2.7 Organización y Presentación de datos Bivariados

2.7.1 Datos bivariados categóricos

Para organizar datos de dos variables categóricas o cualitativas se usan tablas de doble entrada. Los valores de una variable van en columnas y los valores de la otra variable van en filas.

Para hacer esto en **Python** se usa elige la funcion ***crosstab*** de la librería pandas

Ejemplo 2.16

Supongamos que deseamos establecer si hay relación entre las variables tipo de escuela superior y la aprobación de la primera clase de matemáticas que toma el estudiante en la universidad, usando los datos de 20 estudiantes que se muestran abajo:

Est	escuela	aprueba	Est	escuela	aprueba
1	priv	si	11	públ	si
2	priv	no	12	priv	no
3	públ	no	13	públ	no
4	priv	si	14	priv	si
5	públ	si	15	priv	si
6	públ	no	16	públ	no
7	públ	si	17	priv	no
8	priv	si	18	públ	si
9	públ	si	19	públ	no
10	priv	si	20	priv	si

Ejemplo 2.16 (cont)

Los datos estan disponible en academic.uprm.edu/eacuna/eje316.txt

```
Import pandas as pd
#Leyendo los datost")
df=pd.read_table("http://academic.uprm.edu/eacuna/eje316.txt",sep="\s+")
# Haciendo una tabla de clasificacion cruzada para relacionar las variables escuela y
#y si aprueba o no la primera clase matematicas en el Colegio
pd.crosstab(df['escuela'],df['aprueba'],margins=True)
```

aprueba	no	si	All
escuela			
priv	3	7	10
publ	5	5	10
All	8	12	20

*Hay 7 estudiantes que son de escuela privada y que aprueban el examen. Un (7/12*100%=58.33% 0de los que aprueban el examen son de escuela privada*

Ejemplo 2.17.

Los siguientes datos se han recopilados para tratar de establecer si hay relación entre el Sexo del entrevistado y su opinión con respecto a una ley del Gobierno. Los datos estan disponibles en http://academic.uprm.edu/eacuna/eje2biv.csv.

Sexo	Opinion	Conteo
male	si	10
male	no	20
male	abst	30
female	si	15
female	no	31
female	abst	44

Usar **Python** para construir una tabla de contingencia y responder además las siguientes preguntas:

a) ¿Qué porcentaje de los entrevistados son mujeres que se abstienen de opinar?

b) De los entrevistados varones. ¿Qué porcentaje está en contra de la ley?

c) De los entrevistados que están a favor de la ley. ¿Qué porcentaje son varones?

d) De los que no se abstienen de opinar ¿Qué porcentaje son varones?

Solución:

En este caso se usa la función croostab de la librería pandas de Python

```
pd.pivot_table(df,values='conteo',index='Sexo',columns='Opinion',aggfunc=np.
    sum)
      Opinion
Sexo     abst no si
 female   44 31 15
 male     30 20 10
>
```

a) $\frac{44}{150} \times 100 = 29.33\%$

b) $\frac{20}{60} \times 100 = 33.33\%$ (20/60)x100=33.33%

c) $\frac{10}{25} \times 100 = 40.00\%$ (10/25)x100=40.00%

d) $\frac{(10+20)}{(25+51)} \times 100 = \frac{30}{46} \times 100 = 39.00\%$

Cuando se tiene dos variables categóricas se pueden hacer gráficas de barras agrupadas ("bars in clusters") o en partes componentes ("stacked bars") para visualizar la relación entre ellas.

