Análisis de Regresión (2021-3)

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Especialización en Estadística Aplicada

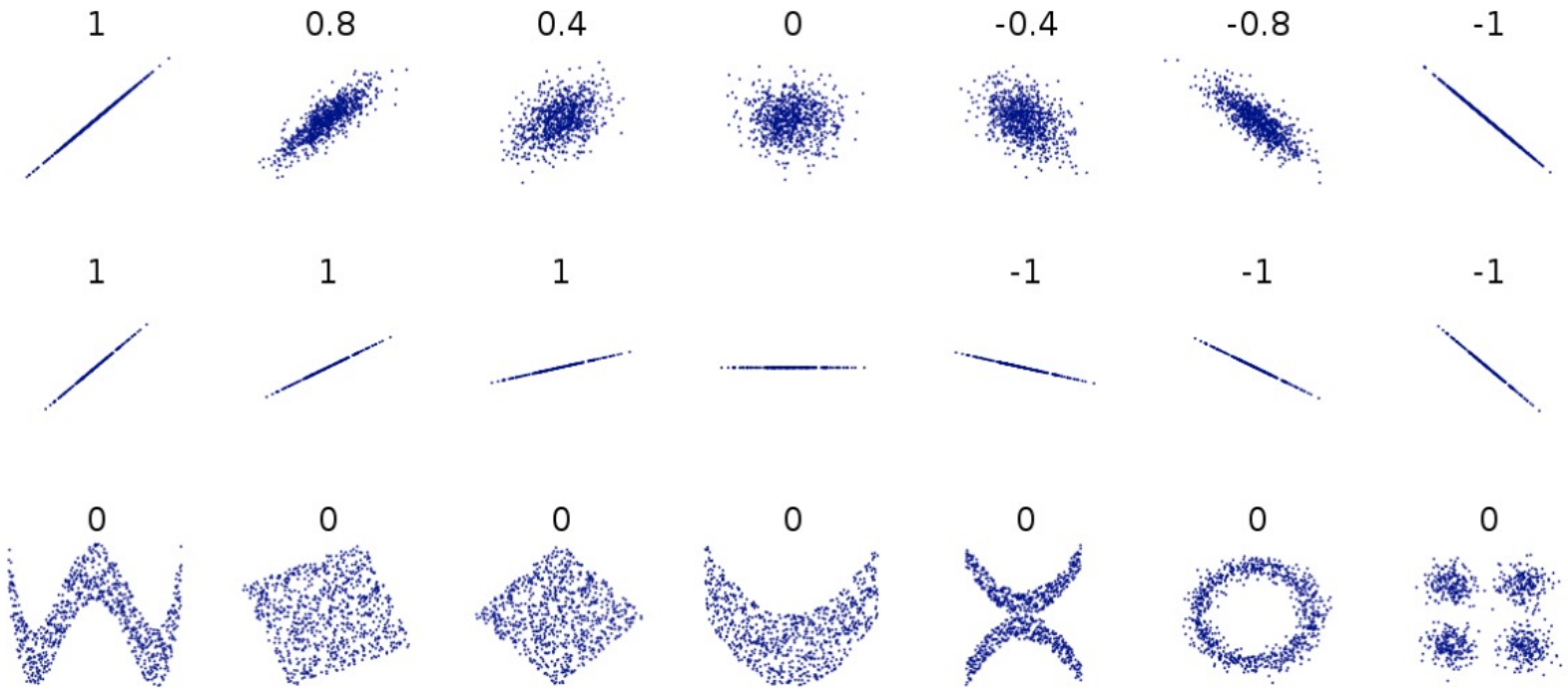
Prof. Sébastien Lozano Forero (slozanof@libertadores.edu.co)

Introducción a la Regresión Lineal Simple

Tabla de contenidos

Modelo de Regresión Lineal Simple

El modelo de regresión lineal simple da cuenta de la relación **lineal** y **causal** entre las variables x, y (x causa a y). Una primera aproximación a esta situación se tendría con la noción de correlación muestral (entendiendo que la misma no establece relación de causalidad) en el sentido de la linealidad.



