Laboratorio de Regresión

Crea un nuevo proyecto y a continuación descarga descarga el modelo de informe usando la siguiente función en la consola:

```
download.file("http://ow.ly/O9PC30bcwZH",
              destfile = "lab1.Rmd")
```

Bateando

En la película Moneyball un entrenador usa la estadística para hacer que un equipo de bajo presupuesto logre ganar. Para hacerlo, usa la estadísticas poco usadas como la habilidad de un jugador para robar una base en vez para predecir mejor la habilidad de anotar entradas que otras más comunes como home runs, carreras impulsadas y el promedio de bateo. Al obtener jugadores que eran excelentes en esas estadísticas poco usadas fue mucho más rentable para el equipo.

En este práctico vamos a ver datos de todos los 30 equipos de ligas mayores de béisbol (para el año que viene voy a ver si consigo datos de fútbol) de EEUU y examinar la relación entre el número de carreras anotadas y examinar la relación linear entre las carreras anotadas en una temporada y otras estadísticas de los jugadores. Nuestro objetivo va a ser resumir las relaciones tanto como gráfica y analíticamente para encontrar que variable, si es que hay alguna, nos ayuda a predecir mejor las entradas anotadas en cada temporada.

Los datos

Carguemos los los datos usando:

```
load(url("https://stat.duke.edu/~mc301/data/mlb11.RData"))
```

Además de las carreras anotadas (`runs`), en conjunto de datos hay siete variables usadas tradicionalmente: bateos (`at_bats`), contactos (`hits`), home runs (`homeruns`), promedio de bateos (`bat_avg`), strikeouts (`strikeouts`), bases robadas (`stolen_bases`) y ganadas (`wins`). También hay tres variables nuevas: porcentaje en bases (`new_onbase`), potencia media (`new_slug`), y en base más potencia media (`new_obs`). Para la primera parte del análisis vamos a usar las siete variables tradicionales. Al final del laboratorio, vas a trabajar con tres nuevas variables por tu cuenta.

```
##Ejercicio 1
¿Qué tipo de gráfico usarias para mostrar las relaciones entre carreras anotadas
y una de las otras variables? Gráfica esta relación usando la variable `at_bats`
como variable predictora. ¿Se ve lineal la relación? Si conocieras el número de
bateos de un equipo estarías cómodo usando un modelo lineal para predecir el
número de carreras anotadas?
```

Si la relación se ve lineal usarías, podemos cuantificar la fuerza de esa relación usando los coeficientes de correlación.

```
mlb11 %>%
  summarise(cor(runs, at_bats))
```

Suma de cuadrados residuales

En esta sección vamos a usar una función interactiva para investigar que a que nos referimos con la "suma de cuadrados residuales". Es necesario que ejecutes la función en la consola y no en el documento de markdown.

Al correr la función también es necesario el conjunto de datos mlb11 esté cargado en el espacio de trabajo.

Para cargarla ejecuta en la consola el siguiente código:

```
source(url("https://raw.githubusercontent.com/StatsWithR/statsr/master/R/plot_ss.R"))
```

Piensa la forma en que describimos la distribución de una sola variable. Recuerda que discutimos características como el centro, la dispersión y la forma. También es útil ser capaz de describir la relación entre dos variables numéricas, tales como `runs` y `at_bats` arriba.

```
##Ejercicio 2

Viendo el gráfico del ejercicio anterior, describe la relación entre las dos
las dos variables. Asegurate de discutir la forma, la dirección y la fuerza
de la relación, así como observaciones inusuales.
```

Tal como usamos la relación entre la media y el desvío estándar para resumir una sola variable, podemos resumir la relación entre estas dos variables encontrando la línea que mejor sigue su asociación. Usen la siguiente función interactiva para seleccionar la linea que crees que hace mejor el trabajo de ir por la nube de puntos.

```
plot_ss(x = at_bats, y = runs, data = mlb11)
```

Luego de ejecutar esta función, les pedirá que hagan clic en dos puntos del gráfico para definir una línea. Una vez que lo hagan, se va a ver la linea especificada en negro y los residuales en azul. Noten que hay 30 residuales, una para cada una de las 30 observaciones. Recuerden que los residuales son la diferencia entre los valores observados y los predichos por la línea:

$$e_i = y_i - \hat{y}_i$$

La forma más común de hacer la regresión lineal es seleccionar la línea que minimiza la suma de los cuadrados de los residuales. Para visualizar los cuadrados de los residuales, podes volver a correr la función del gráfico y añadir el argumento showSquares = TRUE.

```
plot_ss(x = at_bats, y = runs, data = mlb11, showSquares = TRUE)
```

