CI1031 Diseño de Experimentos I Ciclo 2025
Profesor: Ignacio Díaz Oreiro

Laboratorio de regresión lineal

Instrucciones

Deberá crear un reporte con los resultados del laboratorio. El reporte es un documento con un formato uniforme y coherente, en el que se incluye texto explicativo de la secuencia de acciones que componen el trabajo, incluyendo explicaciones de lo que se va realizando, el código R a ejecutar, los resultados y gráficos producto de ese código R, así como el análisis de los resultados y gráficos obtenidos. El reporte debe ser auto explicativo, de forma que pueda entenderlo una persona que no cuenta con el enunciado.

La **primera parte** del laboratorio es un poco más guiada, con instrucciones de acciones que deberá ejecutar en RStudio.

Estas instrucciones que se indicarán con la etiqueta **Pregunta x**, podrán ser respuestas breves de análisis, resultados que genere R ante las instrucciones dadas o gráficos creados en R.

En esta primera parte, aunque el código R esta provisto en el enunciado, inclúyalo en el informe para que se entienda el contexto de las respuestas y gráficos presentados. Además, debe incluir los resultados generados en R, gráficos y respuestas de análisis. Estos elementos deben aparecer de forma fluida en formato de informe. No es válido solo poner las preguntas y respuestas.

En la **segunda y tercera parte** el trabajo será más independiente, y en el que usted deberá desarrollar acciones para las que no se le dará código R. En el reporte a entregar, deberá incluir el código R que se ejecute (sea que se brinde en el enunciado o el que usted diseñe y construya), los gráficos y resultados obtenidos (por ejemplo, la salida de un summary( )) y el análisis respectivo. **Excepcionalmente**, en algunos puntos se le indicará cuando no sea necesario que incluya el código R, pero se le indicará de forma explícita. Nuevamente, el trabajo será guiado por preguntas identificadas con una etiqueta **Pregunta y**.

El trabajo puede entregarse en grupos de dos personas máximo.

Las entregas se realizarán en formato PDF. La letra debe ser Times New Roman, Arial o Cambria. Los tamaños permitidos son 10-12.

Debe incluir una portada de página completa en la que incluya el nombre del curso, del laboratorio, así como los nombres completos de los estudiantes. Esta página será utilizada por el profesor para la calificación y comentarios, en caso de tener que incluir comentarios generales del informe.

El reporte se debe entregar en Mediación Virtual, en un archivo .pdf que pueda ser leído en programas comerciales de uso habitual. Debe verificar que el .pdf que subió a Mediación Virtual contiene los ejercicios resueltos y que el archivo puede abrirse correctamente. En caso de

problemas con el archivo .pdf (no abre correctamente, está corrupto, etc.) se considerará que no entregó la tarea.

Las entregas tardías se penalizarán con un 10% de la nota luego de vencida la fecha y hora de entrega, más un 10% adicional por cada hora de retraso.

Problema

La película Moneyball se centra en la "búsqueda del secreto del éxito en el béisbol". Sigue a un equipo de bajo presupuesto, los Atléticos de Oakland, que creían que las estadísticas infrautilizadas, como la capacidad de un jugador para "embasarse" (llegar a una base), predecían mejor la capacidad de anotar carreras que las estadísticas típicas como jonrones, RBIs (carreras impulsadas) y promedio de bateo. Obtener jugadores que sobresalieran en estas estadísticas infrautilizadas resultó ser mucho más asequible para el equipo.

En este laboratorio, analizaremos los datos de los 30 equipos de las Grandes Ligas de Béisbol y examinaremos la relación lineal entre las carreras anotadas (variable **runs**) en una temporada y otras estadísticas de jugadores. Nuestro objetivo será resumir estas relaciones tanto gráfica como numéricamente para encontrar qué variable, si es que hay alguna, nos ayuda a predecir mejor las carreras anotadas por un equipo en una temporada.

