Cuarta Clase de Análisis de Datos

Prof: Boris Panes
Universidad Del Desarrollo

Septiembre 14, 2024

Durante la clase tendremos la oportunidad de conversar sobre el avance de su proyecto T1 y los detalles del proyecto T2

Regresión Lineal Total

Regresión lineal simple

Presentación de datos
Ecuación de la recta
Definición de componentes
Estimación de coeficientes

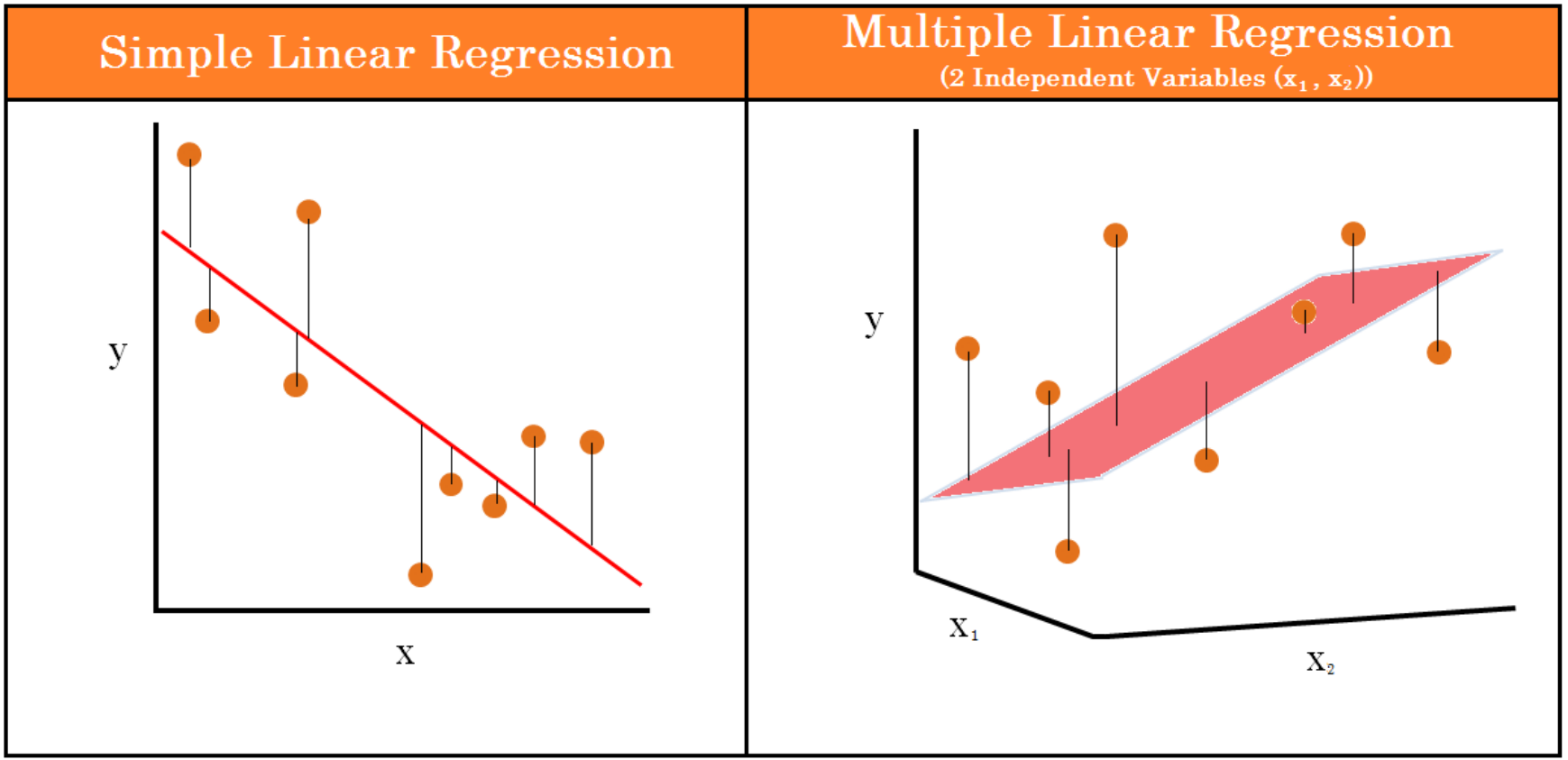
Predicción vs Explicación
Evaluación del Ajuste, Extrapolación
Distribución muestral de los coeficientes
Test estadístico de los coeficientes

