

# Análisis de Regresión (2021-3)



**LOS LIBERTADORES**  
FUNDACIÓN UNIVERSITARIA

## Especialización en Estadística Aplicada

Prof. [Sébastien Lozano Forero](mailto:slozanof@libertadores.edu.co) (slozanof@libertadores.edu.co)

## Introducción a la Regresión Lineal Múltiple

### Tabla de contenidos

- [Modelo de Regresión Lineal Múltiple](#)
- [Resultados Generales](#)
- [Supuestos](#)
- [Validación de Supuestos](#)
- [Influencia y Diagnóstico](#)
- [Ejemplo](#)

### Modelo de Regresión Lineal Múltiple

El modelo de regresión lineal simple da cuenta de la relación **lineal** y **causal** entre las variables  $X_1, X_2, \dots, X_n, Y$  ( $X_1, X_2, \dots, X_n$  causan a  $Y$ ).

Continuamos con el estudio del análisis de regresión considerando, ahora, las situaciones en las que intervienen dos o más variables predictoras o independientes. Este estudio, al que se le conoce como **análisis de regresión múltiple**, permite tomar más factores en consideración y obtener estimaciones mejores que las que son posibles con la regresión lineal simple.

El análisis de regresión múltiple estudia la relación de una variable respuesta con dos o más variables predictoras. Para denotar el número de variables predictoras independientes se suele usar  $p$ . A la ecuación que describe cómo está relacionada la variable respuesta  $Y$  con las variables

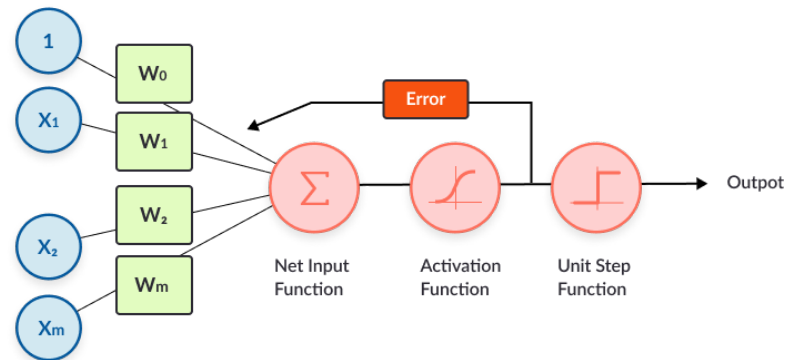
predictoras  $X_1, X_2, \dots, X_p$  se le conoce como **modelo de regresión múltiple** y se escribe de la siguiente forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

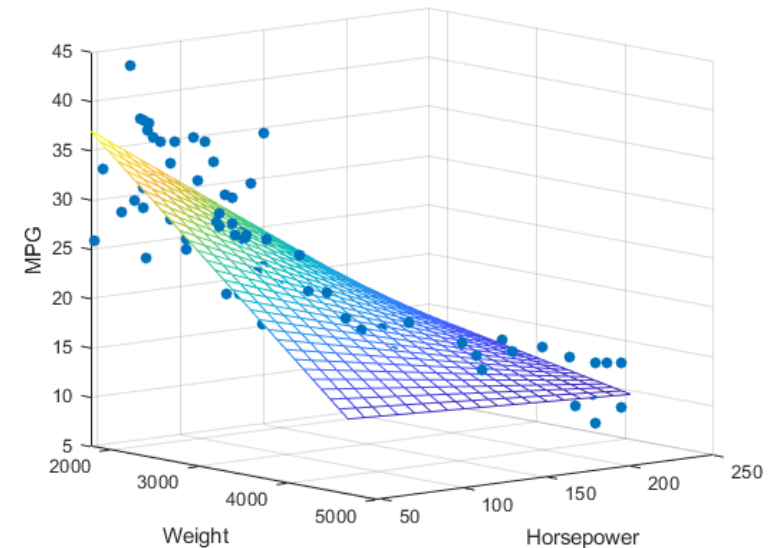
Donde:

- $\beta_0, \beta_1, \dots, \beta_p$  son parámetros poblacionales a ser estimados.
- $\epsilon$  representa un error aleatorio (Nuestra incapacidad para dar cuenta de la realidad tal cuál es)
- $Y$  representa la *variable respuesta*
- $X_1, \dots, X_p$  representan las *variables predictoras*

Si se conocieran los valores de  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ , se podría usar la ecuación (2) para calcular la media de las  $Y$  para valores dados de  $X_1, X_2, \dots, X_p$ .