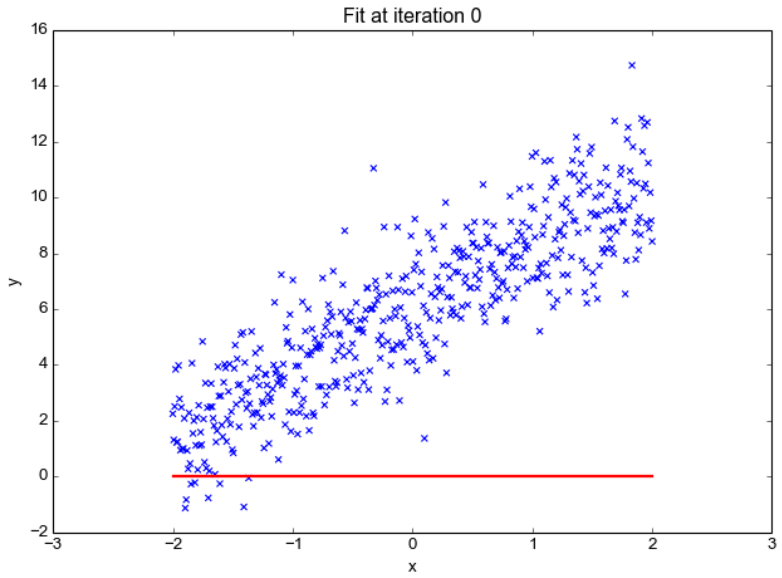
Una vez reconocida esta limitación del modelo, se puede plantear y cuantificar la relación x *causa a* y de la siguiente forma

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \tag{1}$$

Donde:

- β_0 (intercepto) y β_1 (pendiente) son parámetros poblacionales a ser estimados.
- ϵ_t representa un error aleatorio (Nuestra incapacidad para dar cuenta de la realidad tal cuál es
- y_t representa la *variable respuesta*
- x_t representa la *variable predictora*
- $t = 1, 2, \cdots, n$

En esta expresión estamos admitiendo que todos los factores o causas que influyen en la variable respuesta y pueden dividirse en dos grupos: el primero contiene a una variable explicativa x y el segundo incluye un conjunto amplio de factores no controlados que englobaremos bajo el nombre de perturbación o error aleatorio, ϵ, que provoca que la dependencia entre las variables dependiente e independiente no sea perfecta, sino que esté sujeta a incertidumbre. Por ejemplo, en el consumo de gasolina de un vehículo (y) influyen la velocidad (x) y una serie de factores como el efecto conductor, el tipo de carretera, las condiciones ambientales, etc, que quedarían englobados en el error



Algunos elementos a destacar;

- La respuesta Y_i en el i-ésimo ensayo es la suma de dos componentes: (1) el término $\beta_0 + \beta_1 X_i$ y (2) el término aleatorio ϵ_i. Por lo tanto, Y_i es una variable aleatoria.
- Ya que $E\{\epsilon_i\} = 0$, se sigue que $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Por lo tanto, la respuesta Y_i, cuando el nivel de X en el i-ésimo ensayo es X_i, proviene de una distribución de probabilidad cuya media es $E\{Y_i\} = \beta_0 + \beta_1 X_i$ Así, sabemos que la \textbf{función de regresión} para el modelo (1) es:

  $$E\{Y\} = \beta_0 + \beta_1 X$$

  ya que la función de regresión relaciona las medias de las distribuciones de probabilidad de Y para X dado el nivel de X.
- La respuesta Y_i en el i-ésimo ensayo supera o no alcanza el valor de la función de regresión por la cantidad del término de error ϵ_i.
- Los términos de error se asumen que tienen varianza constante σ^2. Esto implica que las respuestas Y_i tienen la misma varianza constante:

  $$\sigma^2\{Y_i\} = \sigma^2$$

  Así, el modelo de regresión (1) asume que las distribuciones de probabilidad de Y tienen la misma varianza σ^2, independientemente del nivel de la variable predictora X.

