Tercera Clase de Análisis de Datos

Prof: Boris Panes
Universidad Del Desarrollo

Septiembre 7, 2024

Durante la clase tendremos la oportunidad de conversar sobre el avance de su trabajo T1, considerando una descripción de los datos

Preguntas específicas sobre set de datos y análisis

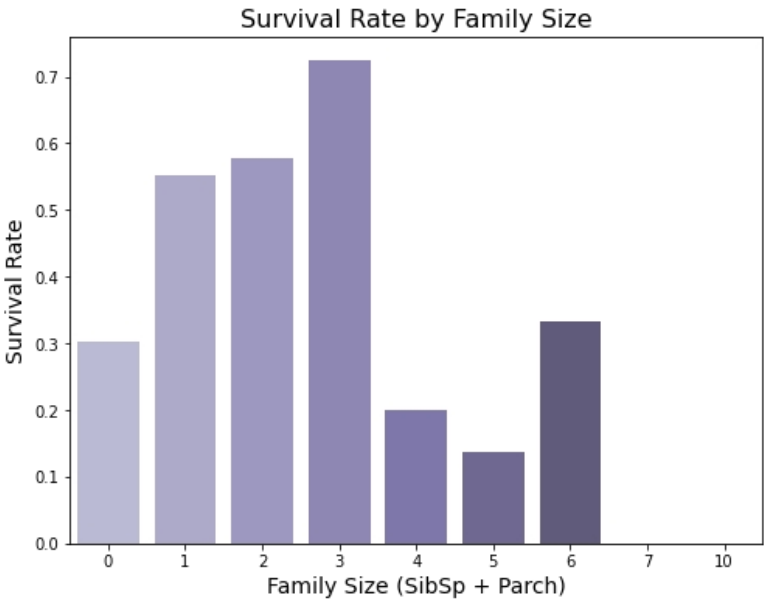
Durante la exploración y análisis básico del set de datos Titanic encontramos lo siguiente

```
In [29]: df["Familysize"] = df["SibSp"] + df['Parch']

In [30]: family_surr = df.groupby('Familysize')['Survived'].mean()

In [31]: family_surr

Out[31]: Familysize
         0     0.303538
         1     0.552795
         2     0.578431
         3     0.724138
         4     0.200000
         5     0.136364
         6     0.333333
         7     0.000000
         10    0.000000
         Name: Survived, dtype: float64
```



Verificar información para aquellos pasajeros con family size de 7 y 10 integrantes

Preguntas específicas sobre set de datos y análisis

El contenido del set de datos del Titanic de hecho contiene entradas con Familisyze>6

```
In [36]: df[df["Familysize"]>=6]
```

Out[36]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Familysize
13	14	0	3	Andersson, Mr. Anders Johan	male	39.000000	1	5	347082	31.2750	S	6
25	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...	female	38.000000	1	5	347077	31.3875	S	6
59	60	0	3	Goodwin, Master. William Frederick	male	11.000000	5	2	CA 2144	46.9000	S	7
68	69	1	3	Andersson, Miss. Erna Alexandra	female	17.000000	4	2	3101281	7.9250	S	6
71	72	0	3	Goodwin, Miss. Lillian Amy	female	16.000000	5	2	CA 2144	46.9000	S	7
119	120	0	3	Andersson, Miss. Ellis Anna Maria	female	2.000000	4	2	347082	31.2750	S	6
159	160	0	3	Sage, Master. Thomas Henry	male	29.699118	8	2	CA. 2343	69.5500	S	10
180	181	0	3	Sage, Miss. Constance Gladys	female	29.699118	8	2	CA. 2343	69.5500	S	10
182	183	0	3	Asplund, Master. Clarence Gustaf Hugo	male	9.000000	4	2	347077	31.3875	S	6

sibsp	# of siblings / spouses aboard the Titanic
parch	# of parents / children aboard the Titanic

La variable sibsp incluye esposo/a y siblings, que en general se refieren a hermanos/as

El grado de parentesco incluye solo los parientes de primer grado

Preguntas específicas sobre set de datos y análisis

In [39]:
```
df[df["Familysize"]==7]
```

Out[39]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Familysize
59	60	0	3	Goodwin, Master. William Frederick	male	11.0	5	2	CA 2144	46.9	S	7
71	72	0	3	Goodwin, Miss. Lillian Amy	female	16.0	5	2	CA 2144	46.9	S	7
386	387	0	3	Goodwin, Master. Sidney Leonard	male	1.0	5	2	CA 2144	46.9	S	7
480	481	0	3	Goodwin, Master. Harold Victor	male	9.0	5	2	CA 2144	46.9	S	7
678	679	0	3	Goodwin, Mrs. Frederick (Augusta Tyler)	female	43.0	1	6	CA 2144	46.9	S	7

In [40]:
```
df[df["Familysize"]==10]
```

Out[40]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Familysize
159	160	0	3	Sage, Master. Thomas Henry	male	29.699118	8	2	CA. 2343	69.55	S	10
180	181	0	3	Sage, Miss. Constance Gladys	female	29.699118	8	2	CA. 2343	69.55	S	10
201	202	0	3	Sage, Mr. Frederick	male	29.699118	8	2	CA. 2343	69.55	S	10
324	325	0	3	Sage, Mr. George John Jr	male	29.699118	8	2	CA. 2343	69.55	S	10
792	793	0	3	Sage, Miss. Stella Anna	female	29.699118	8	2	CA. 2343	69.55	S	10
846	847	0	3	Sage, Mr. Douglas Bullen	male	29.699118	8	2	CA. 2343	69.55	S	10
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	29.699118	8	2	CA. 2343	69.55	S	10

Marco de Referencia Conceptual

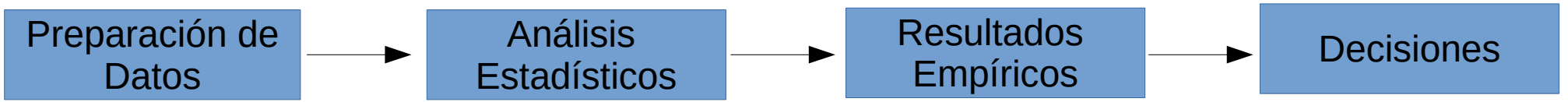
Ciencia de Datos

Descripción

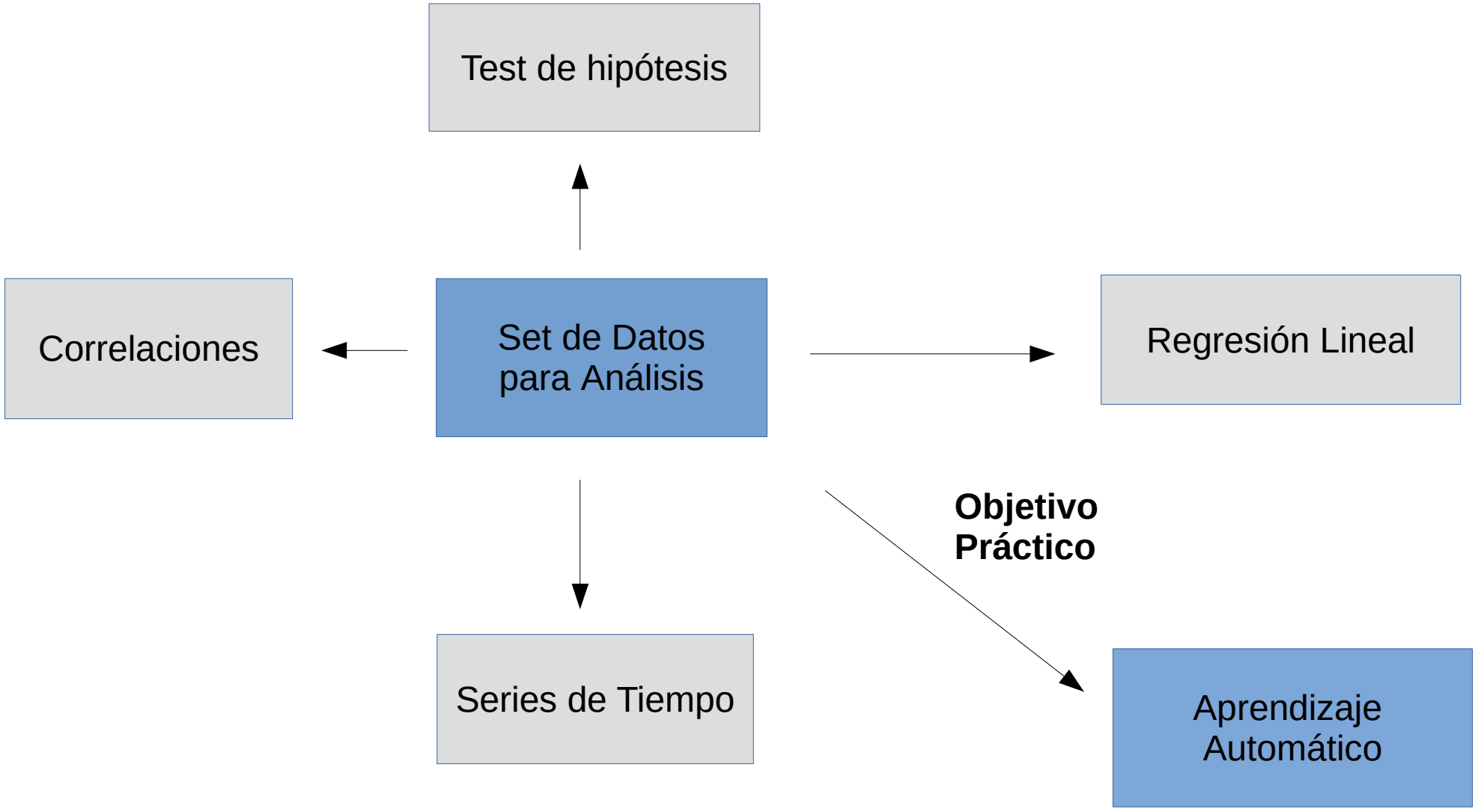
Conjunto de herramientas y metodologías basadas en conceptos estadísticos y probabilísticos que permiten generar un marco conceptual general para el uso de datos para la toma de decisiones