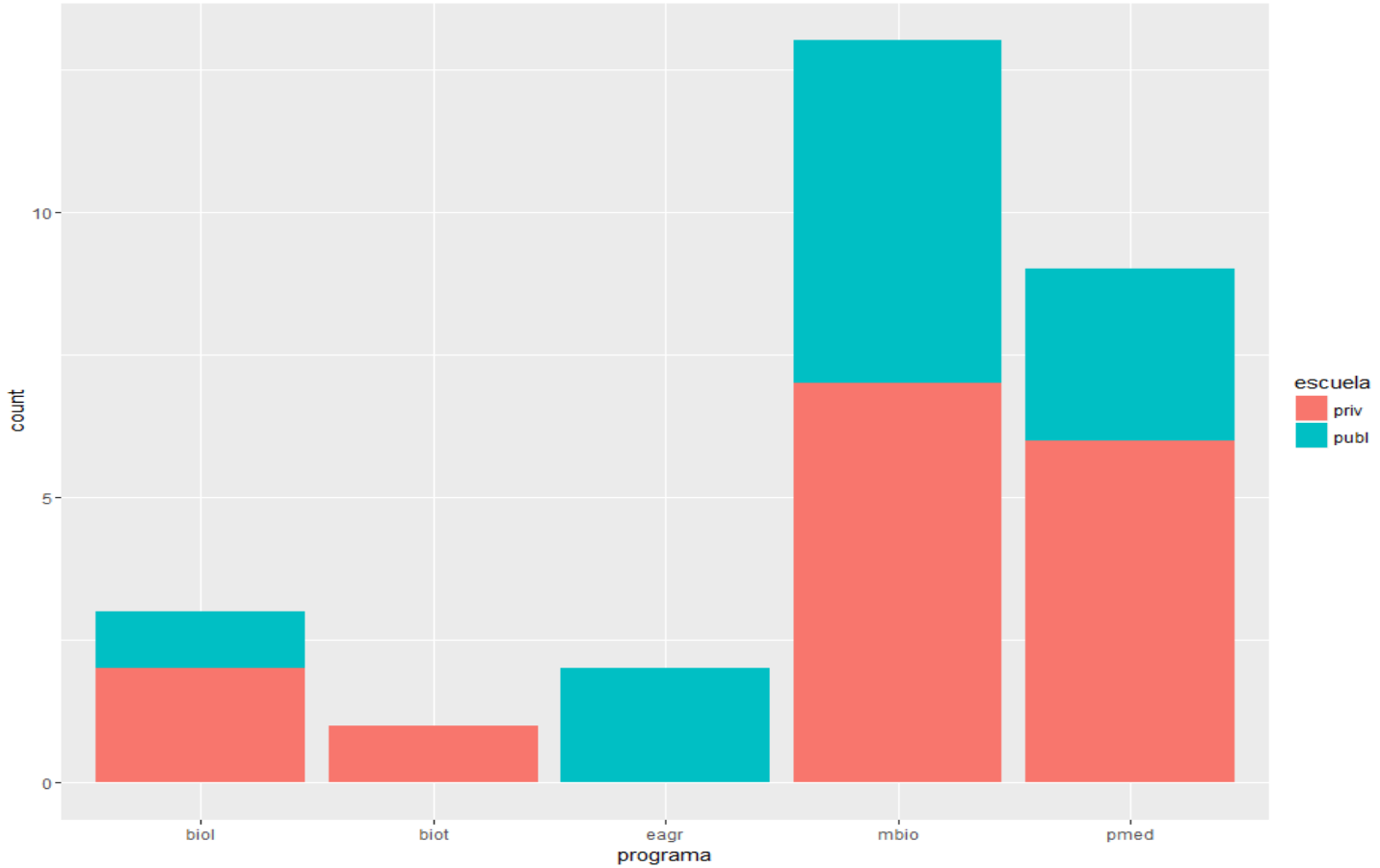
Ejemplo 2.20

Hallar una gráfica de partes componentes para comparar los estudiantes (por programa) según el tipo de escuela de donde proceden, usando datos del ejemplo 2.1.

Solución:

Continuación (Ejemplo 2.20)

Solución:

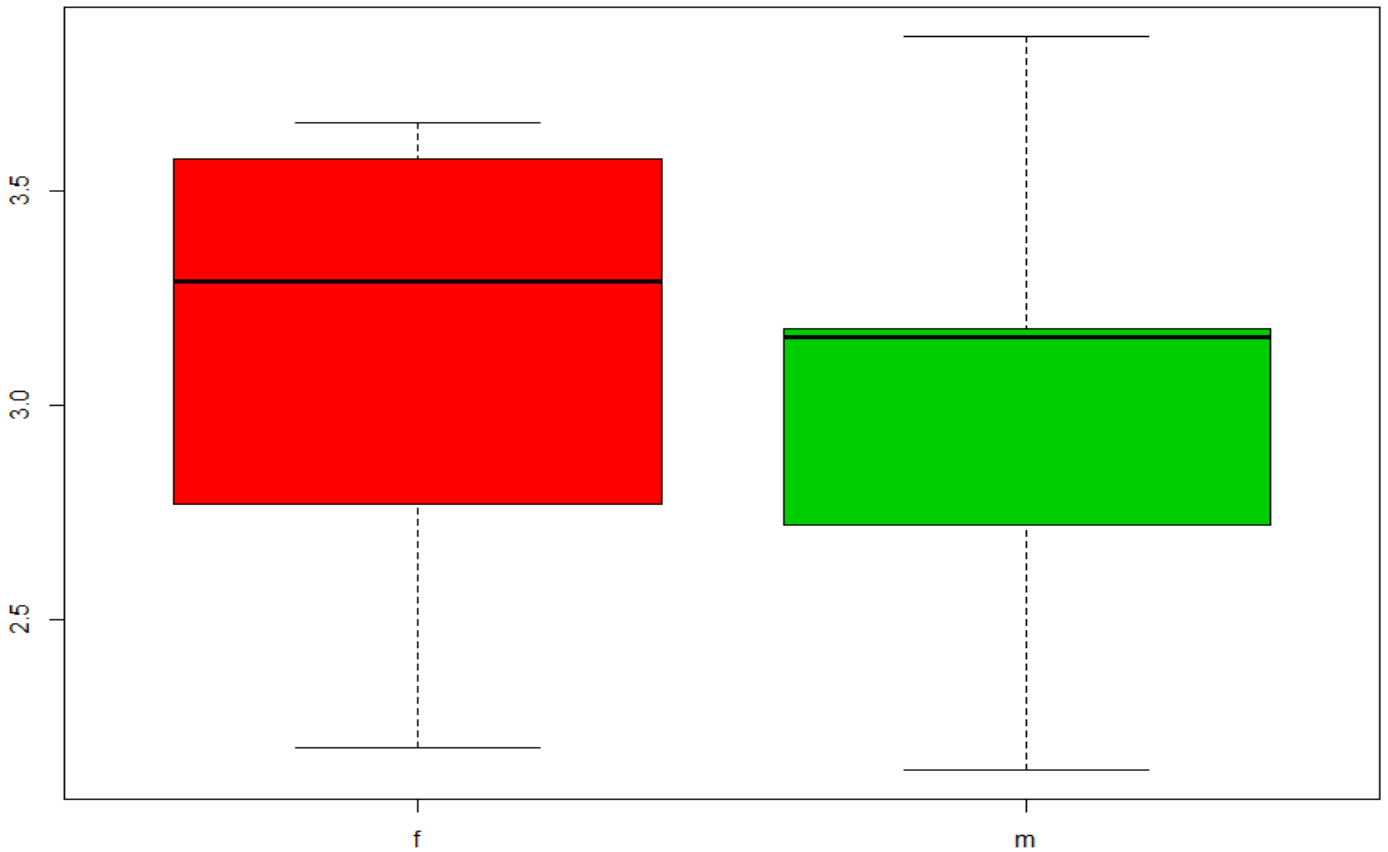


Gráfica de barras en partes componentes para la variable *Programa* según *Escuela*

2.7.2 Conjunto de datos que contienen una variable cualitativa y otra cuantitativa

La forma estándar de presentar los datos es en columnas donde cada columna representa un valor de la variable cualitativa y los valores dentro de cada columna representan valores de la variable cuantitativa. En general el objetivo es comparar los valores de la variable cualitativa según los valores de la variable cuantitativa, esto se lleva a cabo con una técnica llamada *análisis de varianza* .

La gráfica más adecuada para representar este tipo de información es el "Boxplot".


3.5
3.0
2.5
f
m

2.7.3 Datos Bivariados Continuos

Si se quiere representar la relación entre dos variables cuantitativas entonces se usa un diagrama de dispersión ("Scatterplot"). Para obtener un diagrama de dispersión entre dos variables X e Y se usa la función ***plot.scatter de matplolib***

Ejemplo 2.22

Es bien frecuente tener datos de una variable para un período de tiempo (dias, meses o años), estos tipos de datos son llamados series cronológicas o series temporales. Para este tipo de datos se pueden hacer gráficos de barras (aunque éstas son inadecuadas si el período de tiempo es muy grande) y gráficas lineales. Las siguientes gráficas se refieren al número de visitantes a Puerto Rico desde 1950 hasta 1998.