Aspectos comunes en cualquier ejercicio de modelamiento

Regresión lineal múltiple

Factores y variables categóricas
Multicolinealidad, factores de confusión e interacciones

Regresión lineal con uno o mas regresores



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

Contenidos

Modelamiento de Regresión Lineal con un Regresor

- Presentación de Datos
- Ajuste lineal y resultados

Discusión de elementos relacionados con la Regresión Lineal

- Predicción vs Explicación
- Causalidad y Extrapolación
- Evaluación del ajuste obtenido por el módelo lineal
- Coeficiente de determinación R^2
- Distribución de los estimadores

Regresión Lineal Múltiple

- Extensión de la Regresión Lineal Simple
- Factores y variables categóricas
- Multicolinealidad e interacciones entre variables independientes

Datos de Ejemplo: Aplicación Realista

La siguiente tabla de datos muestra el ejemplo del **Capitulo 4 de Stock and Wattson, 2012**

```
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

Flujo para obtener el modelo lineal

```
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

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

Y = **test_scr** = Promedio de Exámenes

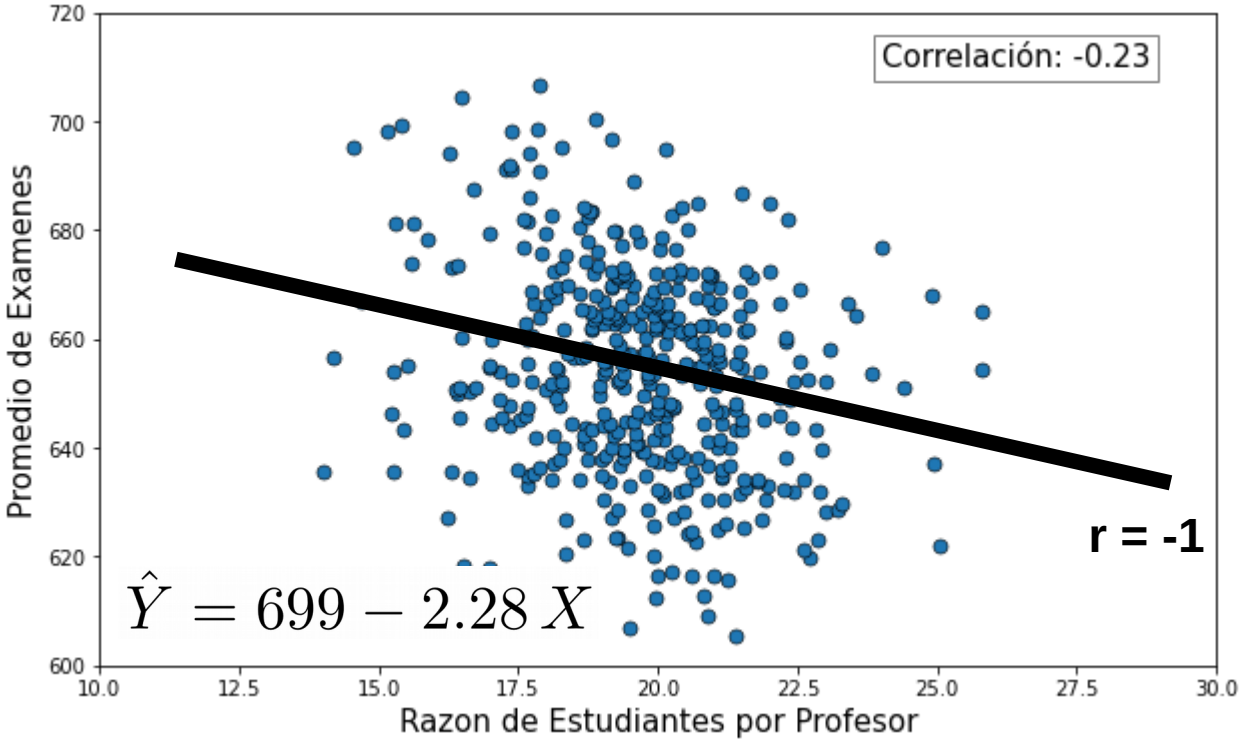
mean(y) = 654.156
mean(x) = 19.64

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

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

Gráfico de dispersión y recta de tendencia



En este gráficos tenemos

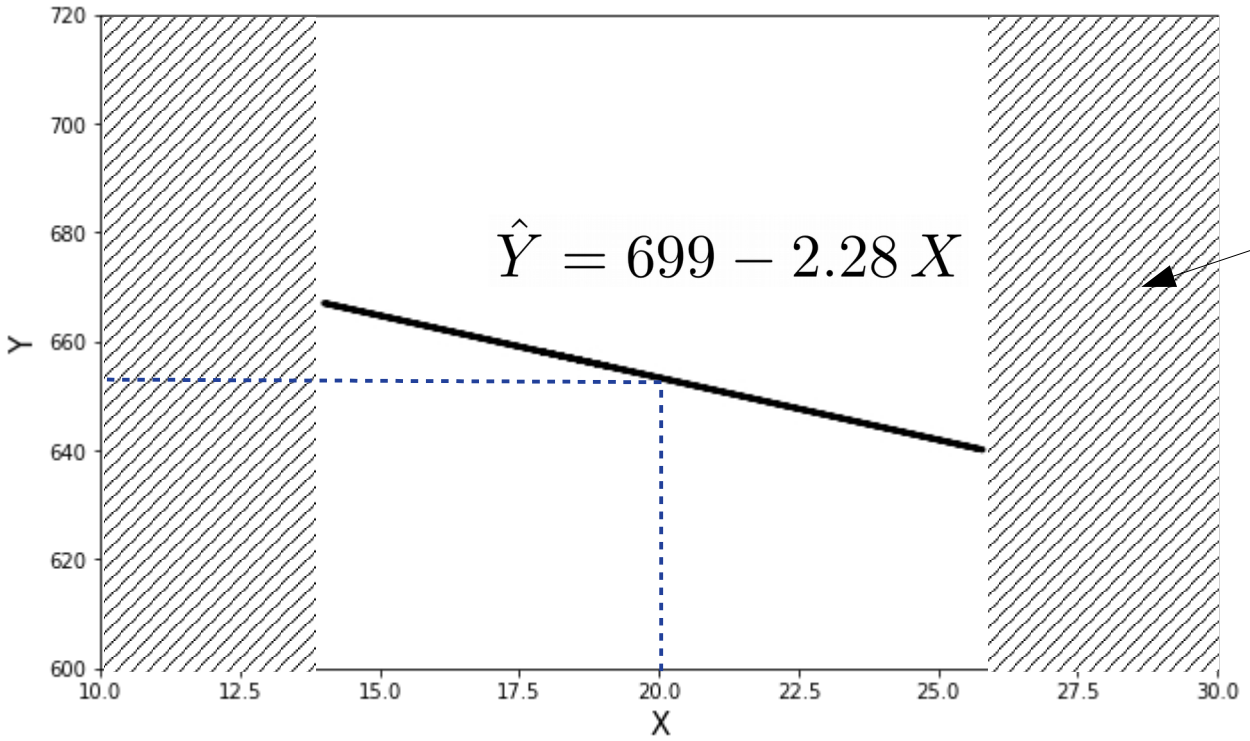
puntos en azul con
r = -0.23

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

Modelo teórico: Gráfico de la recta



Zonas de extrapolación