Primera parte

Se deben cargar los datos de la temporada 2011 de la liga profesional de beisbol que se encuentran en el archivo **beisball.csv**. Puede usar el siguiente código o cargar los datos como considere mejor.

```
beis = (read.csv(file.choose(), header=T, encoding = "UTF-8"))
attach(beis)
```

Nota: Para el código anterior no hay salida que incluir en el reporte.

A partir de aquí los datos se referencian como **beis**.

Además de las carreras anotadas (runs), hay siete variables utilizadas tradicionalmente en el conjunto de datos: at-bats, hits, home runs, batting average, strikeouts, stolen bases, y wins (turnos al bate, hits, jonrones, promedio de bateo, ponches, bases robadas y victorias por lanzador).

También hay tres variables más nuevas: new_onbase, new_slug, new_obs (porcentaje de bateadores que llegan a base, porcentaje de slugging o potencia de bateadores y suma de new_onbase + new_slug), pero estas no las utilizaremos en este trabajo.

Se puede usar la función plot() para mostrar la relación entre la variable **runs** y una de las otras variables numéricas. Trace esta relación usando la variable **at_bats** como predictor.

```
plot(beis$at_bats, beis$runs)
```

Pregunta 1: ¿La relación parece lineal?

También se puede cuantificar la fuerza de la relación con el coeficiente de correlación.

```
cor(beis$runs, beis$at_bats)
```

Pregunta 2: ¿Qué tan fuerte es la correlación entre runs y at_bats?

Suma de residuos al cuadrado

Es útil poder describir la relación de dos variables numéricas, como runs y at_bats , antes mencionadas.

Podemos resumir la relación entre estas dos variables encontrando la línea que mejor sigue su asociación. Utilice la siguiente función interactiva para seleccionar la línea que cree que hace el mejor trabajo para atravesar la nube de puntos.

```
if(!require('statsr')) {
  install.packages('statsr')
  library('statsr')
}
plot_ss(x = at_bats, y = runs, data = beis)
```

Después de ejecutar este comando, se le pedirá que haga clic en dos puntos del gráfico para definir una línea. Una vez que haya hecho eso, la línea que especificó se mostrará en negro y los residuos en azul. Tenga en cuenta que hay 30 residuos, uno para cada una de las 30 observaciones.

Recuerde que los residuos son la diferencia entre los valores observados y los valores predichos por la línea:

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

La forma más común de hacer una regresión lineal es seleccionar la línea que minimiza la suma de los residuos al cuadrado.

Para visualizar los residuos cuadrados, puede volver a ejecutar el comando de trazado y agregar el argumento showSquares = TRUE.

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

Además del gráfico obtenido con plot_ss con showSquares = TRUE, tenga en cuenta que la salida de la función plot_ss le proporciona la pendiente y la intersección con el eje y, así como la suma de los cuadrados.

Pregunta 3. Usando plot_ss, elija una línea que minimice lo más posible la suma de los cuadrados. Para esto ejecute la función varias veces. ¿Cuál fue la menor suma de cuadrados que obtuvo? Para la suma de cuadrados más pequeña que encontró, incluya el gráfico resultante y también los resultados que aparecen en la consola, que presentan los coeficientes de la recta encontrada (pendiente e intersección con el eje y) y el resultado de la suma de cuadrados calculada.

Solo para referencia, la mínima suma de cuadrados que encuentra el modelo lineal (que calcularemos más adelante) es **123721.9** (¡es prácticamente imposible que logre un valor así eligiendo manualmente los puntos!).

El modelo lineal

Es bastante engorroso tratar de obtener la línea correcta de mínimos cuadrados, es decir, la línea que minimiza la suma de los cuadrados de los residuos, a través de prueba y error. En su lugar, se puede usar la función **lm** en R para ajustar el modelo lineal (también conocido como línea de regresión). La sintaxis en R sería:

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

El primer argumento en la función lm es una fórmula que toma la forma y ~ x. Aquí se puede leer que se desea hacer un modelo lineal de runs en función de at_bats (variable predictora).

