

ADVANCED ANALYTICS

Práctica 1

Regresión lineal

Descripción de las tareas a realizar

Regresión lineal: Teoría

En este apartado deberéis responder a algunas cuestiones teóricas que podréis resolver si os habéis leído los apuntes del tema. Los cálculos podéis realizarlos en R si queréis pero son tan sencillos que pueden realizarse a mano.

Pregunta 1.1 (0.7 pts): Supongamos que tenemos los puntos $p_1 = (0,2)$, $p_2 = (1,2)$ y $p_3 = (1,8)$ y queremos hacer una regresión lineal. ¿Cuál sería el valor constante para el modelo base de este conjunto de datos?

Pregunta 1.2 (0.7 pts): Supongamos que tenemos los puntos $p_1 = (0,2)$, $p_2 = (1,2)$ y $p_3 = (1,8)$ y queremos hacer una regresión lineal. Probamos con la recta $y = 3x + 2$. ¿Cuál es el valor para este modelo de la medida SSE?

Pregunta 1.3 (0.7 pts): Supongamos que tenemos los puntos $p_1 = (0,2)$, $p_2 = (1,2)$ y $p_3 = (1,8)$ y queremos hacer una regresión lineal. Probamos con la recta $y = 3x + 2$. ¿Cuál es el valor para este modelo de la medida SST?

Pregunta 1.4 (0.7 pts): Supongamos que tenemos los puntos $p_1 = (0,2)$, $p_2 = (1,2)$ y $p_3 = (1,8)$ y queremos hacer una regresión lineal. Probamos con la recta $y = 3x + 2$. ¿Cuál es el valor para este modelo de la medida R^2 ?

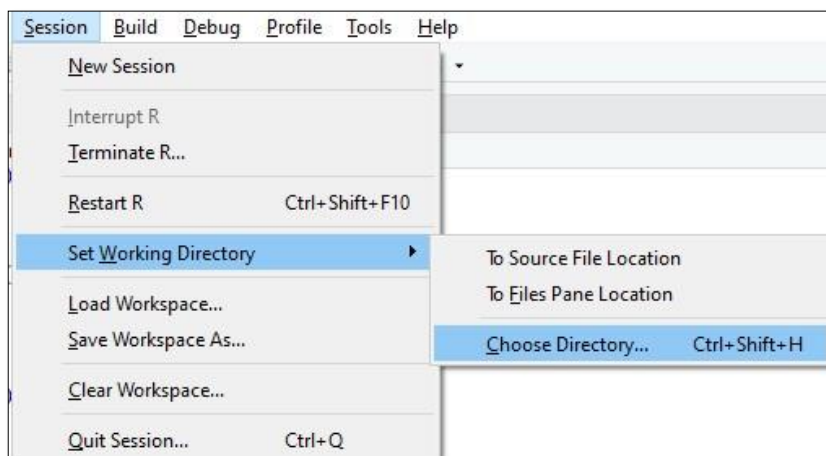
Pregunta 1.5 (0.7 pts): Para un conjunto de 180 registros y 20 variables, hemos conseguido un valor R^2 de 0.81 utilizando 18 de ellas. ¿Cuál es el valor de R^2 ajustado?

Caso práctico de regresión lineal: Moneyball

Del caso Moneyball se han escrito libros (Moneyball: The Art of Winning an Unfair Game, M.Lewis, 2003 es el más conocido) y se ha rodado una película en 2011 protagonizada por Brad Pitt. A principios de la década de 2000, Billy Beane y Paul DePodesta trabajaron para los Oakland Athletics, un equipo de béisbol con recursos económicos limitados. Sin embargo, con el uso de los datos llegaron a ser tan competitivos en la liga regular como otros equipos con un presupuesto mucho mayor. Hasta entonces la selección de jugadores la llevaban a cabo ojeadores que

evaluaban las características de un jugador y decidían si era bueno o no para el equipo. Beane y DePodesta empezaron a seleccionar jugadores en función de sus estadísticas. El conjunto de datos `moneyball.csv`, que podéis encontrar en los recursos del tema, nos servirá para comprender mejor sus métodos.

