

1. Ejercicios de Regresión Múltiple (segunda parte)

1. *Esperanza de vida, parte I.* El conjunto de datos `esperanza2015.txt` contiene 10 variables medidas en 193 países durante el año 2015. Los datos provienen del sitio *Gapminder*, <https://www.gapminder.org/data/>. Estas variables son:
 - **income**: Producto bruto interno por persona, ajustado por diferencias en el poder adquisitivo (en dólares internacionales, precio fijo de 2011, PPP basado en 2011 ICP). Fuente: Gapminder basado en World Bank, A. Maddison, M. Lindgren, IMF y otros. <http://gapm.io/dgdppc>
 - **child**: Número de muertes de niños menores de cinco años por cada 1000 nacidos. Fuente <http://gapm.io/>
 - **life**: (Esperanza de vida al nacer). Número promedio de años que un niño recién nacido viviría si los patrones de mortalidad actuales se mantuvieran iguales. Source: <http://gapm.io/ilex>
 - **gini**: El índice de Gini mide hasta qué punto la distribución del ingreso entre individuos u hogares dentro de una economía se desvía respecto de la distribución perfectamente igualitaria. Un índice de Gini de 0 representa la igualdad perfecta, mientras que un índice de 100 implica desigualdad perfecta. Fuente: <http://data.worldbank.org/indicator/SI.POV.GINI>
 - **energy**: El uso de energía se refiere al uso de energía primaria antes de ser transformada a otros combustibles de uso final, que es igual a la producción local más importaciones y cambios de existencias, menos exportaciones y combustibles suministrados a barcos y aeronaves dedicadas al transporte internacional. Fuente: <http://data.worldbank.org/indicator/EG.USE.PCAP.KG.OE>
 - **dtp3**: Porcentaje de niños de 1 año que han recibido las tres dosis de la vacuna combinada contra la difteria, el tetános y la tos ferina durante el año. Estos datos son del año 2010. Fuente: <http://data.unicef.org/child-health/immunization.html>
 - **school**: Años de educación recibidos, en promedio, por los hombres entre 25 y 34 años, incluyendo educación primaria, secundaria y terciaria. Fuente <https://www.healthmetricsandevaluation.org/>
 - **hdi**: El Índice de Desarrollo Humano es un índice utilizado para clasificar países por nivel de “desarrollo humano”. Contiene tres dimensiones: nivel de salud, nivel educativo y nivel de vida. Combina: esperanza de vida al nacer, años esperados de escolaridad, años promedio de escolaridad y GINI per cápita. Fuente: <http://hdrstats.undp.org/en/indicators/103106.html>
 - **status**: división arbitraria en categorías binarias: “no desarrollado”, si **hdi** está por debajo de 0.75; “developed” otherwise.

El objetivo de este ejercicio es encontrar buenos predictores de la esperanza de vida (media) de un país usando las demás covariables como predictoras. Como los índices de equidad (**gini** y **hdi**) se calculan usando la esperanza de vida, no los utilizaremos en el análisis. Sí usaremos la variable **status**, (que en esta base de datos se construyó a partir de **hdi**, pero podría haberse obtenido de otra forma).

- a) Cargue los datos al R. Observe que algunas variables tienen muchos datos faltantes. Realice scatter plots de las diversas covariables versus **life**: **income**, **child**, **energy**, **gini**, **school**, **hdi**. ¿Cuáles aparecen relacionadas con **life**? Haga un boxplot de **life** versus **status**. Calcule las correlaciones entre las variables. La instrucción del R: `cor(esperanza2015[,c(2:9)],use="pairwise.complete.obs")` permite manejar los datos faltantes y utilizar la mayor cantidad de observaciones disponibles para calcular cada covariable.
- b) Ajuste el modelo lineal simple con **life** como variable a explicar y **child** como covariable (ya que es la que mayor correlación tiene con la variable **life**, exceptuando los índices).

