on 28  degrees of freedom
AIC: 33.779

Number of Fisher Scoring iterations: 5
```

Y en la siguiente celda se imprimen los coeficientes del modelo

```
In [9]: m = 32
coef<-as.vector(coef(modelo))
coef
```

-13.0213468563076 · 0.0951576612956503 · 2.82611259455938 · 2.3786876548125

2)Estime la matriz de varianzas/covarianzas de los estimadores, y concluya si los predictores son individualmente significativos, utilizando el estadístico Z usual, con alfa=0,05.

```
In [10]: varycov<-as.data.frame(summary(modelo)[16])
varycov
```

A data.frame: 4 × 4

	cov.unscaled..Intercept.	cov.unscaled.calificacion_previos	cov.unscaled.promedio_inicial	cov.unscaled.metodo_enseñanza
	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	24.3174487	-0.34624889	-4.57338640	-2.35909128
calificacion_previos	-0.3462489	0.02003740	-0.03692112	0.01491196
promedio_inicial	-4.5733864	-0.03692112	1.59499873	0.42760245
metodo_enseñanza	-2.3590913	0.01491196	0.42760245	1.13328133

```
In [11]: varnosignificativas<-summary(modelo)[12]
varnosignificativas<-as.data.frame(varnosignificativas)
varnosignificativas<-filter(varnosignificativas, abs(coefficients.z.value) <= qnorm(0.975))
varnosignificativas
```

A data.frame: 1 × 4

	coefficients.Estimate	coefficients.Std..Error	coefficients.z.value	coefficients.Pr...z..
	<dbl>	<dbl>	<dbl>	<dbl>
calificacion_previos	0.09515766	0.1415535	0.6722381	0.5014321

Tenemos que la variable califiacion_previos no es significativa usando el estadístico Z con alpha=0.05

3)Dado que en este caso no podemos obtener un coeficiente de determinación como usualmente se obtiene vía ANOVA para el modelo normal, definiremos el siguiente seudo coeficiente de determinación: Número de predicciones correctas/Número total de observaciones. Acá, una observación será predicha como 1 si esta resulta estar en el rango [1, 0,5); y 0 en caso contrario. Luego, la observación predicha será correcta si esta última clasificación corresponde con lo observado. Obtenga entonces el porcentaje de observaciones predichas por el modelo ajustado.

```
In [12]: pred<-predict(modelo,type='response')
predicciones<-ifelse(pred > 0.5, 1,0)
```

```
In [13]: comparacion<-data %>% mutate (comp= predicciones == calificacion_final)
mean(comparacion$comp)
```

0.8125

Segun el indicador que creamos para ver la acertividad de nuestro modelo, tenemos que nuestro modelo acerta en un 81,25% de los casos en los que se puso a prueba

4)Obtenga las Deviances del modelo ajustado completo, y luego en los modelos reducidos sin solamente el método de enseñanza, luego sin solamente las calificaciones de conocimientos previos, y finalmente sin estas dos, en conjunto.

¿Existe, a un nivel alfa=0,05, influencia significativa en la calificación final de estos predictores a nivel individual y en conjunto? Comparar con lo obtenido en 1) e interpretar.

Deviance del modelo ajustado completo

```
In [15]: modelo$deviance
```

25.7792684442628

```
In [17]: modelo_1<-glm(calificacion_final ~ calificacion_previos +
        promedio_inicial , data = data, family=binomial)
summary(modelo_1)
```

```
Call:
glm(formula = calificacion_final ~ calificacion_previos + promedio_inicial,
    family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4085  -0.6882  -0.4677   0.7262   2.4553

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -10.65600    4.05709  -2.627  0.00863 **
calificacion_previos   0.08555    0.13318   0.642  0.52064
promedio_inicial       2.53828    1.18185   2.148  0.03174 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.183  on 31  degrees of freedom
Residual deviance: 31.983  on 29  degrees of freedom
AIC: 37.983

Number of Fisher Scoring iterations: 4
```

Para el modelo sin la variable método de enseñanza la deviance aumenta a 31.98

```
In [18]: modelo_1$deviance
```

31.9829660657946

```
In [22]: modelo_2<-glm(calificacion_final ~ promedio_inicial, data = data, family=binomial)
summary(modelo_2)
```

```
Call:
glm(formula = calificacion_final ~ promedio_inicial, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3672  -0.6854  -0.5262   0.7193   2.4365

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.703      3.671  -2.643  0.00821 **
promedio_inicial    2.840      1.127   2.520  0.01173 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.183  on 31  degrees of freedom
Residual deviance: 32.418  on 30  degrees of freedom
AIC: 36.418

Number of Fisher Scoring iterations: 4
```

Para el caso en el que se mantenía solo la como variable explicativa el promedio incial, la deviance es de 32.418

Para el modelo sin la variable conocimientos previos, la deviance se mantiene casi igual que con todas las variables explicativas.

```
In [21]: modelo_2$deviance
```

26.2531472732633

```
In [26]: modelo_3<-glm(formula = calificacion_final ~ 1 , data = data, family = "binomial" )
summary(modelo_3)
```

```
Call:
glm(formula = calificacion_final ~ 1, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9178  -0.9178  -0.9178   1.4614   1.4614

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6466     0.3722  -1.737   0.0823 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.183  on 31  degrees of freedom
Residual deviance: 41.183  on 31  degrees of freedom
AIC: 43.183

Number of Fisher Scoring iterations: 4
```

Para el modelo saturado la deviance es de 41.183

Para finalizar hay que mencionar que existe solo una variable que resulta en todos los modelos no ser significativa y es la variable de calificacion previas. Además el modelo completo es el modelo con menor deviance y por ende, el modelo más completo