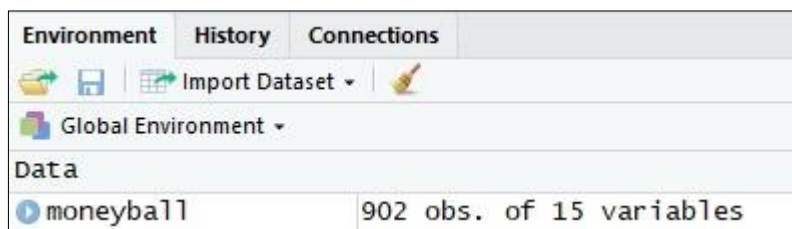
Para ello, antes de nada tenemos que establecer el directorio de trabajo en la ruta donde tengamos el fichero. Podemos hacerlo a través de las opciones de menú de RStudio.



Ahora, carguemos el conjunto de datos:

```
moneyball <- read.csv("baseball.csv")
```

Esta instrucción lee el fichero y lo asigna (`<-` es el operador de asignación y es equivalente a `=`) a una variable a la que hemos llamado `moneyball`. Veréis que os aparece en la sección de entorno de la aplicación



Si hacemos click sobre la variable podremos ver una descripción del conjunto de datos equivalente a utilizar la instrucción `str`

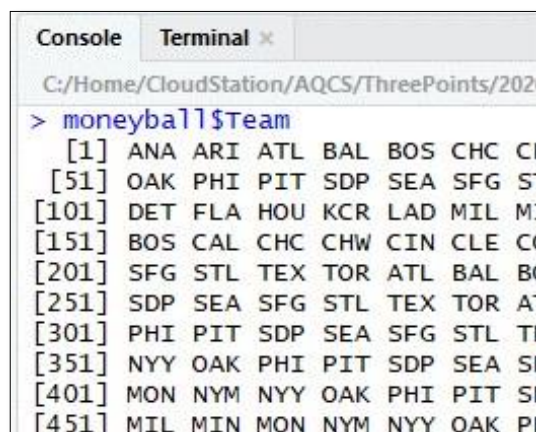
Pregunta 1.6 (0.7 pts): ¿Cuántas de las variables de nuestro conjunto de datos son de tipo 'Factor'? (Variables de texto que representan categorías) `str(moneyball)`

El dataset o conjunto de datos anterior consta de información estadística por temporada de los equipos de beisbol americanos de antes del 2002. Las variables son las siguientes.

- Team
- League
- Year
- Runs Scored (RS)
- Runs Allowed (RA)
- Wins (W)
- On-Base Percentage (OBP)
- Slugging Percentage (SLG)
- Batting Average (BA)
- Playoffs (binary)
- RankSeason
- RankPlayoffs
- Games Played (G)
- Opponent On-Base Percentage (OOBP)
- Opponent Slugging Percentage (OSLG)

Para acceder a una variable, podemos hacerlo utilizando `$` y el nombre de la variable. Por ejemplo, la siguiente instrucción nos listará todos los valores de la variable `Team`

```
moneyball$Team
```



```
Console Terminal x
C:/Home/CloudStation/AQCS/ThreePoints/202
> moneyball$Team
[1] ANA ARI ATL BAL BOS CHC CH
[51] OAK PHI PIT SDP SEA SFG ST
[101] DET FLA HOU KCR LAD MIL M
[151] BOS CAL CHC CHW CIN CLE CO
[201] SFG STL TEX TOR ATL BAL BO
[251] SDP SEA SFG STL TEX TOR AT
[301] PHI PIT SDP SEA SFG STL TE
[351] NYY OAK PHI PIT SDP SEA SI
[401] MON NYM NYY OAK PHI PIT SI
[451] MIL MIN MON NYM NYY OAK PI
```

La función `summary` nos proporciona un resumen de los estadísticos principales para cada variable. Si la cantidad de variables fuese muy grande podemos hacer el `summary` de una sola variable. De igual modo, podemos aplicar directamente ciertos estadísticos a nuestras variables, como por ejemplo la función `quantile`, que nos proporciona información sobre los cuantiles de nuestra variable numérica.