- (i) Evalúe el ajuste obtenido. ¿Tiene sentido el signo del coeficiente de `child` estimado?
 - (ii) Realice el gráfico de residuos versus predichos. ¿Parecen cumplirse los supuestos del modelo lineal para este modelo?
- c) Ajustemos el modelo lineal múltiple con `life` como variable a explicar a través de un polinomio de grado dos en `child`.
- (i) Ajuste el modelo con `child` y `child`² como covariables. Puede ser útil la instrucción `ajuste1bis <- lm(life ~ child + I(child^2))`.
 - (ii) ¿Cómo se compara con la instrucción que le pide al R que arme el polinomio? `aju2 <- lm(life ~ poly(child,2))`. Compare los `summary` de ambos. Calcule los `vif` de las covariables, para estudiar la multicolinealidad. Usar el paquete `rms`, cuya instrucción `vif` funciona en combinación con `poly`. Hacerlo tanto para el modelo ajustado en (i) como en (ii). Hacer un gráfico de residuos versus valores ajustados, para ver si mejoró el ajuste desde el punto de vista de los residuos.
 - (iii) Otra opción para evitar la multicolinealidad consiste en centrar a la covariable primero. Cree la variable `child.cen` centrando a la variable original (restandole la media muestral y dividiéndola por su desvío estándar. Esto también podría hacerse usando estimadores robustos para el centro y la dispersión). Luego ajuste un modelo como en (i) pero con potencias de grado uno y dos de `child.cen`. Compare el `summary` de este modelo ajustado con los dos anteriores. Calcule los `vif` de las covariables.
 - (iv) Compare las predicciones que pueden hacerse con los tres modelos para distintos valores de la covariable `child`. Por ejemplo para `child = 10, 50, 100, 120, 137`. Puede ser útil la siguiente instrucción, `predict(aju2, newdata = data.frame(child = c(10, 50, 100, 120, 137)))`.
 - (v) Finalmente, hacer un scatterplot de los datos. Luego, superponer las curvas ajustadas: la recta y la cuadrática. Esta última, si bien ajuste mejor a los datos, tiene un problema de interpretabilidad, para los valores más altos de la covariable `child`. ¿Por qué? Probar con un polinomio de grado tres, superponiendo la curva al gráfico y también evaluando la significatividad estadística del modelo obtenido. Asegúrese que la falta de significatividad de este modelo no se deba a la presencia de observaciones atípicas comparandolo con el ajuste robusto de un polinomio de grado 3 usando la instrucción `lmRob` del paquete `robust`. Para hacer el gráfico, puede resultar bueno recordar las siguientes instrucciones. ¿Qué guarda el vector `ajuste1bis$na.action`?
- ```
plot(child, life)
points(child[-ajuste1bis$na.action], fitted.values(ajuste1bis), col= "red")
points(child[-ajuste1bisbis$na.action], fitted.values(ajuste1bisbis), col="blue")
```

2. *Esperanza, parte II.* Mejoremos el ajuste incluyendo otras covariables al modelo. Tal vez la esperanza de vida no dependa solamente de las condiciones de la primera infancia.

- a) Ajustar un modelo para explicar a `life` usando las covariables `child`, `income`, `status` y `dtp3.2010`. Evalúe la significatividad de cada coeficiente. Compare con el modelo que solo tiene a `child` como covariable (no el polinomial), ¿debería usar  $R^2$  o con  $R^2_{\text{ajustado}}$  para compararlos?
- b) Si alguna variable no resulta significativa en el ajuste anterior, pruebe el modelo que no la tiene. Compare el ajuste obtenido con el del ítem anterior. Evalúe si son razonables los signos de los estimadores obtenidos. ¿Cómo interpreta el coeficiente que acompaña a `status`? ¿Qué significa? ¿Cuánto aumenta (o disminuye) la esperanza de vida media en un país por ser desarrollado cuando se comparan países con igual ingreso y tasa de mortalidad de 0 a 5 años? ¿Es coherente con lo que uno intuye respecto de la esperanza de vida de los países desarrollados?

- c) Haga un gráfico de residuos versus valores predichos. Lo que observa, ¿es razonable? Y el Q-Q plot de los residuos, ¿es compatible con el supuesto de normalidad de los residuos? Descartemos multicolinealidad: calcular los vif. ¿Hay puntos influyentes que puedan estar cambiando las conclusiones? Comparar con un ajuste robusto, dado por `lmRob`.
- d) Probar agregar `hdi` al modelo anterior (o sea, covariables: `child`, `income`, `status` y `hdi`). ¿Qué pasa con el ajuste? Calcular los vif. En virtud de la descripción de las variables hecha en el ejercicio anterior, es razonable lo que sucede con el ajuste? Dijimos que la esperanza de vida se usa para construir al `hdi`, por eso no es razonable usarla como covariable para explicar la esperanza.
- e) Ajustar el modelo con covariables: `child`, `income`, `status` y `school`. Estudiar la significatividad. ¿Tiene sentido agregarle al modelo ajustado en 2b la covariable `school`?
- f) Para terminar de convencernos de que el modelo ajustado en 2b es un buen modelo, grafique los residuos de ese ajuste versus las covariables excluidas: `hdi`, `school`, `gini`, `dtp3.2010`. ¿Observa alguna estructura en alguno de ellos? Es razonable pensar en incluir `gini` como covariable, pero esto tiene dos problemas. El primero es que la esperanza de vida se usa para definir el índice Gini. El segundo es que este índice está sólo disponible en 57 países, en vez de los 193 países en los cuales está medida la esperanza de vida, y reduciría mucho el tamaño de la muestra sobre la cual basamos nuestras conclusiones. De todos modos, ajústelo como ejercicio y compare con el modelo ajustado en 2b.
- g) Nos quedamos finalmente con el modelo 2b, que explica a `life` con `child`, `income` y `status`.
3. *Discriminación de género, Parte I.* (Del Libro Weisberg [2005]) Los datos contenidos en el archivo `salary.txt` incluyen el salario y otras características de todos los profesores en una pequeña universidad del medio oeste estadounidense, recolectado principios de los años 1980. En el dataset se incluyen todos los profesores titulares o catedráticos (*tenured or tenure track positions*) de la universidad al momento de la recolección de datos, se excluyen a los temporarios. Los datos se recogieron de archivos de personal y consisten en las variables descritas en la tabla más abajo. El objetivo es estudiar esta base de datos para ver si había discriminación por género en esta universidad (en aquel momento).

