UNIVERSIDAD NACIONAL DE COLOMBIA

La enseñanza de la regresión lineal simple usando como mediador una herramienta computacional

Jesus Antonio Villarraga Palomino

Universidad Nacional de Colombia

Facultad de Ciencias

Bogotá, Colombia

2016

La enseñanza de la regresión lineal simple usando como mediador una herramienta computacional

Jesus Antonio Villarraga Palomino

Trabajo de grado presentado como requisito parcial para optar al título de:

Magister en Enseñanza de las Ciencias Exactas y Naturales

Director:

Doctor Ramón Giraldo Henao

Universidad Nacional de Colombia

Facultad de Ciencias

Bogotá, Colombia

2016

A mis familiares.

A mi madre Gilma Palomino por apoyarme en cada proyecto, por sus consejos y la constante motivación; a mis abuelos Marina y Roberto por enseñar con el ejemplo; a Paola Sanabria el gran amor de mi vida y a mis hermanos.

Agradecimientos

Primero y como más importante, a Dios por permitirme llegar a este punto, por llenarme de paciencia y sabiduría.

Al profesor Ramón Giraldo Henao, director de este trabajo de grado, por sus conocimientos, sus orientaciones, su manera de trabajar y el haberme inculcado el rigor académico.

A mi familia, quien me ha apoyado constantemente Diana Palomino, Roberto y Marina son unas de mis motivaciones sin ustedes el presente escrito no sería realidad.

A Felipe Díaz, Adrián, Alberto Santa y Maycol Alirio Rodriguez, por su amistad colaboración y conocimientos, cada momento compartido ha sido edificante y constructivo para la realización de este proyecto y para mi madurez académica.

Resumen

La Regresión lineal simple es una de las herramientas básicas del análisis estadístico. Esta temática se aborda en los cursos de fundamentación estadística de los programas de pregrado. En este trabajo se presenta una propuesta didáctica para su enseñanza basada en el uso de la simulación en R. Las actividades que conforman la unidad didáctica pretenden aportar elementos innovadores al proceso de enseñanza-aprendizaje de la regresión lineal utilizando estrategias computacionales acorde al tema a tratar.

Palabras clave: Regresión lineal, simulación, educación estadística.

Abstract

Simple linear regression is a basic tool in statistical analysis. This methodology is usually teached in basic courses of statistics in many academic undergraduate programs. In this work, we give a didactic proposal for teaching linear simple regression. This one is based on the use of simulations and is implemented in the software R (R development core team,2016). The activities of this didactic unit contribute with innovatory elements to support the teaching-learning process by using computational strategies.

Keywords: Linear regression, simulation, statistical education.

Contenido

Pág.

1. Referente Histórico 5

2. Referente disciplinar 12
2.1 Modelo de regresión lineal simple 12
2.2 Estimación de los parámetros por mínimos cuadrados 13
2.3 Estimación de $\boldsymbol{\sigma 2}$ 15
2.4 Pruebas de hipótesis de la pendiente y de la ordenada al origen. 16
2.5 Estimación de intervalo en la regresión lineal simple 17
2.6 Estimación por máxima verosimilitud 18
2.7 Análisis de los residuales 19
2.8 Enfoque matricial modelo lineal simple 21

3. Referente pedagógico. 24
3.1 Diagramas de dispersión. 26
3.2 Función lineal como regresión lineal. 27
3.3 Simulación como recurso didáctico 28

4. Unidad didáctica 31
4.1 Taller diagnóstico conceptos previos a la regresión lineal 31
4.2 Descripción de la unidad didáctica 41
4.2.1 Objetivo general de la unidad didáctica 41
4.2.2 Objetivos específicos de la unidad didáctica 42
4.2.3 Metodología 42
4.3 Secuencia de actividades 43
4.3.1 Taller regresión lineal de observaciones $\boldsymbol{n = 100}$ y $\boldsymbol{n = 1000}$ 43
4.3.2 Taller regresión lineal 1000 grupos de tamaño 100 48

5. Implementación de la propuesta 51

6. Conclusiones y recomendaciones 55
6.1 Conclusiones 55
6.2 Recomendaciones 56

Lista de figuras

Pág.

Figura 1-1: Ilustración de Galton de la identidad estadística de dos generaciones…….. 5
Figura 2-1: Como se generan observaciones en la regresión. 13
Figura 2-2: Patrones en las gráficas de residuales…… 20
Figura 2-3: Patrones en las graficas de residuales ……21
Figura 3-1: Simulación de 20 líneas de regresión. 28
Figura 4-1: Diagramas de dispersión generado en R. Datos simulados …..……...35
Figura 4-2: Box-plot de los pesos de 210 mujeres. Datos simulados ……..36
Figura 4-3: Histogramas de los pesos por estatura de 210 mujeres Datos simulados ….37
Figura 4-4: Gráficos de residuales. Datos simulados..…..……..44
Figura 4-5: Histogramas de los estimadores de los parámetros del modelo. Datos simulados ……...50
Figura 5-1: Estudiantes ingeniería electrónica desarrollo taller “Regresión lineal simulación de observaciones n=100 y n=1000”.....54

Lista de tablas

Pág.

Tabla 1-1:Principales Modelos Lineales Generalizados, en términos de variable de respuesta, componente, función del enlace y modelo lineal…… 10
Tabla 3-1: Diferencias entre función lineal y regresión lineal 27
Tabla 3-2: Cuadro comparativo SPSS, SAS Y R ……29
Tabla 4-1: Datos generados por simulación en R 33
Tabla 4-2: Valores frecuentes utilizados de la distribución normal para Z……...38

1. Introducción

En el proceso de enseñanza-aprendizaje de los conceptos estadísticos, se evidencian dificultades respecto a la comprensión, interpretación y significado del lenguaje. Esta problemática puede ser producto de prácticas indebidas en el aula que terminan siendo reflejadas en la formación de los estudiantes. Según Batanero et al. (1994) muchos docentes precisan incrementar su conocimiento, no solo en la materia, sino también sobre los aspectos didácticos, pero no profundizan en el estudio de las dificultades y errores que los alumnos encuentran en el aprendizaje dela estadística.

En el caso particular de la enseñanza de la regresión lineal, se reportan numerosas investigaciones en cuanto al proceso enseñanza-aprendizaje de la misma, en las cuales se corrobora la dificultad de abordar este tema de manera significativa. Algunas de estas apuntan a dificultades conceptuales y otras a aspectos pedagógico (Lavalle et al, 2006). Por ejemplo, en cuanto a lo disciplinar Angnelli et al (2009) afirma que a los estudiantes les cuesta diferenciar entre función lineal y regresión lineal, en el análisis de las situaciones, generando así confusiones y dificultades en la enseñanza-aprendizaje de la regresión. Desde el punto de vista pedagógico Lavalle et al. (2006) proponen una reflexión didáctica de la regresión bajo tres aspectos: el contenido, que abarca la naturaleza del concepto de regresión bajo la perspectiva conceptual y procedimental; la enseñanza, con un propósito analítico a la revisión didáctica de los libros y el aprendizaje, identificando los errores y las dificultades en el proceso de enseñanza y aprendizaje. De lo anterior y teniendo en cuenta que la enseñanza de estadístico, puede ser orientada de manera interactiva, generando así espacios de discusión y reflexión, surge la pregunta:

¿Qué características debe tener una estrategia didáctica que permita a los estudiantes universitarios interpretar y aplicar la regresión lineal?

Para dar respuesta a la pregunta anterior, en este trabajo se utiliza la simulación como recurso pedagógico en la construcción de una secuencia didáctica. Según Batanero (2000) la simulación cobra un papel importante en la enseñanza de la estadística, ya que ayuda a diferenciar entre temas experimentales y teóricos. Además, la simulación permite interactuar con diferentes tipos de muestras, lo cual facilita el estudio de las propiedades de los estimadores de los parámetros del modelo.

En este documento se realiza inicialmente una revisión bibliográfica sobre la evolución del concepto de regresión lineal, el cual se presenta como un producto construido a través de la historia y no como un concepto terminado y absoluto. En segunda instancia, se muestran los aspectos disciplinares, resaltando la estimación de los parámetros por mínimos cuadrados, estimación de la varianza, la construcción de intervalos de confianza y un enfoque matricial de la regresión lineal simple. En un tercer momento, se analizan algunos obstáculos epistemológicos y pedagógicos en la enseñanza de la regresión lineal, desde la práctica docente y la interpretación del concepto que le dan los estudiantes. Finalmente se estructura la propuesta para la enseñanza de la regresión lineal, partiendo de una prueba diagnóstico el cual indaga sobre la comprensión de los conceptos previos a la regresión lineal. De esta manera se diseña la una unidad didáctica a partir de talleres que buscan fortalecer la interpretación y un análisis adecuado del diagrama de dispersión, gráficos de residuales, estimación de los parámetros de sus varianzas (teórica y estimada) y de sus distribuciones.

Para la construcción y el diseño de la unidad didáctica se utilizó el software R el cual es uno de los lenguajes más utilizados por la comunidad estadística, siendo además fácil de implementar por ser un software libre y debido a que requiere de elementos de programación, invita a los estudiantes a tomar parte activa en su proceso de aprendizaje, lo cual promueve que al alumno le facilite el cambio de parámetros y muestras, durante los experimentos, para promover un aprendizaje dinámico y participativo.

1. Referente Histórico

En el presente capitulo se realiza una revisión bibliográfica histórica del concepto de regresión lineal simple[1], realizando una síntesis del desarrollo del significado desde su origen con el astrónomo Adophe Lambert hasta su formalización, con los modelos lineales generalizados determinados por los británicos Robert Wedderburn y John Nelder.

Estepa et al (2012) afirma que uno de los personajes más influyentes en el nacimiento de la correlación y la regresión como práctica estadística fue el astrónomo y matemático belga, Lambert Adolphe - Jacques Quetelet (1796-1874), quien señala en uno de sus trabajos la idea del hombre medio, la cual consistió en acumular gran cantidad de datos de hombres, mujeres y niños, para estimar empíricamente las medias y desviaciones típicas de medidas antropométricas las cuales suponía que dependían de varias variables independientes como la edad, sexo, nivel de educación, entre otras. Su originalidad radicó en considerar la dispersión de dichas medidas y haber descubierto que la ley normal ofrecía una descripción aceptable de tal variabilidad. Esta influyente aplicación de la regresión fue publicada por Quetelet en su libro *l´homme et le développement de ses facultés, ou Essai de physique socieale* en 1835 (Estepa et al 2012).

Pero no solo en la biología y en la biometría se observaron avances. Según Hald (2008) en la astronomía el trabajo del francés Augusto Bravais (1811-1863), aportó a la edificación del concepto de correlación, este autor consideró m combinaciones lineales de n errores de distribuciones normales en las medidas de las coordenadas de cuerpos espaciales, Pearson (1857-1956) mostró que Bravais no consideró variables aleatorias correlacionadas, sino consideraba errores independientes.

[1] La regresión lineal tiene como fin representar la asociación entre variables, para cuantificar alguna relación funcional entre estas (Montgomery et al. 2006).

Galton (1822-1911) formalizó el concepto de *correlación y regresión a la media* mediante dos trabajos de variaciones biológicas y herencias, los cuales buscaban dar soporte a la teoría de evolución de Charles Darwin.

De lo anterior Getino (2010) afirma que en su primer experimento Galton distribuye 1875 paquetes de semillas de guisantes a 7 amigos; cada amigo recibió semillas de un tamaño uniforme, los amigos de Galton la sembraban, las cultivaban hasta obtener nuevas semillas y se la devolvían, para que Galton representara gráficamente los pesos de los guisantes generación "hijos" respecto a los pesos de las semillas originales "padre". Galton se dio cuenta que las medias de los pesos de las semillas "hijo" de un tamaño particular de semillas "padre" describía aproximadamente una línea con una pendiente positiva menor que 1. Así dio cuenta de que existía una línea de regresión y una variabilidad constante entre todos los datos.

Según Estepa et al (2012) Galton concluye:

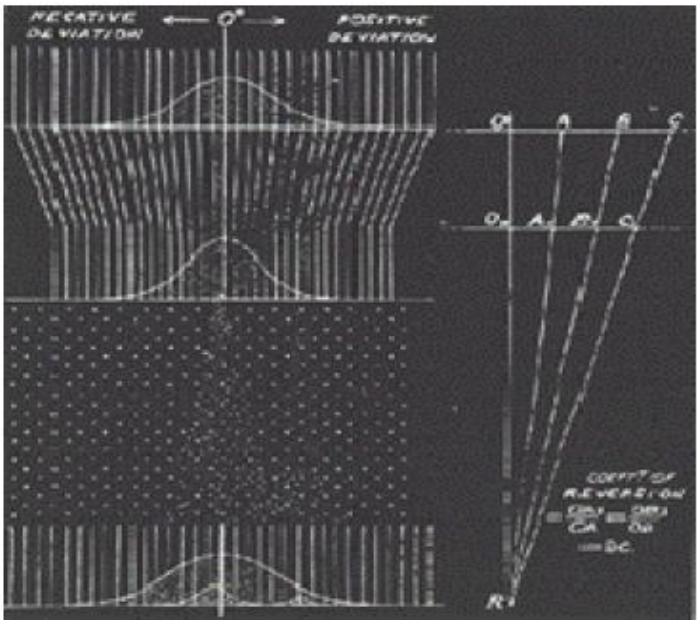
- *El peso medio de las semillas "hijo" es función lineal del peso de las semillas "padres" con una pendiente menor que la unidad, es decir, el peso medio de la progenie se desvía menos de la población media que de los padres. Galton llamó a esta propiedad regresión.*
- *Para cada grupo de semillas "padres", el peso de las semillas "hijas" estaba normalmente distribuido.*
- *La desviación probable del peso de las semillas "hijos" es la misma para todos los grupos y más pequeña que la desviación probable del peso de las semillas padres. Esta es la propiedad que hoy día se conoce como homocedasticidad u homogeneidad de las varianzas.*

Por tanto, los padres de peso $\mathrm{M} + \mathrm{x}$ producen hijos adultos de peso medio $\mathrm{M} + \mathrm{r} \cdot \mathrm{x}$, donde r es la pendiente de la función lineal, la cual esta entre 0 y 1. El peso de los hijos llega ser $\mathrm{M} + \mathrm{r(x)} + \mathrm{y} = \mathrm{M} + \mathrm{z}$ por la variación aleatoria entre hijos del mismo grupo de padres, variación que se observa en el quincunx[2]. La identidad estadística de las dos generaciones significa que $\sigma_x^2 = \sigma_z^2$, en consecuencia $\sigma_x^2 = r^2 \cdot \sigma_x^2 + \sigma_y^2$, o lo que es lo mismo $\sigma_y^2 = \sigma_x^2 \cdot (1 - r^2)$ queda la relación entre la varianza entre grupos

[2] Dispositivo creado por Francis Galton para demostrar el teorema de límite central.

(varianza condicional), la varianza entre el total de la población (varianza marginal) y el coeficiente de reversión (Estepa,1994).

Figura 1-1: Ilustración de Galton de la identidad estadística de dos generaciones



Fuente: Gea (2014)

Galton no había considerado más que la distribución de una medida x, tomada conjuntamente sobre el padre y sus descendencias, pero con posterioridad se da cuenta de la posibilidad de estudiar variaciones conjuntas de medidas biológicas diferentes sobre los mismos individuos. Pearson asegura que hasta ese momento Galton no había alcanzado la idea de correlación negativa.

Casi 10 años después del experimento de guisantes, Galton apoyado por Karl Pearson mide la altura, el peso, fuerza de audición, potencia respiratoria y sentido del color a 205 grupos de padres e hijos, donde concluye que hijos de padres bajos eran ligeramente más altos que ellos, por otro lado, hijos de padres altos eran un poco más bajos que los progenitores, cabe resaltar que Pearson desarrollo el método matemático de asociación entre variables mediante la correlación (Getino, 2010).

En su obra, *Natural Inheritance*, publicada en 1889, Galton muestra un amplio programa de investigación biométrica: estudiar estadísticamente la variabilidad de las medidas físicas de distintas especies, a fin de confirmar matemáticamente el mecanismo de la evolución de Charles Darwin (1809-1882) (Gea, 2014).