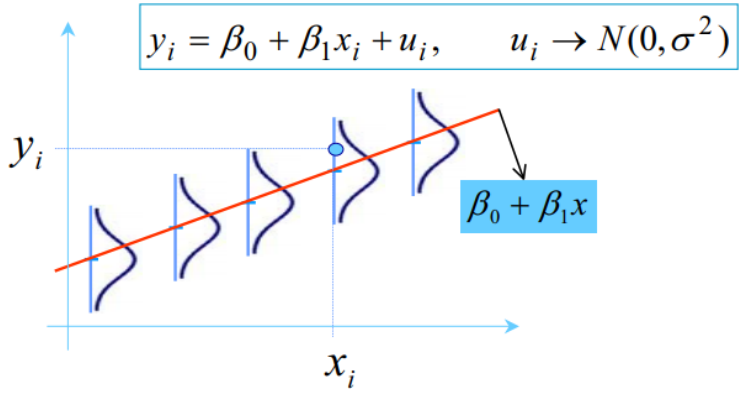
- Los términos de error se supone que no están correlacionados. Dado que los términos de error ϵ_i y ϵ_i no están correlacionados, también lo están las respuestas Y_i y Y_j.

Supuestos

El modelo de regresión lineal simple, como tantísimos otros (casi absolutamente todos) modelos estadísticos, deberá cumplir una serie de supuestos que permitan concluir que el mismo es una buena versión simplificada de la información y, por tanto, tiene sentido usarlo para establecer dichas relaciones ded causación. De esta manera, los principales supuestos para el modelo de regresión lineal simple $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ son

- [S0] El modelo es una buena representación de la realidad (tiene sentido)
- [S1] $E(\epsilon_t) = 0$
- [S2] $Cov(\epsilon_t, \epsilon_s) = 0$ siempre que $i \neq j$
- [S3] $Var(\epsilon_t) = \sigma_t^2 = \sigma^2$ (Homoscedasticidad)
- [S4] x asume por lo menos dos valores
- [S5] [opcional pero recomendado] $y_t \sim N(\mu_t, \sigma^2)$

Típicamente, en todos los modelos estadísticos, la forma e validar los supuestos es a través de los residuos (Hay que tener presente que los errores son variables aleatorios, mientras que los residuos son realizaciones de tales variables). De esta manera se definen los residuos como $r_t = y_t - \hat{y}_t$, donde $\hat{y}_t$ es el valor predicho para y_t por el modelo



Resultados Generales

Frente al modelo dgeneral de Regresión Lineal Simple, se tienen los siguientes resultados generales

Estimación de parámetros

En el modelo $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ se debe realizar un proceso de estimación de parámetros para dotar de sentido al mismo. Es decir, se requiere algún procedimiento para, partiendo de la información de entrada (variables x y y), poder obtener $\hat{\beta}_0$ y $\hat{\beta}_1$.

Para encontrar *buenos* estimadores de los parámetros de regresión β_0 y β_1, empleamos el **método de los mínimos cuadrados**. Para las observaciones (x_i, y_i) para cada caso, el método de los mínimos cuadrados considera la desviación de Y_i de su valor esperado:

$$y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

El método de los mínimos cuadrados requiere que se considera la suma de las n desviaciones al cuadrado. Este criterio se denota por Q:

$$Q = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

De acuerdo con el método de los mínimos cuadrados, los estimadores de β_0 y β_1 son los valores $\hat{\beta}_0$ y $\hat{\beta}_1$, respectivamente, que minimizan el criterio Q para las observaciones muestrales dadas $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$.

puede demostrar para el modelo de regresión (1) que los valores de b_0 y b_1 que minimizan Q para cualquier conjunto particular de datos muestrales están dados por las siguientes ecuaciones simultaneas:

$$\sum y_i = n\beta_0 + \beta_1 \sum x_i \tag{2}$$

$$\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \tag{3}$$

Las ecuaciones (2) y (3) son denominadas **ecuaciones normales**.

Las ecuaciones normales se pueden resolver de forma simultánea para β_0 y β_1:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{4}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5}$$

donde $\bar{x}$ y $\bar{y}$ son las medias de las observaciones x_i y y_i, respectivamente. Estos estimadores, se puede probar, son insesgados, consistentes y eficientes. Adicionalmente, el **Teorema de Gauss-Markov** establece que estos estimados son los mejores estimadores dentro de la categoría de estimadores lineales.

pruebas de hipótesis

En general, va a ser de nuestro interés obtener evidencia empírica de la validez estadística de los parámetros que indexan el modelo. Para esto, se hace necesario introducir pruebas de hipótesis para los parámetros.

