Introducción a la Ciencia de Datos

Guía de trabajos prácticos N°9

Tabla de contenidos

Introducción al modelado de datos (parte 2).

En esta guía, exploramos algunos conceptos más avanzados del **modelado de datos**. Está dividida en tres partes. En la primera y la segunda discutiremos dos aspectos técnicos importantes: los términos de interacción entre covariables y la comparación de modelos. En la tercera, introducimos un tema clave para poder tomar decisiones usando los modelos como herramientas: las variables confundidoras. Mostramos cómo estas variables pueden interferir en las conclusiones que sacamos de un conjunto de datos y cómo hacer para tenerlas en cuenta en un caso sencillo en el que las hemos observado y son parte de nuestro dataset.

Recordamos que en esta materia solo veremos modelos de *regresión*, es decir donde la variable *objetivo* es continua (y además suponemos que está distribuida de manera normal). No discutiremos modelos más generales, como cuando la variable *objetivo* es discreta (se llaman modelos de *clasificación*), o en general cuando la distribución de la *objetivo* no es normal (modelos lineales generalizados).

Empecemos entonces, cargando `modelr` (`stats` se carga automáticamente) y nuestro querido `tidyverse`.

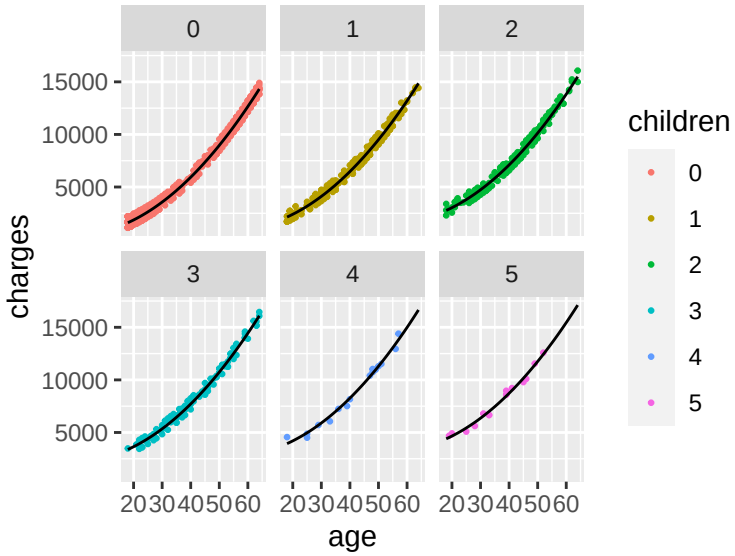
```
library(tidyverse)
library(modelr)
```

Parte 1. Términos de interacción

Vamos a usar nuevamente el dataset de gastos de seguro médico en Estados Unidos (el dataset de `insurance`), pero filtrado para mantener solo clientes no fumadores que no presentan gastos atípicos. Esto mismo lo hicimos en la clase virtual de la que tienen un video.

Vimos durante la clase virtual que podemos obtener un muy buen modelo de la dependencia de los gastos con la edad (si bien después vimos que hay cosas muchísimo mejores), usando a la variable `children` (categórica), en un modelo de fórmula `charges ~ poly(age, 2) + children`.

1. A modo de repaso, repitan este modelo y salvenlo en la variable `mod0` que vamos a usar de comparación. Obtengan el gráfico a continuación. **Tip**: `mutate(children = factor(children))`



Ya vimos que en un modelo de este estilo, los modelos para personas con diferentes cantidad de hijos difieren **únicamente** en un término constante (un *offset*), pero que la forma de la curva es la misma en todos los paneles.

Una forma de flexibilizar este modelo y permitir parámetros del polinomio que cambian con el valor de la variable categórica `children` es usar **términos de interacción**.

> **Importante**
>
> En la mini-sintaxis de los modelos, esto se pone con el signo de multiplicación `*`: `charges ~ poly(age, 2) * children`.