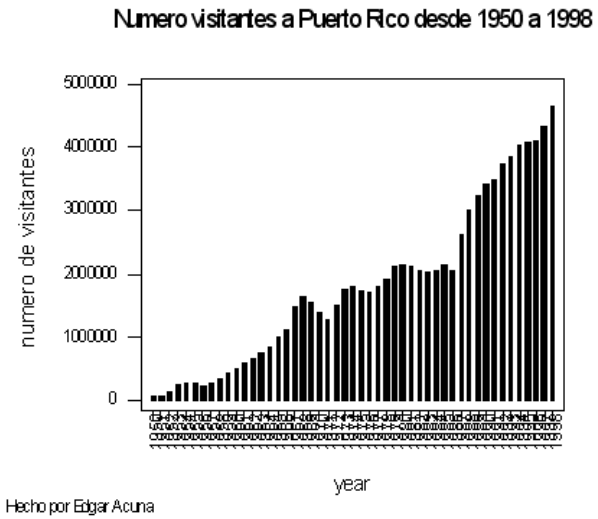


Figura 3.43 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

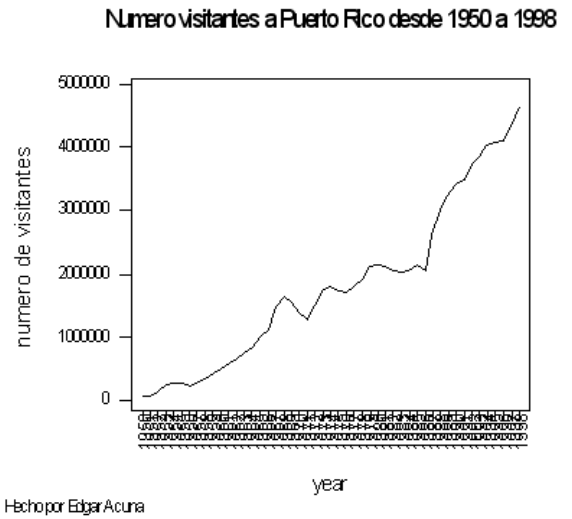


Figura 3.44 Gráfica de barras del número de visitantes a Puerto Rico entre 1950-1998.

Laboratorio 8

2.8 El Coeficiente de Correlación

Llamado también coeficiente de correlación de Pearson, se representa por **r** y es una medida que representa el grado de asociación entre dos variables cuantitativas X e Y.

Se calcula por

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Donde:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \quad , \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \quad \text{y} \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

La correlación varia entre -1 y 1. Un valor de r cercano a 0 indica una relación lineal muy pobre entre las variables. Un valor cercano a 1 indica que hay una buena relación lineal entre la variable y además al aumentar una de ellas la otra también aumenta. Un valor cercano a –1 indica una buena relación lineal pero al aumentar el valor de una de las variables la otra disminuye.

Ejemplo 2.23.

El dueño de una empresa que vende carros desea determinar si hay relación lineal entre los años de experiencia de sus vendedores y la cantidad de carros que venden. Los siguientes datos representan los años de experiencia (X) y las unidades de carros vendidas al año (Y), de 10 vendedores de la empresa.

X(años)	3	4	6	7	8	12	15	20	22	26
Y(ventas)	9	12	16	19	23	27	34	37	40	45

Solución:

Solución: (Ejemplo 2.23.)

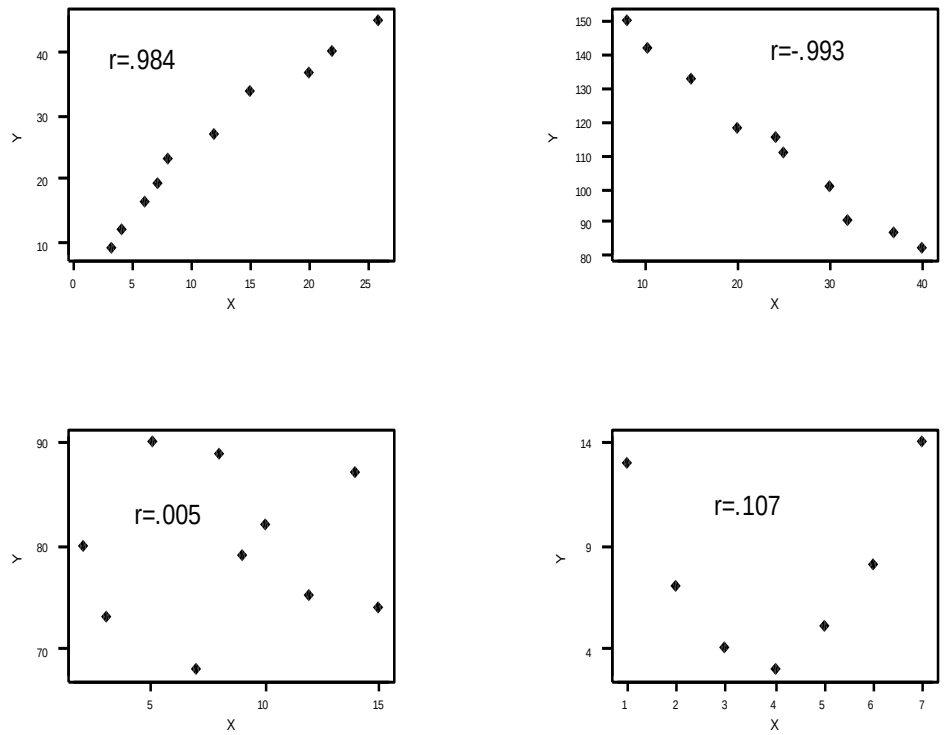
Interpretación:

Existe una buena relación lineal entre los años de experiencia y las unidades que vende el vendedor. Además mientras más experiencia tiene el vendedor más carros venderá. Se puede usar los años de experiencia para predecir las unidades que venderá anualmente a través de una línea recta.

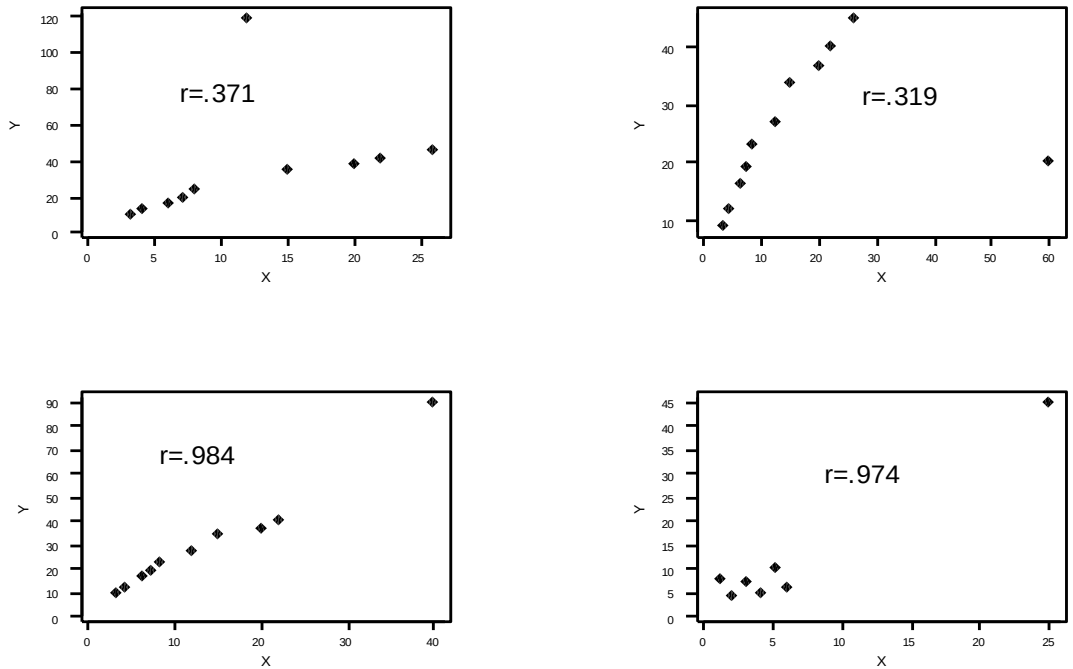
Row	years	ventas	Sxx	Syy	Sxy	r
1	3	9	590.1	1385.6	889.4	0.983593
2	4	12				
3	6	16				
4	7	19				
5	8	23				
6	12	27				
7	15	34				
8	20	37				
9	22	40				
10	26	45				

En **Python**, el coeficiente de correlación se puede obtener usando la función **corr de la librería pandas o corrcoef de nunpy**

Coeficiente de Correlacion para diversos plots



Efecto de valores anormales en el valor de la correlacion



2.9 Una introducción a Regresión Lineal.

La variable Y es considerada como la ***variable dependiente*** o ***de respuesta*** y la variable X es considerada la ***variable independiente o predictora.*** La ecuación de la línea de regresión es:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X,$$

Donde: $\hat{\alpha}$ es el intercepto con el eje Y, y $\hat{\beta}$ es la pendiente de la línea de regresión. Ambos son llamados los coeficientes de la línea de regresión.

Los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son hallados usando el método de mínimos cuadrados, que consiste en minimizar la suma de los errores cuadráticos de las observaciones con respecto a la línea. Las fórmulas de cálculo son:

$$\boxed{\hat{\beta} = \frac{s_{xy}}{s_{xx}}} \quad \text{y} \quad \boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

donde $\bar{x}$ es la media de los valores de la variable X y $\bar{y}$ es la media de los valores de Y.