La recta indica la relación que deberían seguir las medias de los valores de X e Y en términos promedios o en observaciones idealizadas. En principio la recta entre valores sobre los Reales

Reporte sobre resultados parciales

Supongamos que los datos pueden ser representados de manera confiable utilizando el resultado de la regresión lineal, la cual esta resumida por la siguiente ecuación

$$\hat{Y} = 699 - 2.28\,X$$

Interpretación directa y ajustada de las resultados obtenidas (Capitulo 4 de S&W)

La **pendiente de -2.28** significa que un aumento en la proporción estudiantes-maestros en **un estudiante por clase** está, en promedio, asociado a una disminución en las calificaciones del distrito en el examen de **2.28 puntos**. Una disminución de la razón estudiantes-maestros en **dos estudiantes por clase** está, en promedio, asociada a un aumento en las calificaciones de **4.56 puntos [-2 * (-2.28)]**. La pendiente negativa indica que cuantos más estudiantes por maestro (clases más grandes), peor rendimiento en los examenes.

Impacto de las decisiones que se pueden realizar a partir de estos resultados

TABLA 4.1 Resumen de la distribución de las ratios estudiantes-maestros y calificaciones en los exámenes de quinto curso de 420 distritos escolares de California en 1999

			Porcentaje						
	Promedio	Desviación típica	10 %	25 %	40 %	50 % (mediana)	60 %	75 %	90 %
Ratio estudiantes-maestros	19,6	1,9	17,3	18,6	19,3	19,7	20,1	20,9	21,9
Calificación examen	654,2	19,1	630,4	640,0	649,1	654,5	659,4	666,7	679,1

Supongamos que se está contemplando la posibilidad de la contratación de maestros suficientes para permitir una **reducción de la proporción estudiantes-maestros en 2**

¿Cómo afectaría esto a los resultados en los exámenes?