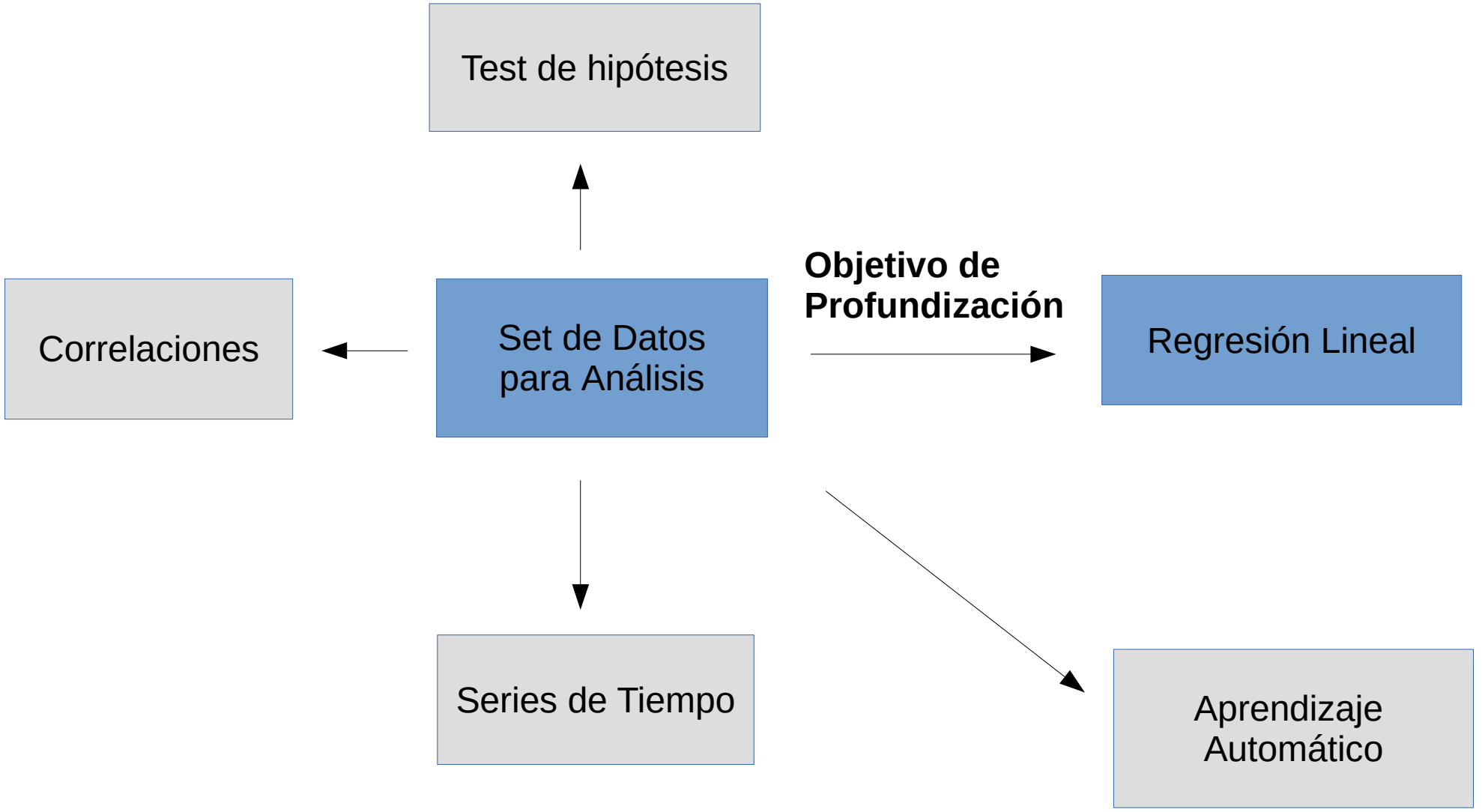
Objetivo

En el mediano a largo plazo es argumentable suponer que uno de los objetivos de la ciencia de datos es establecer un conjunto de etapas de análisis secuenciales y cíclicas que permiten la automatización del proceso de análisis de datos



Regresión Lineal


Test de hipótesis
Correlaciones
Set de Datos para Análisis
Regresión Lineal
Objetivo Práctico
Series de Tiempo
Aprendizaje Automático


Test de hipótesis
Correlaciones
Set de Datos para Análisis
Objetivo de Profundización
Regresión Lineal
Series de Tiempo
Aprendizaje Automático

Regresión Lineal Total

Regresión lineal simple

Ecuación de la recta
Definición de componentes
Estimación de coeficientes

Predicción vs Explicación
Peligros de la interpolación

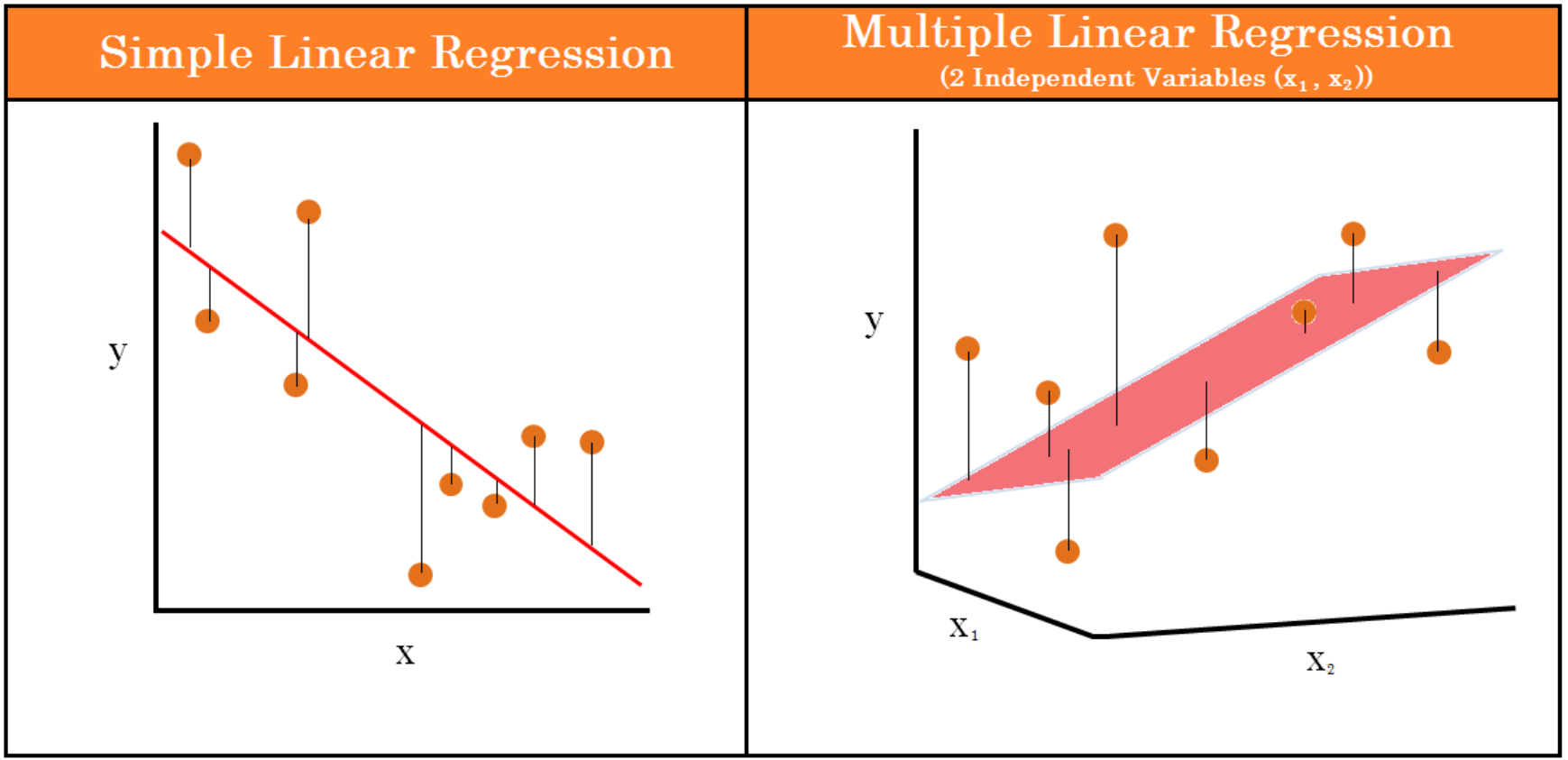
Regresión lineal múltiple

Factores y variables categóricas
Multicolinealidad, factores de confusión e interacciones