2. Usen esta sintaxis en un nuevo modelo, `mod1`, y repitan el gráfico de arriba para el modelo con interacciones. Comparen los dos gráficos. ¿Hay una diferencia notable?
3. Ahora usen la función `summary` con `mod1` para ver los parámetros que tiene este modelo. Intenten entender el significado de cada uno. Vean cuáles son significativos y cuáles no. ¿Tiene sentido esto en vistas de la comparación de los gráficos de los puntos anteriores?
4. Ahora comparen la dispersión de los residuos y el R2 para ambos modelos. Estos valores los pueden sacar del `summary`. Para tener más cifras significativas en el R2, pueden usar `modelr::rsquare`. Si tuvieran que elegir un solo modelo en base a esto, ¿cuál sería?

Ahora vamos a hacer lo mismo con otro dataset. Usaremos `mtcars`, el dataset de automóviles con varias características, como su potencia (`hp`, de *horsepower*) y cantidad de cilindros (`cyl`), y datos sobre su rendimiento `mpg`.

> **Tip**
>
> Pueden usar `plot(mtcars)` para ver todos los datos y todas las relaciones de a pares.

5. Ajusten un modelo para el rendimiento `mpg` en función de la potencia del motor (`hp`) como covariable principal y el peso del vehículo (`wt`) como secundaria. En primer lugar hagan un modelo *sin* interacciones, y guarden el resultado en una variable de nombre `mod0cars`.
6. Hagan una gráfica similar a las anteriores. Recuerden que en el caso de interacciones entre dos variables contínuas, es interesante fijar algunos valores de la variable secundaria. (**Tip**: `wt = seq_range(wt, 4)`)
7. Hagan un gráfico de los residuos y evalúen si el modelo está siendo suficiente para explicar los datos.
8. Ajusten un modelo con interacciones (`mod1cars`) y repitan el gráfico para ver cómo cambia el modelo por clase.

9. Ahora vean con `summary` el valor de los parámetros, y su significancia. Además, comparen el valor en la dispersión de los residuos y el R2 de este modelo con el modelo sin interacciones. A partir de estos valores, ¿qué dirían del modelo con interacción con respecto al modelo sin interacción?

10. Comparen también el gráfico de residuos de ambos modelos.

Parte 2. Comparación de modelos con el test F.

Vimos arriba que en algunos casos, los términos de interacción mejoran el desempeño del modelo, mientras que en otros casos, la mejora es poco importante.

> **Nota**
>
> En general, cuantos más parámetros tenga el modelo mejor será el ajuste de los datos. Entonces, ¿por qué no incluir siempre términos de interacción y listo? La verdad es que además de ajustar bien los datos, nos interesa tener el modelo más simple posible, para poder interpretarlo y comunicar los resultados de manera más directa.

Entonces, necesitamos alguna forma de saber cuándo agregar un parámetro más a un modelo "no vale la pena". En otras palabras, cuándo la mejora del ajuste no es significativa para la cantidad de parámetros que se agregaron. Ya mencionamos el R2 ajustado, que tiene en cuenta el número de parámetros. Ahora vamos a introducir una herramienta mucho más rigurosa para comparar modelos, el test F.

Para poder hacer este test, calculamos para cada modelo que queremos comparar la suma del cuadrados de los residuos, RSS, para cada modelo. Es decir, tenemos que sumar los residuos al cuadrado para cada modelo:

$$RSS = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 \quad ,$$

donde N es el número de puntos del dataset, y_i es el valor de la variable *target* y $f(x_i)$ es la predicción del modelo para ese punto.

Los valores de RSS para cada modelo se commparan teniendo en cuenta el número de parámetros (lo llamamos p) para cada modelo. Construimos el *estadístico de test* (¡jerga!):

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p_1}\right)}{\frac{RSS_2}{N - p_2}} \quad .$$

En esta expresión, N es el número de puntos.

> **Importante**
>
> Podemos ver que **no es lo mismo** qué modelo ponemos como modelo 1 y cuál como modelo 2. Siempre la idea es usar al modelo 1 como el más siemple (el que tendrá RSS mayor).

Este estadístico tiene una distribución conocida bajo la hipótesis de que ambos modelos son igualmente buenos para describir los datos. Se puede calcular un p-valor, que indica la probabilidad de haber obtenido un valor de F igual o más extremo que el observado. Es decir, si esta probabilidad es muy baja, podemos decir que agregar los parámetros adicionales valió la pena y que el nuevo modelo captura mejor los datos.