De acuerdo con las estimaciones, al menos, el recorte en la ratio estudiantes-maestros en una gran cuantía (dos estudiantes por maestro) ayudaría y merecería la pena hacerlo en función de su situación presupuestaria, pero no sería una panacea.

Conceptos generales aplicables tanto a la Regresión Lineal Simple así como la Regresión Multilineal y otros módelos predictivos

Utilización Practica de los Modelos: Explicación y Predicción

En algunos contextos el resultado de la Regresión Lineal (RL) puede ser utilizado para entender aspectos generales respecto a la relación entre dos variables, por lo tanto la atención esta focalizada en la pendiente de la recta. **En estos casos la RL se utiliza como herramienta de explicación**

> **B&B:** Economists want to know the relationship between consumer spending and GDP growth. Public health officials might want to understand whether a public information campaign is effective in promoting safe sex practices. In such cases, the focus is not on predicting individual cases, but rather on **understanding the overall relationship.**

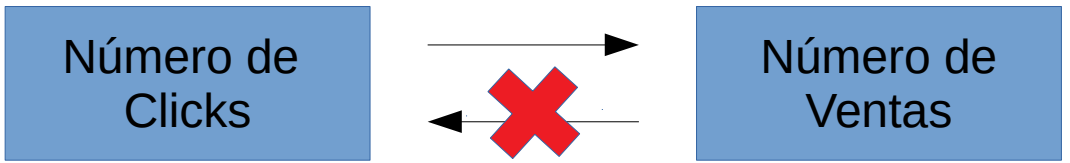
En otros contextos, especialmente considerando escenarios de big data, la RL puede utilizarse para estimar el valor de una variable de objetivo en a partir del valor de una variable de entrada. En este caso el foco esta en el valor explcito de Y obtenido con la RL. **En estos casos la RL se utiliza como modelo predictivo**

> **B&B:** In marketing, regression can be used to predict the change in revenue in response to the size of an ad campaign. Universities use regression to predict students' GPA based on their SAT scores.

Regresión Lineal y Causalidad

En general, la Regresión Lineal por si misma no es suficiente para establecer una **relación de causalidad** entre las variables consideradas. Para esto es necesario entender otros factores y considerar el conocimiento contextual del problema **(especialmente el factor temporal)**.

Un ejemplo muy simple para entender este fenómeno es presentado por **B&B en el Capitulo 4**, donde se presenta un ejemplo en el **contexto de propaganda por Internet**



Considerando el conocimiento existente sobre propaganda en internet, se puede argumentar que la tasa de conversiones es un efecto del número de clicks y no al contrario. Por lo tanto es este conocimiento el que nos permite escoger como **X el Número de Clicks y como Y la tasa de conversiones**

¿Cuál seria la discusión análoga en el ejemplo de Notas vs Tamaño de Clases?

Peligros de la extrapolación

Las predicciones obtenidas a partir de la Regresión Lineal no deben utilizarse fuera del rango de los datos. El modelo es valido en regiones donde existen suficientes datos. Es en estas regiones donde el modelo fue ajustado para obtener valores similares a los observados previamente. **Ejemplo de B&B, Chapter 4**

> As an extreme case, suppose that a model is used to predict the value of a **5,000** square foot empty lot. In such a case, all the predictors related to the building would have a value of 0 and the regression equation would yield an absurd prediction of -**521,900 + 5,000 × (-0.0605)= –$522,202**.

Además, también existe la posibilidad de encontrar puntos sobre el rango de la variable independiente que estan fuera del rango de aplicación realista. Los resultados en estos puntos pueden ser inconsistentes desde un punto de vista de factibilidad de la observación

¿Cuál seria la discusión análoga en el ejemplo de Notas vs Tamaño de Clases?