Diagnóstico de una regresión y supuestos

Sesgos en los análisis

Regresión lineal con uno o mas regresores



Fuente: startups-profit-prediction-using-multiple-linear-regression

Contenido de esta clase

Preguntas que podemos analizar con el estudio de la Regresión Lineal

- Políticas públicas y económicas como punto de partida
- Efecto de tratamientos médicos en pacientes

Continuidad entre conceptos de Estadística, Preparación de Datos y Análisis

- Desde los test de hipótesis a la relación entre valores continuos
- Desde la preparación de datos a tipos de datos útiles para la regresión lineal

Familias distinguibles de formatos de datos

Primeros pasos en el modelamiento de la regresión lineal

- Regresion lineal de un regresor

Contenido de esta clase

Motivación: Preguntas de ejemplo

Fuente: **Introducción a la Econometría , 3° Edición. James H. Stock, Mark W. Watson, 2012.** Desde ahora en adelante solo usaremos **S&W** para esta referencia

¿Qué efecto tiene un cambio en el tamaño de la clase sobre las calificaciones en los exámenes?

¿Qué efecto tiene el aumento de las multas de tránsito en la mortalidad por accidentes?

¿Qué efecto causa un aumento del 1% en el precio de los cigarrillos sobre el consumo de cigarrillos?

¿Qué efectos tiene la etnia sobre sus oportunidades de obtener aprobada una solicitud de hipoteca?

A partir de la investigación de este tipo de casos es posible extrapolar el análisis en situaciones diferentes por analogía.

Continuidad: Efectos entre variables

El **test de hipótesis** plantea la medición de efectos entre variables discretas y otras continuas

test-AB: comportamiento actual comparado con un cambio de condiciones

test de ANOVA: comparación entre múltiples condiciones discretas (ABCD)

test de Chi2: comparación utilizando tablas de contingencia discretas

La **regresión lineal** en general se preocupa de medir efectos entre variables continuas. De todas maneras también existen herramientas para incorporar variables categóricas en el análisis

numero de alumnos comparado con nivel de aprendizaje

precio de los cigarrillos comparado con nivel de consumo

Continuidad: Tipos de datos

La discusión sobre **preparación de datos** nos permitió plantear nuestros **objetivos de largo plazo** en cuanto al tipo de algoritmo hacia el cual nos estamos dirigiendo

> En la práctica, parece recomendable encausar las herramientas de preparación de datos hacia un formato de archivo que funcione como punto inicial de un algoritmo de **aprendizaje automático**

En este segundo bloque del curso vamos a ver si este objetivo es compatible con un análisis mas tradicional dentro del marco de la estadística, como lo es el estudio de **la regresión lineal**

> Para esto sera necesario entender de forma mas detallada el **contenido y contexto de los datos** que se están utilizando
>
> Además, el modelo de regresión lineal sera **analizado en profundidad**, considerando sus reglas básicas así como **extensiones y alcances**

Tipos de datos

En términos generales estamos interesados en datos ordenados en tablas

Datos de sección cruzada: muchos individuos, un solo periodo tiempo

Datos de series temporales: un solo individuo, muchos periodos de tiempo

Datos de panel: muchos individuos, muchos periodos de tiempo

En estas definiciones un individuo se puede referir a una persona, institución, unidad territorial, etc. Por otro lado un tiempo puede ser en cualquier unidad que sea conveniente, como segundos, minutos, semanas, meses, trimestres, años, etc.

¿Que tipos de datos esta utilizando en su proyecto T1?

Discusión de algunos ejemplos

Regresión Lineal Simple

Casos de ejemplo de variables a considerar

Especificación de las variables X e Y

Presentación de datos de trabajo

Modelamiento de Regresión Lineal con un Regresor

Razón de cambio lineal
Ecuación de la recta
Definición de coeficientes
Obtención de valores incógnitos

Discusión de elementos relacionados con la Regresión Lineal

Predicción vs Explicación
Peligros de la interpolación

Regresión Lineal y Predicción

El modelo de regresión lineal simple relaciona una variable X con otra Y

Multas por conducir en estado de ebriedad	Mortalidad en accidentes de transito
Número de alumnos por clase	**Calificación promedio de los cursos**
Duración de una carrera universitaria	Salario obtenido al ejercer la profesión

Podemos notar que en general, al menos en estos tres casos, la variable Y solo se puede medir luego de aplicar la variable X. Por lo mismo podemos decir que la regresión lineal es una herramienta de predicción

Datos de Ejemplo: Aplicación Realista

La siguiente tabla de datos muestra el ejemplo del capitulo 4 mostrado en Scott and Wattson

```
In [12]: df[["district","enrl_tot","teachers","str","testscr"]]
```

Out[12]:

	district	enrl_tot	teachers	str	testscr
0	Sunol Glen Unified	195	10.900000	17.889910	690.799988
1	Manzanita Elementary	240	11.150000	21.524664	661.200012
2	Thermalito Union Elementary	1550	82.900002	18.697226	643.599976
3	Golden Feather Union Elementary	243	14.000000	17.357143	647.700012
4	Palermo Union Elementary	1335	71.500000	18.671329	640.849976
...	...	...	...	...	...
415	Las Lomitas Elementary	984	59.730000	16.474134	704.300049
416	Los Altos Elementary	3724	208.479996	17.862625	706.750000
417	Somis Union Elementary	441	20.150000	21.885857	645.000000
418	Plumas Elementary	101	5.000000	20.200001	672.200012
419	Wheatland Elementary	1778	93.400002	19.036402	655.750000

420 rows × 5 columns

En este caso podemos utilizar

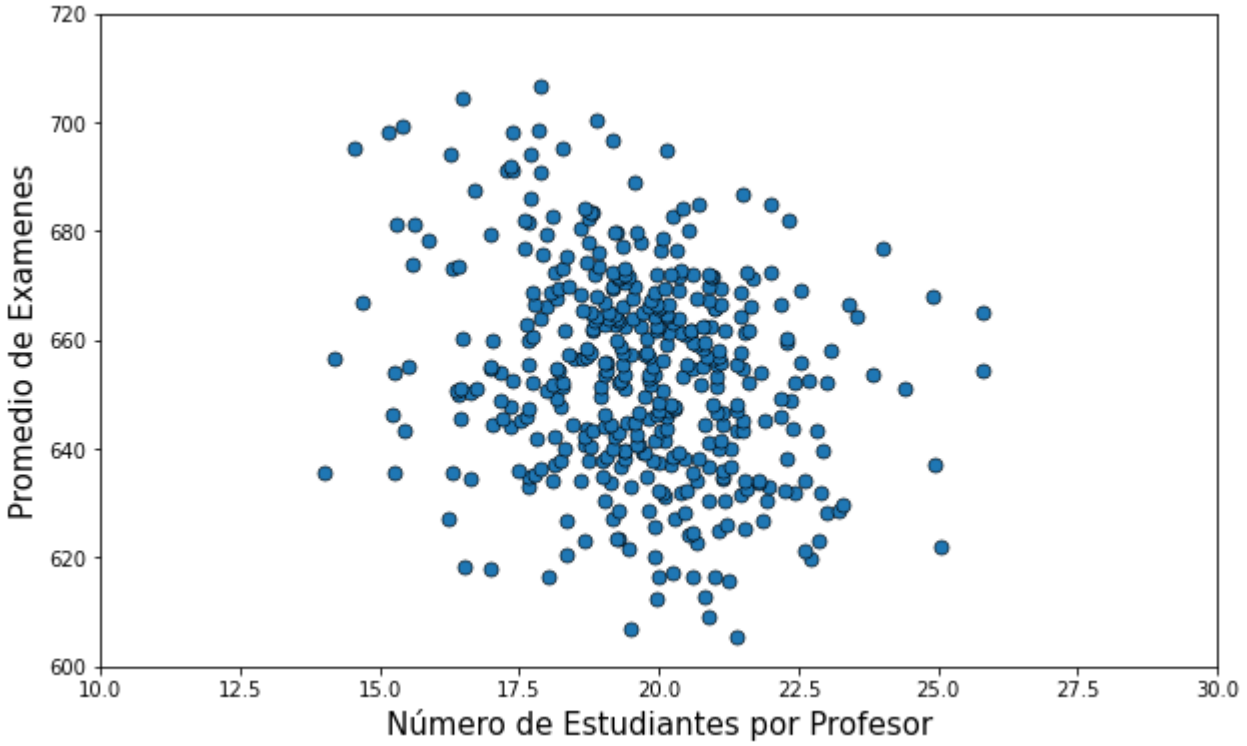
X = **str** = Número de estudiantes por profesor