Considere, la prueba de las hipótesis para β_i:

$$H_0 : \beta_i = \beta_{i0} (\beta_{i0} \text{conocido})$$
$$H_1 : \beta_i \neq \beta_{i0}$$

que tiene como estadístico de prueba:

$$T_{Est} = \frac{\hat{\beta}_1 - \beta_{10}}{s\{\hat{\beta}_1\}} \sim t_{(n-2)}$$

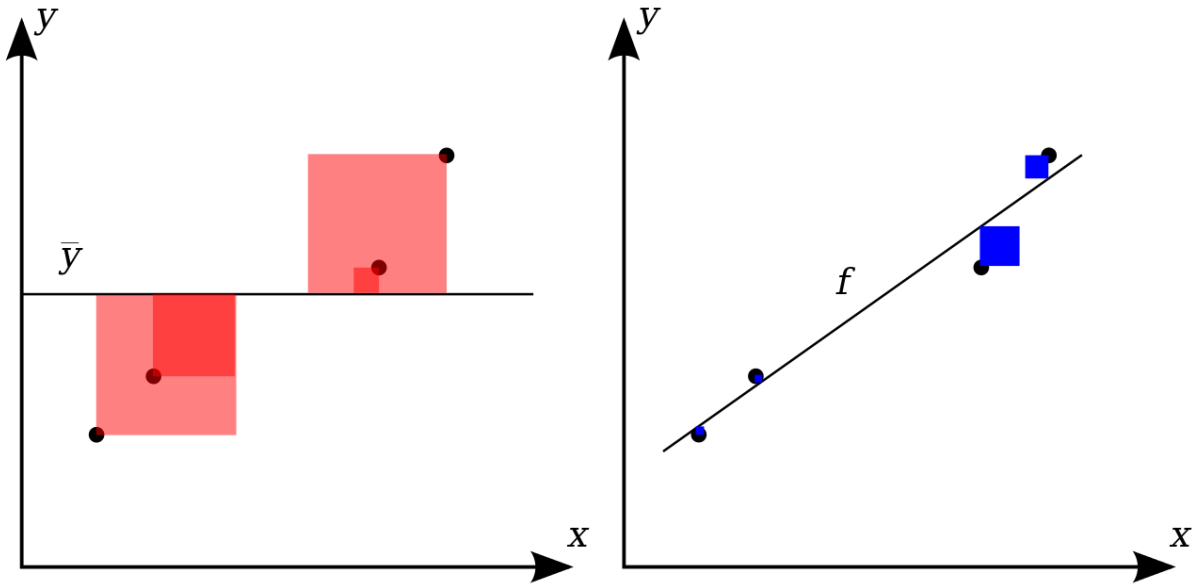
y la regla de decisión para el nivel de significancia α es p-valor $< \alpha$ donde p-valor $= 2P(T > |T_{Est}|)$ con $T \sim t_{(n-1)}$.

Medidas de bondad de ajuste

En la regresión lineal simple se mostró que la suma de cuadrados se podía dividir o particionar en dos componentes: la suma de cuadrados debida a la regresión y la suma de cuadrados debida al error. Esto también aplica a la suma de cuadrados de la regresión múltiple.

Relación entre SCT, SCR y SCE: $SCT = SCR + SCE$, donde

$$SCT = \text{suma de cuadrados total} = \sum (y_i - \bar{y})^2$$
$$SCR = \text{suma de cuadrados debida a la regresión} = \sum (\hat{y}_i - \bar{y})^2$$
$$SCE = \text{suma de cuadrados debida al error} = \sum (y_i - \hat{y}_i)^2$$



Debido a lo complejo de los cálculos de estas tres sumas de cuadrados, es necesario emplear un paquete de software para realizarlos. En los resultados de R, en la parte de análisis de varianza, se presentan los valores: SCT, SCR y SCE.

El valor de la SCT es el mismo en ambos casos debido a que este valor no depende de $\hat{y}$, pero al agregar otra variable (el número de entregas), SCR aumenta y SCE disminuye. Esto tiene como consecuencia que la ecuación de regresión estimada tenga un mejor ajuste a los datos observados.