Pregunta 1.7 (0.7 pts): ¿Cuál es el valor del primer cuartil (25%) de la variable RS?

```
summary(moneyball) summary(moneyball$RS)  
quantile(moneyball$RS)
```

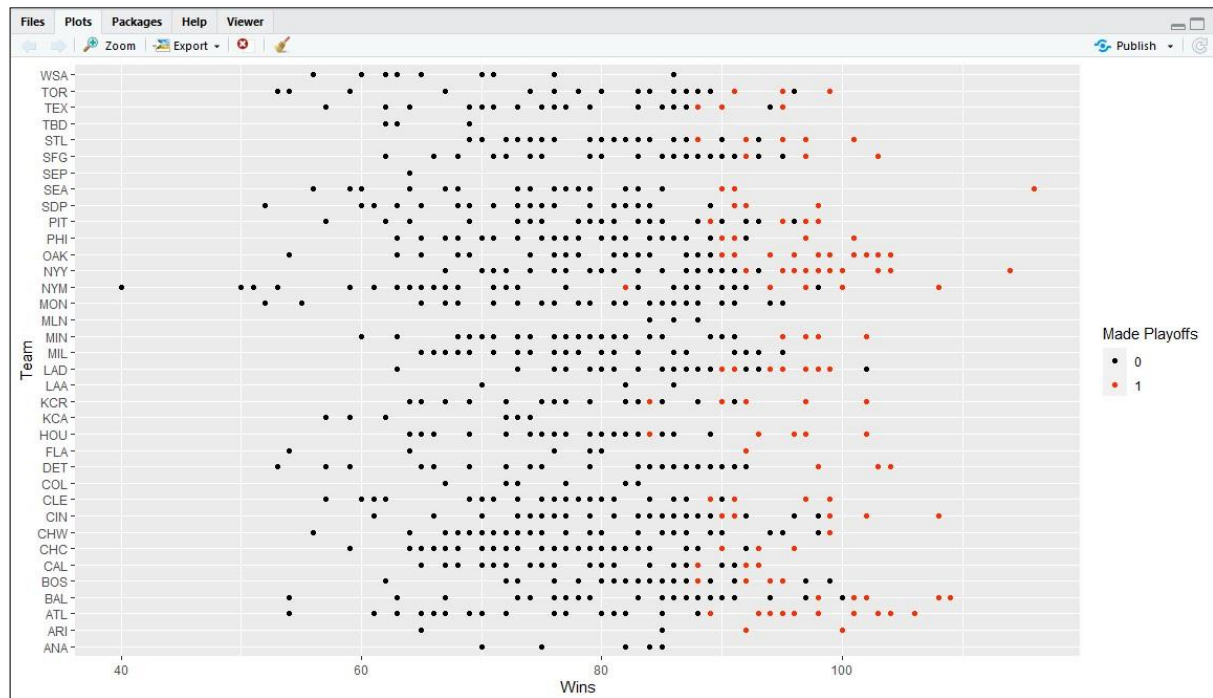
El objetivo de cualquier equipo de beisbol americano es ganar la final de las series mundiales a la cuál acceden los ganadores de los playoffs, a los cuáles acceden a su vez, los mejores de la liga regular (de hecho, hay varias ligas, pero no es relevante). Beane y DePodesta determinaron que su objetivo era llegar a los playoffs y que el método era ganar partidos. Pero eso no era suficiente. ¿Cuántos partidos hace falta ganar para llegar a los playoffs? Llegaron a la conclusión de que necesitaban ganar 95 o más partidos para conseguir entrar en los playoffs.