La salida de lm es un objeto que contiene toda la información del modelo lineal que se acaba de ajustar. Podemos acceder a esta información mediante la función de summary().

```
summary(m1)
```

La salida del summary() consta de varias partes. Primero, la fórmula utilizada para describir el modelo se muestra en la parte superior. Después de la fórmula, aparece información de los residuales. La tabla de "Coeficientes" que se muestra a continuación es clave; su primera columna muestra la intersección y la pendiente del modelo lineal (el coeficiente de at_bats). Con esta tabla, se puede escribir la línea de regresión de mínimos cuadrados para el modelo lineal.

Pregunta 4. Escriba la **fórmula** del modelo lineal obtenido con los valores correctos de pendiente e intersección con el eje y.

La fórmula tendrá el formato siguiente: $\hat{y} = \textit{intersección} + \textit{pendiente} * \text{at\_bats}$.

Otro elemento importante del summary() es el Múltiple R-cuadrado, o simplemente, R^2. El valor R^2 representa la proporción de variabilidad en la variable de respuesta que es explicada por la variable predictora. Para este modelo, el 37.29% de la variabilidad en runs se explica por at_bats.

En regresión lineal simple, el **R2 es el cuadrado del coeficiente de correlación R**.

Finalmente, el **p-value** mostrado corresponde a todo el modelo de regresión lineal. Si el p-value es menor que el nivel de significancia (supongamos 0.05) quiere decir que el modelo de regresión lineal predice mejor los resultados que un modelo con solo intercept.

Ahora se creará un diagrama de dispersión agregando la línea de mínimos cuadrados.

```
plot(beis$runs ~ beis$at_bats)
abline(m1)
```

La función abline traza una línea basada en su pendiente e intersección. Aquí, se utiliza un atajo al proporcionar el modelo m1, que contiene ambas estimaciones de parámetros. Esta línea se puede usar para predecir "y" en cualquier valor de "x". Cuando se realizan predicciones para valores de x que están más allá del rango de los datos observados, se denomina extrapolación y, por lo general, no se recomienda. Sin embargo, las predicciones hechas dentro del rango de los datos son más confiables. También se utilizan para calcular los residuales.

Comprobación de supuestos

Para evaluar si el modelo lineal es confiable, se debe verificar: la relación lineal de ambas variables, la independencia de residuales, la normalidad de los residuales, y la homogeneidad de varianzas. Sin embargo, en este ejercicio no realizaremos esta validación, aunque en la Pregunta 1 se verificó si la relación entre runs y at_bats era lineal usando un diagrama de dispersión y también se calculó el coeficiente de correlación. En este ejercicio no es necesario hacer nada más, pero sí sería necesario en un experimento completo.

Segunda parte

Pregunta 5. Ajuste un nuevo modelo de regresión lineal que use **homeruns** para predecir runs en vez de **at_bats** (el anterior). Muestre el código R usado para crear el modelo y la salida summary() de ese modelo.

Pregunta 6. Usando los coeficientes del summary() anterior, escriba la **ecuación** del modelo (la línea de regresión).

Pregunta 7. ¿Qué nos dice El Múltiple R-Cuadrado de este modelo con homeruns en comparación con el modelo anterior (para at_bats)? ¿Cuál predice mejor los resultados?

Pregunta 8. Ahora que sabe analizar la relación lineal entre dos variables (creando modelos de regresión lineal), investigue las relaciones entre **runs** y cada una de las variables tradicionales (at_bats, hits, homeruns, bat_avg, strikeouts, stolen_bases, y wins). Ya ha generado la información para **at_bats** y **homeruns**.