Discusión de resultados del ejemplo de Notas y Tamaños de Clases

La información predictiva del método esta mayormente contenida en la ecuación de la recta

$$\hat{Y} = 699 - 2.28\,X$$

Donde Y representa el promedio de notas en las pruebas estándar de ingles y matemáticas y X es la razón entre el número de estudiantes y el número de profesores efectivos

En **modo explicativo** podemos ver que el resultado de la RL indica que la tendencia de los datos indica que mientras mas alumnos por profesor menor es la nota promedio

En **modo predictivo** la ecuación de la recta nos permite obtener directamente el valor de la nota promedio dado un numero de alumnos por profesor

El conocimiento sobre como funcionan las clases nos permite confirmar que la variable X dada por el numero de alumnos por clase podría funcionar como **causa** de las notas obtenidas y no al contrario

Además, notamos que la ecuación nos permite **extrapolar** los resultados para valores de X que cubren regiones que van mas alla del dominio de lo factible

Dato histórico cuasi anecdótico

Origen del termino Regresión

Hipótesis 1:

El término regresión viene de un estudio sobre el efecto de **regresión a la media**

En este estudio se muestra como los promedios de las alturas de generaciones subsecuentes tienden a **regresar a la media**

Ejemplo: los hijos de los hombres extremadamente altos tienden a ser mas bajos que sus padres

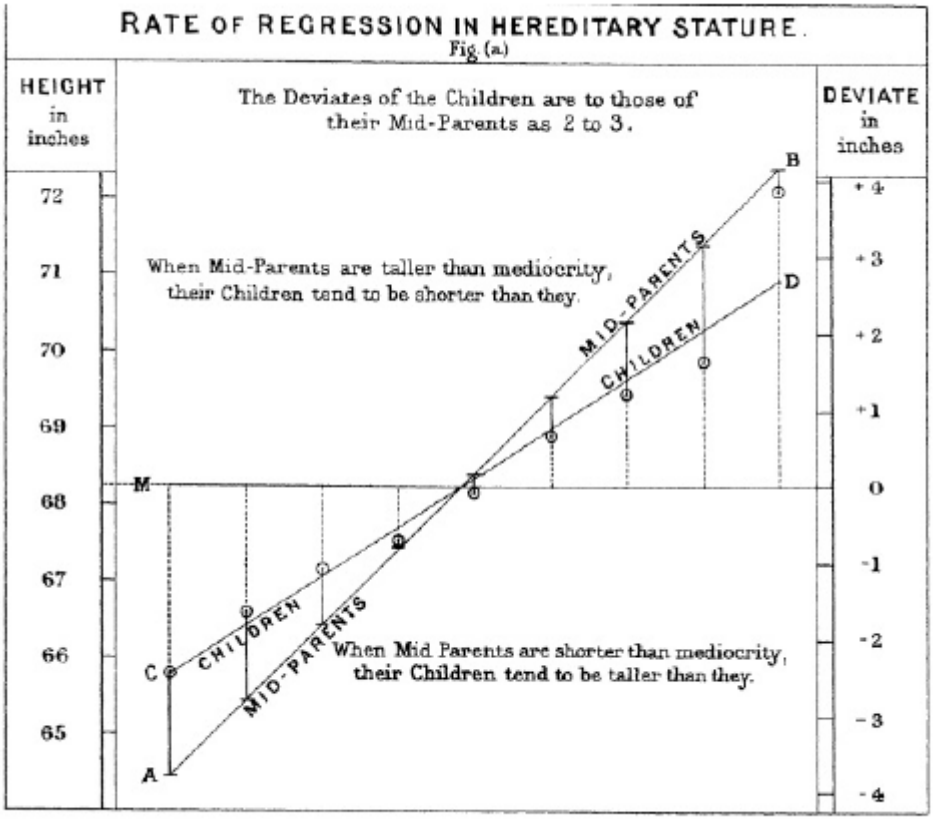


Figure 2-5. Galton's study that identified the phenomenon of regression to the mean

Francis Galton in 1886

Aspectos formales de la Regresión Lineal aplicables en el caso de una y mas variables

Hilo conceptual

Podemos notar que en general siempre es posible encontrar los coeficientes de un modelo lineal considerando una muestra de datos de al menos dos variables

El siguiente paso cuantitativo con respecto al modelamiento utilizando la regresión lineal corresponde a la **evaluación de la comparación entre predicciones y datos observables**

Durante la evaluación de la regresión lineal se considera que en general existe una diferencia entre las predicciones absolutas del modelo y los valores observados. En la práctica se asume que al momento de realizar una observación existe una **contribución de múltiples efectos que introducen cierto nivel de ruido en las mediciones**

Por ejemplo, si suponemos que la relación entre dos variables es **lineal mas ruido normal** podemos calcular el valor típico que tendría una métrica considerando fluctuaciones estadísticas de los datos. **Estos detalles son necesarios** de tener en cuenta cuando **evaluamos los resultados de un modelamiento de los datos**