Gráficamente, podemos validar su conclusión con las siguientes instrucciones en R

```
install.packages("ggplot2") library(ggplot2)  
  
m <- ggplot(moneyball, aes(x = W, y = Team,color = factor(Playoffs)))+ geom_point() +  
scale_color_manual(values = c("#000000", "#FF2D00"), name = "Made Playoffs") m +  
xlab("Wins")
```

Las dos primeras líneas del código anterior instalan el paquete ggplot2 en vuestro entorno (no es necesario si ya lo tenéis) y lo cargan en el sistema de modo que sus funciones estén disponibles.

La tercera sentencia crea el siguiente diagrama de dispersión que veréis en la pestaña Plots de la sección que hemos llamado “Varios”. En él podemos ver para cada equipo (eje vertical), el número de victorias (eje horizontal) que tuvo cada año (puntos) marcando en rojo los años que el equipo accedió a los playoffs.



Con este gráfico parece evidente que obteniendo 95 victorias o más, la probabilidad de acceder a los playoffs es muy alta.

¿Y cómo conseguimos victorias? Un equipo gana un partido si marca más puntos o carreras (RS) de los que encaja (RA).

Creemos una variable “Runs Difference” con la siguiente instrucción.

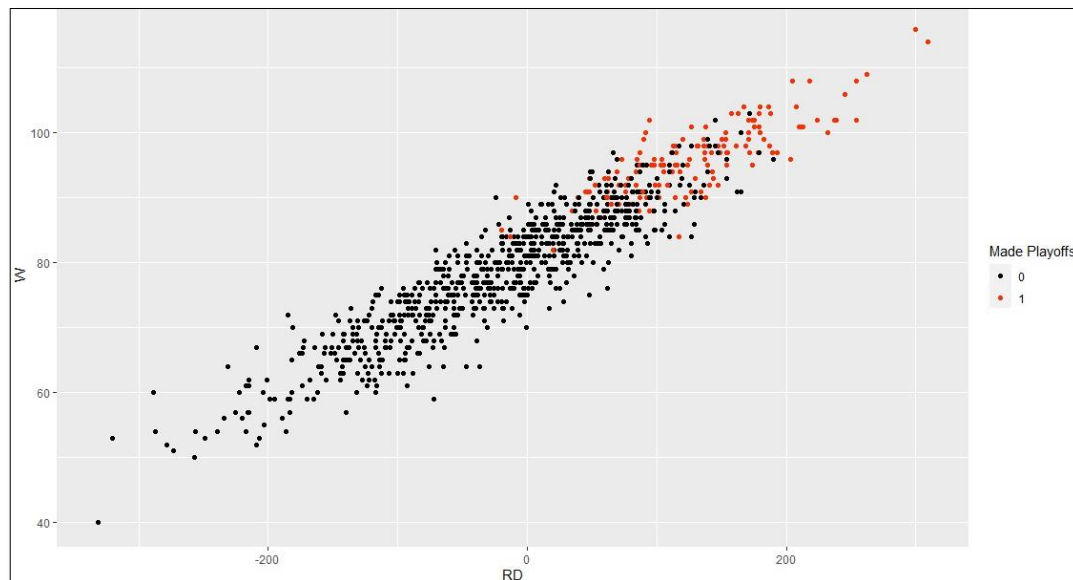
```
moneyball$RD = moneyball$RS - moneyball$RA
```

Pregunta 1.8 (0.7 pts): ¿En qué año se produjo la peor diferencia de carreras? ¿Es decir, en qué año el registro RD es menor?

```
moneyball[which.min(moneyball$RD), ]$Year
```

Con la siguiente instrucción en R podemos ver la gran correlación (bastante obvia, por otra parte) entra la variable que acabamos de crear RD y la variable que indica el número de victorias (W)

```
ggplot(moneyball, aes(x = RD, y = W, color = factor(Playoffs))) + geom_point() +  
  scale_color_manual(values = c("#000000", "#FF2D00"), name = "Made Playoffs")
```



Por lo tanto, parece obvio intentar hacer un modelo de regresión lineal que nos permita estimar la variable *W* a partir de la nueva variable *RD*.