| Variable            | Descripción                                                                                             |
|---------------------|---------------------------------------------------------------------------------------------------------|
| <code>Sex</code>    | 1 para mujeres, 0 para los hombres                                                                      |
| <code>Rank</code>   | 1 si es <i>Assistant Professor</i> , 2 para <i>Associate Professor</i> , y 3 para <i>Full Professor</i> |
| <code>Year</code>   | Número de años en el rango actual                                                                       |
| <code>Degree</code> | Grado más alto obtenido, 1 si Doctor, 0 si Magíster                                                     |
| <code>YSdeg</code>  | Número de años desde que obtuvo el título más alto                                                      |
| <code>Salary</code> | Salario académico anual en dólares                                                                      |

- (a) Plotee un resumen gráfico de los datos, y comente lo encontrado. Incluya boxplots de `Salary` separando por `Sex`, `Salary` separando por `Rank`, y `Salary` separando por los distintos niveles de `Sex` y `Rank` combinados. Lo mismo con `Degree`. Las instrucciones que siguen pueden ser útiles.
- ```
boxplot(Salary ~ Sex)
boxplot(Salary ~ Sex + Rank)
```
- Además, realice scatterplots de las variables continuas, coloreando las observaciones de varones y mujeres en colores distintos.
- (b) Testee la hipótesis de que el salario medio de hombres y mujeres es el mismo. ¿Qué hipótesis alternativa cree que es apropiada? ¿Qué test es razonable usar? ¿Qué test realiza la instrucción `lm` del R? ¿Cuál es la conclusión?
- (c) Testear la hipótesis de que el salario esperado, ajustado por los años en el rango actual, el título más alto obtenido y el número de años desde que la obtención de dicho título, es el mismo para cada

uno de los tres rangos, frente a la alternativa de que los salarios esperados no lo son. Testee para ver si el diferencial de salario debido al sexo es el mismo en cada rango. Recuerde que **Rank** debe ser considerada una variable cualitativa (**factor** en términos del R) para responder correctamente a esta pregunta. Observe que para responder a esta pregunta debe proponer un modelo lineal que tenga (dummies para codificar a) **Rank**, **YSdeg** y **Year**. ¿Puede responder a la pregunta con la salida del **summary** del **lm** directamente? ¿O necesita hacer comparaciones entre grupos a posteriori?

- (d) Para pensar. Si la discriminación se produce en la universidad a la hora de promover la planta docente a rangos más altos, ¿es correcto controlar por rango en el ajuste de regresión lineal, antes de separar por sexos? Ajustar dos funciones de regresión, una que incluya **Sex**, **Year**, **YSdeg** y **Degree**, y en la segunda agregue **Rank**. Resumir y comparar los resultados que se obtienen al dejar de lado los efectos de rango en las inferencias sobre diferencias por sexo en el pago de salario.

4. *Discriminación de género, Parte II.* Usando los datos de salario del dataset *salary.txt*, corrobore que la esperanza condicional del salario ajustada para el modelo que tiene a **Sex**, **Year** y la interacción entre ambos es

$$E(\widehat{Salary|Sex, Year}) = 18223 - 571Sex + 741Year + 169Sex \times Year$$

- (a) Dar los coeficientes del salario medio estimados si la variable **Sex** fuera codificada con el valor 2 para los hombre y el 1 para mujeres.
 (b) Proporcione los coeficientes si la codificación de sexo fuera -1 para hombres y +1 para mujeres.

5. *Discriminación de género, Parte III.* Usando los datos de salario del dataset *salary.txt*,

- a) ajuste el modelo lineal que explica el salario medio con la covariable **Sex**. Verifique que el modelo ajustado resulta ser

$$E(\widehat{Salary|Sex}) = 24697 - 3340Sex$$

donde la variable **Sex** está codificada como uno si el académico es mujer y cero si es hombre. Dé una frase que describa el significado de los dos coeficientes estimados. Según este modelo, en promedio, ¿ganaban más los hombres o las mujeres?