Presente en una tabla la comparación de los valores R^2 de todas estas variables tradicionales que incluya además el valor de correlación y el valor-p de esa variable en el modelo lineal respectivo. Puede usar la tabla siguiente como ejemplo:

Tabla de comparación de variables predictoras de los modelos de regresión lineal simple

Variable predictora	Correlación	p-value del Modelo	R^2
at_bats			
hits			
homeruns			
bat_avg			
strikeouts			
stolen_bases			
wins			

NOTA: No es necesario que incluya el código R de los modelos que construya para hits, bat_avg, strikeouts, stolen_bases y wins. Tampoco es necesario que valide supuestos, aunque sí necesita calcular el coeficiente de correlación para incluirlo en la tabla.

Pregunta 9. ¿Cuál de las siete variables anteriores predice mejor la variable runs y por qué lo considera usted así?

Tercera parte

Ahora realizaremos un análisis de regresión lineal **múltiple**. Para ello crearemos un modelo de regresión lineal con cinco de las variables originales: at_bats, hits, homeruns, bat_avg, y wins.

```
mul <- lm(runs ~ at_bats + hits + homeruns + bat_avg + wins, data = beis)
summary(mul)
```

Pregunta 10. Con base en el resultado anterior, construya la **fórmula** del modelo de regresión lineal que corresponde a estas cinco variables. Recuerde que el formato es:

$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

Pregunta 11. Según los resultados, ¿el modelo de regresión lineal es significativo? ¿Cómo puede determinarlo?

En modelos de regresión lineal múltiple, para determinar la proporción de variabilidad en la variable de respuesta que es explicada por el conjunto de variables predictoras, usaremos el **Adjusted R-squared** (y no el es el Múltiple R-cuadrado, como se realizó en la regresión lineal simple), dado que se debe realizar el ajuste asociado al uso de más de una variable predictora.

Pregunta 12. Según los resultados del summary(mul), ¿cuánto porcentaje de la variabilidad de la variable de respuesta es atribuible al modelo de regresión lineal? ¿Es este modelo de regresión múltiple un mejor predictor que el mejor modelo de regresión lineal simple de la Pregunta 9?

Pregunta 13. Calcule el Factor de Inflación de la Varianza (VIF) para este modelo múltiple. Presente el código R y los resultados, e indique qué se puede concluir de este cálculo de VIF respecto de la multicolinealidad.

Ahora ajuste modelos de regresión lineal múltiple para combinaciones de **cuatro** variables de las cinco variables del modelo anterior (at_bats, hits, homeruns, bat_avg, y wins).

NOTA: Solo haga modelos con cuatro variables, **no** haga de tres variables ni de dos variables que, aunque se podrían hacer y tendría sentido, no son parte de este ejercicio.

Pregunta 14. Construya una tabla donde indique las combinaciones de las cuatro variables utilizadas, así como el valor-p del modelo completo, el R^2 ajustado y los VIF de esas 4 variables. Puede usar la siguiente tabla como ejemplo:

Tabla de comparación de combinaciones de cuatro variables predictoras de los modelos de regresión lineal múltiple

Variables utilizadas en el modelo de regresión lineal múltiple	**p-value del modelo**	**R^2 ajustado**	**VIF de esas variables**
at_bats, hits, homeruns, bat_avg			
at_bats, hits, homeruns, wins			
…			

NOTA: Únicamente incluya el código R para el modelo lineal de cuatro variables at_bats, hits, homeruns, y bat_avg (modelo lm con summary(), vif() asociado y los correspondientes resultados de R). Para los otros modelos que construya para cuatro variables no es necesario que incluya el código R, solamente que indique los valores en la tabla.

Pregunta 15. ¿Considera que alguno de los modelos de cuatro variables de la tabla anterior predice mejor la variable de respuesta que el modelo de cinco variables (analizados en la pregunta 11)? ¿Cuál considera que es el mejor? Debe tomar en cuante que el modelo sea estadísticamente significativo y que no haya multicolinealidad entre las variables.