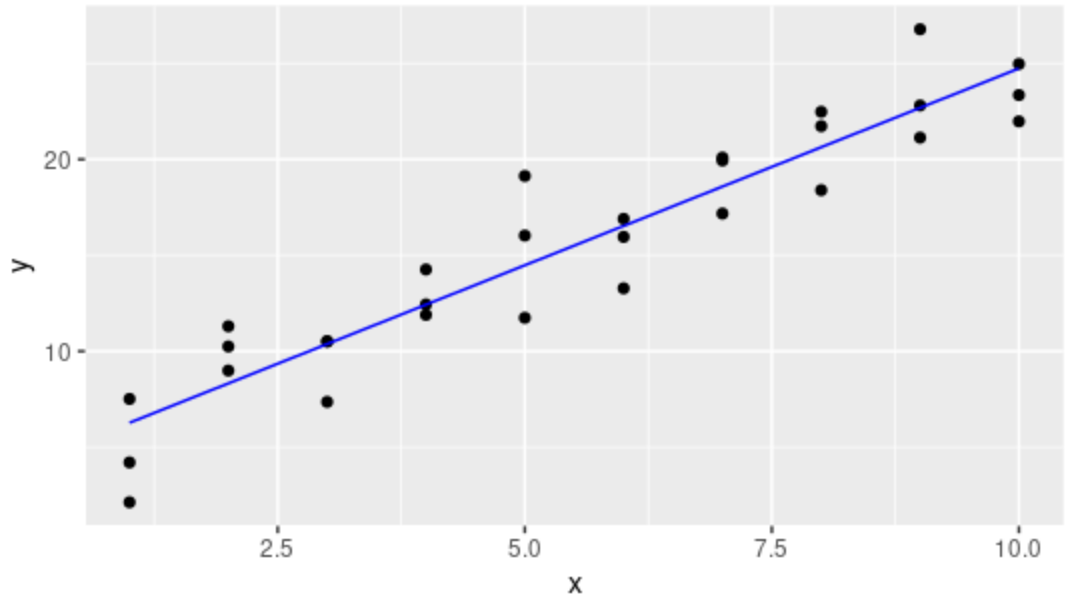
Materia: Introducción a la Ciencia de Datos
Participantes: Crespi, María Sol;Vazquez, Lucía ; Vidal Ramón, Tomás

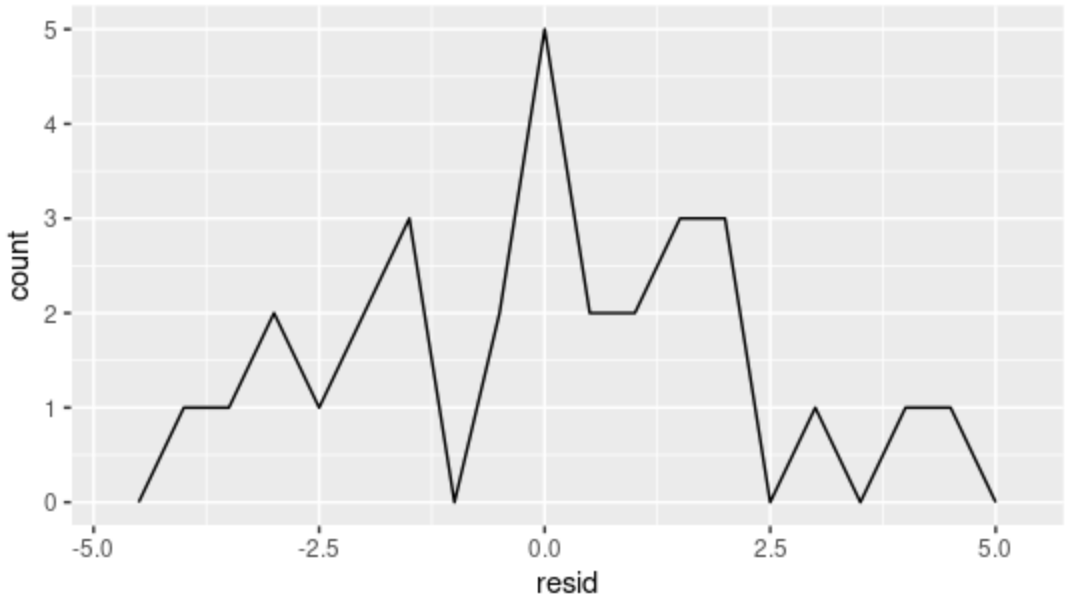
Informe sobre "Test estadísticos"