Regression in neural networks



Desafortunadamente, los valores de estos parámetros no suelen conocerse, es necesario estimarlos a partir de datos muestrales.

Para calcular los valores de los estadísticos muestrales  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ , que se usan como estimadores puntuales de los parámetros  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  se emplea una muestra aleatoria simple. Con los estadísticos muestrales se obtiene la siguiente **ecuación de regresión múltiple estimada**

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p$$

donde  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  son las estimaciones de  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  y  $\hat{Y}$  es el valor estimado de la media de  $Y$ .

## Supuestos

El modelo de regresión lineal múltiple, 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 de causalidad. De esta manera, los principales supuestos para el modelo de regresión lineal simple

$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \epsilon_t$  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]  $X$  es de rango completo
- [S5] [opcional pero recomendado]  $y_t \sim N(\mu_t, \sigma^2)$

Típicamente, en todos los modelos estadísticos, la forma de validar los supuestos es a través de los residuos (Hay que tener presente que los errores son variables aleatorias, 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

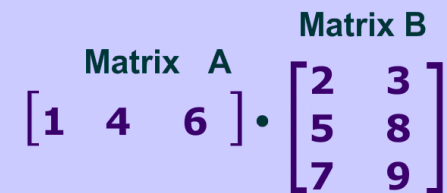
Podemos escribir el modelo de regresión muestral correspondiente al modelo (1) como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \dots, n$$

En notación matricial, este modelo de regresión muestral se representa como

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

donde


$$\begin{matrix} & \text{Matrix A} & & \text{Matrix B} \\ & [1 & 4 & 6] & \cdot & \begin{bmatrix} 2 & 3 \\ 5 & 8 \\ 7 & 9 \end{bmatrix} \end{matrix}$$

© mathwarehouse.com

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Deseamos encontrar el vector de estimadores de mínimos cuadrados,  $\boldsymbol{\beta}$ , que minimiza

$$Q(\boldsymbol{\beta}) = \sum r_i^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Podemos probar que el estimador de mínimos cuadrados de  $\boldsymbol{\beta}$  es

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

siempre que la matriz inversa  $(\mathbf{X}'\mathbf{X})^{-1}$  exista.

## Medidas de bondad de ajuste

### R cuadrado ajustado

Muchos analistas prefieren ajustar  $R^2$  al número de variables independientes para evitar sobreestimar el efecto que tiene agregar una variable independiente sobre la cantidad de la variabilidad explicada por la ecuación de regresión estimada. Siendo  $n$  el número de observaciones y  $p$  el número de variables independientes, el coeficiente de determinación ajustado se calcula como sigue.

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

### Criterios de selección

El AIC se define como:

$$AIC = -2\log\text{Lik} + pk,$$



donde  $\log\text{Lik}$  corresponde al valor de log-verosimilitud del modelo para el vector de parámetros  $\beta$ ,  $p$  es un valor de penalización por el exceso de parámetros y  $k$  corresponde al número de parámetros del modelo.

Se debe recordar siempre que:

- El mejor modelo es aquel que  $\log\text{Lik}$  alto
- El mejor modelo es aquel que AIC bajo

Cuando el valor de penalización  $k = \log(n)$ , entonces el AIC se llama BIC (Schwarz's Bayesian criterion)

## 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  $\beta_i$ :

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

$$H_1: \beta_i \neq \beta_{i0}$$

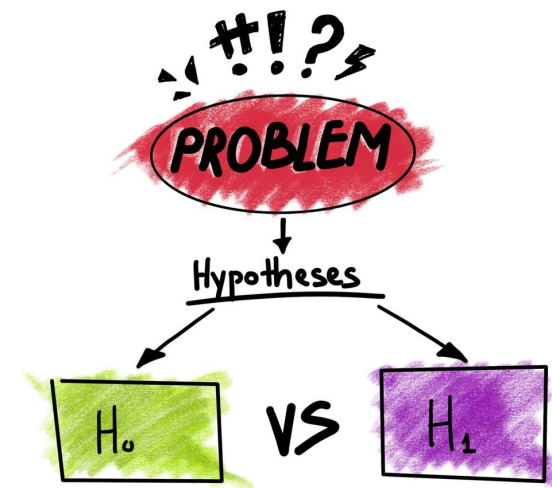
que tiene como estadístico de prueba:

$$T_{Est} = \frac{\hat{\beta}_1 - \beta_{10}}{s\{\hat{\beta}_1\}} \sim t_{(n-p-1)}$$