- b) Un modelo alternativo para estos datos consiste en agregar al modelo anterior la variable explicativa **Year**, el número de años que el empleado lleva trabajando en esta institución. Ajuste ese modelo y verifique que la ecuación obtenida es

$$E(\widehat{Salary|Sex, Year}) = 18065 + 201Sex + 759Years$$

La diferencia importante entre estos dos ajustes es que el coeficiente de **Sex** cambió de signo. Explique cómo pudo pasar esto.

6. *Berkeley Guidance Study.* El *Berkeley Guidance Study* incluyó a niños nacidos en Berkeley, California, entre enero de 1928 y junio de 1929, y luego los midió periódicamente hasta los dieciocho años (Tuddenham & Snyder, 1954). Aparecen en el libro tratados en el libro Weisberg [2005], en el archivo *BGSall.txt*. Las variables registradas aparecen descriptas en la tabla que sigue.

- a) Queremos explicar la altura a los 18, **HT18**, a partir de **HT9** y el factor **Sex**. Realice un scatterplot de **HT18** versus **HT9**, usando un símbolo (o color) distinto para varones y mujeres. Comente la información que el gráfico proporciona sobre la función de regresión adecuada para estos datos. ¿Le parece mejor ajustar una recta que explique a **HT18** por **HT9** por sexo, o una común? ¿O le parece que el vínculo entre ambas no es lineal?

Variable	Descripción
Sex	0 para varones, 1 para mujeres
WT2	Peso (<i>weight</i>) en kg, a los 2 años
HT2	Altura (<i>height</i>) en cm, a los 2 años
WT9	Peso (<i>weight</i>) en kg, a los 9 años
HT9	Altura (<i>height</i>) en cm., a los 9 años
LG9	Circunferencia de pierna en cm., a los 9 años (<i>leg</i>)
ST9	Fuerza en kg., a los 9 años (<i>strength</i>)
WT18	Peso (<i>weight</i>) en kg. a los 18 años
HT18	Altura (<i>height</i>) en cm., a los 18 años
LG18	Circunferencia de pierna en cm, a los 18 años (<i>leg</i>)
ST18	Fuerza en kg, a los 18 años (<i>strength</i>)
Soma	<i>Somatotype</i> , una escala de tipo corporal, desde 1, muy delgado a 7, obeso

- b) Ajuste un modelo con las covariables HT9, el factor **Sex** y la interacción entre ambos. Este modelo, ¿qué tipo de vínculo propone entre las dos pendientes? ¿Y entre las dos ordenadas al origen? Comente el ajuste. ¿Resulta significativo el coeficiente de la interacción? ¿Qué significa esto en términos de las rectas ajustadas? Superponga al gráfico anterior la recta ajustada para cada sexo, de colores distintos. La función **abline** del R, puede ser útil.
- c) Ajuste a las 66 observaciones correspondientes a los varones un modelo lineal simple, para explicar a HT18, a partir de HT9. Ajuste el mismo modelo a las 70 observaciones correspondientes a las mujeres. Compare los coeficientes ajustados por estos dos modelos lineales simples con el modelo ajustado en el ítem anterior. Compare también las significativades de los mismos. Observe que con estos dos modelos lineales simples no puede evaluarse la significatividad estadística de la igualdad entre pendientes. ¿Cómo la evalúa en el modelo con interacciones? Compare el desvío estandar del error estimado en los tres modelos.
- d) Ajuste un modelo con las covariables HT9, el factor **Sex** sin interacción. Este modelo, ¿qué tipo de vínculo propone entre las dos pendientes? ¿Y entre las dos ordenadas al origen? Comente el ajuste obtenido. ¿Qué significa esto en términos de las rectas ajustadas? Superponga al scatterplot de los datos realizado en el primer ítem, la recta ajustada para cada sexo, de colores distintos.
- e) Usando la función **vif** del paquete **rms** del R para calcular los vif de los modelos con y sin interacción. ¿Cuál modelo prefiere, finalmente?
7. En el archivo **exámenes.txt** aparecen las notas del examen final y las notas de dos exámenes preliminares para 22 estudiantes de un curso estadístico.
- a) Realice un scatterplot de las 3 variables. ¿Parecerían ser las notas preliminares buenas predictoras de la nota final?
- b) Ajuste los dos modelos lineales simples para explicar la nota final a partir de las notas parciales. Comente los ajustes obtenidos en términos de la significatividad de las pendientes y el R^2 obtenido. ¿Si tuviera que elegir uno, con cuál de los dos ajustes simples se quedaría? Llamemos a ese modelo A.
- c) Grafique los residuos del ajuste del modelo A versus las notas que **no** fueron incluidas en el modelo A. ¿Observa un vínculo entre ellas? Si agrega esa covariable al modelo A, espera que resulte significativa?
- d) Ajuste el modelo de regresión lineal múltiple para explicar la nota final a partir de las notas parciales. ¿Coincide con lo anticipado? ¿Podría haber anticipado esto a partir de los scatterplots del primer ítem? Calcule la correlación entre las dos notas preliminares.