Tarea 1:

Comenzamos el documento como siempre, importando la librería de tidyverse y la de modelr. Definimos el modelo a partir de la función lm(y~x, data= sim1). A continuación creamos una grilla para calcular las predicciones(geom_line) y así poder graficarlas en conjunto con los otros datos(scatter plot) .



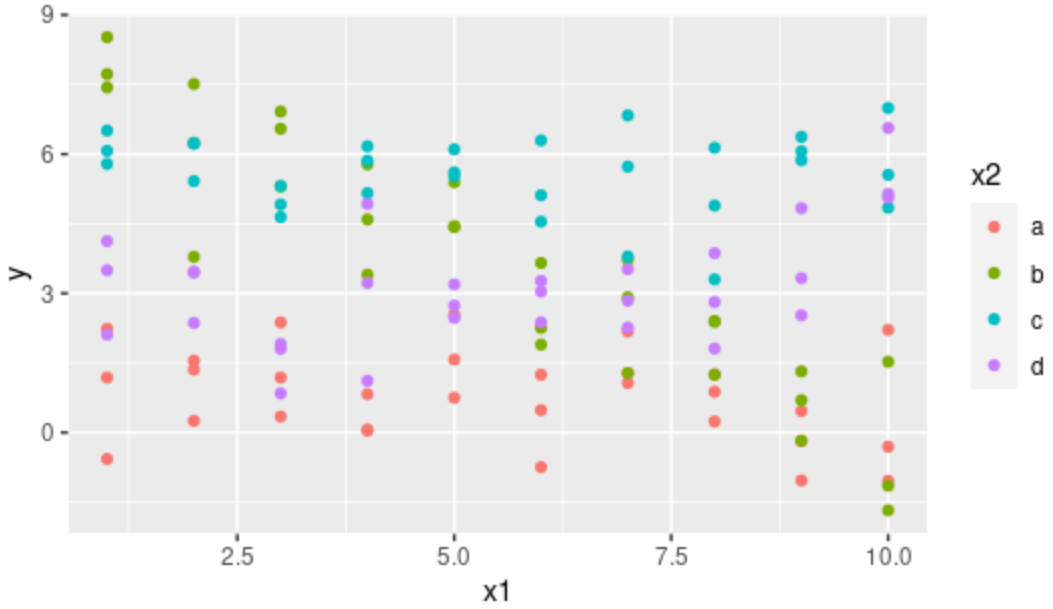
Luego calculamos los residuos con la herramienta add.residuals() y graficamos usando un gráfico de frecuencia(geom_frecuency).



Revisamos el libro de R para poder agregar la columna ruido con el rnorm().Terminando con esta tarea la función summary nos devolvió que la pendiente era muy cercana a cero.