El término **coeficiente de determinación múltiple** indica que mide la bondad de ajuste de la ecuación de regresión múltiple estimada. El coeficiente de determinación múltiple, que se denota R^2, se calcula como $R^2 = \frac{SCR}{SCT}$ o como $R^2 = 1 - \frac{SCE}{SCT}$ y puede interpretarse como la proporción de la variabilidad en el valor de la variable independiente que es explicada por la ecuación de regresión estimada. Por lo tanto, el

producto de este coeficiente por 100, se interpreta como el porcentaje de la variabilidad en y que es explicada por la ecuación de regresión estimada.

Validación de Supuestos

Una vez planteados los supuestos, es necesario ver cómo se validarán en ejemplos prácticos. En esta situación, simularemos los datos para poder validar cada uno de los supuestos.

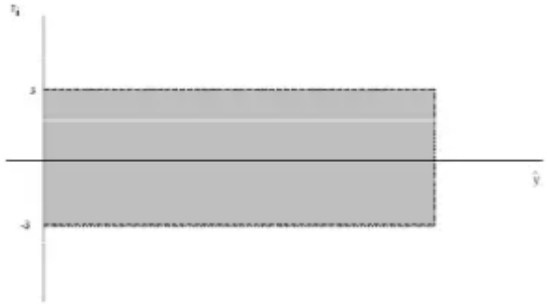


Gráfico satisfactorio de análisis de residuos

Gráficos típicamente insatisfactorios de análisis de residuos

In [ ]:

```
set.seed(123)
x <- rnorm(250, mean=100, sd=5)
y <- 58 + 1.8*x+ rnorm(250, sd=3)
options(repr.plot.width=6, repr.plot.height=6)
plot(x,y, main= "Gráfico de Dispersión")
```

Al haber simulado el modelo $y_t = 58 + 1.8x_t + \epsilon_t$ ($\beta_0 = 58$ y $\beta_1 = 1.8$), debe ser natural que el modelo ajustado sea similar, veamos.

In [ ]:
```
ajuste <- lm(y~x)
options(repr.plot.width=6, repr.plot.height=6)
plot(x,y, main= "Gráfico de Dispersión")
summary(ajuste)
abline(ajuste, col="red")
```

In [ ]:
```
options(repr.plot.width=10, repr.plot.height=10)
par(mfrow=c(2,2))
plot(ajuste)
```

- [S0] El modelo es una buena representación de la realidad (tiene sentido)

El día que exista una prueba de hipótesis que verifique este supuesto, todos todos los que hagamos estadística, nos quedaremos sin trabajo.

- [S1] $E(\epsilon_t) = 0$

In [ ]:
```
residuos <- residuals(ajuste)
mean(residuos)
```

- [S2] $Cov(\epsilon_t, \epsilon_s) = 0$ siempre que $i \neq j$

In [ ]:
```
# install.packages("lmtest")
library(lmtest) # ver https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic
dwtest(ajuste) #H0: No hay autocorrelación en los errores
```

- [S3] $Var(\epsilon_t) = \sigma_t^2 = \sigma^2$ (Homoscedasticidad)

In [ ]:
```
library(lmtest) #https://en.wikipedia.org/wiki/Breusch%E2%80%93Pagan_test
bptest(ajuste) #H0: Homoscedasticidad
```