- e) La base de datos tiene una covariable categórica que se denomina “**econ**”, que es una dummy indicativa de que la carrera en la que está inscripto el estudiante es economía. Hacer un gráfico de las notas finales versus la covariable explicativa incluida en el modelo A, pintando de color azul a las observaciones correspondientes a los estudiantes de economía y en rojo a las restantes. Ajuste el modelo múltiple que tiene la nota preliminar 1 y **econ** como explicativas, en forma aditiva. Evalúe la significatividad obtenida. ¿Resulta la variable significativa? Superponga al modelo las dos rectas ajustadas, pintándolas de los colores respectivos.
- f) ¿Será el modelo aditivo adecuado, o habrá que poner un modelo con interacción entre **parc1** y **econ**? Ajuste este último. Evalúe la significatividad obtenida. ¿Resulta la interacción significativa? Compare con el ajuste del modelo aditivo a través del R^2 , para ello, ¿conviene usar el R^2 o el $R^2_{ajustado}$? Calcule la correlación entre la interacción **parc1** · **econ** y las dos covariables **parc1** y **econ**. ¿Explica esto la significatividad obtenida en el modelo múltiple?
8. *Outliers, influencia y ajuste robusto.* (Basado en un ejercicio del libro Chatterjee y Hadi [2015]). En un estudio de 1976 que explora la relación entre la calidad del agua y el uso de la tierra, Haith (1976) obtuvo las mediciones en 20 cuencas fluviales en Nueva Estado de York que figuran en el archivo **nyrivers.txt**. Una pregunta de interés aquí es cómo el uso de la tierra alrededor de una cuenca fluvial contribuye a la contaminación del agua medida por la concentración media de nitrógeno (**Mghter**). El archivo contiene las siguientes 5 variables medidas en 22 ríos,
- $Y = \text{nitro}$, concentración media de nitrógeno (mg / litro) basada en muestras tomadas a intervalos regulares durante los meses de primavera, verano y otoño.
 - $X_1 = \text{agric}$, porcentaje del área de tierra actualmente en uso agrícola
 - $X_2 = \text{forest}$, porcentaje de tierra forestal
 - $X_3 = \text{residential}$, porcentaje del área de tierra en uso residencial
 - $X_4 = \text{comm}$, porcentaje del área de tierra en uso comercial o industrial
- a) Ajustar el modelo lineal simple que explica a $Y = \text{nitro}$ con la covariable **comm**. Observar si el ajuste resulta significativo. Hacer los gráficos que corresponden al **lm**. Observar que no parece haber ningún problema con el gráfico de residuos versus valores ajustados ni con el **qqplot** de los residuos. No parece haber outliers en estos datos.
- b) Repetir el ítem anterior con la rutina **lmrobdetMM** del paquete **RobStatTM** que ajusta un MM-estimador robusto. Comparar los ajustes obtenidos a través de los valores estimados para los coeficientes. Observar los pesos que el MM-estimador calcula para cada observación y que quedan guardados en el vector **rweights** del objeto **summary(lmrobdetMM)**. Hacer el plot asociado al **lmrobdetMM**. Esta rutina, ¿permite identificar observaciones de alto leverage o influencia (calculadas robustamente)? ¿Hay outliers?
- c) Hacer un scatter plot de los datos y superponer la recta ajustada por mínimos cuadrados. Observar que la observación 5, correspondiente al río Hackensack no tiene alto residuo porque ocurrió el fenómeno de enmascaramiento, la recta de mínimos cuadrados se vio arrastrada por las observaciones 4 y 5 para acercarse a ellas. La observación 5 es de alto leverage, o influencia. Examinando los datos en bruto del archivo **nyrivers.txt**, uno puede identificar fácilmente el río Hackensack, porque tiene un valor inusualmente grande para X_3 (porcentaje de tierra residencial) en relación con el otros valores para X_3 . La razón por la que tiene este valor tan grande es que el río Hackensack es el único río urbano en los datos debido a su proximidad geográfica a la ciudad de Nueva York, con su alta densidad de población. Los otros ríos están en zonas rurales. Aunque el río Neversink es influyente, uno no identifica que responda a un patrón diferente de los demás ríos mirando los datos sin procesar.

d) Ajustar el modelo

$$Y = \beta_0 + \beta_{forest}forest + \beta_{comm}comm + \epsilon \quad (1)$$

con ambos métodos y comparar el ajuste. ¿Coinciden? ¿Con cuál se quedaría?

e) Si uno solamente ajustara el modelo lineal por mínimos cuadrados, ¿se le podría agregar al modelo (1) alguna de las dos covariables restantes? ¿Saca usted la misma conclusión si utiliza el ajuste robusto? ¿Qué modelo, y qué método de estimación propone usted para los datos? Comparar los ajustes de mínimos cuadrados con el de MM-estimadores de regresión del modelo seleccionado, comparando los valores de los coeficientes ajustados bajo una y otra estrategia de estimación.

