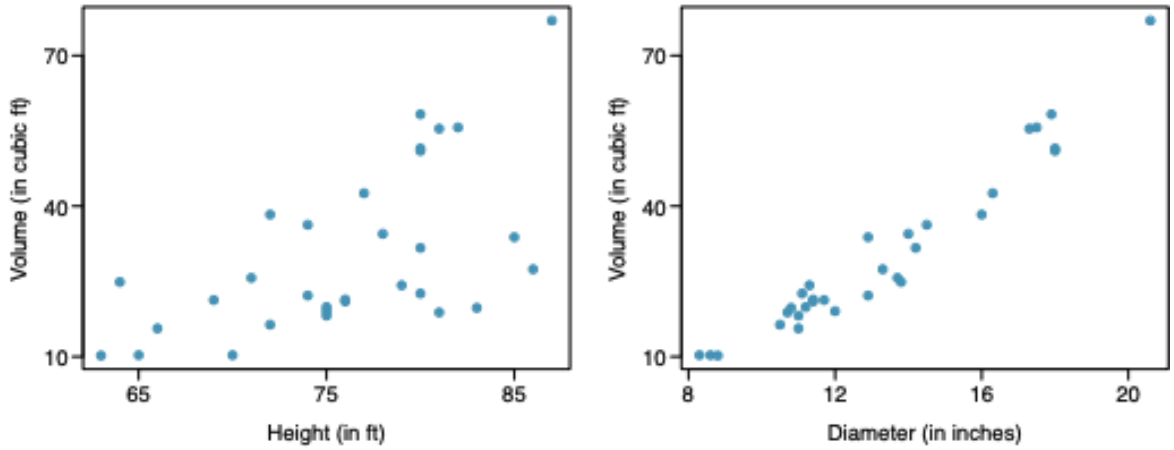
Unidad 5 - Actividad 5A. Introducción a la regresión lineal.
Alumno: **Luis Fernando Izquierdo Berdugo**
Materia: **Estadística**
Fecha de Entrega: **7 de Noviembre de 2024**

Realiza los siguientes ejercicios del capítulo 7 Introducción a la regresión lineal.

Página	Ejercicios a realizar
360	7.12
363	7.23
363	7.24
364	7.27
368	7.36

7.12 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.

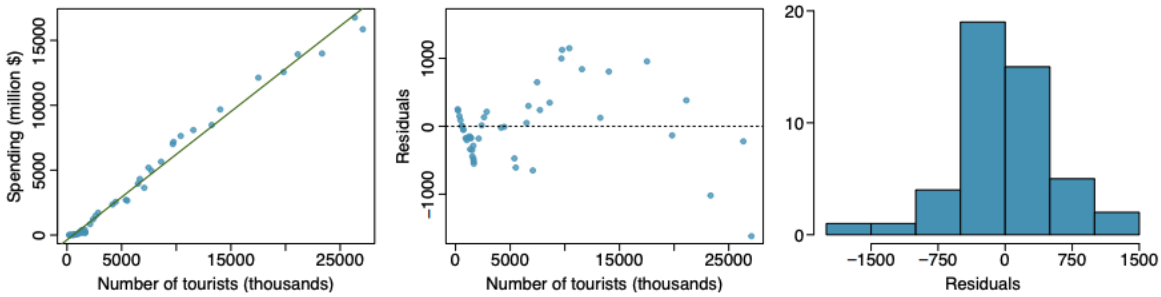


(a) Describe the relationship between volume and height of these trees.
Se observa una relación positiva entre el volumen y grosor de los árboles (mientras más alto el arbol, más volumen tendrá). Es digno de destacar que esta relación es débil ya que los puntos están bastante dispersos.

(b) Describe the relationship between volume and diameter of these trees.
Aquí se puede observer una relación positiva bastante fuerte ya que los puntos están mucho más cercanos a una línea ascendente, por lo cual se sabe que el diámetro del árbol es una característica que predice mejor el volumen de este mismo (mientras mayor el diámetro, mayor el volumen).

