Introducción a la Ciencia de Datos - 2C 2022

Guía de Trabajos Prácticos Nº 6

Tema: *Modelo lineal*

Vamos a trabajar en grupo para volvernos expertxs en el uso de los modelos lineales.

Hagan grupos de tres y asignen un rol a cada integrante en este link. Los roles son:

1. Coder. Lleva adelante el código en la compu; genera los gráficos y la información que se usará para el reporte
2. Reporter. Registra los pasos y va guardando los gráficos o datos necesarios para un reporte. Va generando el reporte en tiempo real.
3. Timer. Tiene la responsabilidad de mantener al equipo a tiempo.

Para cada una de las tareas de la guía tienen **30 minutos**. Al final de cada una de ellas, deben tener un breve reporte, listo para enviar. Pueden cambiar de roles entre una tarea y la otra, pero no es obligatorio. **Tip**: pidan ayuda si se traban, o miren los capítulos 23 y 24 del libro de Wickham.

En sus marcas...

Para empezar, carguen la librería `modelr`. Van a necesitar seguramente las funciones `data_grid`, `add_predictions`, `add_residuals` y `model_matrix`. Pero siéntanse libres de explorar otras funciones de esta librería, que es muy poderosa e interesante.

Tarea 1. Test estadísticos.

En este punto usamos los datos simulados de `modelr` (`sim1`).

1. Ajusten un modelo lineal que relacione la variable `y` con la `x`.
2. Grafiquen los datos junto con las predicciones.
3. Calculen los residuos y grafíquenlos en función de la variable dependiente y dcare la variable x.
4. Exploren los gráficos que se obtienen con `plot(model)`. **Nota:** no esperamos que se entienda todo lo que hay en estos gráficos.
5. Usen `summary(model)` para ver el valor de los coeficientes y de los estadísticos. ¿Podemos decir que la pendiente de la relación es significativamente diferente de cero?
6. Agreguen ruido a los valores de `y` con la función `rnorm()` (vean la docu) y repitan. Aumenten el tamaño del ruido. Encuentren el punto en el que la pendiente ya no es significativa.

Tarea 2. Términos de interacción

El dataset sim3 tiene una variable continua y una categórica. Vamos a usar este conjunto de datos para entender cómo funcionan los términos de interacción en un modelo lineal.

1. Identifiquen las variables categóricas y las variables continuas.
2. Todavía estamos tratando con una cantidad de variables que podemos visualizar en un scatter plot (geom_point). Hagan un gráfico donde se vea el dataset completamente.
3. Estudien cuál es la diferencia entre la fórmula `y ~ x1 + x2` e `y ~ x1 * x2`. Usen `model_matrix` para esto. ¿Cuál de las dos fórmulas produce un modelo con más parámetros?
4. Ajusten ambos modelos y realicen un gráfico que les permita entender la diferencia entre ambos approach. ¿Cuál de los dos creen que es un modelo con "términos de interacción"?
5. Piensen un caso para cada fórmula en el que sería adecuado usar ese modelo.

Tarea 3. Transformación de las variables

Carguen el dataset de diamantes que viene con ggplot2. Vamos a trabajar en esta sección con las variables carat y price. En particular, buscamos encontrar un modelo lineal que nos de una buena aproximación del precio a partir de los quilates de un diamante.

1. Hagan una gráfica que les permita visualizar estas dos variables. Pueden agregar alguna variable categórica adicional, si les parece útil.
2. ¿Les parece que podremos encontrar un buen modelo entre ambas variables? Hagan un modelo con la fórmula `price ~ carat` y grafiquen la predicción.
3. Hagan una gráfica de los residuos del ajuste en función de la predicción o de los quilates. Discutan la forma tendencia que muestra el gráfico y si les parece que es lo que debería ser. **Ayuda**: si se les hace cuesta arriba con las funciones de `modelr`, usen directo `plot(modelo)`, donde modelo es lo que devuelve `lm()`.
4. Prueben otras relaciones entre las dos variables, transformando una o la otra. Inspírense en el gráfico del punto 1, y en lo que se les puede llegar a ocurrir de cómo piensan que cambia el precio de un diamante con su peso.
5. Si no lo hicieron en el punto anterior, hagan un modelo con la fórmula `log2(price) ~ log2(carat)` y analicen sus residuos y el valor de sus coeficientes. ¿Es significativa la relación entre las variables? Comparen con el modelo del punto 2.

Tarea 4. Confounders y modelo causal

Seguimos trabajando con los diamantes. Ahora vamos a ver cómo depende el precio con una variable categórica, como el `color`, el corte (`cut`), o la claridad (`clarity`).

1. Lean la documentación o investiguen para entender cuáles son los cortes o colores más apreciados.
2. Hagan un gráfico de violines o de cajas para ver cómo depende el precio de los diamantes de estas variables. En particular, vean cómo depende la media de estas variables. ¿Es lo que esperaban?
3. Hagan un modelo que relacione al precio con alguna de estas variables. Para esto, construyan una variable dicotómica, que indique si los diamantes son de buena o mala calidad (por ejemplo, los tres mejores cortes y los tres peores), y hagan un ajuste lineal que relacione el precio con esta variable. Analicen el valor de los coeficientes obtenidos y si está en línea con el análisis exploratorio de arriba. **Nota:** en general, confirmar cosas que intuimos en un análisis exploratorio con un modelo estadístico está bien, pero hay que hacerlo con dos conjuntos de datos diferentes, para no estar cayendo en confirmar lo que ya sabemos.
4. Consideren la tarea anterior. Vimos que había una importante dependencia entre el precio y los quilates de un diamante. ¿Puede ser que esto esté entorpeciendo la relación entre el precio y las variables categóricas? Hagan un gráfico que muestre cómo dependen los quilates del corte, del color o de la claridad. ¿Va en línea con la hipótesis que estamos manejando?
5. Obtengan los residuos del ajuste de la tarea anterior (usen `add_residuals`, por ejemplo) y analicen cómo dependen los residuos con las variables categóricas. ¿Tiene sentido lo que vemos?
6. Ahora hagan un modelo que incluya el efecto de los quilates además del de las variables categóricas. ¿Es más adecuado un modelo con interacciones o sin interacciones? ¿Cómo cambia el valor de los coeficientes con respecto a los del punto 3 de esta tarea?