1. Consideremos los datos de PIMA.tr vistos en clase.

   a) En una primera etapa exploratoria de los datos usar la función ggpairs para inspeccionar la relación entre las variables. ¿Qué sugiere?

   b) Ajustar un modelo de regresión logística para predecir diabetes usando todas las variables. ¿Cuánto valen las estimaciones de los coeficientes? ¿Cuáles de estos coeficientes son significativos?

   c) A partir del ajuste anterior construir un intervalo de confianza de nivel asintótico 0.95 para cada uno de los coeficientes del modelo ajustado.

   d) Construir una *tabla de confusión* (confusion matrix o error matrix) que consiste en una tabla de doble entrada en la que cruzan la clasificación observada de los individuos y la predicha por el modelo. Computar el porcentaje de aciertos, que se conoce como la precisión de predicción.

	Predichos	
Observados	0	1
0		
1		

   e) Repetir el ítem anterior sobre la muestra de testeo Pima.te. Estime el error de predicción.

   f) ¿Cuál sería el valor estimado de la probabilidad de tener diabetes para una mujer de la etnia Pima que tuvo 2 embarazos y tiene una concentración de glucosa de 100, una presión sistólica de 70, un valor de piel de triceps de 20, un índice de masa corporal de 26, un valor de la función pedigree de 0.24 y 30 años de edad?

   g) Si se usa el ajuste realizado: ¿cuánto vale la diferencia de log odds cuando la glucosa aumenta en 10 unidades y todas los demás valores permanecen constantes? ¿cuánto vale el cociente de los odds cuando la glucosa aumenta en 10 unidades y todas los demás valores permanecen constantes? Hallar intervalos de confianza de nivel aproximado 0.95 para el parámetro estimado en ambos casos.

   h) Implementar una función que dada la salida de un ajuste del procedimiento glm, compute un intervalo de confianza de nivel aproximado $1-\alpha$ para la probabilidad de éxito en un punto x.

i) Calcular un intervalo de confianza de nivel aproximado 0.95 para la probabilidad del ítem f).

2. Un investigador está interesado en saber como las variables GRE (Graduate Record Exam scores), GPA (grade point average) y prestigio de la institutción de pregrado, influyen en la admisión a una escuela de grado. La variable de respuesta es la admisión o no del candidato.

   La variable de respuesta de este conjunto de datos es la respuesta binaria **admit**, mientras las variables regresoras son: **gre**, **gpa** y **rank**. Trataremos a las variables gre y gpa como continua. La variable rank toma los valores de 1 a 4, donde 1 corresponde a las instituciones de mayor prestigio y 4 a las de menor.

   `https://stats.idre.ucla.edu/stat/data/binary.csv`

   a) Realizar un análisis exploratorio de las variables presentes.

   b) Ajustar un modelo logístico usando todas las variables explicativas. (Tener presente la naturaleza de la variable **rank**. Transformar: rank = factor(rank)).

   De acuerdo, a este ajuste ¿en cuánto aumenta el log odds de la admisión cuando la variable **gre** aumenta en una unidad? ¿Cómo se interpretan las estimaciones de los coeficientes relacionados con la variable **rank**?

   c) ¿Cuánto vale la probabilidad estimada de la admisión para los distintos niveles de la variable **rank** cuando los otros dos predictores toman como valor la media muestral? En consecuencia, ¿Cuál es el valor predicho de la admisión para cada caso?

   d) ¿Cuáles son los valores estimados del cociente de los odds asociados cuando cada variable aumenta en una unidad y el resto permanece constante? Hallar intervalos de confianza de nivel aproximado 0.95 para cada uno de ellos, ¿alguno contiene al 1?

   e) Realizar un análisis secuencial de salida mediante el comando anova de R con la opción test="Chisq".¿Cómo se obtendría cada uno de estos resultados? ¿Cómo se interpretan cada uno de los resultados que arroja esta función?

   f) Comparar el resultado anterior con los que obtendría al realizar

```
# modelo solo intercept
mod1 <- glm(admit ~ 1,  data = datos, family = "binomial")
# modelo intercept + gre
mod2 <- glm(admit ~ gre, data = datos, family = "binomial")
```

```
# modelo intercept + gre + gpa
mod3 <- glm(admit ~ gre + gpa,data = datos, family = "binomial")
# modelo completo
mod4 <- glm(admit ~ gre + gpa + rank, data = datos, family = "binomial")

anova(mod1, mod2, test="LRT")
anova(mod2, mod3, test="LRT")
anova(mod3, mod4, test="LRT")
```

¿Cambia algo en estos resultados de anova si se usa test=“Chisq“?

g) Mediante la instrucción anova compare los tres modelos posibles con solo dos variables (gre+gpa, gre+rank, gpa+rank) con el modelo completo.