Crear un modelo de regresión lineal al que llamaremos `modelW` utilizando la función de R `lm` con la variable objetivo o dependiente *W* y la variable independiente *RD* es tan sencillo como ejecutar el siguiente comando:

```
modelW <- lm(W ~ RD, data = moneyball)
```

Pregunta 1.9 (0.7 pts): ¿Cuál es el valor R^2 de este modelo?

```
summary(modelW)["r.squared"]
```

El modelo anterior es un modelo bastante robusto que nos permite predecir victorias (*W*) a partir de la diferencia de carreras (*RD*).

Pregunta 1.10 (0.7 pts): Si, por lo que vimos anteriormente, necesitamos 95 victorias para pasar a los playoffs, ¿cuántas carreras más debemos hacer de las que debemos permitir?

Es decir, si

$$W = a_0 + a_1 \cdot RD \text{ y}$$

$95 \leq W$, entonces, ¿qué valor mínimo de RD cumple la siguiente ecuación?

$$95 \leq a_0 + a_1 \cdot RD$$

(donde a_0 y a_1 son los coeficientes del modelo generado en el ejercicio anterior que podemos ver con `summary(modelW)`)

IMPORTANTE: el número de carreras debe ser un número entero por lo que la solución es el menor número entero que cumple la desigualdad anterior. Es decir, redondea hacia arriba.

Sabiendo la diferencia de carreras mínima que hay que obtener para llegar a las 95 carreras, la siguiente pregunta es, ¿cómo hago carreras (RS)? ¿y cómo impido que me las hagan (RA)? Veamos los datos que tenemos para predecir RS:

- OBP (On-Base Percentage) es el porcentaje de tiempo que un jugador llega a la base
- SLG (Slugging Percentage) es lo lejos (bases) que llega un jugador en su turno
- BA (Batting Average) es la media de bases conseguidas golpeando la pelota.

La mayoría de los equipos se basaban en BA pero Beane y DePodesta descubrieron lo siguiente

```
modelRS = lm(RS~OBP+SLG+BA, data=moneyball) summary(modelRS)
```

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-788.46	19.70	-40.029	< 2e-16	***
OBP	2917.42	110.47	26.410	< 2e-16	***
SLG	1637.93	45.99	35.612	< 2e-16	***
BA	-368.97	130.58	-2.826	0.00482	**

En el modelo, con un R^2 de 0.93, OBP y SLG son muy significativas y aunque BA también es significativa, tiene un valor negativo lo cual es anti intuitivo e indica un problema de multicolinealidad.

Pregunta 1.11 (0.75 pts): Si repetimos el mismo modelo “modelRS” eliminando la variable BA. ¿Qué valor de R^2 nos proporciona el nuevo modelo?

Del mismo modo, para la predicción de RA utilizaron las variables OOBP y OSLG. Crearemos el modelo “modelRA” RA utilizando esas variables.

Finalmente, para predecir si los Oakland A’s pasarían a los playoffs en el 2002, asumieron que los jugadores que iban a jugar esa temporada (algunos eran nuevos) iban a tener el mismo rendimiento que tuvieron el año pasado si las lesiones les respetaban. Con ello estimaron los siguientes parámetros para el equipo.

- OBP = 0.339
- SLG = 0.430
- OOBP = 0.307
- OSLG = 0.373

Pregunta 1.12 (0.75 pts): ¿Cuál es la predicción para RS del modelo modelRS (summary(modelRS)) si utilizamos los valores anteriores de OBP y SLG?

Pregunta 1.13 (0.75 pts): ¿Cuál es la predicción para RA del modelo modelRA (summary(modelRA)) si utilizamos los valores anteriores de OOBP y OSLG?

Pregunta 1.14 (0.75 pts): Predice finalmente con los parámetros del modelo W (summary(modelW)) y los valores que acabamos de calcular para RS y RA el número de victorias del equipo en 2002. ¿Cuál es ese número de victorias?

NOTA: La solución es el mayor número entero menor que la solución de la predicción. Es decir, redondea hacia abajo.

Supongo que para nadie será una sorpresa que ese valor es superior a las 95 victorias necesarias según los cálculos iniciales para pasar a los playoffs.