Predicción teórica vs observación

La regresión lineal asociado a un set de datos se puede interpretar como un **modelo de la realidad** cuya representación matemática esta dada por la ecuación

$$\hat{Y} = 699 - 2.28\,X$$

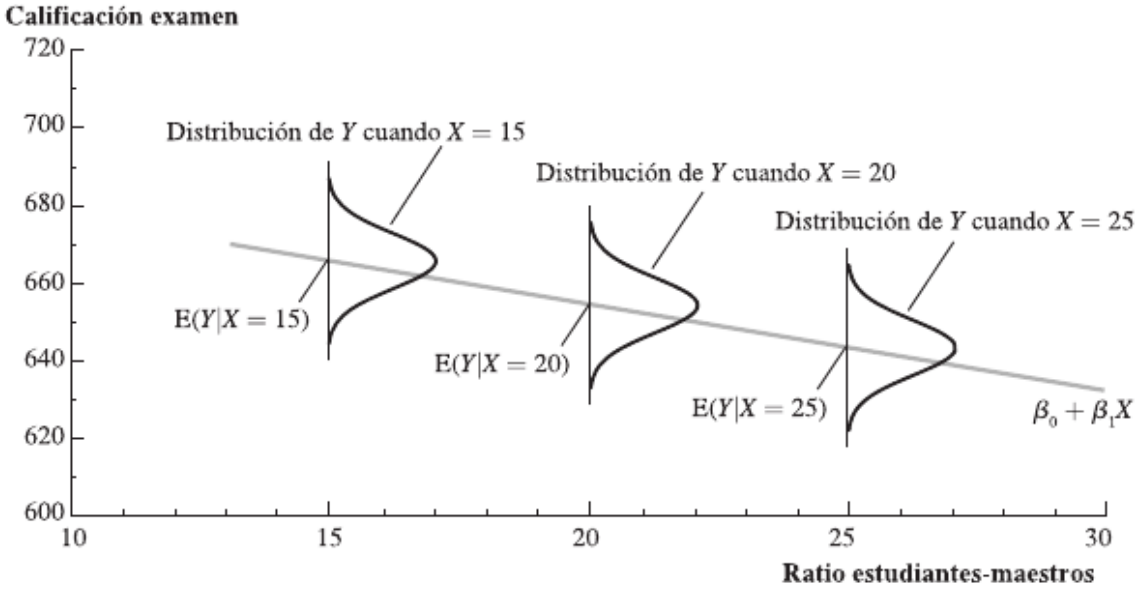
Don X e Y representan las variables observables del problema. Para cada valor de X (número de profesores por estudiante) podemos predecir el valor de Y esperado (promedio de notas de los exámenes.

Bajo la suposición de que esta es una representación correcta podemos preguntarnos sobre el comportamiento de los datos observados.

En general podemos notar que una observación en particular de Y puede estar afectada por una variedad de **factores aleatorios que agregan ruido a la observación.** Este tipo de ruido se puede incorporar en una simulación de los datos considerando que para todo valor de X la variable **Y observado sigue una distribución normal**