Y = **test_scr** = Nota promedio

*Exámenes y tamaño de clases en el año 1999 de 420 distritos escolares de California

teachers = número efectivo de profesores a tiempo completo = número de horas totales de los profesores dividido por el tiempo de una jornada completa

Gráfico de dispersión y Coeficiente de correlación



Es posible notar una ligera tendencia de los datos que indica que **a mayor numero de estudiantes por profesor menor es el promedio de los exámenes**

Dado que los factores son múltiples es esperable que la correlación no sea perfecta

Definamos

X = **str** = Numero de estudiantes por profesor

Y = **test_scr** = Promedio de Exámenes

Versión expandida del coeficiente de **correlación** de pearson

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$$r = \frac{cov(x, y)}{std(x)\,std(y)}$$

el indice i recorre cada fila de la tabla de datos, por lo tanto es una sola sumatoria

aquí, std se refiere a la desviación estándar de cada variable por separado. Al agregar este factor como denominador el valor de r queda restringido entre -1 y 1 (Demostración queda como tarea)

Utilizando una notación más explicita tenemos que

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

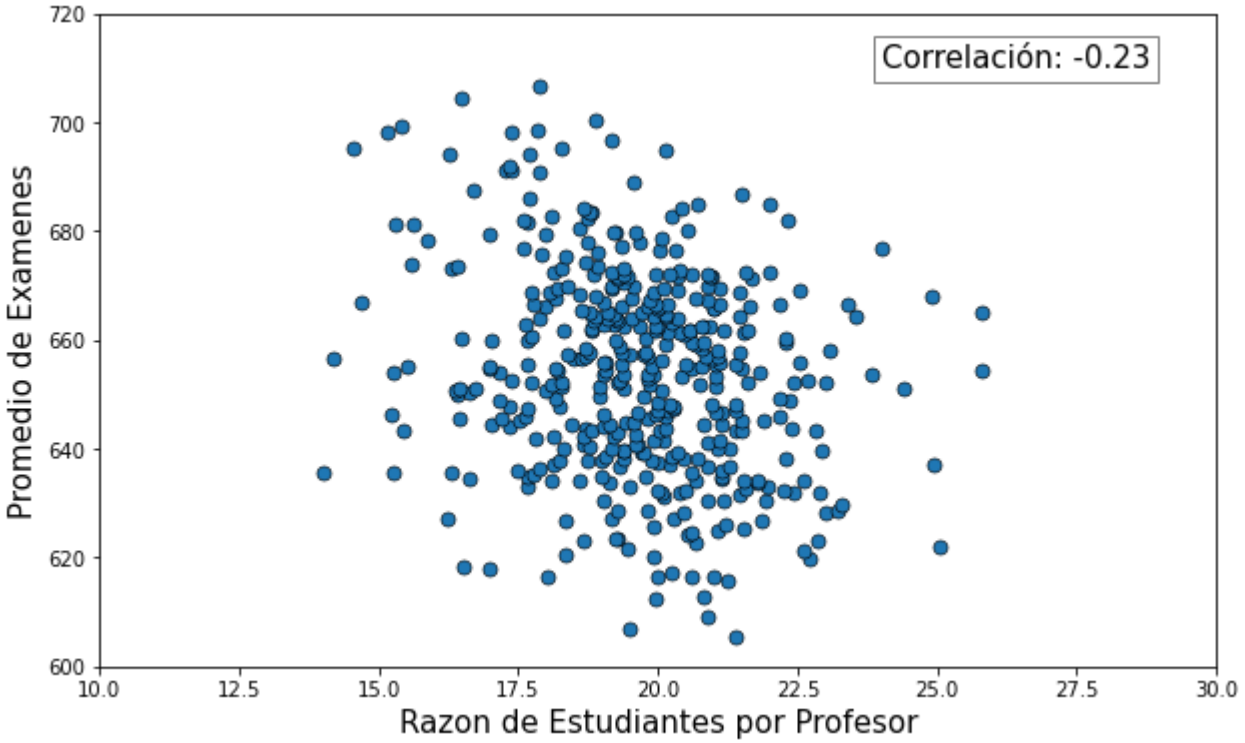
Dado una dependencia lineal

$$Y = b_0 + b_1 X_1$$

para b_0 y b_1 constante

r = 1 si $b_1 > 0$
r = -1 si $b_1 < 0$

Gráfico de dispersión y Coeficiente de correlación

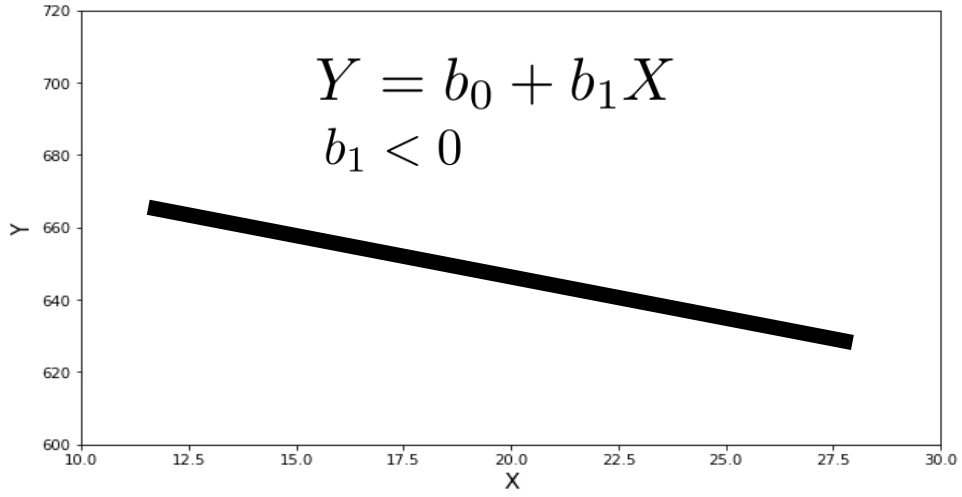


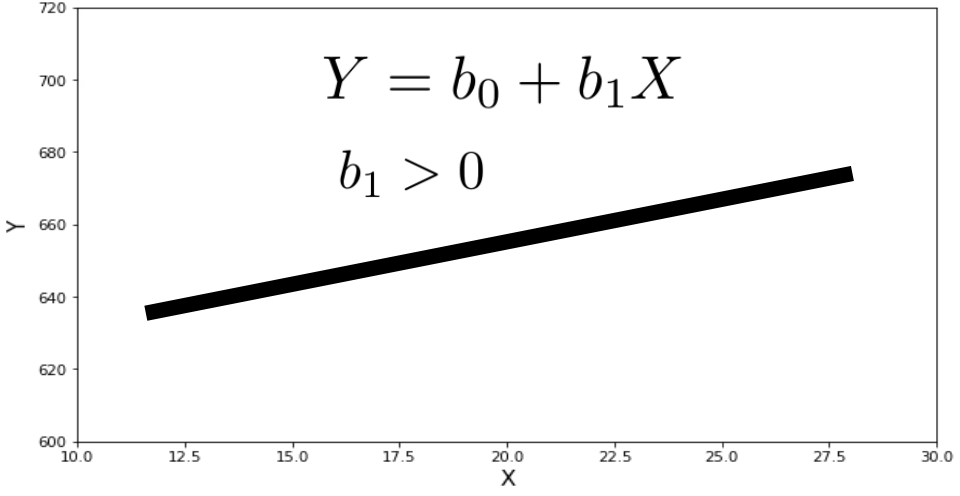
X = **str** = Numero de estudiantes por profesor

Y = **test_scr** = Nota promedio

Un valor de correlación negativo implica que la **recta mas cercana tiene una pendiente negativa, es decir $b_1 < 0$**