9. *Simulación de multicolinealidad* (script `correlacionados.R`). Este es un nuevo ejercicio de simulación, en el que nos enfocaremos en evaluar el impacto de la colinealidad de las covariables.

a) Fije la semilla, con la instrucción `set.seed(456789)`. Genere los siguientes vectores de longitud 50 normales independientes.

```
u1<-rnorm(50)
u2<-rnorm(50)
u3<-rnorm(50)
u4<-rnorm(50)
error<-rnorm(50)
error2<-rnorm(50,sd=0.1)
```

A partir de ellos armaremos las covariables, X_1, \dots, X_5 y la variable respuesta y .

```
x1<-u1
x2<-u2
aa<-c(1,2,3,0)
bb<-c(-1,2,0.5,3)

dd<-c(3,4,5,6)
x3<-aa[1]*u1+aa[2]*u2+error2
x4<-bb[1]*u1+bb[2]*u2+bb[3]*u3+bb[4]*u4
x5<-u1 + 4*u2
y<-dd[1]*x1+dd[2]*x2+dd[3]*x3+dd[4]*x4+error/0.1
```

Calculemos las covarianzas entre ellas. ¿Están muy correlacionadas de a pares?

```
cor(cbind(x1,x2,x3,x4,x5,y))
```

b) Ajuste el modelo lineal que explica a y por x_1 , x_2 y x_5 . Comente el ajuste. ¿Por qué el R estima al coeficiente β_5 por NA? Calcule el rango de la matriz de diseño del problema, que tiene una primer columna de unos y luego a las tres covariables. Se realiza con la instrucción de R:

```
modelo1<-lm(y ~ x1+x2+x5)
summary(modelo1)
X<-model.matrix(modelo1)
dim(X)    #dimension de la matriz
qr(X)$rank #rango, es decir, numero de columnas lin indep
```

O bien podemos calcular los autovalores de la matriz $X^t X$ con

```
eigen(t(X)%*%X)$values #calculamos los autovalores de t(X)*X
```

Vemos que tiene 3 autovalores no nulos y el cuarto vale cero, indicando que las 4 columnas son linealmente dependientes. De hecho, si miramos la forma en la que construimos x_1 , x_2 y x_5 es fácil ver que satisfacen:

$$x_5 = x_1 + 4 x_2,$$

es decir, que son un conjunto de tres variables linealmente dependientes y no se puede ajustar un modelo que tenga a las tres como covariables. Observemos que no nos podemos dar cuenta de esta dependencia solamente mirando las covarianzas o correlaciones de a pares.

- c) Si no fuera por el **error2** que aparece sumando en la definición de x_3 , las variables x_1 , x_2 y x_3 también serían linealmente dependientes. No lo son, pero están muy correlacionadas entre sí. Para constatarlo, calcule el R^2 del modelo lineal que explica a x_3 linealmente a partir de x_1 y x_2 . Compare el **Residual standard error** con el **sd** que pusimos para definir al **error2**. Pruebe modificar el desvío estándar del **error2** agrandándolo y vea si se refleja esta modificación en la estimación del **Residual standard error** del modelo estimado correspondiente. Haga lo mismo con x_4 en lugar de x_3 , calculando el R^2 .
- d) Del ítem anterior, sabemos que las covariables están muy correlacionadas. Ajuste el modelo lineal que tiene a y como respuesta y x_1 , x_2 , x_3 y x_4 .

```
ajusA<-lm(y ~ x1+x2+x3+x4)
summary(ajusA)
```

Calcule los VIF del modelo **ajusA** con la instrucción **vif** de la librería **rms**. Observe que la matriz de diseño de este ajuste tiene rango 5, como la cantidad de columnas, pero que los autovalores de la misma son muy distintos (el cociente entre el mayor y el menor, lo que se conoce como el número de condición de la matriz, es mayor a 10000). Calcule el promedio de los VIF. Mire el **summary** del ajuste obtenido. Evalúe la significatividad de los coeficientes y su desvío estándar estimado.