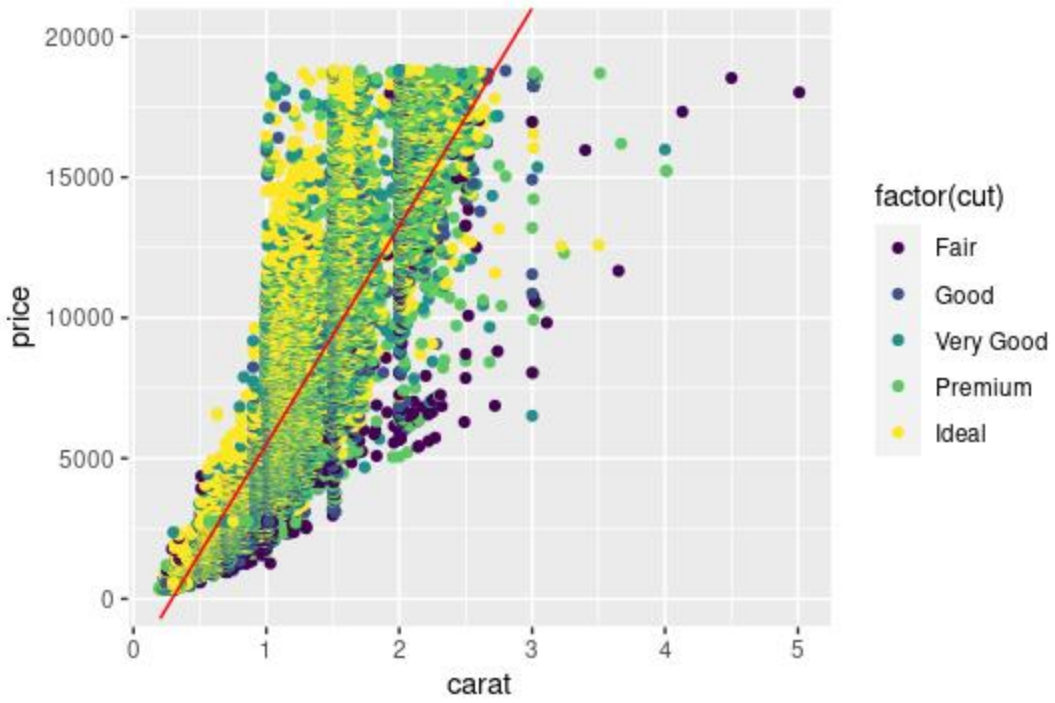
Tarea 2:

En el dataset sim3 la única variable categórica que encontramos es “x2” y las variables “x1” e “y” son continuas.