Gráfico de dispersión y Recta de tendencia





Supongamos que conocemos los valores de b_0 y b_1

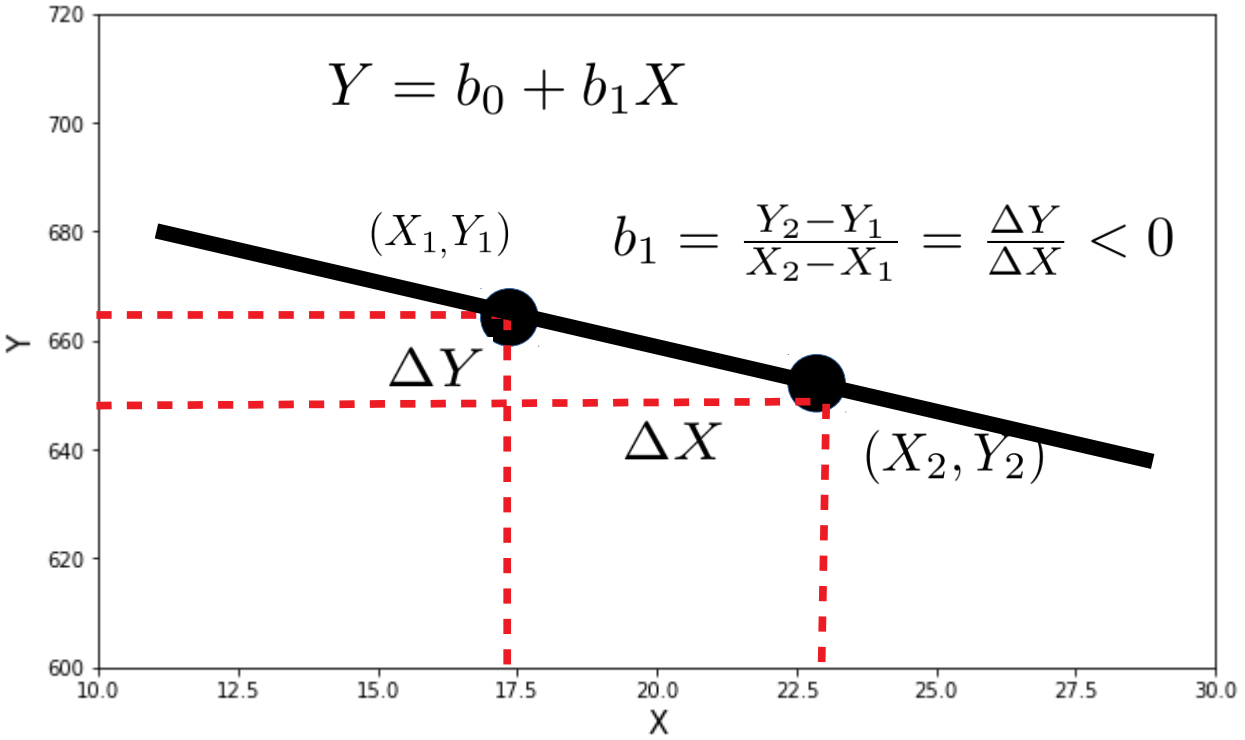
En general, estos valores pueden tener punto decimal, es decir pertenecer a los Reales. También se denominan Numéricos de punto Flotante en programación

Podemos interpretar estos coeficientes con **la pendiente y el intercepto**, es decir

$$b_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$ Pendiente

$$b_0 = Y(X = 0)$$ Intercepto

Gráfico de dispersión y Recta de tendencia



Podemos interpretar estos coeficientes con la pendiente y el intercepto, es decir

$$b_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$b_0 = Y(X = 0)$$

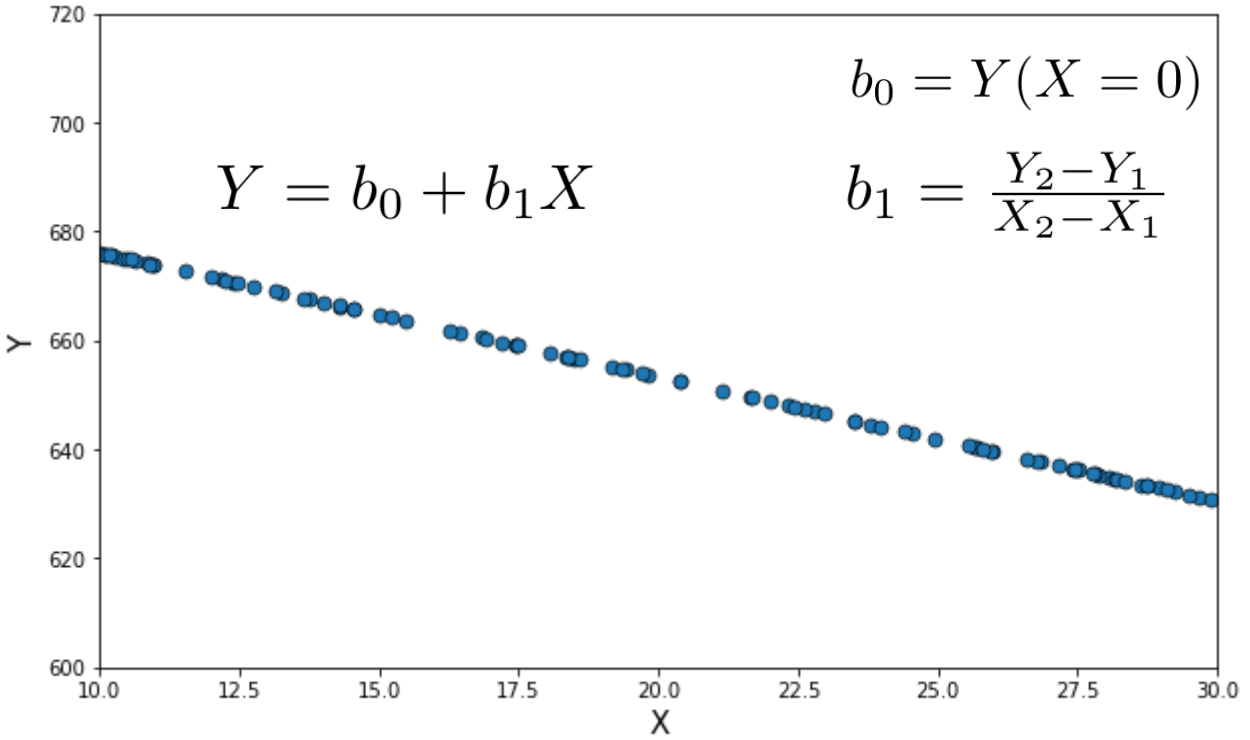
$$\Delta Y = Y_2 - Y_1$$
$$\Delta X = X_2 - X_1$$

El orden de los factores es relevante para identificar el signo de la pendiente

Ejemplo idealizado de puntos sobre una recta

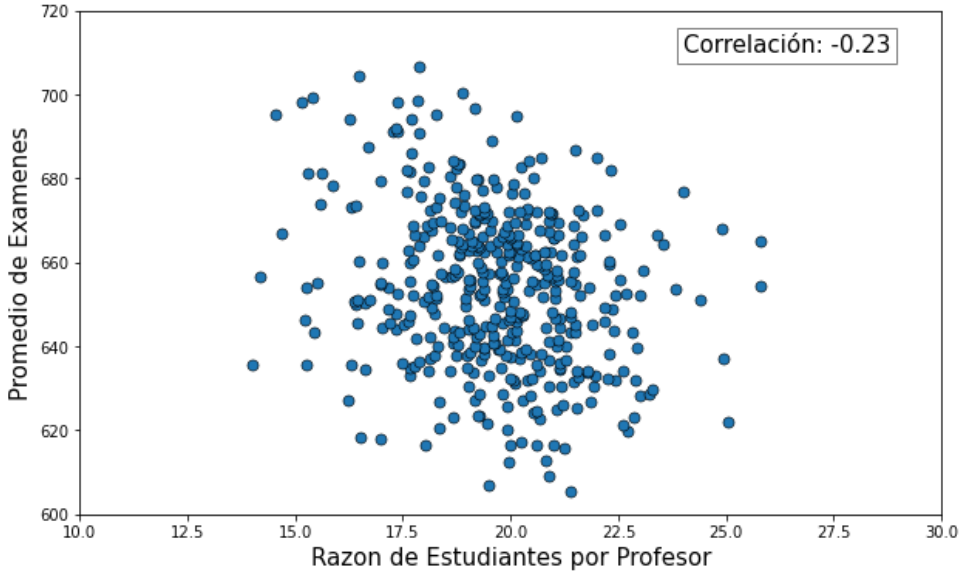
Definamos un set de datos idealizado tal que **Y = b_0 + b_1 X**, con b_0 y b_1 constantes