Predicción teórica vs observación

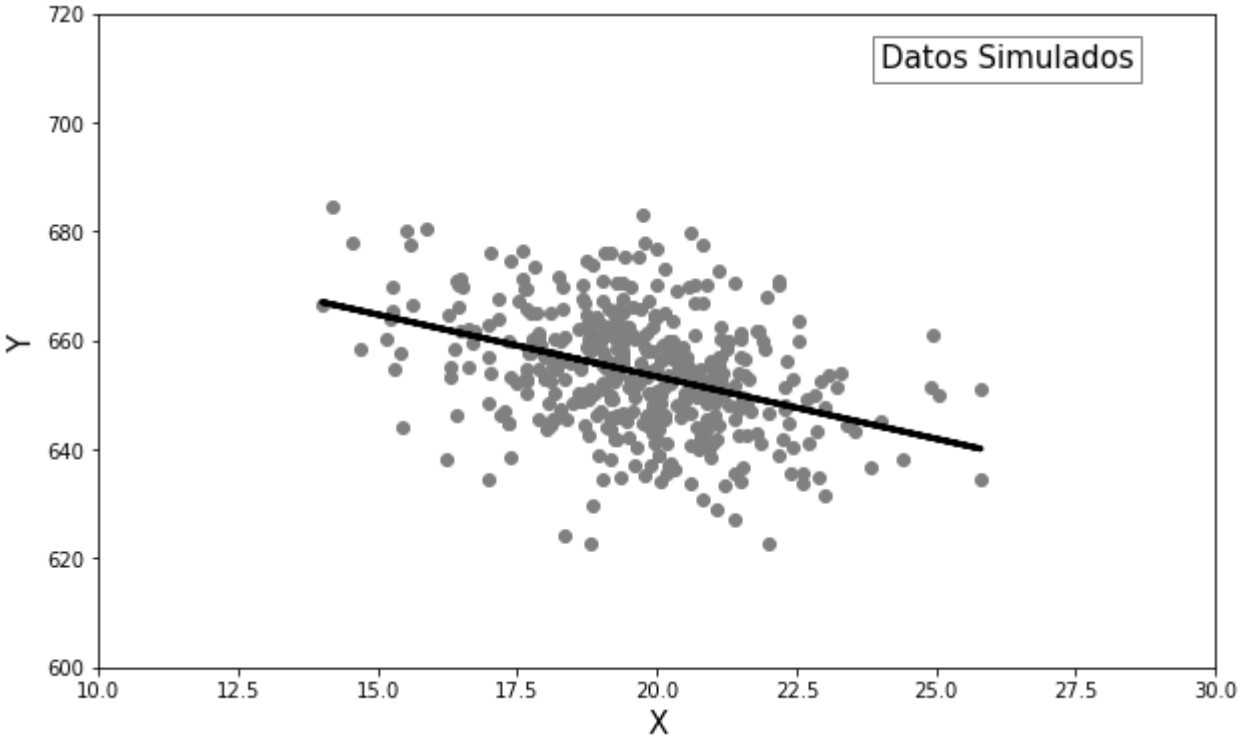
FIGURA 4.4 Las distribuciones condicionales de probabilidad y la recta de regresión poblacional



La figura muestra la probabilidad condicional de las calificaciones en los exámenes para los distritos con tamaño de clases de 15, 20, y 25 estudiantes. La media de la distribución condicional de las calificaciones en los exámenes, dada la ratio estudiantes-maestros, $E(Y|X)$, es la recta de regresión poblacional $\beta_0 + \beta_1 X$. Para un valor dado de X, Y se distribuye en torno a la recta de regresión y el error, $u = Y - (\beta_0 + \beta_1 X)$, tiene una media condicional igual a cero para todos los valores de X.

En general es esperable que los datos observados se alejen de una función lineal estricta puesto que existe la contribución de ruido desde otros factores

Simulación de datos observados a partir de predicción teórica



En general es esperable que los datos observados se alejen de una función lineal estricta puesto que existe la contribución de ruido desde otros factores

Datos simulados a partir de una regresión lineal considerando un efecto de ruido con desviación estándar igual a 10 puntos en unidades comparables con las calificaciones

Análisis de segundo orden

Los siguientes conceptos son aplicables directamente en el caso multilineal

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

Interpretación estadística del error estándar

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

Recordatorio: Métricas para Optimización

Los coeficientes de la recta que deseamos ajustas a los datos son obtenidos a partir de un proceso de minimización de una métrica bien definida

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

RSS = Residual Square Sum o
Suma de Residuos al Cuadrado (SR)

Minimizar con respecto a b_0 y b_1 numéricamente o analíticamente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{cov_{x,y}}{var_x}$$

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

Notamos que los coeficientes obtenidos producto de este proceso de minimización dependen de otras métricas estadística, tales como la media y varianza.

La métrica RSS tiene dimensiones de Y^2 y por lo tanto no esta normalizada

Evaluaciones del Ajuste

Un parámetro relevante para evaluar la regresión corresponde al **Error Stándard de la Regresión (ESR)**, el cual es calculado como

$$\text{ESR} = s_e, \text{donde } s_e^2 = \frac{\sum {e_i}^2}{n-2} \quad \text{y} \quad e_i = Y_i - \hat{Y}_i$$

Las unidades de e_i son las mismas que las de Y_i, por lo tanto el valor de ESR informa sobre el **error medio entre las predicciones y los valores observados**

ESR es el promedio de RSS corregido por los grados de libertad. En general es una métrica que permite comparar entre diferentes modelos, dado que compara directamente los valores obtenidos a partir del modelo con respecto a los valores observados

ESR o RSS representan una métrica de primer orden para discutir la cercanía entre los modelos y los datos observados

Evaluaciones del Ajuste: R^2

Para evaluar el poder predictivo de un modelo es necesario estudiar métricas que permitan evaluar la cercanía entre las predicciones del modelo en comparación con los datos observados. **En particular buscamos una métrica normalizada con extremos claros**