Pearson (1857) realizó un estudio con más de 1000 grupos familiares, observando.

$$\text{altura de hijo} = 85\text{cm} + 0{,}5 \cdot \text{altura del padre (cm)},$$

con lo que concluyó que padres altos tienen tendencia a tener hijos que heredan parte de esta estatura, aunque tienen tendencia a acercarse (regresar) a la media. A pesar que Galton define el término de correlación, es Pearson quien llega a una formalización del concepto considerando la correlación negativa. En la memoria titulada "Regresión, herencia y panmixia" de 1896, Pearson formaliza el coeficiente correlacional. En la misma publicación incluye la distribución del coeficiente de correlación poblacional ρ en caso de una distribución normal bivariante

En la diferenciación entre el coeficiente de correlación muestral y poblacional, afirma que r es el estimador más probable de ρ, en concreto anuncia si demostrarlo, que el valor que maximiza la distribución de probabilidad final que ha obtenido para ρ es el coeficiente de correlación muestral, con lo que anticipa el método de estimación de la máxima verosimilitud que posteriormente desarrollará Fisher. (Gomez, 2009)

Uno de los trabajos más relevantes en el proceso de evolución de la regresión lineal simple, la ofrece Gauss (1777-1855) con su trabajo de mínimos cuadrados, que fue utilizado para predecir la posición del "pequeño planeta" Ceres, descubierto por el astrónomo italiano Giuseppe Piazzi. Gauss calculó la órbita de Ceres a partir de pocas observaciones, en donde evidencio que las variaciones de la información se podían representar mediante una curva acampanada, conocida actualmente como distribución de Gauss o campana de Gauss. Además, observó que el valor estimado se podía hallar haciendo mínima una suma de cuadrados de diferencias con los valores particulares de los datos observados. Sprott (s.f)

Gauss publica por primera vez su teoría de los mínimos cuadrados, en el libro "*Theoria Motus Corporum Celestium"*, en el cual explica que al estimar las k cantidades desconocidas $\theta' = (\theta_1, \theta_2, \dots, \theta_k)$ que determinan la órbita, (llamadas parámetros en la estadística), basado en $n > k$ observaciones $Y' = (y_1, y_2, \dots, y_k)$. No es posible medir θ, directamente; solamente es posible medir funciones de θ, $\xi_i = \xi_i(\theta), i = 1, 2, \dots, n$ si ξ puede observarse sin error, $Y = \xi$, entonces después de observar, Y, ξ sería conocido numéricamente. Se sigue que sería necesario solamente resolver las ecuaciones $Y = \xi$ para θ, lo cual es posible puesto que el sistema de las ecuaciones debe ser consistente. Sin embargo, en la práctica no es posible medir ξ sin error, así las ecuaciones son $Y = \xi + e_i$ donde $e' = (e_1, e_2, \dots, e_n)$ son los errores de las observaciones Y (errores observacionales), y Y, ξ y e son vectores $n \cdot 1$ sprott (s.f).

Mayorga y Soto (1988) Afirman que entre 1821 y 1826 Gauss presenta varios escritos bajo el nombre de *"theoria combinationis obsevationum erroribus minimis obnoxiae"* dentro de los cuales se resalta:

- *Al descubrir una desigualdad general aplicable a cualquier distribución de probabilidad continua unimodal alrededor de la moda, abandona el supuesto de normalidad.*
- *La determinación de los estimadores insesgados de los β, que además resultan ser de mínima varianza.*
- *La obtención de la expresión para determinar la suma de cuadrados residuales a partir de los parámetros estimados.*
- *La presentación de un procedimiento para incluir otros parámetros al modelo, sin tener que modificar los ya estimados.*
- *La generación de una forma aproximada para estimar la desviación estándar de los errores y la deducción del error estándar de tal estimador.*

El método de los mínimos cuadrados también fue trabajado en Francia por Adrien-Marie Legendre (1752-1833) quien no fue el primero en utilizarlo, pero si el primero en publicarlo en su libro *Nouvelles Méthodes pour la Determination des orbites des cométes* 1805. Gauss reclamó en 1806, afirmando que hacía 12 años atrás él ya había trabajado este método.

Fisher (1890-1962) contribuye a la regresión lineal por medio del método de estimación de máxima verosimilitud el cual pretende obtener el estimativo del parámetro seleccionado aquel que maximiza la probabilidad de los datos observados. En 1925 Fisher probó que los estimadores eran los mejores asintóticamente normales (bajo ciertas condiciones de regularidad):

$$\hat{\theta} \rightarrow N\left(\theta, I^{-1} - \hat{\theta}\right),$$

donde $I^{-1} - \hat{\theta}$ es la información de Fisher, que representa la mínima varianza y $NI(\theta)$ recoge la información acerca de θ contenido en la muestra (Yánez, 2002).

El trabajo de Fisher fue conocido por Pearson, quien en 1915 le publicó un artículo sobre la distribución muestral del coeficiente de correlación en la revista *Biométrika,* que estaba bajo su dirección. Pearson era reconocido por sus investigaciones en estadística y en genética y por cuenta propia, realizó un estudio cooperativo en sus laboratorios, con el que publicó en 1917 lo que consideraba correcciones al artículo de

Fisher, sin notificárselo previamente. La razón es que Pearson no había entendido el método de máxima verosimilitud que Fisher utilizaba y lo llamó inferencia inversa, algo que este último había evitado deliberadamente. En este tiempo Fisher era un desconocido y se sintió agredido por los comentarios de Pearson (Figueroa 2002).

Una de las técnicas de regresión más utilizadas actualmente en medicina es la regresión logística. La cual estudia el efecto de múltiples variables explicatorias sobre una variable respuesta categórica, ya sea dicotómica o con más de dos categorías en escala ordinal. En este caso no se puede aplicar el modelo de regresión lineal múltiple por no cumplirse el supuesto de continuidad y de distribución normal de la variable dependiente. (Taucher,1999).

Fisher sugiere en 1938 el uso de esta técnica para analizar datos binarios. El término "logit" fue introducido por Berkson (1899-1982) en 1944 para designar esta transformación y sus trabajos popularizaron la utilización de la regresión logística. (Molinero, 2004).

Pero quizás el principal difusor de la regresión logística fue Roxbee (1961-1965) en 1970 con su libro *The Analysis Binary Data*. El cual establece una base sistemática para el análisis de datos binarios y en particular para el estudio de como la probabilidad de éxito depende de variables explicativas. Roxbee (1970).

Uno de los avances más importantes en el campo de la regresión de los últimos años, corresponde a la introducción de los modelos lineales generalizados (GLM), por los estadísticos británicos Nelder (1924-2010) y Wedderburn (1947-1975) en 1972, que unifican toda la teoría existente en cuanto a modelos probit y modelos logísticos, con los modelos lineales basados en la distribución normal. (Molinero, 2004).

Tabla 1-1: Principales Modelos Lineales Generalizados, en términos de variable de respuesta, componente, función del enlace y modelo lineal.

Naturaleza de la variable de repuesta	Sistemático	Aleatorio	Función de enlace	Modelo lineal	
• Numérica cuantitativa	• Numérico • Categórico • Mixto	-Normal	• Identidad	• Regresión Lineal • ANOVA • ANCOVA	ML
• Categórica binaria -No agrupado	• Mixto • Categórico	-Binomial (1) -Bernoulli	• Logit • Logit	• Regresión Logística • Análisis logit	MLG

Naturaleza de la variable de repuesta	Sistemático	Aleatorio	Función de enlace	Modelo lineal	
-Agrupada		-Binomial (n)	• Probit	• Regresión probit	
• Categórica politómica -No agrupado -Agrupado	• Mixto • Categorico	-Multinomial	• Logit Generalizado	• Regresión Logística Multinomial • Regresión logit multinomial	MLG
• Recuento	• Mixto	-Poisson	• Logarítmica	• Regresión de Poisson	MLG
• Frecuencia	• Categórica	-Poisson	• Logaritmica	• Analisis loglineal	MLG

Fuente: López y Ruiz (s,f)

Otra posible generalización consiste en reducir restricciones sobre los errores manteniendo sin embargo la propiedad de Normalidad, es decir, contemplando errores no independientes o heterogéneos. Esta generalización, que permite dotar de estructura a la variabilidad de los errores del modelo, da lugar a los modelos mixtos. (Badiella, 2011)

Los modelos mixtos siguen una estrategia lógica propia de muchos otros tipos de modelos estadísticos por la que se trata de describir la relación entre una variable de respuesta y una o varias “variables explicativas”. Este tipo de modelo es muy frecuente en biología, ecología, ciencias forestales etc. Ya que los datos que se presentan en estas áreas tienen una estructura agregada en distintos niveles que pueden estar ordenados o no jerárquicamente, también se pueden encontrar efectos fijos y aleatorios. (Seoane, 2014).

2. Referente disciplinar

Los conceptos descritos en esta sección son tomados de Montgomery et al (2006), Canavos (1988) y Cardona et al (2013). Aunque no se usan en el capitulo 4 algunos temas de regresion como la estimación por máxima verosimilitud y el enfoque matricial, son incluidos por completez del referente disciplinar.

2.1 Modelo de regresión lineal simple

El estudio de regresión lineal simple es una técnica estadística para indagar y modelar la relación entre variables. Esta técnica tiene numerosas aplicaciones en ingeniería, ciencias físicas, químicas, economía y ciencias biológicas entre otras. Este modelo con un solo regresor x que tiene una relación con una variable de respuesta y, donde la relación es una línea recta.

$$y = \beta_0 + \beta_1 x + \varepsilon\,,$$

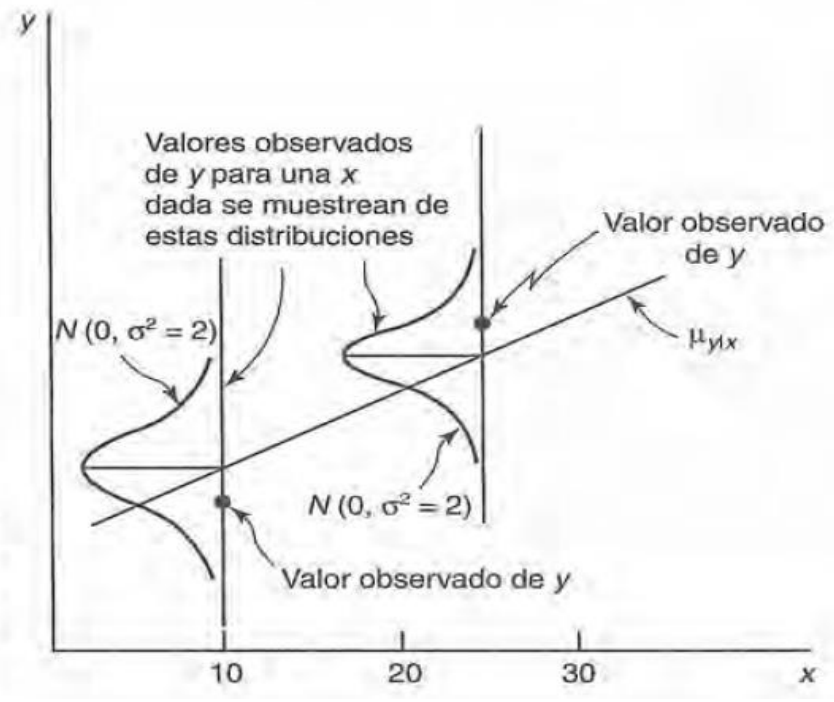
donde la ordenada al origen β_0 es la media de la distribución de la repuesta y cuando $x = 0$ y la pendiente β_1 es el cambio de la media de la distribución de y producida por un cambio unitario de x y ε es un componente aleatorio de error, se supone que los errores tienen promedio cero y varianza σ^2 desconocida.

Se considera y como una variable aleatoria con la que hay una distribución de probabilidad de y para cada valor posible de x. La media y la varianza de esta distribución son.

$$E(y|x) = \beta_0 + \beta_1 x \;\text{ y }\; Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2.$$

Así, el verdadero modelo de regresión $\mu_{y|x} = \beta_0 + \beta_1 x$ es una línea recta de valores promedios, esto es, la altura de la línea de regresión en cualquier valor de x no es más que el valor esperado de y para x.

Figura 2-1: Como se generan observaciones en la regresión.



Fuente: Montgomery et al (2006)

2.2 Estimación de los parámetros por mínimos cuadrados.

Para obtener los estimadores de mínimos cuadrados de los parámetros β_0 y β_1, se generalizará un conjunto de datos consistente en n pares $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, donde los valores de y son las observaciones de la variable aleatoria respuesta. El método de mínimos cuadrados considera la desviación de las observaciones Y_i de su valor medio de y determina los valores de β_0 y β_1 que minimiza la suma de los cuadrados de estas desviaciones. La $i - ésima$ desviación o error es.

$$e_i = y_i - (\beta_0 + \beta_1 x_i).$$

La suma de los cuadrados de los errores es

$$\sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Los estimadores, por mínimos cuadrados, de β_0 y β_1, que se designaran por $\widehat{\beta_0}$ y $\widehat{\beta_1}$, deben satisfacer.

$$\frac{\partial \sum {e_i}^2}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right) = 0\,,$$

y

$$\frac{\partial \sum {e_i}^2}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i\right)x_i = 0\,.$$

Al simplificar y distribuir la suma en estas ecuaciones, se tiene.

$$n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i\,,$$

$$\widehat{\beta_0}\sum_{i=1}^{n} x_i + \widehat{\beta_1}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i\, x_i\,.$$

Las anteriores ecuaciones son llamadas ecuaciones normales de mínimos cuadrados. Su solución nos ofrece.

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}\,,$$

y

$$\widehat{\beta_0} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n}(x_i)^2}{n}}\,.$$

Los valores dados por las anteriores ecuaciones son aquellos que minimizan la suma de cuadrados de los errores.

Dados los estimadores de mínimos cuadrados $\widehat{\beta_0}$ y $\widehat{\beta_1}$ para la intersección y la pendiente, respectivamente, la recta de regresión estimada para el modelo $y = \beta_0 + \beta_1 x + \varepsilon$ es.

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_i\,,$$

donde $\hat{y}_i$ es el estimador para la media de las observaciones y_i la cual corresponde al valor x_i de la variable de predicción, la diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$ se llama residual. Matemáticamente, el $iésimo$ residual es.

$$e_i = y_i - \hat{y}_i = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i\right), \qquad i = 1, 2, \dots, n$$

2.3 Estimación de σ^2

La varianza σ^2 de la variable respuesta es igual a la varianza del error y esta es constante para todos los valores de la variable de predicción. En general, dado que el valor de σ^2 no se conoce, puede obtenerse un estimador de éste a partir de los estimados de mínimos cuadrados $\widehat{\beta_0}$ y $\widehat{\beta_1}$. Dado que cada $\hat{y}_i$ estima la media de $\hat{Y}_i$, la diferencia de $y_i - \hat{y}_i$ representa la desviación de $\hat{Y}_i$ con respecto a su propia media. La suma de los cuadrados de estas diferencias, dividida entre una constante apropiada, es la forma en la que se determina una varianza. Pero estas diferencias son los residuos; por lo tanto, la suma de los cuadrados de los residuales divida entre $n - 2$ es el estimador de σ^2.

Se pierden dos grados de libertad al tener que estimar los dos parámetros de β_0 y β_1 antes de obtener $\hat{y}_i$. El estimador de σ^2 se denota como S^2 y está dado por

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^{n} e_i^2}{n - 2}.$$

El estimador S^2 recibe el nombre de varianza residual la cual es una medida absoluta de qué tan bien se ajusta la recta estimada de regresión a las medidas de las observaciones de la variable de respuesta. Por lo tanto, en general entre más pequeño sea el valor de S^2 es un estimador no sesgado de σ^2 con tal que la forma del modelo de regresión sea la correcta.