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

## 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.



@luminousmen.com

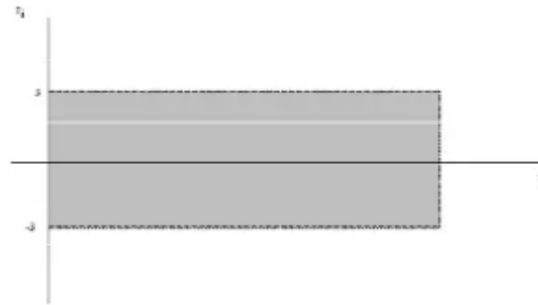


Gráfico satisfactorio de análisis de residuos



Gráficos típicamente insatisfactorios de análisis de residuos

```
In [ ]: set.seed(123)
x <- rnorm(250, mean=10, sd=5) # Distribución normal
z <- rexp(250, rate=0.1) #Distribución exponencial
w <- rbinom(250, 100,0.25) #Distribución Binomial
y <- 58 + 1.8*x+4.3*z +5*w + rnorm(250, sd=3)
head(data.frame(x=x,y=y,z=z,w=w))
```

Al haber simulado el modelo  $y_t = 58 + 1.8x_t + 4.3z_t + 5w_t + \epsilon_t$  ( $\beta_0 = 58, \beta_1 = 1.8, \beta_2 = 4.3$  y  $\beta_3 = 5$ ), debe ser natural que el modelo ajustado sea similar, veamos.

```
In [ ]: ajuste <- lm(y ~ x + z + w) #tarea: glm
summary(ajuste)
```

```
In [ ]:
```

```
In [ ]: ww<- rnorm(250, mean=15, sd=2)
ajuste2 <- lm(y~x+z+ww)
summary(ajuste2)
```

```
In [ ]: options(repr.plot.width=12, repr.plot.height=12)
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]  $Cov(\epsilon_i, \epsilon_j) = 0$  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]  $X$  es de rango completo

```
In [ ]: cor(data.frame(x=x, z=z, w=w)) #Matriz de correlaciones
```

```
In [ ]: # install.packages("car")
library(car)
vif(ajuste) #Autovalores de la matriz de correlaciones
```

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

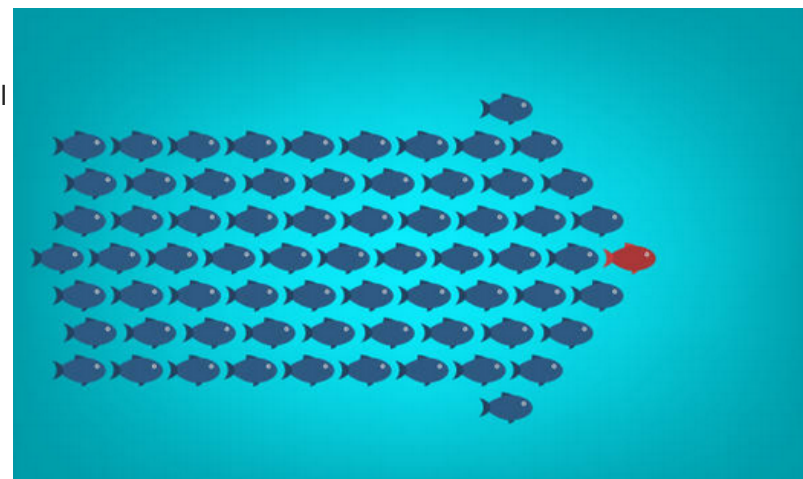
```
In [ ]: # install.packages("tseries")
library(tseries)#https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test
jarque.bera.test(residuos) #H0: Normalidad
```

## Influencia y Diagnóstico

### Problemas con los errores

#### Posibles problemas y controles de diagnóstico

- Es posible que los errores no se distribuyan normalmente o que no tengan el mismo varianza: qqnorm puede ayudar con esto. Puede que esto no sea demasiado importante en muestras grandes.
- La varianza puede no ser constante. También se puede abordar en una gráfica de  $X$  frente a  $e$  u otra tendencia indica varianza no constante.
- Observaciones influyentes. Qué puntos "afectan" la línea de regresión lo mas?
- Valores atípicos: puntos en los que el modelo realmente no encaja.  
Posiblemente errores en la transcripción de datos, errores de laboratorio, ¿quién sabe? Debiera ser reconocido y (con suerte) explicado.





## Tipos de residuos

- Residuos ordinarios:  $\epsilon_i = Y_i - \hat{Y}_i$ . Estos miden la desviación del valor predicho del valor observado, pero su distribución depende de una escala desconocida,  $\sigma$ .
- Residuos estudiados internamente ( `rstandard` en R):

$$r_i = \epsilon_i / SE(\epsilon_i) = \frac{\epsilon_i}{\hat{\sigma} \sqrt{1 - H_{ii}}}$$

- Arriba,  $H$  es la matriz de "sombrero"  $H = X(X^T X)^{-1} X^T$ . Estos son casi  $t$  distribuidos, excepto  $\hat{\sigma}$  que depende de  $\epsilon_i$ .
- Residuos estudiados externamente ( `rstudent` en R):

$$t_i = \frac{\epsilon_i}{\hat{\sigma}_{(i)} \sqrt{1 - H_{ii}}} \sim t_{n-p-2}.$$

Estos son exactamente  $t$  distribuidos, por lo que conocemos su distribución y puede usarlos para pruebas, si lo desea.

- La cantidad  $\hat{\sigma}_{(i)}^2$  es el MSE del modelo ajustado a todos los datos excepto al caso  $i$  (es decir, tiene  $n - 1$  observaciones y  $p$  características).
- Numéricamente, estos residuos están altamente correlacionados, como era de esperar.