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.
Como se explicó previamente, el diámetro es una característica que tiene una relación mucho más fuerte con el volumen, por lo que se escogería este. En los gráficos observamos una menor dispersión del volumen frente al diámetro, lo cual indica que puede proporcionar una predicción más precisa del volumen.

7.23 Tourism spending. The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year. Three plots are provided: scatterplot showing the relationship between these two variables along with the least squares fit, residuals plot, and histogram of residuals.



(a) Describe the relationship between number of tourists and spending.
Se puede observer una relación positiva entre el número de turistas y el gasto turístico en el diagrama de dispersión. La relación se ve bastante lineal ya que los puntos están bastante cercanos a la línea de ajuste por mínimos cuadrados.

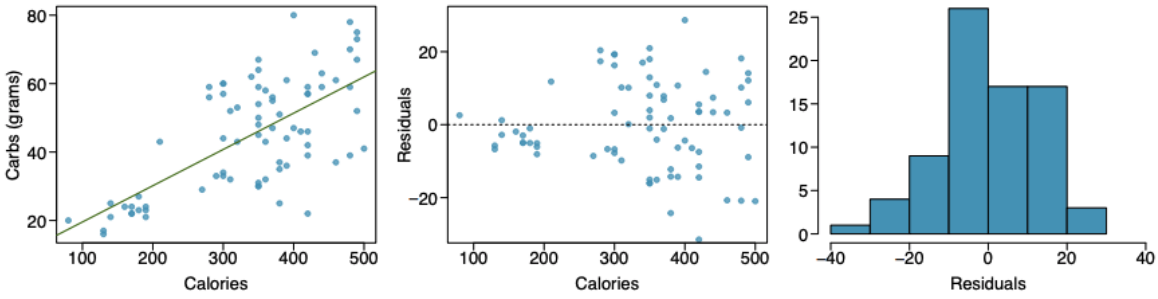
(b) What are the explanatory and response variables?
La variable explicative es el número de turistas y la variable de respuesta es el gasto turístico, ya que, se espera que el gasto turístico dependa del número de turistas.

(c) Why might we want to fit a regression line to these data?
Para modelar la relación ente el número de turistas y el gasto turístico, esto permitiría predecir el ingreso que dejarían los turistas en varios periodos de tiempo.

(d) Do the data meet the conditions required for fitting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.
Los datos parecen cumplir las condiciones para ajustar una línea de regresión de mínimos cuadrados, cuyas condiciones son:

- Linealidad: En el primer gráfico se observa una relación lineal clara.
- Distribución normal de los residuos: En el tercer gráfico se observa una distribución simétrica de los residuos con cierta dispersión, lo cual la hace factible.
- Varianza constante: En el segundo gráfico se observa una distribución aleatoria y sin patrones evidentes de los residuos.
- Independencia de los residuos: No hay información suficiente para evaluarlo.

7.24 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content.



(a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.
Se observa una relación positive en el primer gráfico, por lo cual, a mayor cantidad de calorías se espera una mayor cantidad de carbohidratos. Sin embargo, esta relación es dispersa, por lo cual podríamos decir que es una relación débil.