Cuando se obtiene una recta de regresión por el método de mínimos cuadrados, surgen algunas propiedades:

$$1. \sum_{i=1}^{n} e_i = 0$$

$$2. \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

$$3.\ \sum_{i=1}^{n} x_i\, e_i = 0.$$

2.4 Pruebas de hipótesis de la pendiente y de la ordenada al origen.

Si se desea probar la hipótesis que la pendiente es igual a una constante las hipótesis correspondientes son.

$$H_0: \beta_1 = \beta_{10}$$

$$H_1: \beta_1 \neq \beta_{10}\ .$$

En donde sea especificado una alternativa bilateral los errores e_i son distribuidos normal e independientemente $(0, \sigma^2)$, las observaciones y_i son distribuidos normal e independientemente $(\beta_1 + \beta_0 x_i,\ \sigma^2)$. Ahora, $\widehat{\beta_1}$ es una combinación lineal de las observaciones, de modo que $\widehat{\beta_1}$ es distribuido normalmente con promedio β_1 y varianza $\frac{\sigma^2}{S_{xx}}$, usando la media y la varianza de $\widehat{\beta_1}$ por consiguiente, el estadístico.

$$Z_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \text{ donde } S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} (x_i)^2}{n}\ ,$$

se distribuye $N(0,1)$ si es cierta la hipótesis nula $H_0: \beta_1 = \beta_{10}$. Si se conociera σ^2se podría usar Z_0. Comúnmente se desconoce σ^2, ya se ha vistos que MS_{Res} es un estimador insesgado de σ^2.

$$t_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \text{ donde } MS_{Res} = \frac{\sum_{i=1}^{n} y_i^2}{n-2}\ .$$

Sigue una distribución t_{n-2} si es cierta la hipótesis nula $H_0: \beta_1 = \beta_{10}$. La cantidad de grados de libertad con t_0 es igual a la cantidad de grados de libertad asociados con MS_{Res}. Así, la razón t_0 es el estadístico con que se prueba $H_0: \beta_1 = \beta_{10}$. El procedimiento de prueba calcula t_0 y compara su valor observado de acuerdo con $t_0 = \frac{\widehat{\beta_1} - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$ con el punto porcentual de $\frac{\alpha}{2}$ superior de t_{n-2} de la distribución $(t_{n-2,n-2})$ este procedimiento rechaza la hipótesis nula sí.

$$|t_0| > t_{n-2,n-2}\ .$$

El denominador del estadístico t_0 se llama con frecuencia error estándar estimado o error estándar de la pendiente, esto es:

$$se(\widehat{\beta_1}) = \sqrt{MS_{Res}/S_{xx}}\,.$$

por lo anterior, se ve con frecuencia a $t_0$escrito de la forma.

$$t_0 = \frac{\widehat{\beta_1} - \beta_{10}}{se(\widehat{\beta_1})}.$$

Para probar la hipótesis acerca de la ordenada al origen.

$$H_0: \beta_0 = \beta_{00}$$

$$H_1: \beta_0 \neq \beta_{00}\,.$$

Se podría usar el estadístico de prueba.

$$t_0 = \frac{\widehat{\beta_1} - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}},$$

en donde $se(\widehat{\beta_0}) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$ es el error estándar de la ordenada al origen. La hipótesis nula $H_0: \beta_0 = \beta_{00}$ se rechaza si $|t_0| > t_{n-2,n-2}$

2.5 Estimación de intervalo en la regresión lineal simple

Además de las estimados puntuales β_0, β_1 y σ^2, se pueden obtener estimaciones por medio de intervalo de confianza para esos parámetros. El ancho de dichos intervalos es una medida de calidad general de la recta de regresión. Si los errores se distribuyen en forma normal e independiente, entonces la distribución de muestreo tanto de.

$$\frac{\widehat{\beta_1}-\beta_1}{se(\widehat{\beta_1})} \qquad \frac{\widehat{\beta_0}-\beta_0}{se(\widehat{\beta_0})}.$$

Es t, con $n-2$ grados de libertad. Así un intervalo de confianza de $100(1-\alpha)$ por ciento para la pendiente β_1 se determina.

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}se(\hat{\beta}_1),$$

y un intervalo de confianza de $100(1-\alpha)$ por ciento para la ordenada de origen β_0 es.

$$\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2} se(\hat{\beta}_0) .$$

Un intervalo de confianza de $100(1-\alpha)$ por ciento para σ^2 es.

$$\left[\frac{(n-2)MS_{res}}{\chi^2_{\frac{\alpha}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{res}}{\chi^2_{1-\frac{\alpha}{2},n-2}} \right].$$

2.6 Estimación por máxima verosimilitud.

Se tiene los datos (y_i, x_i), $i = 1,2,\dots,n$ si se supone que los errores en el modelo de regresión son distribuidos normal e independientemente, las observaciones y_i en estas muestras son variables aleatorias normal e independientemente distribuidas con promedio $\beta_1 + \beta_0 x_i$, y varianza σ^2, la función de verosimilitud o de posibilidades se determina con la distribución conjunta de las observaciones.

$$L(y_i, x_i, \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(y_i - \beta_0 - \beta_1 x_i \right)^2 \right]$$

$$= (2\pi\sigma^2)^{\frac{-n}{2}} \exp\left[-\frac{1}{2\sigma^2} \left(y_i - \beta_0 - \beta_1 x_i \right)^2 \right].$$

Los estimadores de máxima posibilidad son los valores que pueden tomar los parámetros, por ejemplo $\widehat{\beta_0}, \widehat{\beta_1}$ y $\hat{\sigma}^2$ que maximizan a L, o lo que es lo mismo, a $ln\, L$. Así,

$$\ln L(y_i, x_i, \beta_0, \beta_1, \sigma^2) = -\left(\frac{n}{2}\right) \ln 2\pi - \left(\frac{n}{2}\right) \ln \sigma^2 = -\left(\frac{1}{2\sigma^2}\right) \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2 .$$

Los estimadores de posibilidad máxima, $\widehat{\beta_0}, \widehat{\beta_1}$ y $\hat{\sigma}^2$ deben satisfacer.

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right) = 0 ,$$

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right) x_i = 0 ,$$

y

$$\frac{\partial \ln L}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{n}{2\hat{\sigma}^4}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 .$$

La solución de la ecuación anterior determina los estimadores de máxima posibilidad:

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 x ,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n},$$

Los estimadores de máxima posibilidad de la ordenada al origen y la pendiente $\hat{\beta}_0$ y $\hat{\beta}_1$, son idénticos a los obtenidos con los mínimos cuadrados.

2.7 Análisis de los residuales

Montgomery et al (2006). Un residual es la desviación entre los datos y el ajuste, la cual también funciona como una medida de variabilidad de la variable de respuesta que no explica el modelo de regresión. El análisis de los residuales es una forma eficaz de descubrir diversos tipos de inadecuación del modelo.

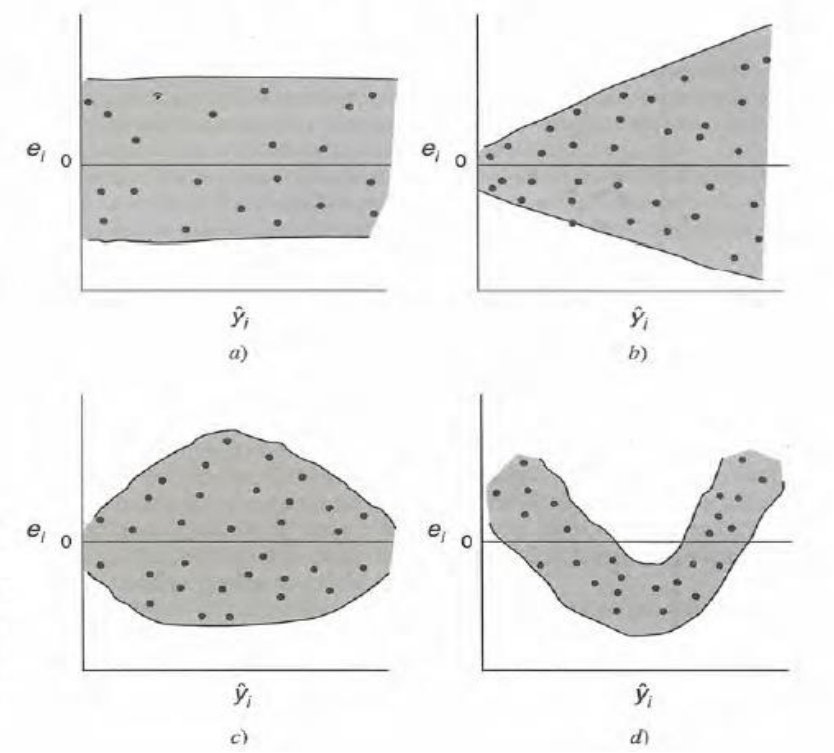
Los residuales tienen varias propiedades importantes. Tienen media cero y su varianza promedio aproximada se estima con:

$$\frac{\sum_{i=1}^{n}(e_i - \hat{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res} .$$

Los residuales no son independientes, ya que los n residuales solo tienen $n-p$ grados de libertad asociados a ellos.

El análisis gráfico de los residuales es una forma muy efectiva de investigar la adecuación del ajuste de un modelo de regresión. La gráfica de residuales en función de x es usada para detectar algunos tipos frecuentes de inecuación del modelo, si los residuales se pueden encerrar en una banda horizontal entonces indica que el modelo es adecuado.

Figura 3-2: Patrones en las gráficas de residuales



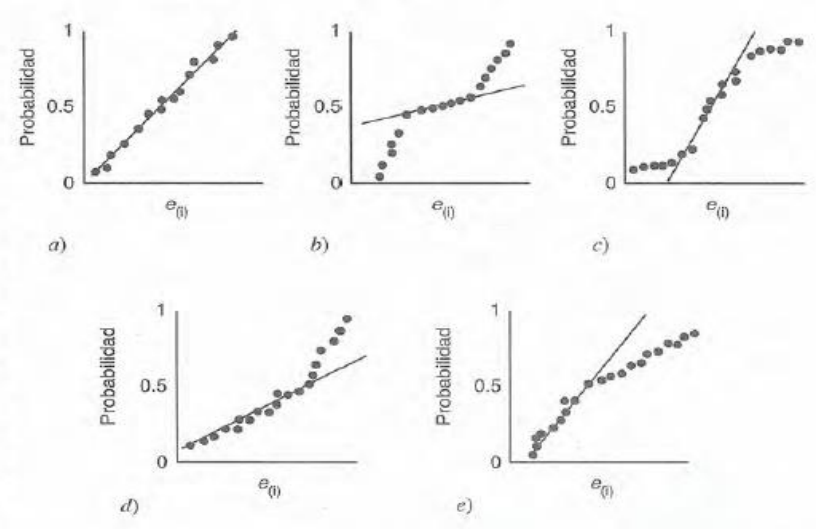
Fuente: Montgomery et al (2006)

La gráfica b indica que la varianza de los errores no es constante y que la varianza es función creciente de y, la gráfica c presenta una distribución de doble arco, esto sucede cuando y es una proporcion entre 0 y 1, la última curva se concluirá que el modelo lineal no representa adecuadamente la relación entre las variables y entonces se pensaria en modelos curvilíneos o de regresión multiple.

La mayoria de las gráficas de residuales que se obtienen con el uso de programas estadísticos muestran una version estandarizada de los residuales. Estandarizar una variable aleatoria significa restarle su media y dividir el resultado entre su desviación estandar.

Un método muy sencillo de comprobar la suposición de normalidad es trazar una gráfica de probabilidad normal de los residuales. Es una gráfica diseñada para que al graficarse la distribución normal acumulada parezca una linea recta. Sean $e_1 < e_1 < \cdots < e_n$ los residuales ordenados en orden creciente. Si se grafican e_i en función de probabilidad acumulada $P_l = \frac{\left(i-\frac{1}{2}\right)}{n}, i = 1,2, \ldots, n$ en el papel de la probabilidad normal, los puntos que resulten deberian estar aproximadamente sobre una linea recta.

Figura 3-3: Patrones en las gráficas de residuales



Fuente: Montgomery et al (2006)

La gráfica (a) representa una gráfica de probabilidad normal ideal, las gráficas (b) y (c) representan distribuciones con cola gruesas y delgadas respectivamente, (d) representa una distribución asimétrica positiva y la (e) una asimétrica negativa.

2.8 Enfoque matricial modelo lineal simple

El uso del álgebra de matrices proporciona un medio conveniente para el análisis de regresión de modelos lineales, en forma especial de aquellos que contienen más de una variable de predicción.

Para los n pares $(x_1, Y_1),\ (x_1, Y_1), \dots (x_n, Y_n)$, el siguiente modelo lineal simple.

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, \qquad i = 1,2, \dots n$$

Es decir

$$Y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$Y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = \beta_0 + \beta_1 x_n + \varepsilon_n\,.$$

Son n ecuaciones lineales para las que $Y_1, Y_2, \dots Y_n$ son las observaciones de la respuesta para los correspondientes valores fijos $x_1, x_2, \dots x_n$ de la variable de predicción, $\varepsilon_1, \varepsilon_2, \dots \varepsilon_n$ son los errores aleatorios no observables y β_0 y β_1 son los parámetros a estimar. Se definen las matrices.

$$Y = \begin{bmatrix} Y_1 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix} X = \begin{bmatrix} 1\ x_1 \\ 1\ x_2 \\ \vdots \\ 1\ x_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Entonces

$$\begin{bmatrix} Y_1 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1\ x_1 \\ 1\ x_2 \\ \vdots \\ 1\ x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_2 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}.$$

Como resultado se tiene que el modelo lineal simple puede expresarse en la notación de matrices.

$$Y = X\beta + \varepsilon$$

Si se supone el caso de la teoría normal, entonces ε es un vector de variables aleatorias normales tales que.

$$E(\varepsilon) = 0$$

$$Var(\varepsilon) = \sigma^2 I\ ,$$

donde σ^2 es la varianza del error, común a todos, e $\boldsymbol{I}$ es la matriz de identidad correspondiente.

$$X'X = \begin{bmatrix} 1 & 1 & \dots \\ x_1 & x_2 & \dots \end{bmatrix} \begin{bmatrix} 1\ x_1 \\ 1\ x_2 \\ \vdots \\ 1\ x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \dots \\ x_1 & x_2 & \dots \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_I \\ \sum X_I Y_I \end{bmatrix}.$$

Entonces

$$(X'X)\beta = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 n & \beta_1 \sum x_i \\ \beta_0 \sum x_i & \beta_1 \sum x_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_I \\ \sum X_I Y_I \end{bmatrix} = X'Y.$$

Donde

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Es el vector que contiene los estimadores para obtener los estimadores de mínimos cuadrados de β_0 y β_1

$$(X'X)^{-1}(X'X)\beta = (X'X)^{-1}X'Y$$

$$I\beta = (X'X)^{-1}X'Y$$

$$\beta = (X'X)^{-1}X'Y\,.$$

La anterior es la expresión matricial para obtener los estimadores de mínimos cuadrados β_0 y β_1.

3. Referente pedagógico.

La unidad didáctica diseñada tuvo como apoyo el análisis de algunos planes de estudio de universidades públicas y privadas ubicadas en la ciudad de Bogotá. para esto se tuvo en cuenta los programas de ingeniería en las cuales se incluía el concepto de regresión lineal en su curso de estadística. Lo anterior con el objetivo de identificar los pre-requisitos y la bibliografía utilizada por estas universidades para enseñar la regresión lineal simple

Se revisó la información anteriormente mencionada en los syllabus[3] de los cursos de estadística y probabilidad e inferencia, de 4 instituciones de educación superior: Universidad Libre de Colombia, Fundación Universitaria Agraria de Colombia, Universidad Distrital Francisco José de Caldas y la Fundación Universitaria los Libertadores. De dicha revisión se encontró:

- En la Universidad Distrital Francisco José de Caldas presenta el concepto de regresión lineal después de estudiar los conceptos básicos de estadística descriptiva, teoría de probabilidad (incluyendo modelos de distribución discretos y continuos) y teorías de estimación. La sección de análisis de regresión y correlación se estudia en la última semana académica abordando la regresión por el método de mínimos cuadrados.
- En la Facultad de Ingeniería de la universidad Libre de Colombia, en el plan académico se imparten las asignaturas de estadística I y estadística inferencial. La regresión lineal es abordada en la asignatura de estadística inferencial después de los temas de pruebas de hipótesis y análisis de varianza. Cuando se estudia la regresión y la correlación inicia con los temas de mínimos

[3] Los programas de los cursos de estadística I y II se descargaron de la página oficial de cada uno de las universidades mencionadas

cuadrados, regresión lineal (supuestos del modelo, error estándar, análisis de correlación, intervalos de confianza, análisis de varianza) y su aplicación en Excel y R.

- Las fundaciones universitarias UNIAGRARIA y los Libertadores abordan la regresión lineal en la asignatura de estadística II, garantizando los temas de distribución muéstrales, la inferencia estadística, las pruebas de hipótesis y el análisis de correlación.