Noten que la salida de la función `plot_ss` les provee con la pendiente y la ordenada al origen de su línea así como también la suma de cuadrados.

```
## Ejercicio 3
Usando plot_ss, elijan una línea que haga un buen trabajo minimizando la suma
de cuadrados. Corre la función varias veces. ¿Cuál es la menor suma de cuadrados
que obtuvieron? ¿Cómo se compara con sus las de sus vecinos?
```

Es bastante incómodo tratar de obtener la línea de mínimos cuadrados, es decir la línea que minimiza la suma de cuadrados, a través de prueba y error. En vez de eso podemos usar la función `lm` en R para ajustar nuestro modelo de de regresión lineal (también conocido como línea de regresión).

```
m1 <- lm(runs ~ at_bats, data = mlb11)
```

Los primeros dos argumentos de la función `lm` tiene la forma de `y ~ x`. Aquí puede ser leído como que queremos hacer un modelo lineal de las carreras `runs` como función de `at_bats`. El segundo argumento le especifica a R que debe buscar en el `data.frame` para buscar las variables `runs`y `at_bats`.

La salida de `lm` es un objeto que contiene toda la información que necesitamos saber sobre el modelo lineal que acabamos de ajustar. Podemos acceder a la información usando función `summary`.

```
summary(m1)
```

Vamos a interpretar la salida parte por parte. Primero, la formula usada para describir el modelo se muestra arriba. Luego de la formula, encontramos un resumen de cinco números de los residuales. La tabla de "Coeficientes" que se muestra a continuación es clave; la primer columna muestra la ordenada al origen y el coeficiente de `at_bats`. Con esta tabla podemos escribir la ecuación de regresión mínimos cuadrados para la línea de regresión del modelo linear:

$$\hat{y} = -2789.2429 + 0.6305 \times at\_bats$$

Lo último que vamos a discutir del resumen de la salida es la R-Cuadrado múltiple o simplemente R^2. El valor de R^2 representa la proporción de variabilidad en la variable de respuesta que es explicada por la variable explicatoria. Para este modelo, el 37.3% de la variabilidad de las entradas está explicada por `at_bats`.

```
## Ejercicio 4

Ajusta un modelo que use homeruns para predecir runs. Usando lo estimados del
salida de R, escribe la ecuación de recta de regresión. ¿Qué nos dice la pendiente
en el contexto de la relación entre el éxito de un equipo y sus homeruns?
```

Predicción y errores de predicción

Vamos a crear un diagrama de dispersión con la regresión de mínimos cuadrados superpuesta.

```
ggplot(data = mlb11, aes(x = at_bats, y = runs)) +
  geom_point()+
  stat_smooth(method = "lm", se = FALSE)
```

Aquí añadimos literalmente una nueva capa encima del gráfico. `stat_smooth` crea la lía ajustando un modelo linear. También puede mostrarnos el error estándar asociado con nuestra línea, pero lo eliminamos por ahora.

Esta línea puede ser usada para predecir y con cualquier valor de x. Cuando se hacen predicciones fuera del rango de los datos observados, se refiera a ella como extrapolación y no es recomendado. Sin embargo las predicciones hechas dentro del intervalo de los datos son más confiables. Y son usadas para calcular los residuales.

```
## Ejercicio 5

Si un manager de equipo viese la línea regresión de mínimos cuadrados y no los
datos verdaderos ¿Cuántas entradas predeciría para un equipo con 5579 at_bats
```

Diagnósticos del Modelo

Para comprobar que un modelo lineal es confiable, necesitamos probar (1) linealidad, (2) residuales cercanos a normales, y (3) variabilidad constante.

Linealidad: Ya comprobamos si la relación entre entradas y bateos es lineal usando un diagrama de dispersión. Deberíamos verificar esta condición con un gráfico de los residuales vs valores ajustados (predichos).

```
ggplot(aes(x = .fitted, y = .resid), data = m1) +
  geom_point()+
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Valores ajustados") +
  ylab("Residuales")
```

Observa que nuestro objeto del modelo `m1` puede servir también como datos porque dentro tiene guardado los valores ajustados $\hat{y}$ y también los residuales. También nota que estamos usando código nuevo. Luego de crear

el diagrama de dispersión (primeras dos líneas de código), agregamos una línea rayada horizontal en $y = 0$ (para ayudarnos a ver si los residuales están distribuidos alrededor de 0), también lo usamos para ajustar las etiquetas de los ejes para que sean más informativas.

```
## Ejercicio 6

¿Hay algún patrón aparente en el gráfico de residuales? ¿Qué indica sobre
la linealidad de la relación entre runs y at-bats?
```