```
In [ ]: options(repr.plot.width=4, repr.plot.height=4)
plot(resid(ajuste), rstudent(ajuste), bg='blue')
```

```
In [ ]: options(repr.plot.width=4, repr.plot.height=4)
plot(rstandard(ajuste), rstudent(ajuste), bg='blue')
```

Influencia de una observación



Otras gráficas proporcionan una evaluación de la "influencia" de cada observación. Por lo general, esto se hace eliminando un caso completo ( $y_i, x_i$ ) del conjunto de datos y reacondicionamiento del modelo.

- En esta configuración,  $\cdot_{(i)}$  indica que la observación  $i$ -ésima no se utiliza para ajustar el modelo.
- Por ejemplo:  $\hat{Y}_{j(i)}$  es la función de regresión evaluados en los predictores de observación  $j$ -ésimo PERO los coeficientes  $(\hat{\beta}_{0(i)}, \dots, \hat{\beta}_{p(i)})$  estaban en forma después de eliminar  $i$ -ésimo caso de los datos.
- Idea: si  $\hat{Y}_{j(i)}$  es muy diferente de  $\hat{Y}_j$  (usando todos los datos) entonces  $i$  es un punto influyente, al menos para estimando la función de regresión en  $(X_{1,j}, \dots, X_{p,j})$ .
- También podría ver la diferencia entre  $\hat{Y}_{i(i)} - \hat{Y}_i$ , o cualquier otra medida.
- Hay varias medidas estándar de influencia.

DFFITS

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{H_{ii}}}$$

- Esta cantidad mide cuánto cambia la función de regresión en el  $i$ -ésimo caso / observación cuando el  $i$ -ésimo caso / observación es eliminada.
- Para conjuntos de datos pequeños / medianos: el valor de 1 o más es "sospechoso" (RABE). Para un conjunto de datos grande: valor de  $2\sqrt{(p+1)/n}$ .
- R tiene sus propias reglas estándar similares a las anteriores para marcar una observación tan influyente.

```
In [ ]: 2*sqrt(4/250)
```

```
In [ ]: plot(dffits(ajuste), pch=23, bg='orange', cex=0.5, ylab="DFFITS")
```

In [ ]:

Distancia de Cook

La distancia de Cook mide cuánto toda la función de regresión cambia cuando se elimina el caso  $i$ -ésimo.

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\hat{\sigma}^2}$$

- Debe ser comparable a  $F_{p+1, n-p-1}$ : si el " $p$ -valor" de  $D_i$  es del 50 por ciento o más, entonces el caso de  $i$ -ésimo es probablemente influyente: investigar más a fondo. (RABE)
- Nuevamente, **R** tiene sus propias reglas similares a las anteriores para marcar una observación tan influyente.
- ¿Qué hacer después de la investigación? No es una respuesta fácil.

In [ ]:

```
plot(cooks.distance(ajuste), pch=23, bg='orange', cex=0.5, ylab="Cook's distance")
```

DFBETAS

Esta cantidad mide cuánto cambian los coeficientes cuando el  $i$ -ésimo caso se elimina.

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}_{(i)}^2 (X^T X)^{-1}_{jj}}}$$

Para conjuntos de datos pequeños / medianos: el valor absoluto de 1 o mayor es "suspicaaz". Para un conjunto de datos grande: valor absoluto de  $2/\sqrt{n}$ .

In [ ]:

```
2/sqrt(250)
```

In [ ]:

```
plot(dfbetas(ajuste)[, 'x'], pch=23, bg='orange', cex=0.5, ylab="DFBETA (x)")
```

```
In [ ]: plot(dfbetas(ajuste)[,'z'], pch=23, bg='orange', cex=0.5, ylab="DFBETA (z)")
```

#### Valores atípicos

La definición esencial de un *valor atípico* es un par de observación  $(Y, X_1, \dots, X_p)$  que no sigue el modelo, mientras que la mayoría de las otras observaciones parecen seguir el modelo.

- Valor atípico en *predictores* : los valores de  $X$  de la observación pueden estar fuera de la "nube" de otros valores  $X$ . Esto significa que puede ser extrapolando su modelo de manera inapropiada. Los valores  $H_{ii}$  pueden ser se utiliza para medir qué tan "atípicos" son los valores de  $X$ .
- Valor atípico en *respuesta*: el valor  $Y$  de la observación puede ser muy lejos del modelo ajustado. Si los residuales studentizados son grandes: la observación puede ser un valor atípico.



#### Valores atípicos en $X$

Una forma de detectar valores atípicos en los *predictores*, además de observar los valores reales en sí mismos, es a través de sus valores de apalancamiento, definidos por

$$\text{leverage}_i = H_{ii} = (X(X^T X)^{-1} X^T)_{ii}.$$

```
In [ ]: plot(hatvalues(ajuste), pch=23, bg='orange', cex=0.5, ylab='Hat values')
```

Hasta aquí, parece terrible hacer estudios de influencia y diagnóstico. La matemática es compleja y ésta es un área todavía en estudio, por lo tanto, las cosas pueden cambiar en los próximos años. Sin embargo, no debe generar preocupación el uso de toda esta información para determinar cuáles puntos deben ser influyentes o atípicos de forma desmedida.

```
In [ ]: summary(influence.measures(ajuste))
```

## Ejemplo

Para este ejemplo retomaremos el conjunto de datos estudiado en la clase 3. Esta es la 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
```

```
In [ ]: head(base, 10)
```

## Ejercicio 1

Construya la matriz de correlaciones de entre las variables cuantitativas de la base de datos. Qué variables parecen razonables para usar en un modelo de regresión?

**Respuesta.**

```
In [ ]: cor(base[,c("age", "performance", "seniority", "pay")])
```

```
In [ ]: # install.packages("corrplot")
library(corrplot)
options(repr.plot.width=8, repr.plot.height=8)
corrplot(cor(base[,c("age", "performance", "seniority", "pay")]), method="circle")
```

## Ejercicio 2

Construya Figuras que permitan una comparación en la variable pay, indexada por género con las variables cualitativas de la base de datos. Qué variables parecen razonables para usar en un modelo de regresión?

**Respuesta.**

```
In [ ]: library(gridExtra)
library(ggplot2)
options(repr.plot.width=16, repr.plot.height=12)
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(angle=45))
fig4 <- ggplot(base, aes(x=factor(performance), y=pay, col=gender))+geom_boxplot()
grid.arrange(fig2,fig3,fig4, ncol=2, nrow=2)
```

## Ejercicio 3

Conluya si tiene sentido suponer que la formación académica influye en el pago.

```
In [ ]: levels(base$education)
```

**Respuesta.**

```
In [ ]: ajuste2 <- lm(pay ~ factor(gender) + factor(education), data=base)
summary(ajuste2)
```

## Ejercicio 5

Incluya la variable age en el modelo. Tiene sentido el modelo obtenido?

**Respuesta.**

```
In [ ]: ajuste3 <- lm(pay~gender + age + factor(education), data=base)
summary(ajuste3)
```

## Ejercicio 6

Qué hace la línea de código `base[which(dffits(ajuste) > 0.5),]` ? Qué puede interpretar de esto?