De lo anterior se puede concluir que en general se aborda el concepto de regresión lineal contemplando como temas específicos el análisis de residuales, prueba de hipótesis sobre los parámetros y el análisis de la tabla ANOVA. Todo esto con el objetivo común de identificar los escenarios de la aplicación de la regresión.

Los textos más citados por estas universidades que sirven de apoyo bibliográfico son:

- Walpole, R., Myers, R. & Myers, S. (2007). Probabilidad y Estadística para Ingeniería y Ciencias. Grupo editorial Person Educación. México.
- Freund, E. & Simon, G. (1994). Estadística Elemental. Grupo editorial Person Educación. México.
- Mendenhall, W., Beaver, R. & Beaver, B. (2010) Introducción a la Probabilidad y Estadística. Grupo Editorial Iberoamérica. México

El texto “Probabilidad y Estadística para Ingeniería y Ciencias” de Walpole et al. (2007) inicia el capítulo de regresión lineal destacando la necesidad de encontrar relaciones entre variables, para hallar un método de pronóstico. Este acercamiento se realiza por medio de un ejemplo en donde se encuentra la relación existente entre la reducción de sólidos y la demanda química de oxígeno en 50 observaciones y con el mismo ejemplo aborda el método de estimación por mínimos cuadrados e intervalos de confianza para los parámetros.

El texto “Estadística Elemental” de Freund & Simon (1994) define el ajuste de curva, por el método de mínimos cuadrados. Dichos conceptos son aclarados mediante ejemplos en diferentes contextos. En la sección de análisis de regresión se abordan las inferencias acerca de los coeficientes de regresión, incluyendo intervalos de confianza e intervalos de pronósticos.

El capítulo de regresión lineal y correlación del libro “Introducción a la Probabilidad y Estadística” de Mendenhall et al (2010), inicia con la explicación del modelo lineal simple, del método de mínimos cuadrados y del análisis de varianza. Estos conceptos son aplicados a un ejemplo cuyo contexto se centra en la calificación de un examen de matemáticas y nota final en cálculo, para 10 estudiantes de primer año de universidad. Dichos datos son graficados en un diagrama de dispersión donde se observa la tendencia lineal de los datos, además, explica cómo hacer cálculos de regresión por medio del programa *“MINITAB”*.

Batanero et al (1994) afirma que los errores y dificultades de los estudiantes en el aprendizaje de la estadística, no se presentan de un modo aleatorio con frecuencia es posible encontrar regularidades, ciertas asociaciones con variables propias de las tareas propuestas, de los estudiantes o de las circunstancias presentes o pasadas. Estos obstáculos son identificados como ontogénicos, didácticos y epistemológicos.

Teniendo en cuenta lo anterior se identifican estos tres obstáculos, en la adquisición del concepto de regresión lineal simple en estudiantes de pregrado, estos, se ven reflejados en las características del concepto que presenta el alumno.

3.1 Diagramas de dispersión.

Un primer acercamiento que se hace al estudio de la regresión se realiza mediante la interpretación de los diagramas de dispersión. Ya que es el instrumento gráfico más usado, para observar la relación entre dos variables. Con él se puede establecer una interpretación del coeficiente de correlación o modelo de regresión.

En este sentido se expresa Estepa (2008) indicando que la mayoría de los estudiantes no tienen inconvenientes en la lectura de las coordenadas en los puntos de los diagramas de dispersión, pero si existe una dificultad moderada en la asociación del coeficiente de correlación a un diagrama de dispersión. La interpretación de correlación es más sencilla para el estudiante cuando los puntos del diagrama de dispersión, tienen una tendencia evidente.

Teniendo en cuenta lo anterior Gea (2014) afirma que el alumno se limita a corroborar la asociación según un subconjunto de datos, obviando la tendencia global de los mismos, además, se les dificulta comprender que el coeficiente de correlación no tiene propiedad transitiva; es decir, dos variables A y B pueden estar correlacionadas, así como B y C sin estar relacionadas las variables A y C.

3.2 Función lineal como regresión lineal.

En el aprendizaje de la regresión lineal podemos encontrar dificultades que no dependen de la falta del manejo de conceptos estadísticos del estudiante. Por ejemplo, confundir la regresión lineal con la función lineal, este error puede derivar de las practicas docente inadecuadas en el aula de clase, debido a que posiblemente el único modelo utilizado para realizar la introducción del tema de regresión, es la comparación de la función lineal con la regresión lineal como recurso didáctico.

Teniendo en cuenta lo anterior Agnelli et al (2009) considera que la regresión lineal debe ser abordada mostrando los límites de aplicación de la función lineal, para ello se debe clarificar el uso operacional que tiene la función lineal en el marco estadístico y evidenciar el empleo que tiene el concepto determinístico cuando es usado para la modelización de fenómenos aleatorios.

Algunas dificultades que se pueden evidenciar al no marcar el límite entre una función lineal y la regresión lineal son:

- Cuando se presentan situaciones de regresión cuyos valores se muestran en una tabla de la misma forma que se presenta para graficar una función, se tiende a confundir la correspondencia unívoca y la dependencia casual.
- Al presentar la ecuación de la recta de regresión y asociarla a un conjunto de datos, no se puede reproducir la tabla a partir de la expresión analítica, esto se debe a la posibilidad que no exista una relación funcional entre los datos y se presenta por medio de una función.

Con respecto a la dificultad anteriormente mencionada, Agnelli et al (2009) establece las siguientes diferencias entre funcion lineal y regresión lineal.

Tabla 3-1: Diferencias entre función lineal y regresión lineal

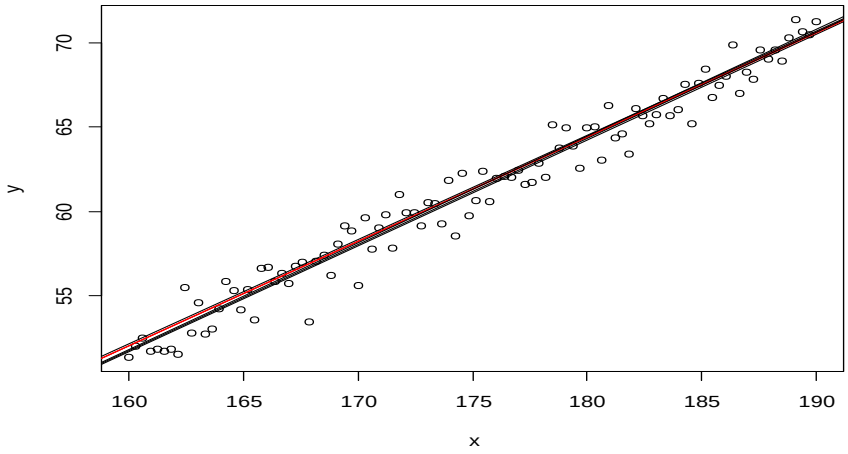
Función lineal	Regresión lineal
La variación de x e y, es determinista	La variación de x e y, es aleatoria
El cálculo de los parámetros es independiente de los puntos seleccionados sobre la recta.	La estimación de los parámetros depende de los puntos seleccionados.
Dada una tabla de valores x e y se obtiene una recta y dada esta es posible reproducir la tabla	Dada la tabla de los valores x e y se obtiene una recta, pero ella no permite reproducir la tabla

3.3 Simulación como recurso didáctico

La enseñanza y el aprendizaje de la estadística se han visto influenciadas en las últimas décadas por el uso de computadores. Su capacidad de almacenar y manejar datos junto a la facilidad para representarlos gráficamente, hacen que sea una herramienta atractiva para la práctica pedagógica de la estadística (Corredor 2000).

Actualmente se cuenta con varias investigaciones que resaltan los beneficios de la simulación como recurso didáctico en el aula, entre ellas se encuentra la de Mills (2002) el cual afirma que la simulación permite a los estudiantes, experimentar con muestras aleatorias de una población con parámetros conocidos con el propósito de aclarar conceptos abstractos difíciles y los teoremas estadísticos. Por ejemplo, se puede demostrar la distribución muestral de la pendiente de la regresión lineal, al generar 20 o más líneas de regresiones de una población, para compararla con su modelo poblacional mediante un gráfico.

Figura 3-1: Simulación de 20 líneas de regresión.



Ejercicios como el anterior permiten enseñar el concepto de normalidad, varianza de los estimadores, varianza estimada de los estimadores, distribución de los estimadores, varianza de la estimación.

Cabe resaltar que la aplicación de la simulación como recurso didáctico, se muestra en varios temas de la estadística con resultados satisfactorios en la apropiación del concepto. Corredor (2000) concluye que las ganancias en el conocimiento probabilístico son más fuertes para los estudiantes que trabajan por medio de simulación que para los estudiantes expuestos a ejercicios de análisis de datos.

La simulación permite condensar el experimento en el tiempo, en el espacio y operar con el experimento simulado para obtener conclusiones válidas para el experimento original. Además, proporciona un método “universal” para obtener una estimación de la solución de los problemas probabilísticos. La simulación cobra un papel importante ya que ayuda al estudiante, conocer las diferencias entre probabilidad experimental y la teórica (Batanero 2000).

La International Association for Statistical Education (IASE) reconoce los recursos tecnológicos como herramientas que implican cambios metodológicos en la enseñanza de la estadística, cuyo efecto se ve reflejado en el aprendizaje del estudiante. Actualmente son numerosos los simuladores y software estadísticos que se utilizan como mediadores didácticos en el proceso de enseñanza-aprendizaje de la estadística. Los cuales tiene gran acogida por ser software´s libres

Salas (2008) establece las características que deberían tenerse en cuenta a la hora de elegir un programa estadístico como mediador didáctico.

Tabla 3-2: Cuadro comparativo de los softwares estadísticos más usados en términos de la facilidad de manejo y calidad de los gráficos.

Aspectos	***programas estadísticos***		
	SPSS	***SAS***	***R***
Amigabilidad con el usuario	Permite acceder a todas sus opciones mediante un menú de funciones	Requiere de conocer sintaxis del programa	Requiere de conocer sintaxis y comandos de programación propios del programa
Manipulación de datos	Trabaja físicamente sobre el archivo, modificándolo cada vez que se realice un análisis	Emplea sintaxis y procedimientos que afectan a los datos en una memoria virtual, cualquier modificación de los datos, no es guardada físicamente en el archivo.	Emplea sintaxis y procedimientos que afectan a los datos en una memoria virtual, cualquier modificación de los datos, no es guardada físicamente en el archivo.
Calidad de Gráficos	Ofrece gráficos difíciles de personalizar	Requiere de diferentes paquetes, la sintaxis	Requiere de una sintaxis sencilla y no

Aspectos	***programas estadísticos***		
	SPSS	***SAS***	***R***
		para gráficas se puede considerar complicada	necesita gran variedad de paquetes, la calidad visual es excelente.
Variedad de análisis estadístico	Ofrece un amplio rango de procedimientos estadísticos, sin embargo, cuando se requiere de procedimientos específicos (la necesidad de definir el rango posible de valores para los parámetros de un modelo no lineal) poca versatilidad	Ofrece un rango más amplio en el análisis estadístico, esto se debe a que la empresa invierte cerca del 20% de sus utilidades en investigación e innovación del software	Por ser un software desarrollado por estadísticos que trabajan en diferentes instituciones a nivel mundial, ofrece algoritmos modernos y robustos, además, de varios paquetes continuamente desarrollados.

Teniendo en cuenta la tabla 2.2 se trabajará el programa R para este estudio, como instrumento mediador computacional para la enseñanza de la regresión lineal simple, en específico la simulación como método de estudio.

4. Unidad didáctica

Zabala (2000) afirma que una secuencia didáctica es un conjunto de actividades ordenadas, estructuradas y articuladas para la consecución de unos objetivos educativos que tienen un principio y un final conocido tanto para el profesorado como por el alumnado. De igual forma Montserrat (2010) define la secuencia didáctica como la articulación de diversas actividades de enseñanza y aprendizaje para conseguir un determinado objetivo.

Teniendo en cuenta la definición de secuencia didáctica, nos apoyamos para su realización en un taller diagnóstico que busca, establecer el nivel real de los estudiantes en cuanto a los conceptos estadísticos previos de la regresión lineal, para detectar posibles errores que puedan dificultar el logro de los objetivos planteados.

4.1 Taller diagnóstico conceptos previos a la regresión lineal

Para el diseño de la unidad didáctica que se presenta en este trabajo, además del estudio y contrastación de la simulación como recurso didáctico, en la regresión lineal simple, se tuvieron en cuenta los resultados y el análisis del taller diagnóstico “conceptos previos a la regresión (Anexo 1)” a un grupo de estudiantes de ingeniería inscritos a un curso de estadística inferencial.

La prueba está compuesta por 15 preguntas que indagan sobre los conocimientos de los estudiantes respecto a medidas de posición, medidas de dispersión, interpretación gráfica, probabilidad y estadística inferencial. Para medir estos conceptos en el estudiante, el taller diagnostico se realizó a través de un contexto cotidiano, donde se simuló (Anexo 2) el peso y la estatura de 210 estudiantes mujeres de la facultad de ingeniería de la Universidad Nacional de Colombia.

Se aplicó a 15 estudiantes de la fundación universitaria los libertadores de Bogotá D.C y su objetivo era identificar las habilidades de los estudiantes para analizar y argumentar gráficas estadísticas, probabilidad condicional e inferencia.

Para el análisis de la prueba diagnóstica, las preguntas se clasificaron en 4 niveles teniendo en cuenta la temática abordada por el curso de estadística y probabilidad, el cual es prerrequisito para abordar el tema de regresión lineal.

- **Análisis descriptivo:** Nivel compuesto por tres preguntas que indagan en el estudiante la identificación de los conceptos de población, muestra, variable aleatoria, medidas de centralización, localización y dispersión, en un contexto determinado.
- **Representación gráfica:** Exige análisis, elaboración e interpretación de graficas estadísticas tales como diagramas de dispersión, box-plot e histogramas.
- **Probabilidad:** Analiza la interpretación del concepto de probabilidad usando la distribución normal.
- **Inferencia:** requiere una elaboración más compleja para su planteamiento y solución donde se propone justificar o refutar inferencia basadas en razonamiento estadístico a partir de los resultados del ejercicio propuesto.

A continuación, se presenta la clasificación y el análisis de cada pregunta, todas ellas referidas a la información de la siguiente tabla y su contextualización, los datos de la tabla fueron generados mediante simulación en R (Anexos 2).

En la Facultad de ingeniería de la Universidad Nacional de Colombia se les preguntó a 210 estudiantes mujeres, sobre su peso y estatura, con el fin de analizar el porcentaje de mujeres con bajo nivel de peso.

La siguiente tabla muestra los datos de 210 mujeres ordenadas de acuerdo a su estatura (menor a mayor).