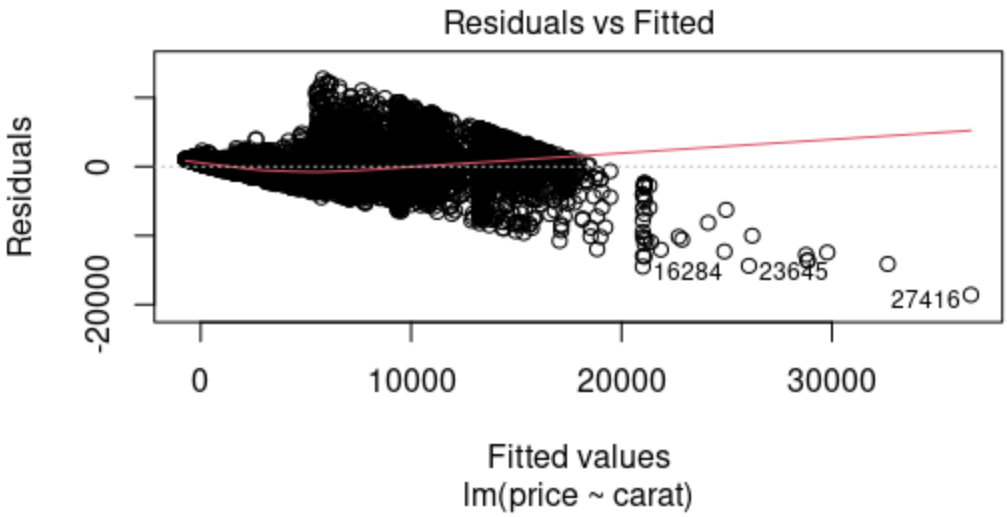


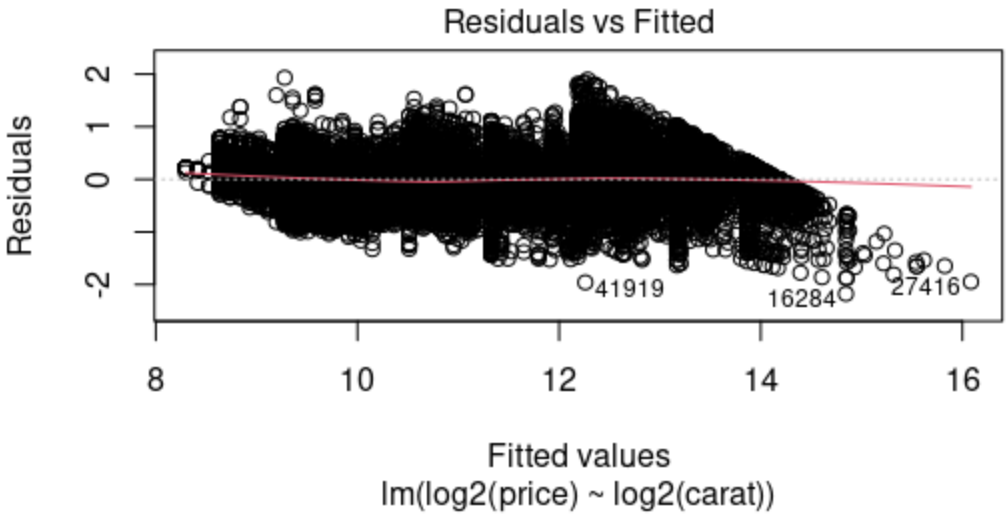
Después de usar model_matrix(), concluimos que la fórmula que da más parámetros es y~ x1 *x2. Luego de haber realizado los gráficos, vimos que el modelo con “términos de interacción” es también la fórmula anterior, ya que una variable depende de la otra.
FALTA 5

Tarea 3:



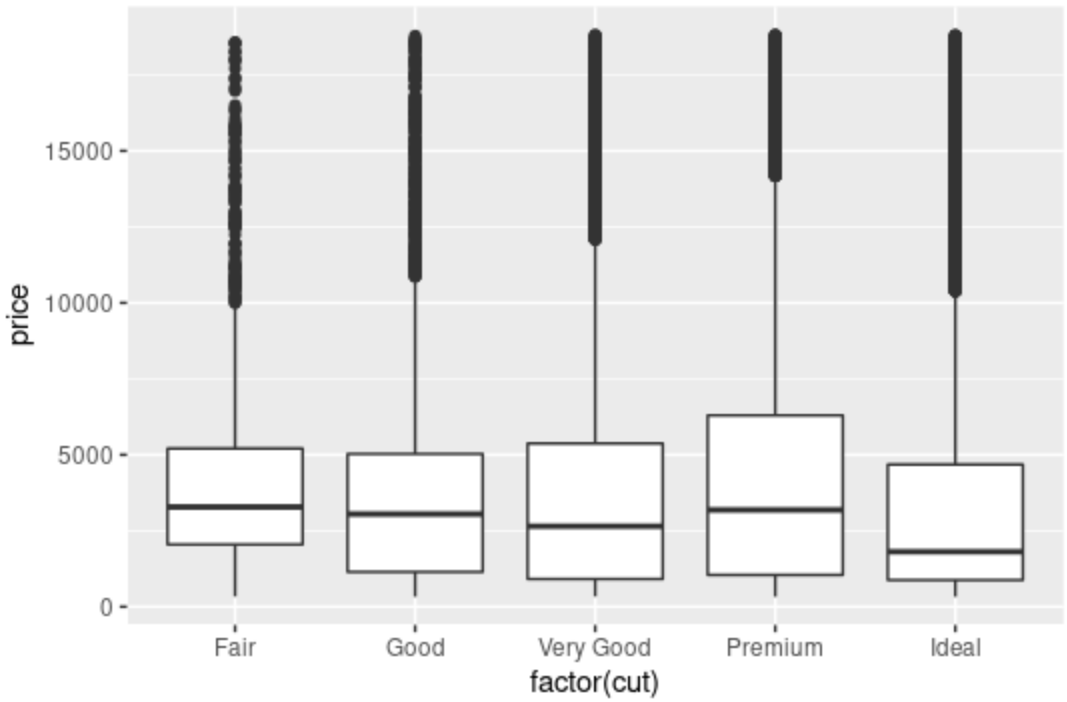
Este gráfico resume el punto 1 y 2 de la tarea 3, donde vemos un geom_point de dos variables (“price” y “carat”) con sus predicciones y categorizadas por “cut”.





Después de comparar los últimos dos gráficos, nos dimos cuenta que el último tiene la línea mucho más cerca del cero a lo largo de las función (log2(price)~log2(carat)).

Tarea 4:



La media era lo que esperábamos, ya que los extremos tienen coherencia al ser “Fair” la categoría más baja.