	X	Y
0	19.553094	654.318945
1	12.456354	670.499512
2	14.566432	665.688536
3	29.998506	630.503407
4	19.126944	655.290569
5	11.137600	673.506272
6	27.716985	635.705275
7	16.932035	660.294960
8	21.958118	648.835492
9	16.372438	661.570840
10	25.148118	641.562292



Para este set de datos idealizado el valor del intercepto es **b_0 = 698.9** y la pendiente es **b_1 = -2.28**. Para un valor de X es posible saber exactamente el valor de Y

Gráfico de dispersión y Recta de tendencia



El **valor negativo de la correlación** indica que la recta mas cercana debería tener **b_1 negativo**

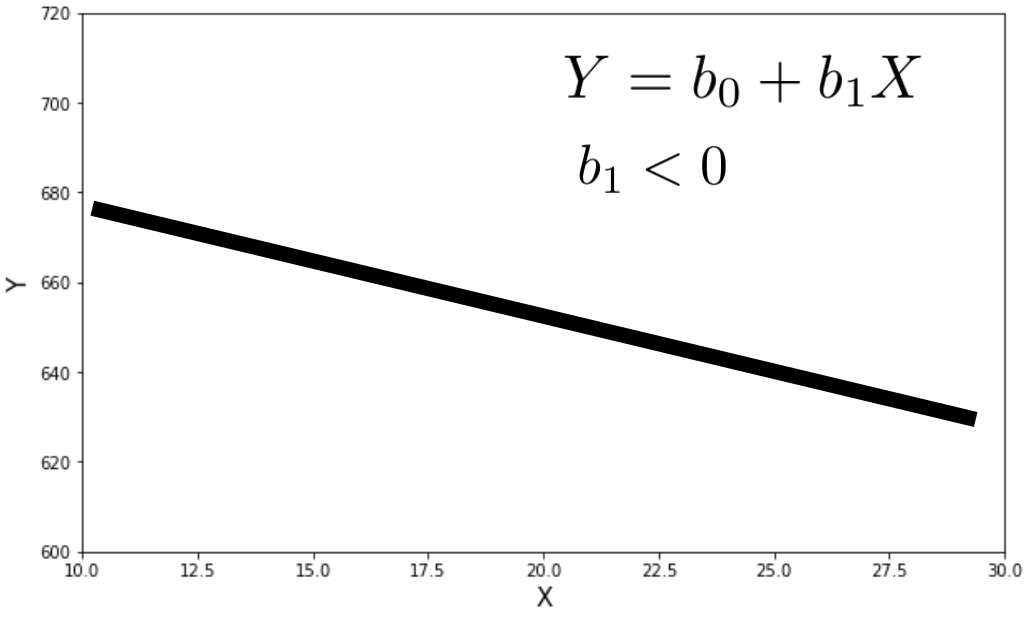
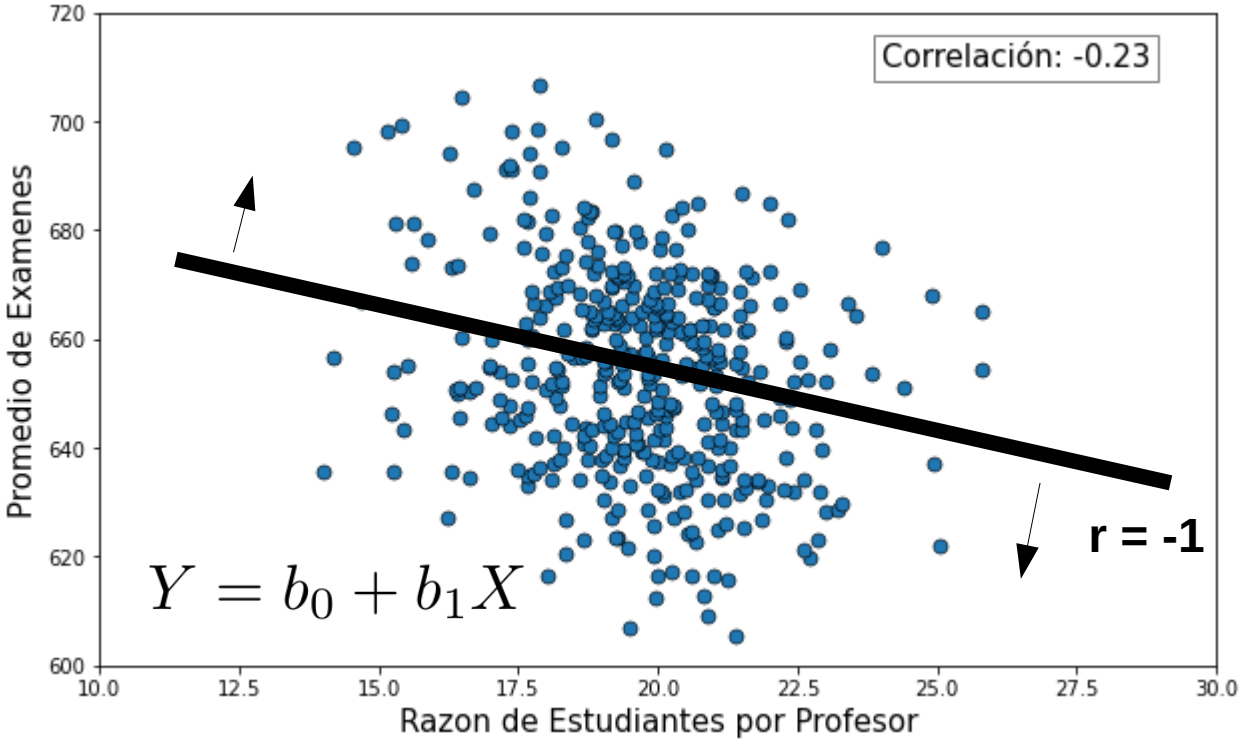


Gráfico de dispersión y Recta de tendencia



En este gráfico tenemos

puntos de datos en azul con r = -0.23

y

puntos sobre la recta con r = -1 y $b_1 < 0$

Dado que los puntos azules no siguen una recta de forma exacta, en principio no podemos acceder de manera directa los valores de b_0 y b_1

Valores de ajuste y términos de error

Dado un conjunto de puntos nos interesa encontrar los valores de b_0 y b_1 que nos permitan encontrar la recta que refleje de mejor forma la tendencia de los datos. Para esto es necesario estimar los valores de estos coeficientes a partir de los datos.

Por definición vamos a considerara el valor de los **estimadores de estos coeficientes** usando un acento circunflejo sobre el nombre del coeficiente. Asi, la recta que expresa la mejor tendencia esta dada por

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Una vez encontrados los valores de los estimadores tenemos acceso a todos los valores que viven sobre la recta. Para accederá a los valores del set de datos debemos agregar una corrección que llamaremos **términos de error** (residuals en Bruce and Bruce)

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

$$e_i = Y_i - \hat{Y}_i$$

Valores de ajuste y términos de error

Ejemplo de Bruce and Bruce que muestra todas las componentes discutidas anteriormente en un mismo gráfico

Debemos considerar

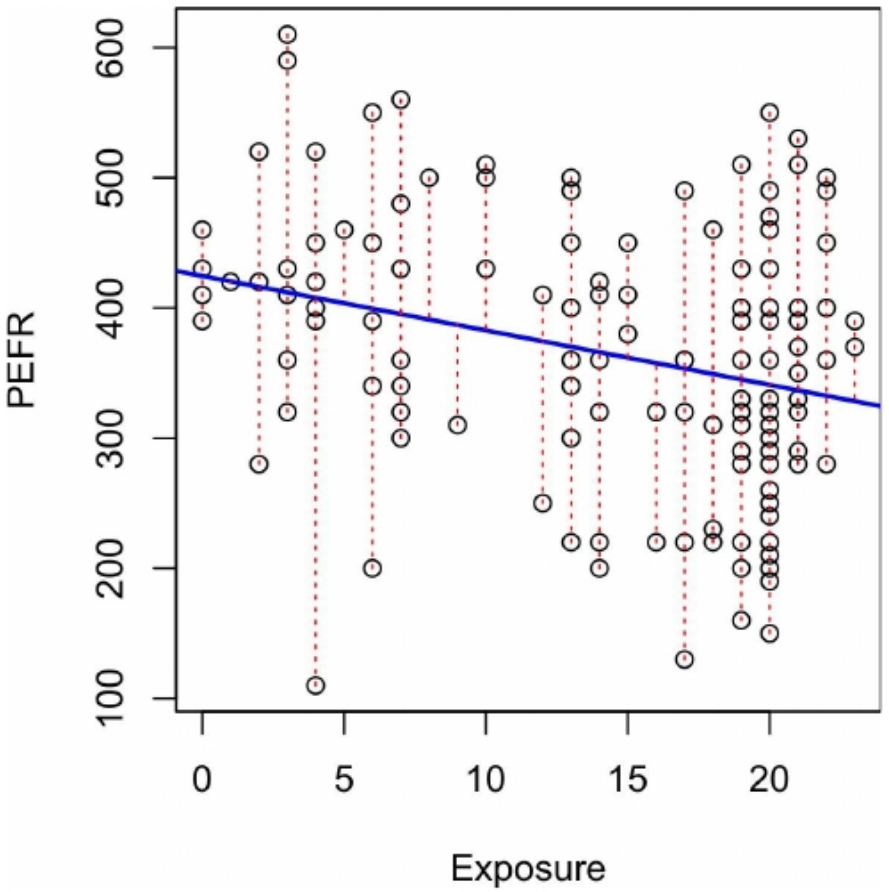
X = Exposure to cotton dust
Y = PEFR = Capacidad Pulmonar