Tabla 4-1: Datos del peso 210 de mujeres según su estatura. Generados por simulación en R

Estatura [mts]	peso [Kg]										$\mu_{peso\|esta}$
1,6	51.11	52.21	51.48	51.48	53.63	51.03	53.08	52.14	52.32	52.57	51.20
	52.02	52.59	50.61	52.78	52.66	53.94	53.28	50.73	51.70	51.75	
	54.18	51.97	51.09	51.97	51.91	51.81	52.44	52.16	53.55	49.91	
1,65	55.18	56.35	53.78	55.68	53.43	56.34	57.29	56.32	55.07	54.54	53.79
	52.71	54.36	56.01	54.88	55.35	54.63	55.20	55.00	54.09	53.68	
	54.87	54.24	54.64	56.53	56.27	54.03	55.17	54.55	55.31	57.44	
1,7	56.74	58.98	57.22	57.29	58.41	56.74	57.63	58.49	57.89	58.15	57.80
	58.85	58.82	59.32	56.89	57.90	60.66	58.09	59.14	58.36	58.04	
	58.65	60.63	56.86	58.46	58.47	57.48	57.11	57.83	56.18	58.80	
1,75	62.58	60.49	62.60	61.07	61.62	60.38	61.44	61.60	60.51	62.31	60.55
	61.12	61.31	62.43	61.18	61.38	62.40	62.53	61.63	59.85	61.96	
	60.43	62.58	60.38	60.78	61.05	61.16	61.37	61.52	62.60	59.66	
1,8	64.48	63.52	62.63	62.31	65.01	64.10	64.98	65.60	65.41	63.15	64.80
	63.03	64.37	66.51	64.96	62.52	64.08	64.89	63.83	62.88	62.72	
	65.61	63.55	62.33	65.40	64.70	66.17	65.68	65.25	64.46	62.76	
1,85	66.10	68.25	69.45	68.73	67.44	68.30	69.45	66.31	68.27	66.55	67,71
	66.99	68.53	68.24	67.86	68.05	66.41	66.67	70.86	68.18	66.75	
	67.12	68.93	66.24	67.22	69.20	68.31	68.45	68.66	68.80	67.05	
1,9	70.59	71.55	69.97	71.00	68.68	70.14	71.25	69.44	71.11	71.80	72.20
	70.72	69.67	69.88	69.92	70.60	71.49	69.28	73.11	70.23	69.84	
	69.93	70.76	69.67	70.49	71.38	70.99	69.14	69.39	71.29	70.37	

$\mu_{est|peso}$: Promedio de peso ideal de mujeres de acuerdo a su estatura

Sección 1: Análisis descriptivo

1. ***Con la información anterior responda.***

A. ¿Cuál es la población de interés?

B. ¿Cuál es el tamaño de la muestra para cada estatura?

C. ¿Cuáles son las variables consideradas?

Esta pregunta indaga por la población y tipos de variables consideradas en el ejercicio, 11 estudiantes afirman que existen 7 poblaciones una por cada estatura dada, mientras que 4 estudiantes coinciden en que la población de interés es el peso de las estudiantes de ingeniería de la Universidad Nacional de Colombia.

2. Completa la siguiente tabla.

	1,60	***1,65***	***1,70***	***1,75***	***1,80***	***1,85***	***1,90***
Mínimo							
Cuartil (Q1)							
Mediana							
Media							
Cuartil (Q3)							
Máximo							
Desviación estándar							
Coeficiente de variación							

El objetivo de esta pregunta busca el uso compresivo de las diferentes medidas de posición, centralización y dispersión. Se pudo evidenciar que el 100% de los estudiantes identifican y hallan correctamente las diferentes medidas de posición y de dispersión.

3. Interprete el coeficiente de variación.

Esta pregunta indaga por la interpretación del coeficiente de variación como una medida de dispersión, que describe la cantidad de la variabilidad con relación a la media de cada peso teniendo en cuenta la estatura.

Los 15 estudiantes afirman que los datos muestran una variación baja, es decir, están agrupados respecto a la media, son homogéneos.

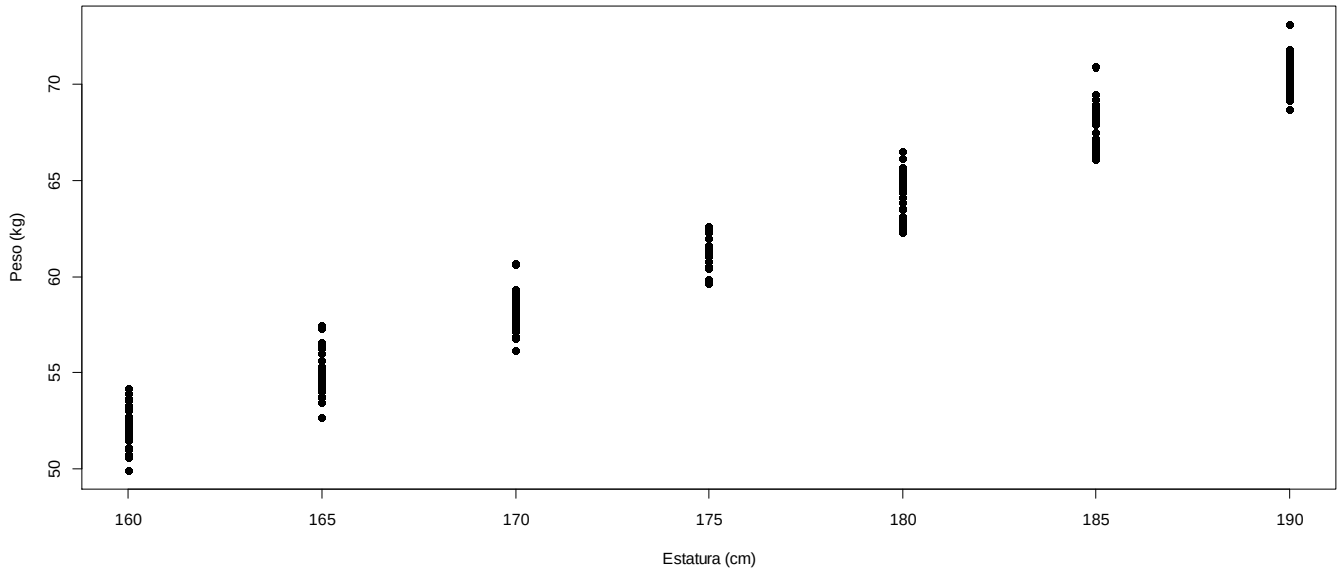
Sección 2: Representación estadística.

4. ***Realice un gráfico de dispersión.***
5. ***¿Qué puede concluir del anterior diagrama? Interprete la relación entre las dos variables***

La pregunta 4 y 5 se realizó con el fin de comprobar la construcción apropiada del diagrama de dispersión y la posible relación entre las variables. Los 15 estudiantes concuerdan en que exististe una relación lineal, entre las variables peso y estatura, además, afirman que la relación lineal es directamente proporcional, esto se podría asumir por la forma ascendente de la línea recta.

6. ***Identifique la media del peso por estatura en la gráfica y trace una línea recta por cada una de ellas.***

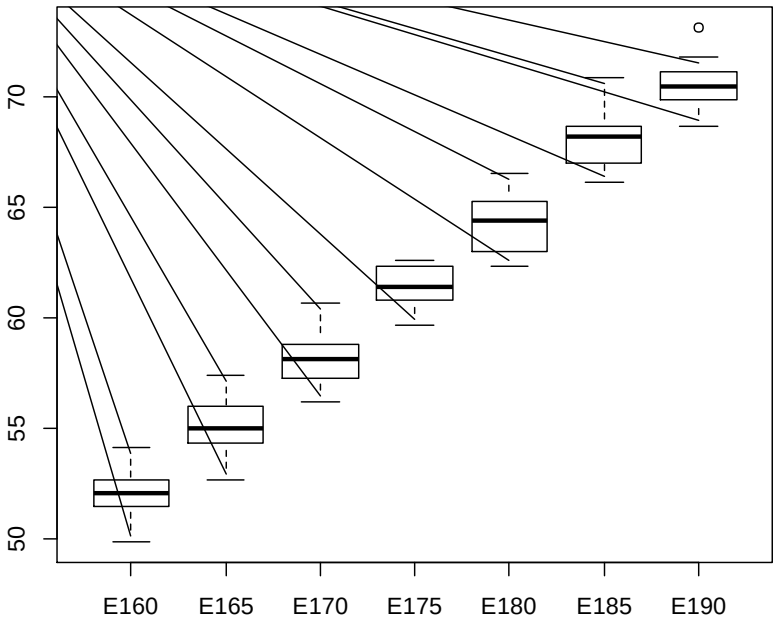
Figura 4-1: Diagrama de dispersión generado en R. Datos simulados



Se analiza en este ítem la ubicación de la media de cada peso respecto a la estatura, además, de trazar la línea recta que pasa por cada una de las medias aritméticas. Se evidencio que el 100% de los estudiantes no tuvo dificultad para el desarrollo de este punto

De acuerdo con el box plot del peso de las 70 mujeres encuestadas conteste las preguntas 7 y 8

Figura 4-2: Box-plot de los pesos de 210 mujeres. Datos simulados.



7. ¿Hay evidencia de datos atípicos? Justifique

8. Haga un comparativo del valor de la mediana y de los cuartiles Q_1 y Q_3. Concluya

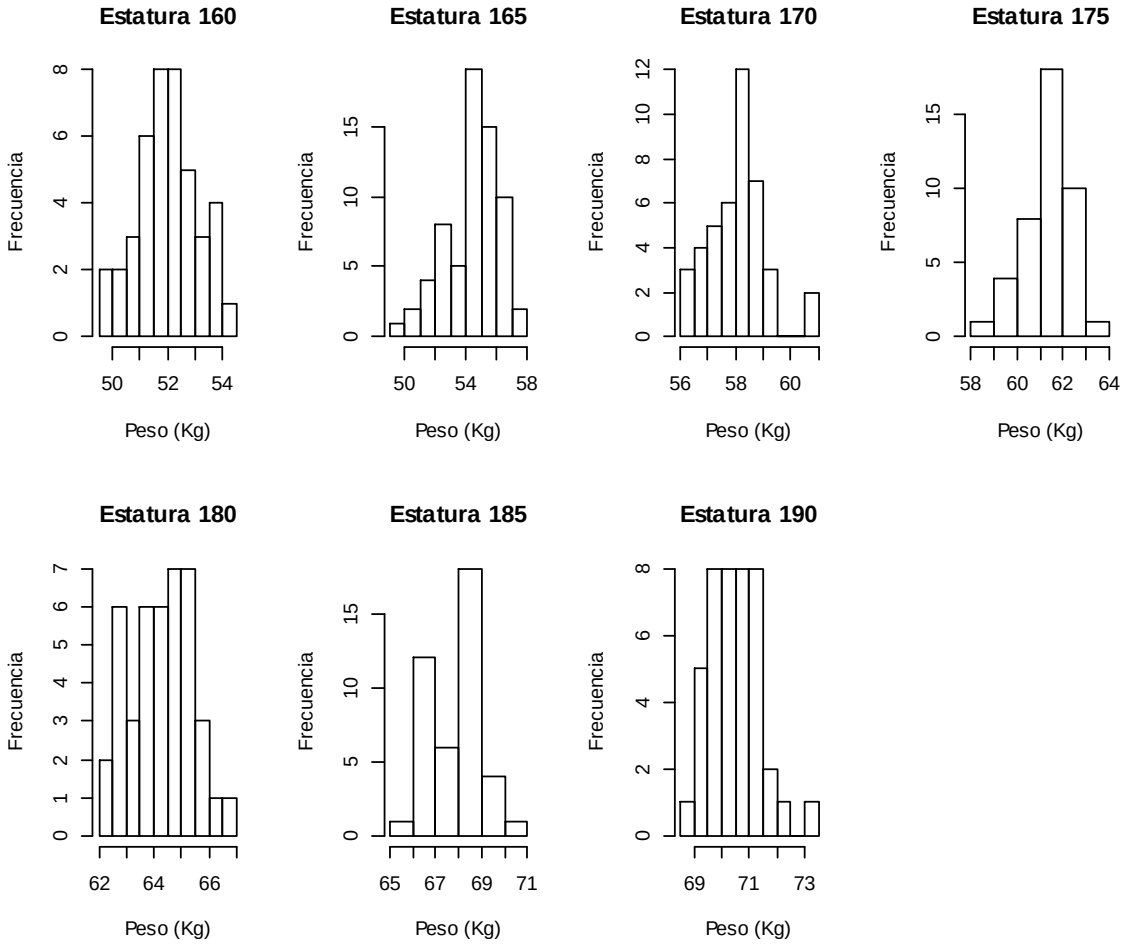
Las preguntas 7 y 8 exigen la interpretación de diagrama de cajas y bigotes, identificación de datos atípicos y análisis de rango intercuartilico, esto con el fin de analizar características de dispersión y simetría de los datos.

Los 15 estudiantes identifican que el dato atípico se encuentra en la distribución de pesos de la estatura 1,90. Respecto a la pregunta 8, los estudiantes evidencian que las mujeres que miden 1,80 metros, tienen una distribución de pesos asimétricas, pues afirman que la parte inferior de la caja es mayor que la parte superior; lo que indica que los pesos comprendidos entre el 25% y el 50% de la población están más dispersos que entre el 50% y el 75%. Un análisis análogo realiza para las estaturas de 1,65 y 1,85 metros, aunque en este último aclaran que el bigote inferior es más corto que el superior; por ello el 25% de los pesos bajos están más concentrados que el 25% de

los pesos considerados altos. Las demás distribuciones de peso las consideran simétricas.

De acuerdo con los histogramas del peso de las 70 mujeres encuestadas conteste las preguntas 9 y 10

Figura 4-3: Histogramas de los pesos por estatura de 210 mujeres. Datos simulados



9. ***¿Hay evidencia de asimetría en la distribución de los pesos? Indique en cuales estaturas.***
10. ***¿Se puede asumir normalidad en la distribución de los pesos? Indique en cuales estaturas.***

Las preguntas 9 y 10 se basan en los histogramas de peso de cada una de las estaturas con el fin de analizar simetría y normalidad de cada uno de los pesos.

Tan solo 3 estudiantes consideran la distribución de los pesos de la estatura 160 cm como una distribución simétrica y normal, 8 estudiantes consideran que existe normalidad y simetría en la distribución de pesos de la estatura 175 cm y los 15 estudiantes coincidieron que las estaturas 180, 185 y 190 cm no son simétricas ni normales, tan solo un estudiante aclara que el tamaño de muestra puede influir en la forma del histograma.

Sección 3: Probabilidad

11. ***Calcule las siguientes probabilidades, usando la distribución normal de probabilidad con $\sigma = 1$ y $\mu_{peso|estatura}$ según corresponda en cada caso.***

A. $\boldsymbol{P(peso < 52.59Kg|1,60\ mts)}$

B. $\boldsymbol{P(peso > 57,29\ Kg|1,70mts)}$

C. $\boldsymbol{P(peso \geqslant 69,67\ Kg|1,90mts)}$

Esta pregunta tenía como objetivo calcular las probabilidades de distribución normal para así recordar al estudiante que dicha distribución tiene una forma acampanada y un solo pico en el centro de la distribución, ya que las medidas de tendencia central se localizan en este pico. Los 15 estudiantes contestaron de forma correcta, cabe resaltar que para la solución de este punto fue necesario realizar un ejemplo preliminar para recordar el manejo de la tabla de la distribución normal.

Sección 4: Inferencia

Considere la hipótesis nula de que el peso promedio de las estudiantes es de 65 kilogramos, contra la hipótesis alternativa de que es diferente a 65 kilogramos, teniendo en cuenta la tabla y una desviación estándar de 7,4 Kg. Responda las preguntas del 12 al 14

Tabla 4-2: Cuantilas de la distribución normal estándar, según el nivel de significancia α.

$1-\alpha$	$\alpha/2$	$Z_{\alpha/2}$
0,90	***0,05***	***1,645***
0,95	***0,025***	***1,96***
0,99	***0,005***	***2,575***

12. De acuerdo con la información anterior analice cada enunciado.

- ***A. La prueba con un nivel de significancia del 5% para la hipótesis del encuestador se rechaza***
- ***B. La prueba con un nivel de confianza del 90% para la hipótesis del encuestados no es concluyente***

Esta pregunta establece pruebas de hipótesis referentes a las medias, partiendo del planteamiento de hipótesis, interpretación de zona de aceptación y rechazo de la respectiva hipótesis nula, además, del cálculo y análisis.

En el ítem A, el 100% de los estudiantes coincidieron en que se rechaza la hipótesis del encuestador, sin embargo, se evidencio dificultad en la interpretación del ejercicio y en la extracción de los datos. En este ejercicio fue necesario la orientación del docente, para aclarar el concepto de nivel de significancia y qué papel juega en el desarrollo del mismo.

En el ítem B, se observó que la frase “no es concluyente” es un obstáculo en la interpretación del ejercicio, fue necesario aclarar que dicha frase, es otra forma de decir “no se rechaza”, sin embargo, se evidenció una mejor interpretación en el contraste de las dos hipótesis, podríamos afirmar que esto se debe a la socialización y retroalimentación del ítem anterior.

13. ¿Cuántos intervalos de confianza para la media, dada la información disponible podría encontrar? Realiza el cálculo y escribe si observa alguna tendencia.

Esta pregunta busca identificar la cantidad de intervalos posibles que se podía hallar, dada la información de la tabla, especificando el nivel de confianza, además, de construir intervalos que contuvieran el parámetro a estimar y su posible tendencia.

De los 15 estudiantes que realizaron la prueba, 9 afirman que pueden realizar un intervalo de confianza de la media del peso dada la estatura, 6 estudiantes afirman que por cada estatura se pueden realizar 3 intervalos de confianza, teniendo en cuenta, los niveles de confianza del 90%, 95% y 99%.