**Respuesta.**

```
In [ ]: base[which(dffits(ajuste) > 0.5),]
```

```
In [ ]: summary(influence.measures(ajuste))
```

Hay una observación que resalta por la diferencia que el modelo prediciría su salario.

## Ejercicio 7

Monte el modelo saturado (con todas las variables). Algún cambio que llame su atención?

**Respuesta.**

```
In [ ]: levels(base$jobtitle)
```

```
In [ ]: ajuste3 <- lm(pay~factor(jobtitle) + age + factor(performance) + factor(education) + factor(department) + factor(seniority), data=base)
summary(ajuste3)
```

## Ejercicio 8

Teniendo el modelo anterior, cuáles son las áreas que mejor pagan?

- (a) Marketing Associate
- (b) Software Engineer
- (c) Manager
- (d) Graphic Designer

**Respuesta.**

- (b) Manager

Considerando la diferencia en el  $R^2$  de más de 80% entre el modelo más sencillo y el modelo más complejo, siendo género incluido en ambos, quiere decir que hay incidencia importante sobre la variabilidad de la información en las otras variables.

## Ejercicio 9:

Basado en lo anterior, considere las siguientes afirmaciones.

I. Después de tener en cuenta job title, education, performance y age, la proporción de la diferencia salarial atribuible únicamente al género es pequeña.

II. Existe evidencia de que la discriminación salarial entre hombres y mujeres se debe únicamente al género.

III. Hay motivos para creer que podría haber una cantidad desproporcionada de mujeres en trabajos peor pagados como marketing, mientras que podría haber más hombres en trabajos mejor pagados, como gerente.

Elija la respuesta correcta:

- (a) I es correcto, II y III son incorrectos.
- (b) II es correcto, I y III son incorrectos.



(c) I es incorrecta, II y III son correctas.

(d) I y III son correctos, II es incorrecto.

**Respuesta.**

(d) I y III son correctos, II es incorrecto.

## Ejercicio 10:

Establezca la cantidad de hombres y mujeres en los niveles de seniority y en jobtitle. Algo que llame su atención?

```
In [ ]: counts <- table(base$seniority, base$gender)
counts
```

```
In [ ]: counts <- table(base$jobtitle, base$gender)
counts
```

```
In [ ]: ggplot(base, aes(x=factor(jobtitle), y=pay, col=gender))+
  geom_boxplot()+
  theme(axis.text.x = element_text(angle = 30, vjust = 0.5, hjust=1))
```

## Ejercicio 11:

Conclusiones del caso?

**Respuesta.**

Nuestra conclusión inicial (basada en el modelo de regresión simple) es equivocada, no podemos concluir que exista discriminación por motivos de género. Lo anterior, ya que, como se vio recientemente, el salario también resulta una función de la posición de trabajo. Por tanto, el problema que puede ser entedido de forma exógena como discriminación, realmente se debería entender como la cantidad de mujeres que están presentes en algunas posiciones. En general, hay muchas más mujeres que hombres en posiciones muy mal pagas (Marketing Associate) y muchos más hombres que mujeres en posiciones muy bien pagas (Software Engineer y Manager). Por tanto, la recomendación hacía la forma de contratar mujeres para ciertas posiciones debería ser revisado.

In [ ]:

In [ ]:

## Transformación de variables

Idea central: Reducir el sesgo de los supuestos del modelo de regresión lineal que puedan presentar nuestros datos

[https://www.statisticshowto.com/box-cox-](https://www.statisticshowto.com/box-cox-transformation/#:~:text=A%20Box%20Cox%20transformation%20is,a%20broader%20number%20of%20tests.)

[transformation/#:~:text=A%20Box%20Cox%20transformation%20is,a%20broader%20number%20of%20tests.](https://www.statisticshowto.com/box-cox-transformation/#:~:text=A%20Box%20Cox%20transformation%20is,a%20broader%20number%20of%20tests.)