- e) Ajuste el modelo lineal que tiene a y como respuesta y x_1 , x_2 , y x_4 .

```
ajusB<-lm(y ~ x1+x2+x4)
summary(ajusB)
```

Calcule los VIF del modelo **ajusB** con la instrucción **vif** de la librería **rms**. Compare con lo obtenido en el ítem anterior. Calcule los autovalores de la matriz de diseño en este caso, y el cociente entre el menor y el mayor. Mire el **summary** del ajuste obtenido. Evalúe la significatividad de los coeficientes y su desvío estándar estimado.

- f) El objetivo de este ítem es probar que la multicolinealidad de las covariables no tiene ningún impacto sobre los valores predichos. Para eso vamos a predecir las observaciones bajo los dos modelos, **ajusA** y **ajusB**, y luego calculamos la suma de cuadrados de las diferencias entre y y los predichos por el **ajusA**. Del mismo modo, lo hacemos con el modelo **ajusB**.

```
predichosA<-predict(ajusA)
sum((predichosA-y)^2)
```

```
predichosB<-predict(ajusB)
sum((predichosB-y)^2)
```

Para hacer una comparación que no sea especialmente ingenua, generamos nuevas variables y y covariables bajo el mismo esquema descrito en el punto (a). Luego comparamos los valores predichos con los dos modelos ajustados la primera vez. Tomamos la suma de los cuadrados de los residuos para hacer esta comparación. Defina los vectores **sumaA** y **sumaB** inicialmente compuestos de 1000 ceros. Luego, con un bucle **for**, repite con un contador i que se mueva entre 1 y 1000 lo siguiente en cada iteración

- genere 50 nuevos datos en las mismas condiciones que los originales, para y , x_1 , x_2 , x_3 y x_4

- para estos nuevos datos, calcule los valores predichos por el modeloA y el modeloB
- calcule la suma de los cuadrados de los residuos obtenida para cada uno de los dos modelos con estos nuevos datos, y guárdela en la coordenada i -ésima de **sumaA** y **sumaB**.

Así obtendrá dos vectores **sumaA** y **sumaB** de longitud 1000. Comparar los valores obtenidos y guardados en **sumaA** con los valores guardados en **sumaB**. Concluir si las predicciones bajo ambos modelos son (o no) muy distintas entre sí, por ejemplo realizando un boxplot de ambos vectores en la misma escala o contando cuántas de las coordenadas de **sumaA** son mayores que las respectivas coordenadas de **sumaB**. El siguiente código en R puede ser útil.

```
set.seed(123456)
sumaA<-rep(0,1000)
sumaB<-rep(0,1000)

for(i in 1:1000){
  u1<-rnorm(50)
  u2<-rnorm(50)
  u3<-rnorm(50)
  u4<-rnorm(50)
  error<-rnorm(50)
  error2<-rnorm(50,sd=0.1)

  x1<-u1
  x2<-u2
  aa<-c(1,2,3,0)
  bb<-c(-1,2,0.5,3)

  dd<-c(3,4,5,6)
  x3<-aa[1]*u1+aa[2]*u2+error2
  x4<-bb[1]*u1+bb[2]*u2+bb[3]*u3+bb[4]*u4
  y<-dd[1]*x1+dd[2]*x2+dd[3]*x3+dd[4]*x4+error/0.1

  predichosA<-predict(ajus,newdata = data.frame(x1=x1,x2=x2,x3=x3,x4=x4))
  sumaA[i]<-sum((predichosA-y)^2)

  predichosB<-predict(ajusmejor,newdata = data.frame(x1=x1,x2=x2,x3=x3,x4=x4))
  sumaB[i]<-sum((predichosB-y)^2)
}

boxplot(sumaA, sumaB)
sum(sumaA>sumaB)
```

- g) El objetivo de este ítem es probar que la multicolinealidad de las variables afecta mucho a los estimadores de los coeficientes, especialmente a sus varianzas estimadas, complicando la interpretación de los coeficientes y la evaluación de su significatividad. Para verlo, repetiremos 1000 veces la generación de 50 datos en las condiciones originales y el ajuste de los modelos A y B.