En cuanto a la tendencia de intervalo de confianza, se socializaron las diferentes percepciones y se concluyó, que entre más nivel de confianza los intervalos son más

amplios por tal motivo es más probable que se encuentre el valor del parámetro. A continuación, se muestra 3 intervalos de confianza para la estatura de 1,65 metros.

$$\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$52{,}13 - 1{,}645 * 0{,}51 \leq \mu \leq 52{,}13 - 1{,}645 * 0{,}51$ $(51.296\ , 52.976)$	Intervalo nivel de confianza del 90%
$52{,}13 - 1{,}96 * 0{,}51 \leq \mu \leq 52{,}13 - 1{,}96 * 0{,}51$ $(51.135\ , 53.137)$	Intervalo nivel de confianza del 95%
$52{,}13 - 2{,}575 * 0{,}51 \leq \mu \leq 52{,}13 - 2{,}575 * 0{,}51$ $(50.821\ , 53.451)$	Intervalo nivel de confianza del 99%

14. Estime un intervalo de confianza para la proporción de mujeres de 1.60 mts a 1.90 mts con peso por debajo del ideal dada su estatura.

El objetivo de esta pregunta era evidenciar la construcción de un intervalo de confianza para el parámetro p. A continuación, se presenta una de las soluciones de intervalo de confianza presentada por los estudiantes, con un nivel de confianza del 99%.

$$IC = \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC = 0{,}1428 \pm 2{,}575\sqrt{\frac{0{,}1428(0{,}8571)}{210}}$$

$$(0.080\ , 0.205).$$

4.2 Descripción de la unidad didáctica

La unidad didáctica que se diseñó está compuesta por 2 talleres los cuales parten de un problema situacional, donde se presenta la regresión lineal simple como medio de solución. Como característica principal se usó la simulación en R como herramienta didáctica para la construcción de las actividades.

La construcción del taller 1 se plantea con una simulación de regresión para $n = 100$ y $n = 1000$ observaciones, en este taller se busca un acercamiento analítico a la tabla del ANOVA, a las gráficas de los residuales, al diagrama de dispersión, predicciones del modelo, varianza estimada de los estimadores y varianza de los estimadores, asimismo se busca identificar diferencias del modelo muestral cuando el tamaño de la muestra varia significantemente.

Para la realización del taller 2 se plantea una simulación de regresión de 1000 grupos de tamaño 100, esto con el fin de analizar la normalidad de $\widehat{\beta_0}$ y $\widehat{\beta_1}$ y la media aritmética de los mismos.

Las actividades fueron diseñadas de tal modo que no se encontrara ninguna definición o enunciado formal relativo a la regresión lineal simple, esto con el fin de no dar privilegio a la memorización de conceptos y darle prioridad a un análisis de la regresión más significativo a partir de los talleres propuestos.

4.2.1 Objetivo general de la unidad didáctica

- Interpretar el concepto de normalidad en regresión, varianza de los estimadores, varianza estimada de los estimadores, distribución de los estimadores, varianza de la estimación.

4.2.2 Objetivos específicos de la unidad didáctica

- Representar y analizar gráficamente la asociación de variables por medio de diagramas de dispersión.
- Identificar la importancia de los residuales en un modelo de regresión lineal simple.
- Comparar las propiedades de los estimadores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ para una muestra de tamaño 100 y una de tamaño 1000.
- Predecir nuevas observaciones de y a partir del modelo poblacional.
- Interpretar la normalidad de los estimadores a partir de 1000 grupos de tamaño 100.
- Hallar e interpretar la media aritmética y el error estándar de las estimaciones de β_0 y de β_1 a partir de 1000 grupos de tamaño 100.

Como materiales en el desarrollo de esta unidad didáctica, es necesario que cada estudiante cuente con su guía de trabajo, además, de un ordenador por estudiante previamente instalado el software R.

4.2.3 Metodología

La metodología para desarrollar estas actividades está comprendida en 5 fases, para las cuales se hace necesario un manejo óptimo en R, además del interés constante del estudiante.

El primer momento se enmarca en la lectura de la guía de trabajo individualmente e interpretación y orientación del docente, posteriormente cada estudiante procede a desarrollar el *taller Regresión Lineal simulación de observaciones* $n = 100$ *y* $n = 1000$ con la ayuda del computador utilizando R, como tercera fase se realizará un trabajo cooperativo dentro del marco de resultados y análisis del primer taller, esto con el fin de garantizar el pensamiento crítico y analítico del estudiante, frente a la interpretación de cada concepto trabajado.

La cuarta y quinta fase consiste respectivamente en dar solución al taller *Regresión Lineal 1000 grupos de tamaño 100* y socializar en forma grupal los resultados de dicha actividad.

El docente realizará un acompañamiento permanente frente al desarrollo de los talleres por el alumno, además, tendrá un rol activo en el proceso enseñanza-aprendizaje. En la

fase de trabajo colaborativo, el docente será un mediador cognitivo, haciendo preguntas que verifique el conocimiento de los estudiantes, siempre indagando el ¿porque?, ¿para qué? y ¿qué significa? De las respuestas del estudiante.

4.3 Secuencia de actividades

4.3.1 Taller regresión lineal de observaciones $n = 100$ y $n = 1000$

Con esta actividad se espera que el estudiante resuelva situaciones problemas, relacionados con la interpretación de diagrama de dispersión, residuales, estimación de los parámetros de la regresión lineal, comparar las propiedades de los estimadores cuando varia su muestra significativamente. En este taller se plantea actividades que permita al estudiante simular datos de una situación concreta, como lo es el peso dada la estatura.

Responda las preguntas de acuerdo a la siguiente información.

Se realizó un estudio para determinar la relación que hay entre la estatura y el peso, como resultado del estudio se obtuvo el siguiente modelo lineal

$$y = -49 + 0.63x + \varepsilon$$

1. Simule una muestra de tamaño n=100 del modelo población anterior, con el siguiente código en R.

```
########## SIMULACIÓN REGRESIÓN OBSERVACIONES N=100 ############
set.seed(8)     # fijar semilla en 8
sigma=1
n=100
x=seq(160, 190, length=n)
y= -49 + 0.63*x + rnorm(n,0,sigma)
plot(x,y)
cbind(x,y)
reg=lm(y~x)
summary(reg)
anova(reg)
```

2. ¿Qué conclusiones podría sacar a partir del diagrama de dispersión de la relación peso y estatura?

_____

_____

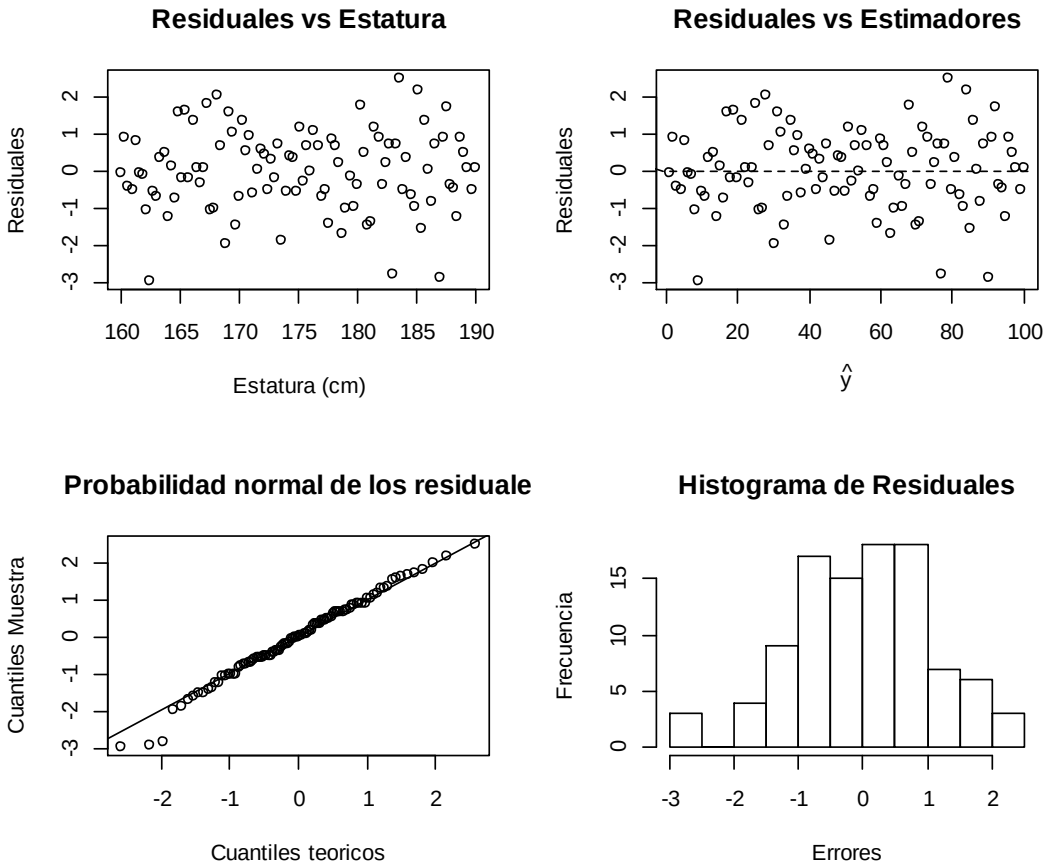
3. ¿Qué puede concluir del análisis descriptivo de los residuales? **(summary(reg$residuals))**

_____

_____

De acuerdo con los siguientes gráficos responda las preguntas 4, 5 y 6

Figura 4-4: Gráficos de residuales. Datos simulados.



4. ¿Se puede afirmar que en el modelo existe la presencia de un efecto lineal? Justifique.

5. ¿Se puede afirmar que los residuos tienen varianza constante? Justifique.

6. ¿los residuales son normales? Explique su respuesta.

7. Registre el valor de los betas que se ha obtenido y su error estándar (típico). ¿Ha obtenido una buena estimación?

Estimadores	Estimación	Error estándar
β_0		
β_1		

8. ¿Qué prueba usaría para evaluar las hipótesis $H_0: \beta_1 = 0$ y $H_0: \beta_0 = 0$?

9. ¿Usando un nivel de significancia $\alpha = 0.05$ que se puede concluir?

_____

_____

_____

10. Contraste la hipótesis de que la pendiente de la recta es igual a cero ¿Qué se puede concluir a partir de esta prueba de hipótesis?

_____

_____

_____

11. Analice el modelo si $\hat{\beta}_0 = 0$

_____

_____

12. ¿los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de mínimos cuadrados son estimadores insesgados de los parámetros de modelo?

_____

_____

13. ¿Cómo afecta el tamaño muestral a las estimaciones?

_____

_____

14. Una aplicación importante del modelo de regresión es predecir nuevas observaciones y que correspondan a un nivel especificado de la variable regresora x. realice las predicciones de los pesos para las estaturas dadas. Utilice el siguiente código en R.

```
###############PREDICCIONES DEL MODELO ####################
nuevos_pesos<-data.frame(x=seq(160,190,by=5))
predict(reg,nuevos_pesos)
```

Estatura (mts)	Peso (Kg)
1,60	
1,65	
1,70	
1,75	
1,80	
185	
190	

15. Realice la varianza de los estimadores y varianza estimada de los estimadores con el siguiente código y registre dichos resultados en la tabla.

VARIANZA DE LOS ESTIMADORES Y VARIANZA ESTIMADA DE LOS ESTIMADORES

```
varB_0=1*(sum(x^2))/(100*sum((x-mean(x))^2))
varestB_0=var(errores)*(sum(x^2))/(100*sum((x-mean(x))^2))
varB_1=1/sum((x-mean(x))^2)
varestB_1=var(errores)/sum((x-mean(x))^2)
```

$var(\widehat{\beta_0})$	$\widehat{var}(\widehat{\beta_0})$	$var(\widehat{\beta_1})$	$\widehat{var}(\widehat{\beta_1})$

16. Realice la práctica anterior con n=1000 y responda.
    a) Compare las estimaciones $\widehat{\beta_0}$ y $\widehat{\beta_1}$ del ejercicio n=100 y n=1000 ¿Qué puede concluir?

	$n = 100$	$n = 1000$
$\widehat{\beta_0}$		
$\widehat{\beta_1}$		

4.3.2 Taller regresión lineal 1000 grupos de tamaño 100

Como se observa en el taller anterior se trabaja el concepto de regresión, en una muestra de tamaño 100, identificando las propiedades de los estimadores e interpretando las gráficas de los residuales, a continuación, se presentará en el siguiente taller el concepto de regresión para 1000 grupos de tamaño 100, buscando como objetivo principal el análisis de la distribución de los estimadores y su normalidad.

Responda las preguntas con la siguiente información.

Se tomaron 1000 muestras de tamaño 100 a estudiantes de diferentes universidades de la ciudad de Bogotá donde se les preguntó la estatura y su respectivo peso.

1. Simule una muestra de 1000 grupos de tamaño 100 con el siguiente código en R teniendo en cuenta que el modelo población es $y = -49 + 0.63x + \varepsilon$.

```
#######SIMULACIÓN REGRESIÓN 1000 GRUPOS DE TAMAÑO 100 #########
sigma=1
x=seq(160,190,length=100)
y=-49+0.63*x+rnorm(100, 0, sigma)
plot(x,y,xlab="Estatura (metros)",ylab="peso (Kg)")
reg=lm(y~x)
reg
names(reg)
summary(reg)
anova(reg)
reg$xlevels
x=seq(160,190,length=100)
B_0=NULL
B_1=NULL
##NULL es como tener un vector de nada y lo voy llenando
for(i in 1:1000)
```

```
{
 y=-49+0.63*x+rnorm(100, 0, sigma)
 reg=lm(y~x)
 reg$coefficients
 B_0[i]=reg$coefficients[[1]]#muestra solo el intercepto
 B_1[i]=reg$coefficients[[2]]#muestra solo la pendiente
  }
```

2. ¿Qué conclusiones podría sacar a partir del diagrama de dispersión de la relación estatura y peso?

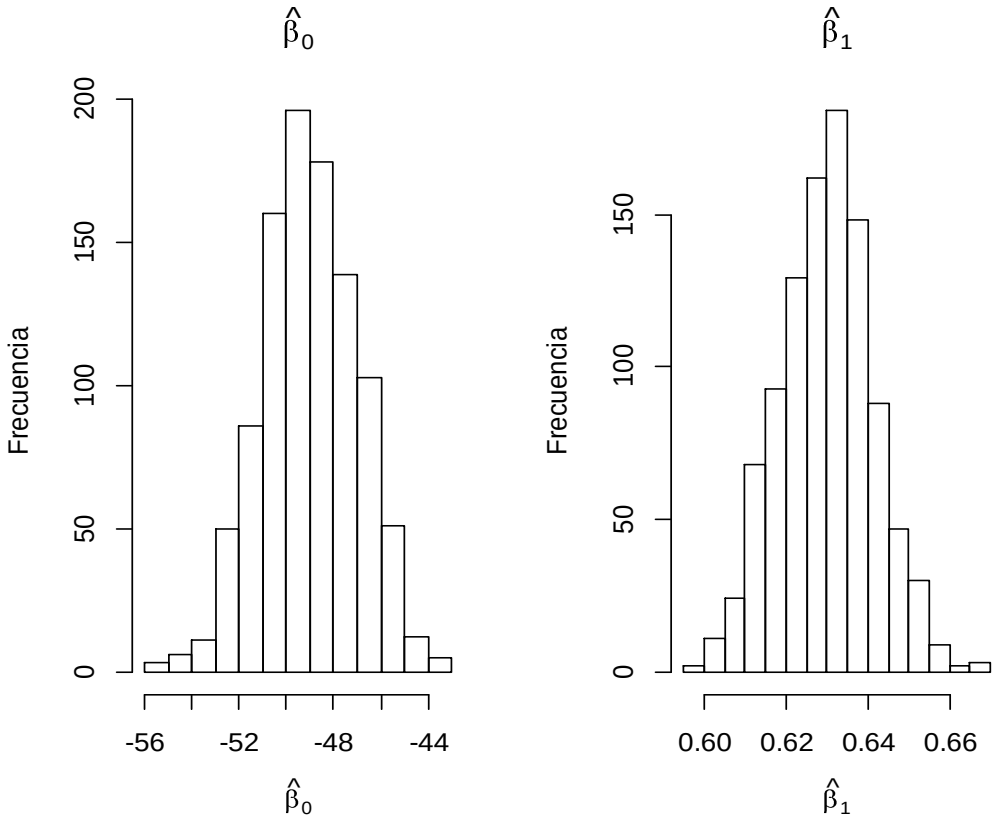
3. Genere mil estimaciones de los parámetros β_0 y β_1 ¿Cómo se puede interpretar cada una de estas estimaciones?