- [S4] x asume por lo menos dos valores

Estadística Descriptiva

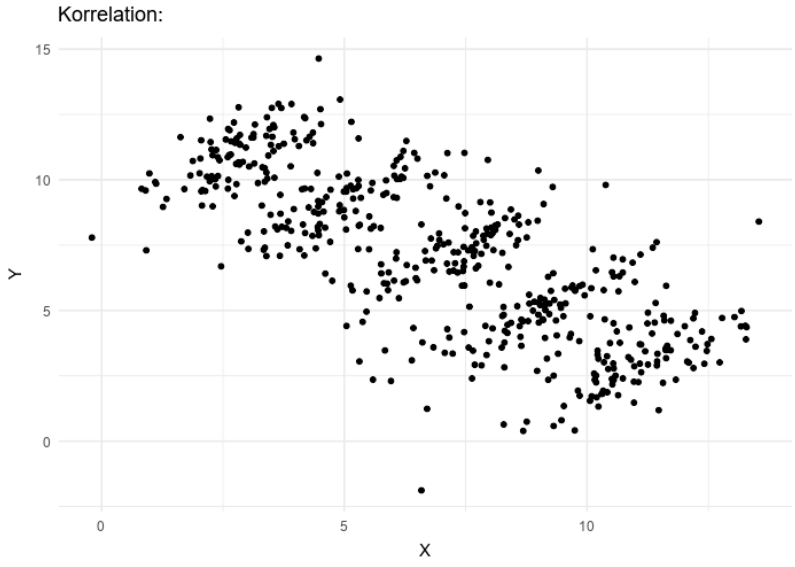
- [S5] [opcional pero recomendado] $y_t \sim N(\mu_t, \sigma^2)$

In [ ]:
```
# install.packages("tseries")
library(tseries)#https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test
jarque.bera.test(residuos) #H0: Normalidad
```