Residuales cercanos a normales: para comprobar este supuesto podemos ver un histograma.

```
ggplot(aes(x = .resid), data = m1) +
  geom_histogram(binwidth = 25) +
  xlab("Residuales")
```

o un gráfico de probabilidad normal de los residuales

```
ggplot(aes(sample = .resid), data = m1) +
       stat_qq()
```

Noten que la sintaxis para hacer un gráfico de probabilidad normal es algo diferente de lo que están acostumbrados a ver: pusimos que `sample` (muestra) es igual a los residuales en vez de `x` y usamos el método estadístico qq, que quiere decir "quantile-quantile" (cuantil-cuantil), otro nombre usado normalmente para los gráficos de probabilidad normal.

```
## Ejercicio 7

Basados en el histogramo y el gráfico de probabilidad normal ¿se cumple con
el supuesto de normalidad de los residuos?
```

Variabilidad constante:

```
## Ejercicio 8

Basados en los gráficos de residuales vs ajustados ¿Se cumple con el supuesto
de variabilidad constante?
```

Regresión Múltiple

La regresión múltiple es una extensión de la regresión simple. Cuando se tienen dos o más variables explicatorias es un modelo de regresión múltiple. Vamos a probar cuanto explica el modelo lineal cuando tenemos bateo y homeruns como variable regresoras.

```
m2 <- lm(runs ~ at_bats + homeruns, data = mlb11)
```

Como vemos, la función es la misma y la formula se escribe de forma similar. Pero agregamos un `+` por cada variable explicatoria que agregamos.

Revisemos el nuevo modelo.

```
summary(m2)
```

Como en el modelo simple tenemos arriba la formula que produjo el modelo. Debajo tenemos 5 números que resumen los residuales. Luego la tabla de coeficientes, primero la ordenada al origen ("Intercept"), después la pendiente de `at_bats` y luego la pendiente de `homeruns`. Además nos dice si las pendientes y ordenadas al origen son significativamente distintas de 0. Abajo del todo, aparecen 3 datos: la suma de residuales del modelo y sus grados de libertad, luego el R^2 seguido de $\bar{R}^2$ (R cuadrado ajustado) recordemos que si añadimos variables regresoras al modelo por azar puede aumentar R^2 y $\bar{R}^2$ corrije esto penalizando por

cada nueva variable regresora que agregamos. Luego tenemos el estadístico F que prueba la significancia del modelo completo.

En este caso la ecuación que nos define la nueva recta o mejor dicho plano de ajuste está dada por:

$$\hat{y} = -1613.99 + 0.3761 \times at\_bats + 1.5167 \times homeruns$$

El diagnóstico del modelo se prueba de igual manera que en el caso de regresión lineal simple, pero además hay que ver que no hay colinealidad entre las variables predictoras. Esto lo podemos ver de la misma forma que vemos si hay linealidad entre carreras y bateo, solo que cambiando la variable.

```
ggplot(mlb11, aes(x = at_bats, y = homeruns)) +
  geom_point()
```

```
## Ejercicio 9

Vean el gráfico que hicimos arriba. ¿Parece haber alguna relación entre bateo y
homeruns?
```

Por su cuenta

Elijan una de las otras siete variables tradicionales de `mlb11` además de bateos que piensen que puede ser un buen predictor de entradas. Creen un diagrama de dispersión de las dos variables y ajusten un modelo lineal. A ojo, ¿parece lineal la relación?

¿Cómo se compara con la relación entre entradas y bateos? Usen el valor de R^2 de los resúmenes de los dos modelos para comparar. ¿La variable que elegiste parece predecir mejor las carreras que bateos? ¿Cómo te das cuenta?

Ahora que podes resumir la relación lineal entre dos variables, investiga la relación entre carreras y las otras cinco variables tradicionales. ¿Cuál de las variables predice mejor las carreras? Apoya tus conclusiones usando los métodos gráficos y numéricos que discutimos (por brevedad, solo incluye la salida de la mejor variable, no las cinco).

Ahora examina las tres nuevas variables. Estas son las estadísticas usados por el autor de Moneyball para predecir el éxito de un equipo. En general, ¿tienen mejor o peor efectividad prediciendo carreras que las viejas variables? Explica usando los métodos gráficos y numéricos apropiados. De las diez variables que analizamos ¿Cuál parece ser la mejor predictora de carreras? Usando la limitada información que sabes de estas estadísticas de béisbol ¿tienen sentido tus resultados?

Prueba los diagnósticos de modelo para la regresión con la variable que decidiste que es la mejor predictora de carreras.

Crea un modelo de regresión lineal múltiple entre las dos variables que mejor predicen los resultados.

Prueba los diagnósticos para este modelo y que no haya colinealidad entre las variables predictoras.