4. Halle e interprete la media aritmética error estándar de las estimaciones β_0 y β_1. Utilice los códigos **mean(B_0)** , **mean(B_1), sd(B_1)** y **sd(B_0)**

Estimadores	Media aritmética	Error estándar
$\widehat{\beta_0}$		
$\widehat{\beta_1}$		

Los siguientes gráficos muestran los histogramas de las estimaciones de los parámetros β_0 y β_1.

Figura 4-5: Histogramas de los estimadores de los parámetros del modelo.



5. ¿Hay evidencia de asimetría y de no normalidad en la distribución de las estimaciones de β_0 y β_1?

_____

_____

6. Construya 4 diagramas de dispersión con su respectiva recta de regresión con el siguiente código.

```
sigma=1
x=seq(160,190,length=100)
y=-49+0.63*x+rnorm(100, 0, sigma)
par(mfrow=c(2,2))
plot(x,y,xlab="Estatura (metros)",ylab="peso (Kg)")
abline(reg, col = 'blue')  #cambie el color para las 4 gráficas
```

Socialice con el grupo el resultado obtenido.

5. Implementación de la propuesta

A continuación, se describe la aplicación del proyecto desde la descripción de las acciones realizadas en cada taller, las dificultades y el aprendizaje encontrados durante el avance.

Como se mostró en la propuesta de trabajo de grado, esta unidad didáctica se dirigió a estudiantes universitarios de ingeniería que manejaban conceptos previos a la regresión lineal. Esta unidad se implementó en 15 estudiantes de ingeniería electrónica de la Fundación Universitaria los Libertadores. Su principal característica radicaba que eran estudiantes no repitentes (estadística y probabilidad II), lo cual nos garantizaba un grupo homogéneo en cuanto a conocimientos previos.

En la implementación del taller “Regresión lineal simulación de observaciones $n = 100$ y $n = 1000$”, respondieron individualmente las preguntas usando como principal herramienta el computador. Después de esto todos los estudiantes del curso, discutieron la veracidad de las respuestas, lo cual les permitió establecer aciertos y errores.

De lo anterior se puede identificar las siguientes dificultades y aciertos de aprendizaje que ofreció el primer taller.

Dificultades.

- Dificultad para asociar que los datos simulados podrían pertenecer a un contexto de la vida real, en este caso específico peso estatura.
- La interpretación del error estándar de los estimadores $\widehat{\beta_0}$ y $\widehat{\beta_1}$. Se observó poca argumentación en el punto 7 del taller, podríamos asociar esta dificultad, como un concepto previo poco manejado.

- El tiempo fue un factor que impidió un análisis más significativo y más profundo, en el trabajo colaborativo, este espacio permitió fortalecer conceptos, pero demandaba más de una sesión de 2 horas.

Aprendizaje.

- El concepto predicción fue intuitivo en el ejercicio 14, el estudiante identificó acertadamente que el modelo de regresión aportaba las predicciones de nuevas observaciones y que corresponden a un nivel especificado de la variable regresora x.
- Por medio del ejercicio 16 el estudiante comprende que los estimadores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ tienen una mejor aproximación al parámetro cuando el tamaño de la muestra es grande, en este caso $n = 1000$.
- El trabajo colaborativo permitió un espacio de fortalecimiento de conceptos teóricos, las afirmaciones de los estudiantes, permitían un espacio de aclaración de percepciones, donde ellos mismo participaban como entes formadores.
- El ejercicio 3 permitió por medio de sus gráficas, resaltar de la utilidad de los residuales en la regresión lineal, el estudiante asocio le comportamiento de las gráficas a estimadores no sesgados y con mínima varianza de los parámetros.

En la implementación del taller “Regresión lineal 1000 grupos de tamaño 100” además de garantizar el concepto de normalidad de los estimadores, se buscaba evitar la dificultad del taller “Regresión lineal simulación de observaciones n=100 y n=1000” en cuanto al tiempo en la fase de trabajo colaborativo y el manejo de R como mediador en la simulación.

Dificultades.

- Interpretación del error estándar de las mil muestras, esto se evidencio en la poca argumentación de los estudiantes, además, se observó debilidades conceptuales en las medidas de dispersión.

Aprendizaje.

- Interpretación de diagrama de dispersión, cabe resaltar que dentro del desarrollo de la unidad didáctica y las explicaciones magistrales se reforzó el análisis de gráficas, lo anterior con el objetivo de observar cuando dos variables están relacionadas.
- Al realizar las mil estimaciones de los parámetros β_0 y β_1 los estudiantes comprendieron que tenía información de mil rectas muéstrales, lo cual ayudo a fortalecer el concepto de estimación de parámetro y estimación de la recta poblacional.
- Por medio del punto 5 del taller 2 los estudiantes acogieron el termino de normalidad de los estimadores, intuyendo que a medida que el tamaño de la muestra aumenta, los estimadores convergen hacia sus verdaderos valores, esta aclaración es apoyada por el punto 16 del taller 1.
- El punto 6 muestra de una forma más interactiva el comportamiento de 4 muestras de las 1000 con su respectiva línea de regresión.

Figura 5-1: Estudiantes ingeniería electrónica desarrollo taller “Regresión lineal simulación de observaciones n=100 y n=1000”

Figura 5-2: Estudiantes ingeniería electrónica desarrollo unidad didáctica

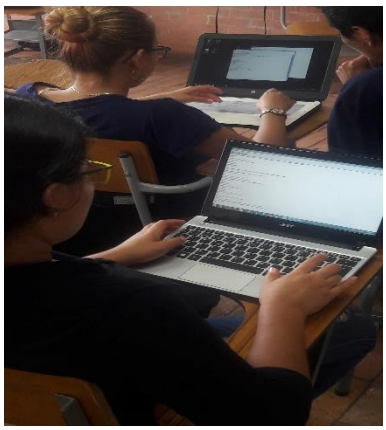


Figura 5-3: Estudiantes ingeniería electrónica desarrollo taller “1000 grupos de tamaño 100”

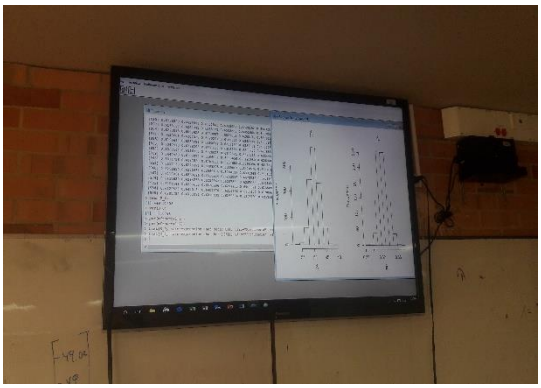


Figura 5-4: Estudiantes ingeniería electrónica trabajo colaborativo

6. Conclusiones y recomendaciones

6.1 Conclusiones

La construcción de esta unidad didáctica, permitió fortalecer los conocimientos disciplinares frente a la regresión lineal, además, de consolidar la simulación en R, como herramientas didácticas a la hora de transmitir estos conocimientos, los logros alcanzados en este trabajo se presentan a continuación.

El análisis del proceso histórico del concepto de regresión lineal, nos permitió presentar este término, como un resultado de mejoramiento a través del tiempo y no como un producto terminado e incuestionable. Además, esta visión histórica fue una herramienta de contextualización y humanización de la regresión, muchos de los obstáculos epistemológicos que se presentaron en la construcción del concepto, se podrían relacionar con las dificultades que muestran los estudiantes, que se derivan en la complejidad de la regresión.

Se identificaron como dificultades en la compresión de la regresión lineal, la interpretación de independencia lineal, mediante la asignación del coeficiente de correlación a una gráfica de dispersión, dentro de la revisión bibliográfica, también, se observó que un error cotidiano en la práctica docente, es la comparación de la función lineal con la regresión lineal como recurso didáctico. Lo anterior nos permitió, anticiparnos a las posibles dificultades que se pueden presentar en los estudiantes, lo cual contribuyó en el diseño de la unidad didáctica.

El *“taller diagnostico conceptos previos a la regresión”* mostró la importancia de los temas anteriores a este concepto, los cuales son indispensables en el análisis e interpretación de la regresión lineal; la mayoría de las deficiencias de los estudiantes, se evidenciaron en la

inferencia estadística, donde la socialización de respuestas con el grupo fue necesaria, para aclarar cálculos y conceptos referidos a la estimación puntual y por intervalo.

6.2 Recomendaciones

Esta propuesta consideró el estudio de dificultades en relación al proceso de enseñanza-aprendizaje en la regresión lineal, por tal motivo se sugiere a los docentes investigar sobre posibles obstáculos que se pueden presentar al estudiante, en la construcción del concepto, además, es necesario indagar sobre el nivel de apropiación de la estadística descriptiva e inferencial de los estudiantes, en lo posible aplicar la prueba diagnóstica (anexo 1), ya que dicha prueba está incluida en un mismo contexto de la unidad didáctica.

Teniendo en cuenta lo extensa de la unidad didáctica, se sugiere al docente que la desarrolle en varias sesiones, esto con el fin de profundizar en los conceptos y en analizar las respuestas de los estudiantes en el trabajo cooperativo. En la finalización de cada taller el docente puede plantear una evaluación didáctica, para fines pertinentes.

Esta unidad didáctica es un punto de partida para elaborar otras actividades, por ejemplo, se puede solicitar al estudiante que modifique el código y diseñe otra simulación, pero con un contexto diferente, con la virtud de cambiar tamaños de muestras y desviación estándar.

A. Anexo: Prueba diagnóstica

. Taller Diagnóstico Conceptos Previos a la Regresión Lineal

Nombre: _____ semestre: _____ fecha: __

Responda las preguntas de acuerdo con la siguiente información.

En la facultad de ingeniería de la Universidad Nacional de Colombia se les preguntó a 210 estudiantes mujeres, sobre su peso y estatura, con el fin de analizar el porcentaje de mujeres con bajo nivel de peso.

La siguiente tabla muestra los datos de 210 mujeres ordenadas de acuerdo a su estatura (menor a mayor).

Estatura [mts]	peso [Kg]										$\mu_{peso\|esta}$
1,6	51.11	52.21	51.48	51.48	53.63	51.03	53.08	52.14	52.32	52.57	51.20
	52.02	52.59	50.61	52.78	52.66	53.94	53.28	50.73	51.70	51.75	
	54.18	51.97	51.09	51.97	51.91	51.81	52.44	52.16	53.55	49.91	
1,65	55.18	56.35	53.78	55.68	53.43	56.34	57.29	56.32	55.07	54.54	53.79
	52.71	54.36	56.01	54.88	55.35	54.63	55.20	55.00	54.09	53.68	
	54.87	54.24	54.64	56.53	56.27	54.03	55.17	54.55	55.31	57.44	
1,7	56.74	58.98	57.22	57.29	58.41	56.74	57.63	58.49	57.89	58.15	57.80
	58.85	58.82	59.32	56.89	57.90	60.66	58.09	59.14	58.36	58.04	
	58.65	60.63	56.86	58.46	58.47	57.48	57.11	57.83	56.18	58.80	
1,75	62.58	60.49	62.60	61.07	61.62	60.38	61.44	61.60	60.51	62.31	60.55
	61.12	61.31	62.43	61.18	61.38	62.40	62.53	61.63	59.85	61.96	
	60.43	62.58	60.38	60.78	61.05	61.16	61.37	61.52	62.60	59.66	
1,8	64.48	63.52	62.63	62.31	65.01	64.10	64.98	65.60	65.41	63.15	64.80
	63.03	64.37	66.51	64.96	62.52	64.08	64.89	63.83	62.88	62.72	

Estatura [mts]	peso [Kg]										$\mu_{peso\|esta}$
	65.61	63.55	62.33	65.40	64.70	66.17	65.68	65.25	64.46	62.76	
1,85	66.10	68.25	69.45	68.73	67.44	68.30	69.45	66.31	68.27	66.55	67,71
	66.99	68.53	68.24	67.86	68.05	66.41	66.67	70.86	68.18	66.75	
	67.12	68.93	66.24	67.22	69.20	68.31	68.45	68.66	68.80	67.05	
1,9	70.59	71.55	69.97	71.00	68.68	70.14	71.25	69.44	71.11	71.80	72.20
	70.72	69.67	69.88	69.92	70.60	71.49	69.28	73.11	70.23	69.84	
	69.93	70.76	69.67	70.49	71.38	70.99	69.14	69.39	71.29	70.37	

ANALISIS DESCRIPTIVO _____

1. Con la información anterior.

a. ¿Cuál es la población de interés?
R/: _____

b. ¿Cuál es el tamaño de la muestra para cada estatura?
R/:_____

c. ¿Cuáles son las variables consideradas?
R/:_____

2. Completa la siguiente tabla.

	1,60	1,65	1,70	1,75	1,80	1,85	1,90
Mínimo							
Cuartil (Q1)							
Mediana							
Media							
Cuartil (Q3)							
Máximo							
Desviación estándar							
Coeficiente de variación							

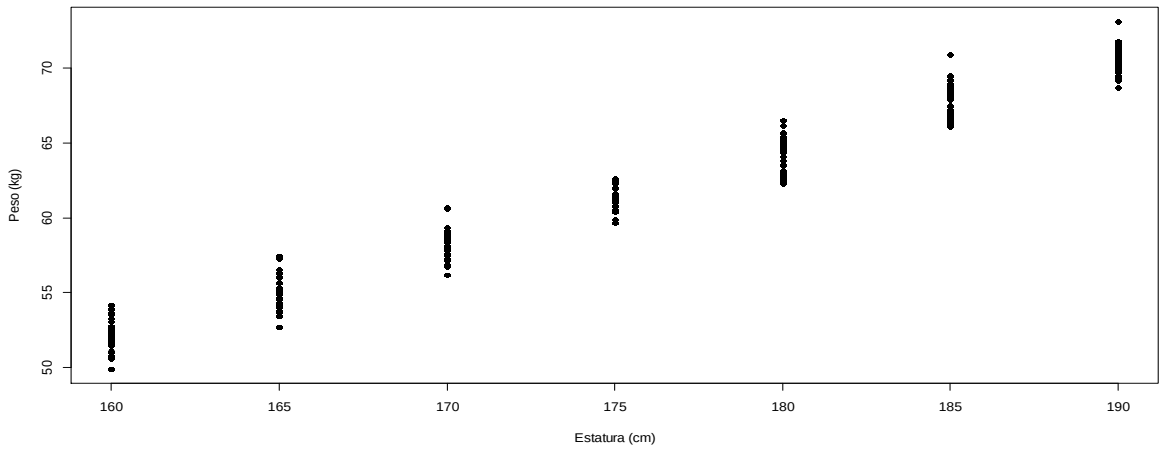
3. Interprete el coeficiente de variación.

REPRESENTACIONES GRÁFICAS

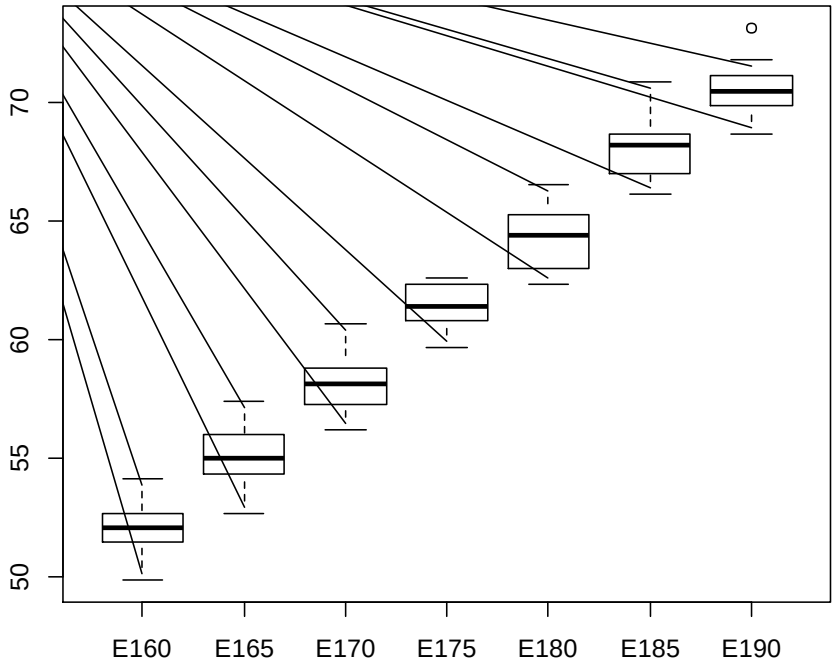
4. Realice un gráfico de dispersión.