Términos representados

(X_i, Y_i) Puntos del set de datos

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ Puntos sobre la recta

$e_i = Y_i - \hat{Y}_i$ Tamaño del error



Estimación de los coeficientes

El acercamiento estándar para estimar los valores de b_0 y b_1 pone su foco de atención en la **expresión para los errores**

$$e_i = Y_i - \hat{Y}_i$$
$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

Se puede argumentar que la recta que es capaz de minimizar alguna **medida proporcional al error total o promedio** es aquella que representa de mejor forma los datos observados

Un **estimador consistente del error total** esta dado por la **suma de los cuadrados de los errores** o **Residual Sum of Squares (RSS)** siguiendo Bruce and Bruce

$$\text{RSS} = \sum_{i=1}^{N} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

El método para obtener los coeficientes a partir de este estimador se denomina estimador de **Mínimos Cuadrados Ordinarios (MCO)** en S&W o **Least Square Regression (LSR)** en B&B

Estimación de los coeficientes

En esta ocasión solo reportaremos los resultados de este proceso de optimización. Pero en principio una de las formas mas directas es utilizando la noción de derivada de los cursos iniciales de Cálculo

$$\mathrm{RSS} = \sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Minimizar con respecto a b_0 y b_1 numéricamente o analíticamente

$$\beta_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{cov_{x,y}}{var_x}$$

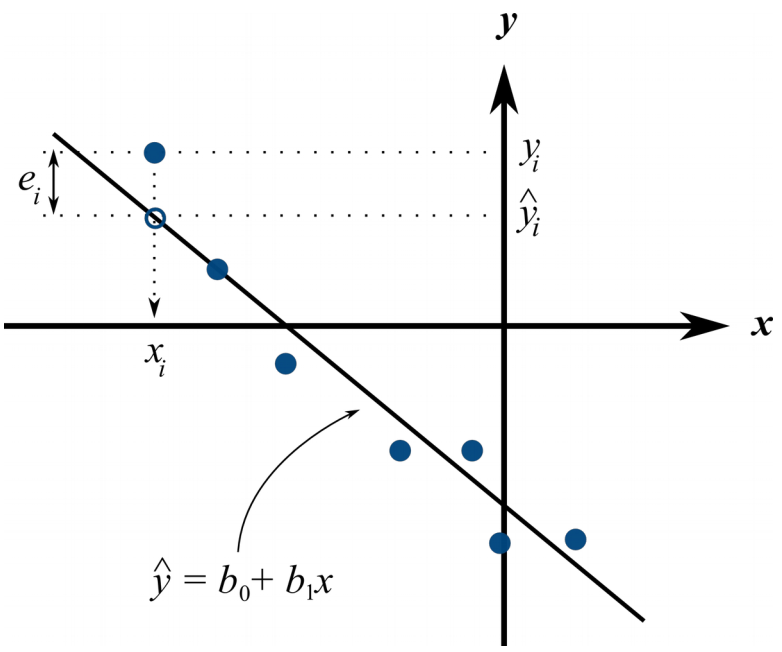
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

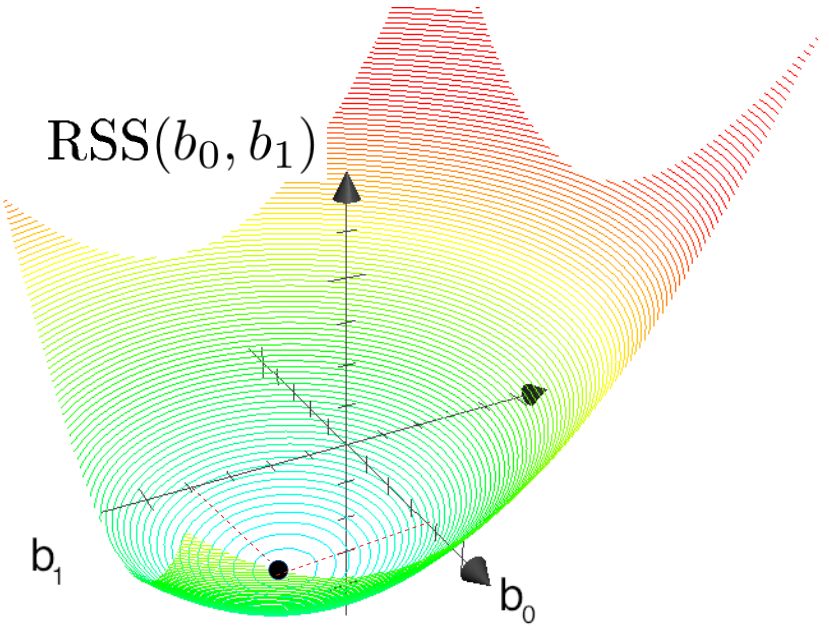
Por lo tanto, podemos ver que los valores de los coeficientes se pueden obtener directamente a partir de las entradas de la tabla de datos

Veamos que obtenemos a partir de los datos de promedios y tamaños de clases

Estimación de los coeficientes

Visualización gráfica de la obtención de los coeficientes de la recta (https://learnche.org/)





Los valores de b_0 y b_1 que minimizan la distancia a los puntos de los datos se pueden interpretar como los valores que minimizan la función de costo RSS(b_0, b_1)

Ejemplo de Notas y Tamaños de las Clases

```
In [12]: df[["district","enrl_tot","teachers","str","testscr"]]
```

Out[12]:

	district	enrl_tot	teachers	str	testscr
0	Sunol Glen Unified	195	10.900000	17.889910	690.799988
1	Manzanita Elementary	240	11.150000	21.524664	661.200012
2	Thermalito Union Elementary	1550	82.900002	18.697226	643.599976
3	Golden Feather Union Elementary	243	14.000000	17.357143	647.700012
4	Palermo Union Elementary	1335	71.500000	18.671329	640.849976
...	...	...	...	...	...
415	Las Lomitas Elementary	984	59.730000	16.474134	704.300049
416	Los Altos Elementary	3724	208.479996	17.862625	706.750000
417	Somis Union Elementary	441	20.150000	21.885857	645.000000
418	Plumas Elementary	101	5.000000	20.200001	672.200012
419	Wheatland Elementary	1778	93.400002	19.036402	655.750000