In [ ]:

In [ ]:

```
x <- runif(1000, min=10, max=50)
y <- log(34 + 10*x + rnorm(1000, sd=30))
plot(x,y)
```

In [ ]:

```
ajuste <- lm(y~x)
summary(ajuste)
plot(x,y)
abline(ajuste)
```

In [ ]:

```
options(repr.plot.width=7, repr.plot.height=7)
par(mfrow=c(2,2))
plot(ajuste)
```

In [ ]:

```
plot(x,y)
```

In [ ]:

```
plot(x,exp(y))
```

```
In [ ]: y_exp <- exp(y)
ajuste <- lm(y_exp~x)
summary(ajuste)
plot(x,y)
abline(ajuste)
```

```
In [ ]: par(mfrow=c(2,2))
plot(ajuste)
```

```
In [ ]: library(MASS)

#create data
y=c(1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 6, 7, 8)
x=c(7, 7, 8, 3, 2, 4, 4, 6, 6, 7, 5, 3, 3, 5, 8)

#fit linear regression model
model <- lm(y~x)
summary(model)
plot(x,y)
```

```
In [ ]:
```

```
In [ ]: #find optimal lambda for Box-Cox transformation
bc <- boxcox(y ~ x)
lambda <- bc$x[which.max(bc$y)]

lambda
```

```
In [ ]: #fit new linear regression model using the Box-Cox transformation
y_trans <- y^{lambda-1}/lambda
y
y_trans
# new_model <- lm(((y^lambda-1)/lambda) ~ x)
```

```
In [ ]: plot(x,y_trans)
```

```
In [ ]: ajuste2 <- lm(y_trans~x)
summary(ajuste2)
```

```
In [ ]: initech = read.csv("initech.csv")
initech
```

```
In [ ]: plot(initech$years, initech$salary)
```

```
In [ ]: plot(initech$years, initech$salary)
ajuste <- lm(salary~years, data=initech)
summary(ajuste)
abline(ajuste)
```

```
In [ ]: par(mfrow=c(2,2))
plot(ajuste)
```

```
In [ ]: bc <- boxcox(initech$salary~initech$years)
lambda <- bc$x[which.max(bc$y)]
lambda
```

```
In [ ]: y_trans <- log(initech$salary)
plot(initech$years, y_trans)
```

```
In [ ]: ajuste_bc <- lm(y_trans~years,data=initech)
summary(ajuste_bc)
plot(initech$years, y_trans)
abline(ajuste_bc)
```

```
In [ ]: par(mfrow=c(2,2))  
plot(ajuste_bc)
```

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

```
In [ ]: library(lmtest) # ver https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson\_statistic  
dwtest(ajuste_bc) # H0: No hay autocorrelación en los errores
```

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

```
In [ ]: library(tseries)  
jarque.bera.test(residuales)
```

```
In [ ]: brain <- read.csv('headbrain.csv')
```

```
In [ ]: head(brain)
```

```
In [ ]: plot(brain$Head.Size.cm.3., brain$Brain.Weight.grams.)
```

```
In [ ]: ajuste_brain <- lm(Brain.Weight.grams.~Head.Size.cm.3., data=brain)  
summary(ajuste_brain)  
plot(brain$Head.Size.cm.3., brain$Brain.Weight.grams.)  
abline(ajuste_brain)
```

```
In [ ]: par(mfrow=c(2,2))
```

```
plot(ajuste_brain)
```

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

```
In [ ]: library(lmtest) # ver https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson\_statistic
        dwtest(ajuste_brain) # H0: No hay autocorrelación en los errores
```

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

```
In [ ]: library(tseries)
        jarque.bera.test(residuales)
```

```
In [ ]: x_trans <- log(brain$Head.Size.cm.3.)
        y_trans <- log(brain$Brain.Weight.grams.)
```

```
In [ ]: options(repr.plot.width=14, repr.plot.height=7)
        par(mfrow=c(1,2))
        plot(brain$Head.Size.cm.3., brain$Brain.Weight.grams.)
        plot(x_trans,y_trans)
```

```
In [ ]: ajuste2_brain <- lm(y_trans~x_trans)
        summary(ajuste2_brain)
```

```
In [ ]: residuales <- residuals(ajuste2_brain)
        mean(residuales)
```

```
In [ ]: library(lmtest) # ver https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson\_statistic
        dwtest(ajuste2_brain) # H0: No hay autocorrelación en los errores
```

```
In [ ]: library(lmtest) #https://en.wikipedia.org/wiki/Breusch%E2%80%93Pagan\_test  
bptest(ajuste2_brain) #H0: Homoscedasticidad
```

```
In [ ]: library(tseries)  
jarque.bera.test(residuales)
```

```
In [ ]:
```