5. ¿Qué puede concluir del anterior diagrama? **Interprete la relación entre las dos variables** R/:

6. Identifique la media del peso por estatura en la gráfica y trace una línea recta por cada una de ellas.



De acuerdo con el box plot del peso de las 70 mujeres encuestadas conteste las preguntas 7 y 8



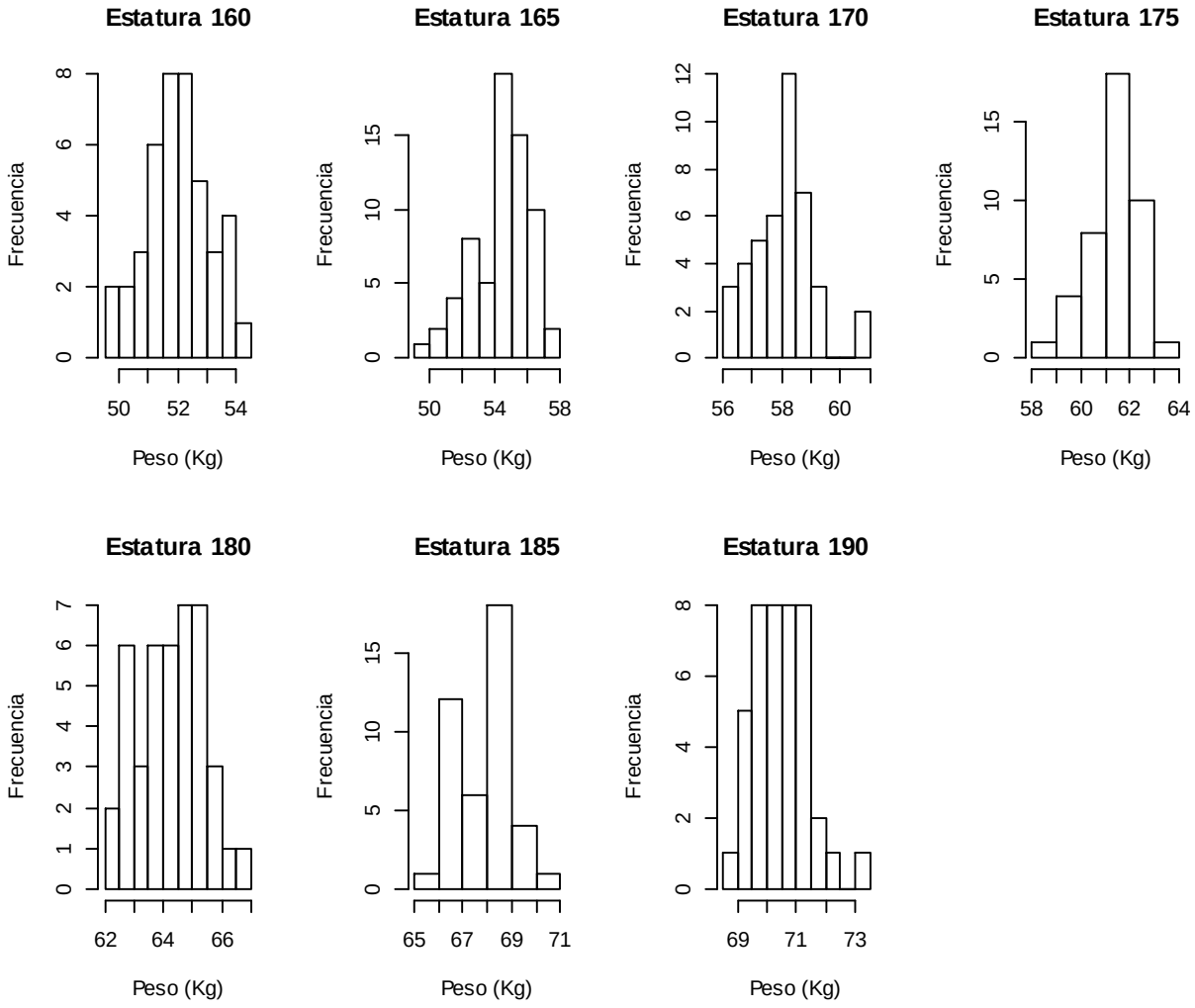
7. ¿Hay evidencia de outliers? Justifique

_____

8. Haga un comparativo del valor de la media y de los cuartiles Q_1 y Q_3. Concluya

_____

_____



9. ¿Hay evidencia de asimetría en la distribución de los pesos? Indique en cuales estaturas.

_____

_____

10. ¿Se puede asumir normalidad en la distribución de los pesos? Indique en cuales estaturas.

_____

_____

PROBABILIDAD_____

11. Calcule las siguientes probabilidades, usando la distribución normal de probabilidad con $\sigma = 1$ y $\mu_{peso|esta}$ según corresponda en cada caso.

a. $P(peso < 52.59Kg|1{,}60\ mts) =$ _____

b. $P(peso > 57{,}29\ Kg|1{,}70mts)=$ _____

c. $P(peso \geqslant 69{,}67\ Kg|1{,}90mts)=$ _____

INFERENCIA _____

Considere la hipótesis nula de que el peso promedio de las estudiantes es de 65 kilogramos, contra la hipótesis alternativa de que es diferente a 65 kilogramos, teniendo en cuenta la tabla y una desviación estándar de 7,4 Kg

$1-\alpha$	$\alpha/2$	$Z_{\alpha/2}$
0,90	0,05	1,645
0,95	0,025	1,96
0,99	0,005	2,575

12. De acuerdo con la información anterior escriba falso (f) o verdadero (v) según corresponda.
    C. La prueba con un nivel de significancia del 5% para la hipótesis del encuestador se rechaza __
    D. La prueba con un nivel de confianza del 90% para la hipótesis del encuestados no es concluyente __
    E. El valor p de la prueba corresponde a 0,5% __

13. ¿Cuantas pruebas de hipótesis respecto a la media podría hacer? R/:_____

14. ¿Cuántas pruebas de hipótesis respecto a La varianza podría hacer? R/:_____

B. Anexo: Códigos en R usados.

```
############ SIMULACIÓN DATOS PRUEBA DIAGNÓSTICA ##################
estatura=seq(160, 190, by=5)
peso_observado=c(54.37, 56.72, 62.83, 65.67, 67.47, 70.88, 74.46)
plot(estatura, peso_observado)
regresion=lm(peso_observado~estatura)
summary(regresion)
n=40
estatura=seq(160, 190, by=5)
peso_simulado=matrix(0, nrow=7, ncol=n)
for (i in 1:n)
    {
    peso_simulado[,i]=-49+0.63*estatura+rnorm(7, 0, 1)
    }
round(peso_simulado,2)
apply(peso_simulado, 1, summary)
apply(peso_simulado, 1, sd)
matplot(estatura, peso_simulado, col=1, pch=16, xla="Estatura (cm)", ylab="Peso (kg)")
```

```
#########SIMULACIÓN REGRESIÓN OBSERVACIONES N=100########
set.seed(8)     # fijar semilla en 8
sigma=1
n=100
x=seq(160, 190, length=n)
y= -49 + 0.63*x + rnorm(n,0,sigma)
plot(x,y)
cbind(x,y)
reg=lm(y~x)
summary(reg)
anova(reg)
 ################## GRAFICO DE RESIDUALES#########################
plot(reg$residuals, main="Residuales vs Estimadores",ylab="Residuales", xlab=expression(hat(y)))
abline(h=var(reg$residuals),lty=3)
abline(h=-var(reg$residuals),lty=3)
abline(h=mean(reg$residuals),lty=2)
par(mfrow=c(2,2))
plot(x,reg$residuals,main="Residuales vs Estatura",ylab="Residuales", xlab="Estatura")
abline(h=mean(reg$residuals),lty=4)
plot(yesti,reg$residuals, main="Residuales vs estimados", ylab="Residuales", xlab=expression(hat(y)))
abline(h=mean(reg$residuals),lty=4)
qqnorm(reg$residuals,ylab="Cuantiles Muestra", xlab="Cuantiles teoricos")
qqline(reg$residuals)
hist(reg$residuals,main="Histograma de Residuales", ylab="Frecuencia", xlab=expression(hat(y)))
###############PREDICCIONES DEL MODELO ##########################
nuevos_pesos<-data.frame(x=seq(160,190,by=5))
predict(reg,nuevos_pesos)
VARIANZA DE LOS ESTIMADORES Y VARIANZA ESTIMADA DE LOS ESTIMADORES varB_0=1*(sum(x^2))/(100*sum((x-mean(x))^2))
varestB_0=var(errores)*(sum(x^2))/(100*sum((x-mean(x))^2))
```

```
varB_1=1/sum((x-mean(x))^2)
varestB_1=var(errores)/sum((x-mean(x))^2)
```

```
#####SIMULACION REGRESION 1000 GRUPOS DE TAMAÑO 100 ##############
sigma=1
x=seq(160,190,length=100)
y=-49+0.63*x+rnorm(100, 0, sigma)
plot(x,y,xlab="Estatura (metros)",ylab="peso (Kg)")
reg=lm(y~x)
reg
names(reg)
summary(reg)
anova(reg)[2]
x=seq(160,190,length=100)
B_0=NULL
B_1=NULL
eeB_0=NULL
eeB_1=NULL
##NULL es como tener un vector de nada y lo voy llenando
for(i in 1:1000)
{
  y=-49+0.63*x+rnorm(100, 0, sigma)
  reg=lm(y~x)
  reg$coefficients
  B_0[i]=reg$coefficients[[1]]#muestra solo el intercepto
  B_1[i]=reg$coefficients[[2]]#muestra solo la penddiente
  eeB_0[i]=summary(reg)$coeff[3]
  eeB_1[i]=summary(reg)$coeff[4]
    }
B_0
B_1
mean(B_0)
var(B_0)
```

```
sd(B_0)
mean(B_1)
var(B_1)
sd(B_1)    # Compararlo con la media de eeB_1 y con la raíz de la ecuación 13.13 de Canavos
par(mfrow=c(1,2))
hist(B_0, main=expression (hat(beta)[0]), ylab="Frecuencia", xlab=expression (hat(beta)[0]))
hist(B_1, main=expression (hat(beta)[1]), ylab="Frecuencia", xlab=expression (hat(beta)[1]))
sd(B_1)
lines(density(B_0))
hist(B_0,freq=FALSE)
lines(density(B_0))
```

Bibliografía

Agnelli, H., Konic P., Peoarelli S., Zon, N & Flores , P. (2009). La Función lineal obstáculo para la enseñanza de la regresión lineal. *Revista Ibeoramericana de Educación Matemática*, (17), 52-61

Badiella, Ll. (2011). Modelos lineales generalizados mixtos algunos casos Prácticos. *X Congreso Galego de estadística e investicacion de operaciónes.* Pontevedra, 1-2

Batanero , C., Diaz J., Green, D., Holmes, P, y Vallecillos , A. (1994). Errors and difficulties in understanding elementary statistical concepts. *Internation Journal of Mathematics Education in Science and Technology*, 25 (4), 527-537.

Batanero, C. (2000). ¿Hacia donde va la educación estadística? *Blaix* 15 (2), 2-13.

Batanero, C. (2001). Aleatoriedad, Modelización, Simulación. *Actas de las X Jornada sobre el Aprendizaje y la Enseñanza de las Matemáticas .* 119-130.

Behar, R. & Grima P. (2001). Mil y una dimesiones del aprendizaje de la estadistica. *Estadistica Española*, 148 (43), 190-205.

Taucher, E. (1999). Análisis de supervivencia. *Taucher E. Bioestadística. Santiago de chile:* Editorial Universitaria, 257-266.

Canavos, G. (1988). Probabilidad y estadística: Aplicaciones y métodos . México: McGraw Hill.

Cardona, D., Gonzalez, J., Rivera, M. & Cárdenas, E. (2013). Inferencia estadística-modelo de regresión lineal simple. *Documento de investigación. Universidad de Rosario.*

Estepa Castro, A. (1994). *Concepciones iniciales sobre asociaciones y su evolución como consecuencia de una enseñanza basada en el uso de de Ordenadores.* Universidad de Jaén.

Estepa, A. (2008). Interpretación de los Diagramas de Dispersión por estudiantes de Bachillerato. *Enseñanza de las ciencias,* 26(2), 257- 270.

Estepa, A., Gea M., Cañada de la Fuente, G., & Contreras, J. (2012). Algunas notas históricas sobre la correlación y regresión y su uso en el aula . *Números revistas de didáctica de didáctica de las matemáticas,* 81, 5-14*.*

Figueroa G., (2002) Ronald Aymer Fisher, Apuntes de Historia de las Matemáticas. 1, (3).

Freund, E. & Simon, G. (1994). *Estadística elemental*. Grupo editorial Person Educación. México

Hald, A. (2008). *A History of Parametric Statistical Inference from Benoulli to fisher 1713 to 1935*. Springer Science & Business media.

Serrano, G. & Gea, M. (2014). *La correlación y regresión en bachillerato análisis de libros de texto de Teoría texto y del conocimiento de futuros profesores.* Universidad de la Granada.

Villegas. M (2009). Karl Pearson, el creador de la estadística matemática. *Historia de la Probabilidad y la Estadística.* Universidad de Huelva, 351-356.

Lavalle A., Micheli E., Rubio N. (2006). Análisis didáctico de regresión y correlación para la enseñanza. *Revista Latino Americana de Investigación en Matemática Educativa*, 9(3),383-406.

Mayorga, H., Soto, O. (1988). El Analisis de Regresión: perspectiva Historica. *Revista Colombiana de Estadística*, 9, 1-15.

Mendenhall, W., Beaver, R. & Beaver, B. (2010) *Introducción a la Probabilidad y Estadística.* Grupo Editorial Iberoamérica. México

Mills, J. (2002). Using Computer Simulation Methods to teach Statistics: A Review of the literature. *Journal of statistics Education*, 10(1), 1-20

Molinero, L. (2004). Historia razonamiento estadistico. *Publicado por Asociación de la sociedad española de hipertensión y la liga española para la lucha de hipertención arterial*. 23(10), 8

Getino, T. (2010). Importancia de Darwin en el desarrollo de la estadística moderna. *Estadistica española*, 52(175), 375-381.

Montserrat, F. (2010). *Secuencias didácticas. Aprendizaje y evaluación de competencias.* Mexico: Pearson

Montgomery, D., Peck, E. & Vining, G. (2006). *Introducción al análisis de regresión lineal.* Mexico: Compañía editorial continental

R Development Core Feam (2016). A language and environment for statistical computing. R Foundation for Statistical computing., Vienna,Australia. ISBN 3-900051-07-0, URL http: www.R-project.org.

Sprott, D. (s.f.). Contribuciones de Gauss a la Estadística . *Elementos números* , 5-6.

Salas, C. (2008). ¿Por qué comprar un paquete estadístico si existe R? *Ecología Austral , 18 (2),* 223-231.

Cox, D. & Snell E. (1989). *The Analysis of Binary Data*. Routledge, Chapman & Hall, Incorporated, Bocaraton.

Truran, J. (1997). Understanding of Association and Regression by First Year Economics Students from Two Different Countries as Revealed in Responses to the Same Examination Questions. Research papers on stochastics educations from, 205-212.

Canal, S. (2000). La Estadística una Ciencia del Siglo XX. R.A. Fisher, el Genio. *Revista colombiana de estadistica.* 23(2), 1-14.

Walpole R, Myers R, Myers S. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias*. Grupo editorial Person Educación. México

Zabala, V. (2000). *La práctica educativa. Como enseñar.* Mexico: Graó, Barcelona