420 rows × 5 columns

$$Y = 699 - 2.28\,X$$

X = **str** = Numero de estudiantes por profesor

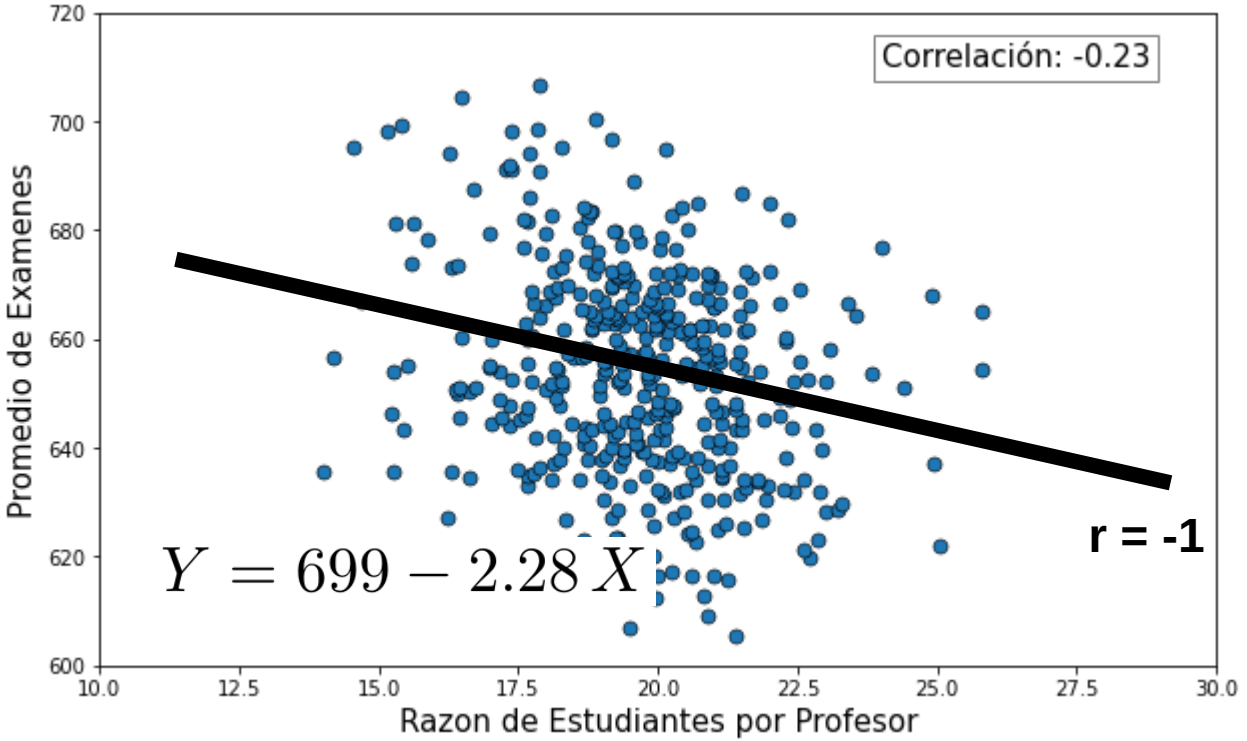
Y = **test_scr** = Promedio de Exámenes

mean(y) = 654.156
mean(x) = 19.64

cov(x,y) = -8.15
variance(x) = 3.57

b_1 = -8.15/3.57 = -2.28
b_0 = 654.15 - (-2.28) x 19.64
= 699.03

Gráfico de dispersión y Recta de tendencia



En este gráficos tenemos

puntos en azul con
r = -0.23

y

puntos sobre la recta con
r = -1 y $b_1 < 0$

Considerando solo la información entregada por los puntos azules (datos iniciales) es posible obtener directamente una tendencia usando una recta, cuyos coeficientes dependen de los estimadores estadísticos de covarianza y varianza, al igual que la correlación

Relación entre pendiente de la recta y correlación

Desde los datos obtenemos

mean(y) = 654.156
mean(x) = 19.64

cov(x,y) = -8.15
variance(x) = 3.57
variance(y) = 362

std(x) = 1.89
std(y) = 19.03

b_1 = -8.15/3.57 = -2.28
b_0 = 654.15 - (-2.28) x 19.64
= 699.03

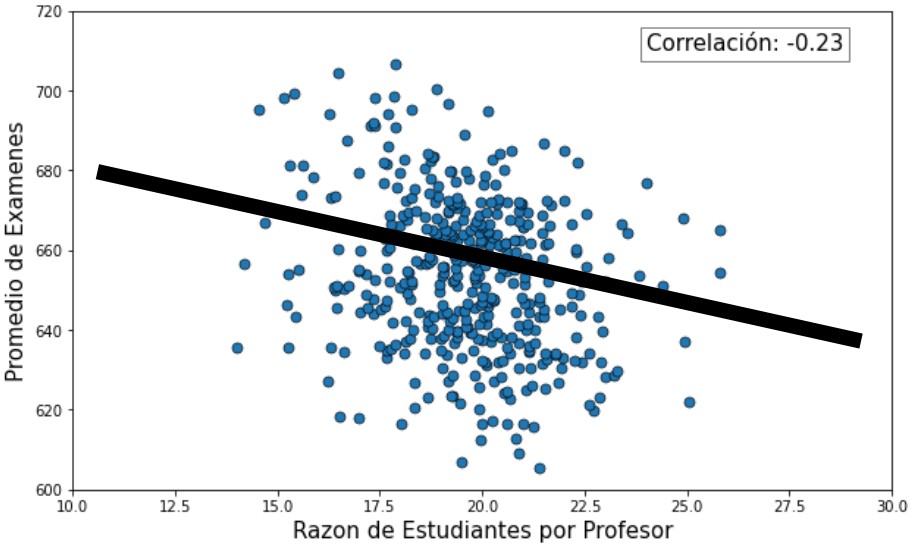
r = -8.15/(1.89*19.03) = -0.23

$$\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\longrightarrow \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

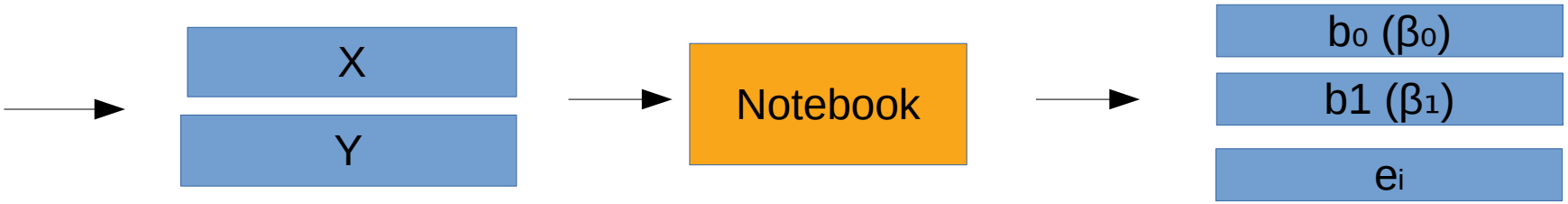
$$\hat{r} = \frac{cov(x,y)}{std(x)std(y)}$$

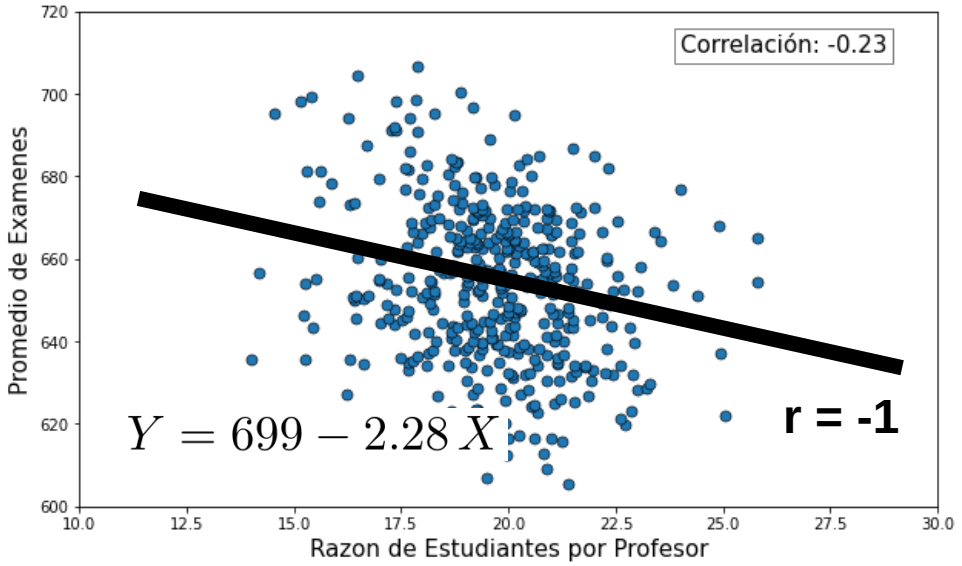


Terminología

	Bruce and Buce	Scott and Watson
X	Independent Variable Feature Attribute	Variable Independiente Regresor
Y	Dependent Variable Target Outcome	Variable Dependiente Regresando
b_0 (β_0)	Intercept, Y(X=0)	Intercepto de la recta
b1 (β_1)	Regression Coeficient Slope	Pendiente de la recta
e_i	Residuals	Termino de Error

Interpretación y discusión del método





Con la información de los coeficientes de la recta, obtenida automáticamente a partir de los datos, podemos realizar un par de **discusiones** relacionadas con la **utilidad** de la recta obtenida y su **interpretabilidad**

Notebook básico de regresión lineal

Notebook

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [42]:
```
df = pd.read_csv("./datos/caschool.csv", sep=",")
```

In [46]:
```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
```

In [73]:
```
t = LinearRegression().fit(df[['str']], df[['testscr']])
print(t.coef_[0][0], t.intercept_[0])
```

```
-2.2798082870234673 698.9329523279793
```

Análisis de segundo orden

Evaluaciones del Ajuste: cuantificaciones de la distancia entre la recta y los valores observados u otros estimadores, como la media

El error estándar de la regresión

Distribución muestral de los estimadores MCO

Estimadores con intervalos de confianza, que permiten generar una cuantificación de la incertidumbre en la predicción

Los supuestos de mínimos cuadrados

Se puede mostrar que el proceso de minimización genera los valores del modelo subyacente a los datos, cuando se cumplen ciertas condiciones de las variables