La salida del `summary` muestra el valor de este test en comparación con un modelo simple, constante (es decir, un único parámetro, la ordenada al origen, si el modelo que se usa lo tiene; la constante es cero si el modelo no tiene ordenada al origen).

11. Calculen el valor del estadístico F para el modelo `mod_cars0`, comparado con un modelo de una sola variable, la constante. Comparen con la salida de `summary(mod_cars0)`.

> **Tip**
>
> Después de calcular los residuos con `add_residuals`, pueden usar `mutate` para hacer el cálculo que les permite encontrar el estadístico F.

Ahora bien, ¿qué pasa si queremos comparar dos modelos ajustados entre sí, en lugar de cada uno con un modelo simple? R trae una función llamada `anova` (del inglés *analysis of variance*, análisis de la varianza), que sirve exactamente para eso. Se usa `anova(mod0, mod1, ...)`, donde ponemos todos los modelos a comparar.

12. Usen esta función para comparar cada uno de los modelos hechos en la parte 1 con un modelo constante. Para eso, creen un modelo constante en cada caso, `modbase <- lm(data=df, formula=charges ~ 1)`, en el caso de `insurance`. Comparen con la salida de `summary` y con los cálculos que hicieron en el punto 11, para entender cada una de las columnas de la **tabla de anova** que devuelve la función `anova`.

13. Ahora hagan lo mismo comparando los modelos con o sin interacciones en cada uno de los casos que vimos arriba. ¿Cómo resulta el resultado del test (el p-valor) en un caso y el otro? ¿Se condice con la conclusión que habíamos obtenido en la parte 1?

Parte 3. Variables confundidoras

Cuando queremos tomar decisiones basadas en análisis como los que venimos haciendo, tenemos que estar seguros de no estar confundiendo efectos. Por ejemplo, es bien conocido que "la correlación no implica causalidad". Esto quiere decir que si vemos una relación entre dos variables, esto no implica necesariamente que una variable sea **la causa** la otra. Por ejemplo, puede existir una tercera variable (llamada confundidora) que esté causando ambas variables originales, por lo que vemos una relación entre ellas.

Por lo tanto, si queremos ver la intensidad del efecto de una variable sobre la otra, tenemos que tener en cuenta, en la medida de lo posible, potenciales variables confundidoras. Esto se dice **controlar por tal variable confundidora**.

Vamos a ver un ejemplo práctico con el datasets de diamantes (`diamods`) que contiene datos sobre diamantes, varias características y sus precios.

Carguemos el dataset.

```
data(diamonds)
```

13. Exploren el dataset e identifiquen las variables categóricas que existen. ¿Cuántas filas tiene el dataset? ¿Cuántas variables? Vean `?diamonds` para entender los valores de las variables categóricas, y el orden "natural" que tienen.

14. Nos interesa entender de qué depende el precio (`price`) de un diamante. Podemos hacer un boxplot del precio para algunas de las variables categóricas, por ejemplo, la claridad o el color. Evalúen la evolución de la mediana en función de esta variable. ¿Es lo que esperaban?

> **Nota**
>
> Este es un típico ejemplo de la paradoja de Simpson. Una variable confundidora que no se tiene en cuenta no solo impide ver un patrón, sino que simula un patrón **inverso** al verdadero.

En este caso, hay una variable que es muy importante para la determinación del precio de un diamante, que es su peso (sus quilates, `carat`)

15. Hagan una gráfica del precio vs. los quilates. Describan la relación que ven de forma cualitativa. Ahora, intenten ajustar la relación entre precio y quilates de la mejor manera posible con un modelo lineal de regresión.

> **Tip**
>
> Recuerden que pueden modificar las covariables e incluso la target; prueben, por ejemplo, `log(price)` vs. `log(carat)`.

Hagan todos los gráficos y reporten todos los valores de métricas necesarios para mostrar que el ajuste funciona razonablemente bien.

16. Una vez que tengan un modelo satisfactorio entre el precio y los quilates, calculen los residuos y grafiquen un boxplot de los residuos para alguna de las variables categóricas del punto 14. ¿Cómo es ahora la relación? ¿Se condice con lo que esperaríamos?

17. Con este resultado bajo la manga, realicen un modelo (con o sin interacciones) que les permita cuantificar la relación entre el precio de los diamantes y el color o la claridad.