- Para ello, defina para el modeloA
 - cuatro vectores para guardar los coeficientes estimados: **coef.beta1A**, **coef.beta2A**, **coef.beta3A**, **coef.beta4A**
 - cuatro vectores para guardar los p-valores obtenidos para cada coeficiente, **pval.beta1A**, **pval.beta2A**, **pval.beta3A**, **pval.beta4A**
- todos inicialmente valiendo cero en cada coordenada,

- Para ello, defina para el modeloB
 - tres vectores para guardar los coeficientes estimados: `coef.beta1B`, `coef.beta2B`, `coef.beta4B`
 - tres vectores para guardar los p-valores obtenidos para cada coeficiente, `pval.beta1B`, `pval.beta2B`, `pval.beta4B`
- todos inicialmente valiendo cero en cada coordenada,

Luego, con la ayuda de un `for` (o de la manera que lo entienda mejor),

- genere 50 observaciones nuevas siguiendo la estructura original
- ajuste el modeloA y el modeloB a estas nuevas observaciones
- guarde los coeficientes $\hat{\beta}_1$ a $\hat{\beta}_4$ en el vector correspondiente para el modeloA y también los p-valores obtenidos
- guarde los coeficientes $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_4$ en el vector correspondiente para el modeloB y también los p-valores obtenidos
- repita los items anteriores 1000 veces

Realice un boxplot (en la misma escala) con los valores de los coeficientes β_1 estimados en las 1000 simulaciones en `modeloA` y `modeloB`. Repita para los restantes coeficientes. Observe que el modelo B no tiene coeficiente β_3 . Compare los boxplots obtenidos en cuanto a la variabilidad. Calcule las varianzas muestrales de ambos y compárelas. Calcule la cantidad de simulaciones cuyos p-valores son menores a 0.05 para ambos modelos. Concluya que la colinealidad en los predictores afecta la interpretación de los coeficientes estimados por el modelo y también la significatividad de los mismos.

```
coef.beta1A<-rep(0,1000)
coef.beta2A<-rep(0,1000)
coef.beta3A<-rep(0,1000)
coef.beta4A<-rep(0,1000)
coef.beta1B<-rep(0,1000)
coef.beta2B<-rep(0,1000)
coef.beta4B<-rep(0,1000)
```

```
pval.beta1A<-rep(0,1000)
pval.beta2A<-rep(0,1000)
pval.beta3A<-rep(0,1000)
pval.beta4A<-rep(0,1000)
pval.beta1B<-rep(0,1000)
pval.beta2B<-rep(0,1000)
pval.beta4B<-rep(0,1000)
```

```
set.seed(123456)
for(i in 1:1000){
  u1<-rnorm(50)
  u2<-rnorm(50)
  u3<-rnorm(50)
  u4<-rnorm(50)
  error<-rnorm(50)
  error2<-rnorm(50,sd=0.1)
```

```
x1<-u1
x2<-u2
aa<-c(1,2,3,0)
```

```

bb<-c(-1,2,0.5,3)

dd<-c(3,4,5,6)
x3<-aa[1]*u1+aa[2]*u2+error2
x4<-bb[1]*u1+bb[2]*u2+bb[3]*u3+bb[4]*u4
y<-dd[1]*x1+dd[2]*x2+dd[3]*x3+dd[4]*x4+error/0.1

# ajustamos ambos modelo con los nuevos datos
# y guardamos sus coeficientes estimados y
# los p-valores obtenidos
modeloA<-lm(y ~ x1+x2+x3+x4)
coef.beta1A[i]<-as.numeric(coef(modeloA)[2])
coef.beta2A[i]<-as.numeric(coef(modeloA)[3])
coef.beta3A[i]<-as.numeric(coef(modeloA)[4])
coef.beta4A[i]<-as.numeric(coef(modeloA)[5])
#pvalores
pval.beta1A[i]<-summary(modeloA)$coefficients[2,4]
pval.beta2A[i]<-summary(modeloA)$coefficients[3,4]
pval.beta3A[i]<-summary(modeloA)$coefficients[4,4]
pval.beta4A[i]<-summary(modeloA)$coefficients[5,4]

modeloB<-lm(y ~ x1+x2+x4)
coef.beta1B[i]<-as.numeric(coef(modeloB)[2])
coef.beta2B[i]<-as.numeric(coef(modeloB)[3])
coef.beta4B[i]<-as.numeric(coef(modeloB)[4])
pval.beta1B[i]<-summary(modeloB)$coefficients[2,4]
pval.beta2B[i]<-summary(modeloB)$coefficients[3,4]
pval.beta4B[i]<-summary(modeloB)$coefficients[4,4]

}

boxplot(coef.beta1A,coef.beta1B)
boxplot(coef.beta2A,coef.beta2B)
boxplot(coef.beta4A,coef.beta4B)

var(coef.beta1A)
var(coef.beta1B)

var(coef.beta2A)
var(coef.beta2B)

var(coef.beta4A)
var(coef.beta4B)

sum(pval.beta1A<0.05)
sum(pval.beta1B<0.05)

sum(pval.beta2A<0.05)
sum(pval.beta2B<0.05)

```

```
sum(pval.beta4A<0.05)  
sum(pval.beta4B<0.05)
```

Referencias

Chatterjee, S., y Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Weisberg, S. (2005). *Applied linear regression* (3rd. ed. ed.). John Wiley & Sons.