Algunas Consideraciones

Paradoja de Simpson

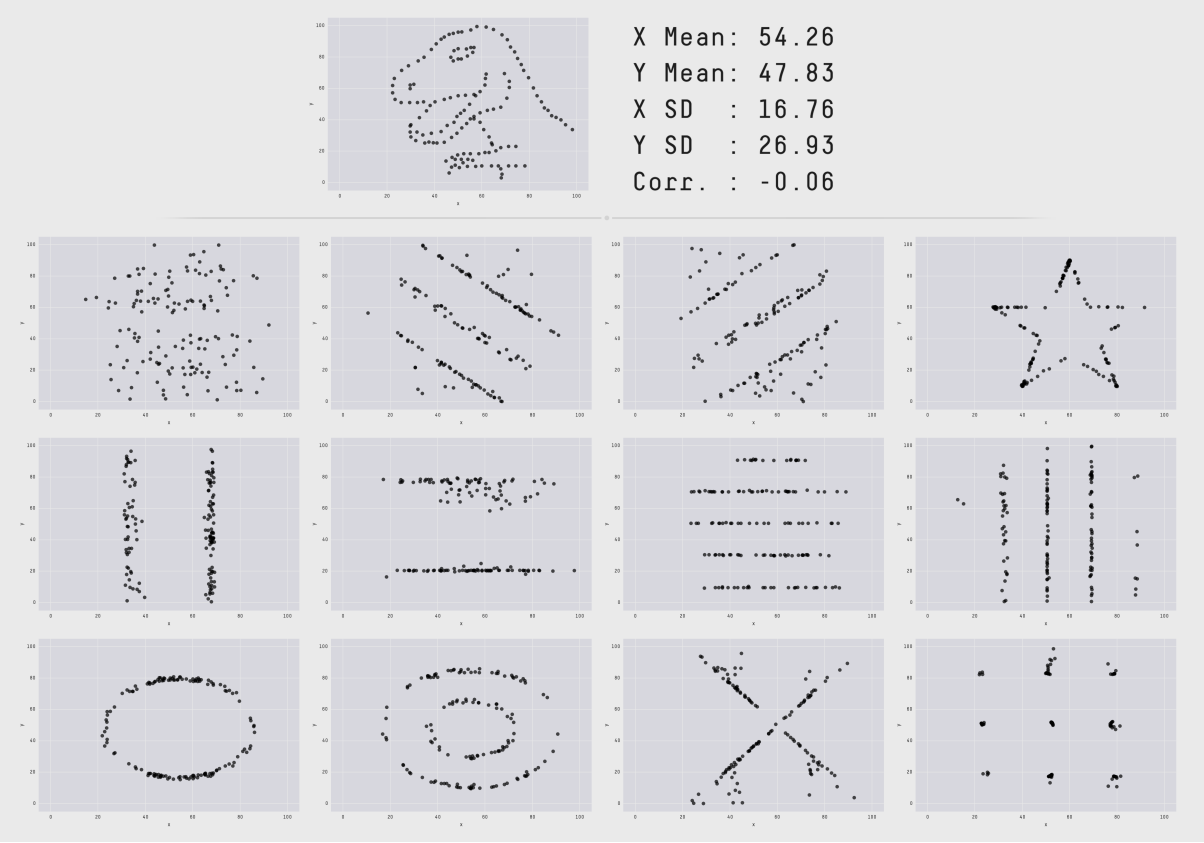
Lectura recomendada aquí

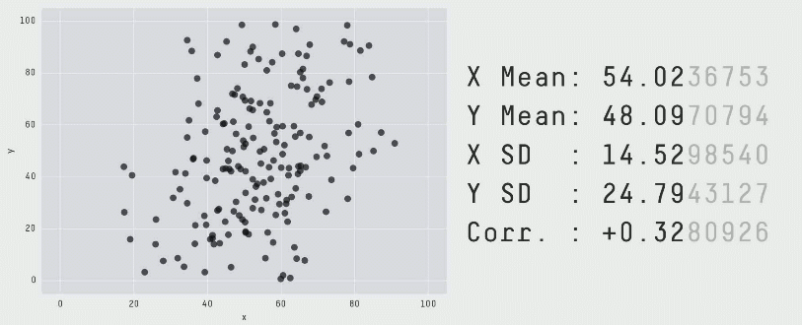


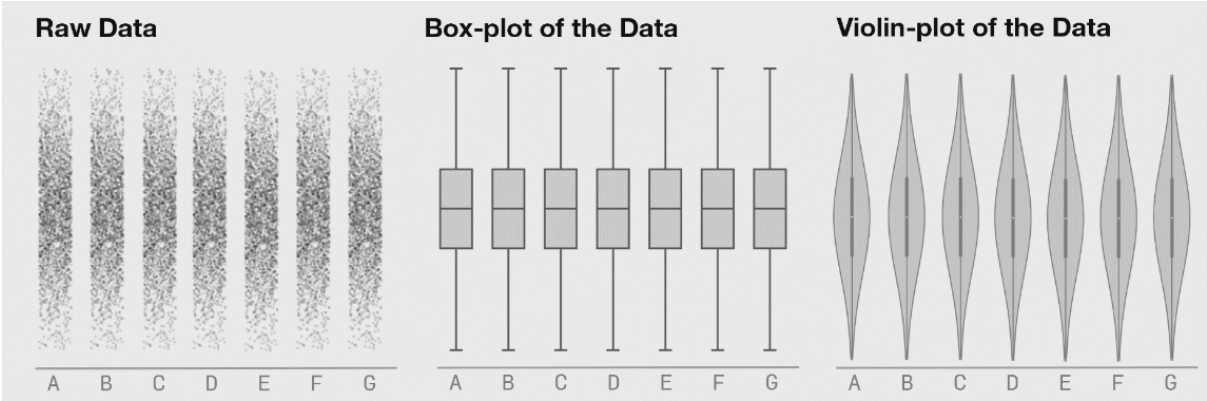
Conjunto de datos de Anscombe

Lectura recomendada aquí

Estadísticas de Resumen







Ejemplo de Aplicación

En este caso, estableceremos una comprensión básica de las estadísticas necesarias para la regresión lineal y luego introduciremos la regresión lineal comenzando con 2 parámetros. Asuntos como la interpretación de los coeficientes y la comprensión de varias métricas para evaluar correctamente el rendimiento del modelo, serán presentados.

Para este ejemplo de aplicación se escogió una base de datos que da cuenta de la distribución de pagos en una empresa ¿Hay discriminación por género?

In [ ]:
```
base <- read.csv('glassdoordata.csv')
head(base)
```

Las variables son

- **job title**: Título del trabajo (e.g. “Graphic Designer”, “Software Engineer”, etc);
- **gender**: Hombre o mujer;
- **age**: edad;
- **performance**: en escala del 1 al 5, 1 siendo el más bajo y 5 siendo la más alta;
- **education**: niveles de educación (e.g. "College", "PhD", "Masters", "Highschool");
- **department**: diferentes departamentos (e.g. "Operations", "Management", etc);
- **seniority**: en escala del 1 al 5, 1 siendo el más bajo y 5 siendo la más alta;
- **income, bonus**: Expresados en dólares

Pequeña transformación (feature engineering)

In [ ]:
```
base$pay <-  base$income + base$bonus
head(base,10)
```

Ejercicio 1

Construya un boxplot comparando los salarios entre hombres y mujeres

Respuesta.