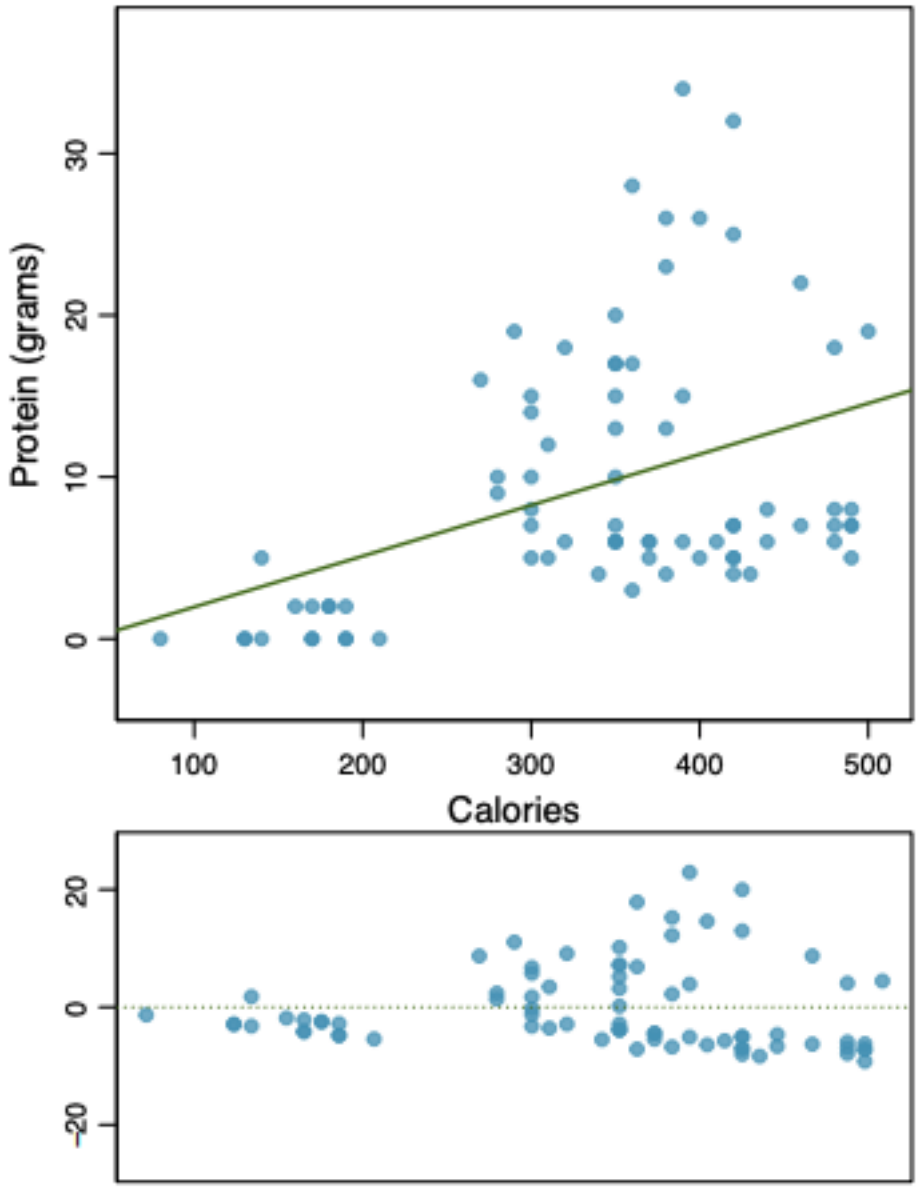
(b) In this scenario, what are the explanatory and response variables?
La variable explicative es el número de calorías y la variable de respuesta es la cantidad de carbohidratos, ya que se espera que la cantidad de carbohidratos se pueda predecir con la cantidad de calorías.

(c) Why might we want to fit a regression line to these data?
Para modelar la relación entre estas dos características, lo cual podría ayudar a estimar la cantidad de carbohidratos de un producto sabiendo únicamente su cantidad de calorías.

(d) Do these data meet the conditions required for fitting a least squares line?
Los datos parecen cumplir las condiciones para ajustar una línea de regresión de mínimos cuadrados, cuyas condiciones son:

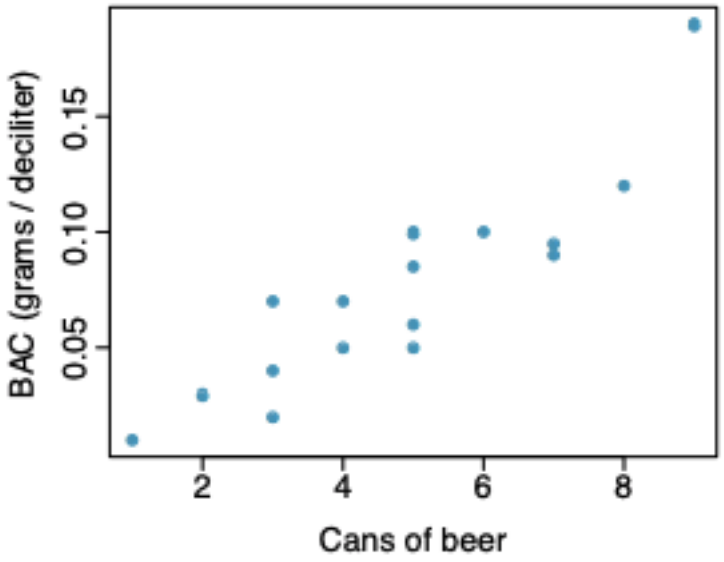
- Linealidad: En el primer gráfico se observa una relación lineal moderada, es aceptable.
- Distribución normal de los residuos: En el tercer gráfico se observa una distribución casi simétrica de los residuo, aunque no son totalmente normales, sin embargo, parecen ser aceptables para ajustar un modelo.
- Varianza constante: En el segundo gráfico se observa una distribución aleatoria y sin patrones evidentes de los residuos.
- Independencia de los residuos: No hay información suficiente para evaluarlo.

7.27 Nutrition at Starbucks, Part II. Exercise 7.24 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.



La relación entre el contenido de proteinas y las calorías de los productos es positiva, por lo cual, a medida que aumentan las calorías también aumenta el contenido de proteínas, sin embargo, esta relación no es perfectamente líneal ya que los datos están distribuidos alrededor de la línea ascendente.
Observando el gráfico de residuos, se nota cierta dispersión y un patrón leve, lo cual puede indicar que los errores no son homogéneos en toda la línea de regresión, que a su vez indica que la relación podría no ser perfectamente líneal, por lo cual un modelo lineal simple no paree ser apropiado debido a la falta de aleatoriedad en los residuos.

7.36 Beer and blood alcohol content. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The scatterplot and regression table summarize the findings.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0127	0.0126	-1.00	0.3320
beers	0.0180	0.0024	7.48	0.0000

(a) Describe the relationship between the number of cans of beer and BAC.
La relación entre la cantidad de cervezas consumidas y el contenido de alcohol en la sangre parece ser positiva (mientras más cervezas se consuman, mayor será el alcohol en la sangre), lo cual se puede observar en la dispersión ascendente del gráfico.

(b) Write the equation of the regression line. Interpret the slope and intercept in context.

$$BAC = -0.0127 + 0.0189(Cervezas)$$

(c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion.
La hipótesis nula H0 sería que no existe relación entre la cantidad de cervezas consumidas y el contenido de alcohol en la sangre.
La hipótesis alternativa H1 dice que existe una relación positiva entre la cantidad de cervezas consumidas y el contenido de alcohol en la sangre.

Como el valor de p para la pendiente de la regresión es 0.0000, se tiene evidencia suficienta para rechazar la hipótesis nula y aceptar la hipótesis alternativa.

(d) The correlation coefficient for number of cans of beer and BAC is 0.89. Calculate R^2 and interpret it in context.

$$R^2 = (0.89)^2 = 0.7921$$

El valor de R^2 indica que aproximadamente el 79.21% de la variación en el contenido de alcohol en la sangre se explica debido con el número de cervezas consumidas, por lo cual, el modelo de regresión lineal con la cantidad de cervezas como variable explicativa tiene un ajuste bastante fuerte.

(e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?
Es bastante probable que la relación no sea tan fuerte fuera del estudio, ya que este no toma en cuenta factores adicionales que podrían influir en el contenido de alcohol en la sangre, como podrían ser la duración del consumo, la marca de cerveza que están tomando, género, peso, hábitos de consumo, etc.