2.8 Una introducción a Regresión Lineal (cont)

Interpretación de los coeficientes de regresión:

La pendiente $\hat{\beta}$ se interpreta como el cambio promedio en la variable de respuesta Y cuando la variable predictora X se incrementa en una unidad adicional.

El intercepto $\hat{\alpha}$ indica el valor promedio de la variable de respuesta Y cuando la variable predictora X vale 0. Si hay suficiente evidencia de que X no puede ser 0 entonces no tendría sentido la interpretación de .

El coeficiente de detreminacion R^2

Es una medida de la bondad de ajuste de la linea a los datos. Varia entre 0 y 100% y mientras mas ce acerca a 100% mejor es el ajuste de la línea a los puntos y probablemente sea buena la predicción.

Se puede calcular elevando la correlacion al cuadrado y multiplicando por 100

Ejemplo 2.25.

Supongamos que se desea establecer una relación entre la nota que un estudiante obtiene en la parte de aprovechamiento matemático de ingreso (CEEB) y el Promedio académico al final de su primer año de universidad (GPA). Se toma una muestra de 15 estudiantes y se obtiene los siguientes datos:

Est	CEEB	GPA
1	425	2.81
2	495	2.56
3	600	2.92
4	610	3.18
5	612	2.51
6	648	3.43
7	652	2.72

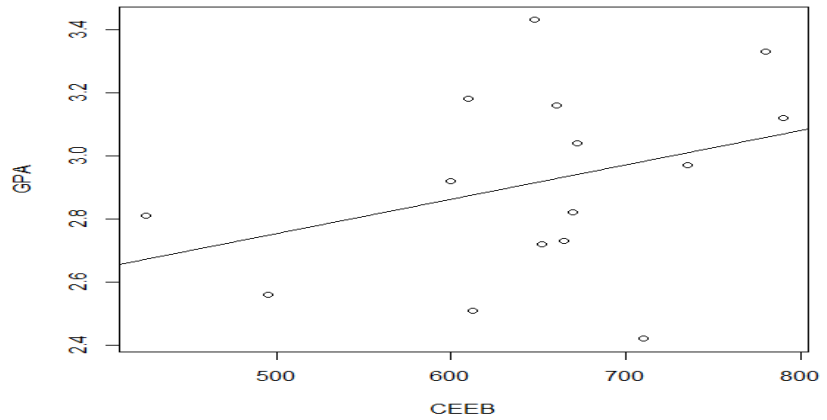
Est	CEEB	GPA
8	660	3.16
9	665	2.73
10	670	2.82
11	720	3.04
12	710	2.42
13	735	2.97
14	780	3.33
15	790	3.12

Obtener el diagrama de dispersión de los datos, la ecuación de la línea de regresión y trazar la línea encima del diagrama de dispersión.

Los datos estan disponibles en academic.uprm.edu/eacuna/eje1reg.txt

Solución (Ejemplo 2.25.)

La variable independiente es CEEB y la variable dependiente es GPA. La gráfica es:



***Interpretación**: El coeficiente de determinación es .121 y como la pendiente de la línea de regresión es positiva resulta ser que la correlación es .348035 esto indica una pobre relación lineal entre las variables CEEB y GPA. O sea que es poco confiable predecir GPA basado en el CEEB usando una línea.*

La ecuación de la línea de regresión esta dada por

GPA=2.209878+ .001087 CEEB))

Interpretación: *La pendiente 0.00109 indica que por cada punto adicional en el College Board el promedio del estudiante subiría en promedio en 0.00109, o se podría decir que por cada 100 puntos más en el College Board el promedio académico del estudiante subiría en .109. Por otro lado, si consideramos que es imposible que un estudiante sea admitido sin tomar el College Board, podemos decir que no tiene sentido interpretar el intercepto.*

Predicción

Uno de los mayores usos de la línea de regresión es la predicción del valor de la variable dependiente dado un valor de la variable predictora. Esto se puede hacer fácilmente sustituyendo el valor dado de X en la ecuación.

Por ejemplo, supongamos que deseamos predecir el promedio académico de un estudiante que ha obtenido 600 puntos en la parte matemática del examen de ingreso. Sustituyendo x =600 en la ecuación de la línea de regresión se obtiene Y=2.21+.00109*600=2.21+.654=2.864. Es decir que se espera que el estudiante tenga un promedio académico de 2.86.

Laboratorio 9