In [ ]:
```
# install.packages("ggplot2")
library(ggplot2)
options(repr.plot.width=6, repr.plot.height=6)
ggplot(base, aes(x=gender, y=pay, col=gender))+geom_boxplot()
```

Ejercicio 2

Establezca por medio de una prueba t-studente si hay diferencia significativa entre los grupos

Respuesta.

In [ ]:
```
library(tidyverse)
x <- base %>% filter(gender == 'Male') %>% select(pay)
y <- base %>% filter(gender == 'Female') %>% select(pay)
t.test(x,y,alternative ="two.sided")
```

Ejercicio 3

Construya un boxplot de pay, con respecto a las siguientes variables: seniority, education, jobtitle and performance. También realice un scatterplot de pay vs. age. Algo que llame su atención?

Respuesta.

In [ ]:
```
library(gridExtra)
```

```
options(repr.plot.width=8, repr.plot.height=8)
fig1 <- ggplot(base, aes(x=factor(seniority), y=pay))+geom_boxplot()
fig2 <- ggplot(base, aes(x=factor(education), y=pay))+geom_boxplot()
fig3 <- ggplot(base, aes(x=factor(jobtitle), y=pay))+geom_boxplot()+theme(axis.text.x = element_text(angle = 30, vjus
fig4 <- ggplot(base, aes(x=factor(performance), y=pay))+geom_boxplot()
grid.arrange(fig1,fig2,fig3,fig4, ncol=2, nrow=2)
```

In [ ]:

```
options(repr.plot.width=6, repr.plot.height=6)
ggplot(base, aes(x=age, y=pay))+geom_point()
```

Ejercicio 4

Construya las mismas figuras del punto anterior pero esta vez mediado por género

Respuesta.

In [ ]:

```
library(gridExtra)
options(repr.plot.width=8, repr.plot.height=8)
fig1 <- ggplot(base, aes(x=factor(seniority), y=pay, col=gender))+geom_boxplot()
fig2 <- ggplot(base, aes(x=factor(education), y=pay, col=gender))+geom_boxplot()
fig3 <- ggplot(base, aes(x=factor(jobtitle), y=pay, col=gender))+geom_boxplot()+theme(axis.text.x = element_text(angl
fig4 <- ggplot(base, aes(x=factor(performance), y=pay, col=gender))+geom_boxplot()
grid.arrange(fig1,fig2,fig3,fig4, ncol=2, nrow=2)
```

Ejercicio 5

Construya un modelo de regresión lineal entre las variables pay y age

Respuesta.

In [ ]:

```
ajuste <- lm(pay~age, data=base)
options(repr.plot.width=4, repr.plot.height=4)
ggplot(base, aes(x=age, y=pay))+stat_smooth(method="lm", se=FALSE)+geom_point()
summary(ajuste)
```

Ejercicio 5

Valide los supuestos del modelo anterior.

Respuesta

In [ ]:
```
options(repr.plot.width=8, repr.plot.height=8)
par(mfrow=c(2,2))
plot(ajuste)
```

In [ ]:
```
residuales <- residuals(ajuste)
mean(residuales)
```

In [ ]:
```
dwtest(ajuste) #H0: No hay autocorrelación en los errores
```

In [ ]:
```
bptest(ajuste) #H0: Homoscedasticidad
```

In [ ]:
```
jarque.bera.test(residuos) #H0: Normalidad
```

Ejercicio 6

Conclusiones del caso. ¿En efecto, hay evidencia que sugiera discriminación por género?

Respuesta.

Utilizando las herramientas que tenemos a mano, parece sensato pensar en un modelo de regresión para resolver esta pregunta. Por tanto, ajustemos uno.

In [ ]:
```
ajuste_final <- lm(pay~gender, data=base)
summary(ajuste_final)
```

La validación del modelo es una **TAREA** que recomiendo enormemente. Este modelo da la sensación que la respuesta a la pregunta "Hay algún

tipo de discriminación por género" es sí. Sin embargo, en la clase 5, retomaremos este caso y lo estudiaremos con más detalle.