Uno de estas métricas es el **coeficiente de ajuste R^2,** el cual se define como la proporción entre la varianza explicada y la varianza total

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad e_i = Y_i - \hat{Y}_i \longrightarrow SE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \qquad ST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \longrightarrow R^2 = \frac{SE}{ST}$$

Cuando b_1 es cero, tenemos que b_0 = <Y> y por lo tanto las predicciones de la regresión lineal estan dadas por un solo valor igual a <Y>. En este caso **R^2 = 0** y se entiende que el **poder predictivo del modelo es nulo**. Por otro lado, en el caso hipotético que las predicciones sean todas iguales a los valores observados **SE = ST** y **R^2 = 1**. En este caso se entiende que el **poder predictivo es máximo**

Evaluaciones del Ajuste

Otra forma de calcular R^2 se puede obtener a partir de la suma de residuos al cuadrado (SR)

$$e_i = Y_i - \hat{Y}_i \longrightarrow \begin{matrix} SR = \sum_{i=1}^{n} {e_i}^2 \\ ST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \end{matrix} \longrightarrow R^2 = 1 - \frac{SR}{ST}$$

En particular, es posible mostrar que **$R^2 = r^2$** cuando consideramos el caso particular de una regresión lineal de un regresor

$$\mathrm{Var}(\hat{Y}) = \beta_1^2 \mathrm{Var}(X)$$

$$\mathrm{Var}(\hat{Y}) = \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)}$$

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} = \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)} = r^2$$

Para un modelamiento basado en una regresión lineal podemos demostrar que el R^2 de la regresión es igual al coeficiente de correlación al cuadrado

Comentarios sobre el ajuste

ESR y R^2 para el ejemplo de las escuelas y su interpretación

El valor de **ESR = 18.6 y R^2 es 0.05**. Esto indica que por un lado la regresión lineal solo explica un 5% de la variación observada en los datos, lo cual sugiere que el resto de la variación debería ser explicado por otros factores relacionados al problema. Además podemos notar que el ESR es bastante elevado, mucho mayor que la tasa de cambio unitaria de la nota de los exámenes, por ejemplo, lo cual indica que las predicciones obtenidas por la regresión no serán muy precisas.

El valor de ESR es útil para comparar entre modelos

Distribución muestral de los estimadores MCO

Los coeficientes de la Regresión Lineal Simple, b_0 y b_1 son obtenidos a partir de una muestra particular de los datos. Por lo tanto el valor de los coeficientes puede variar dependiendo de la muestra considerada. Suponiendo que tenemos acceso a múltiples muestras aleatorias independientes es posible estimar la **distribución muestral de los coeficientes**

De la misma forma que la distribución muestral del promedio de una variable aleatoria se puede obtener a partir de la técnica de bootstrapping en el caso de la distribución de los coeficientes se puede aplicar la misma lógica.

$\bar{X}$

$\hat{\beta}_1$

→ Bootstrapping o Teórema del Limite Central →

$$\mathrm{E}(\bar{X}) = \mu_X \ , \ \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \ , \ \bar{X} \sim N(\mu_X, \sigma_{\bar{X}})$$

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 \ , \sigma_{\hat{\beta}_1} \quad , \ \hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1})$$

Es interesante notar que el mismo tipo de analogía y relaciones entre los valores medios y los coeficientes permiten obtener variados detalles sobre las distribuciones buscadas

Coordinación de Proyectos

Bases para Proyecto T2

Objetivos: Ejemplo directo de modelamiento de datos usando Regresión Lineal Simple. Aplicar el flujo de análisis de datos, desde la selección y limpieza de datos hasta el calculo y visualización de una Regresión Lineal Simple.

Se aplican las mismas reglas generales de T1. En particular, la presentación debería tener entre 10 y 15 laminas y durar entre 15 y 30 minutos

El algoritmo del notebook, hilo conductor de la presentación y esquema del video debería considerar los siguientes pasos

Seleccionar datos (csv de kaggle u otros)
Preparación de datos (opcional)
Selección de columnas X e Y considerando contexto causal entre las variables
- Chequeo del contenido y distribución de las variables
- Gráfico de dispersión
- Planteamiento del modelo de Regresión Lineal para ajustar tendencia de X e Y
- Calcular explícitamente los valores de los coeficientes

Discusión de resultados

Calendario y Evaluaciones

TRIM.	FECHA	HORA
TRIM.2	sábado, 24 de agosto de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 31 de agosto de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 7 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 14 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 28 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 5 de octubre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 19 de octubre de 2024	11.20 - 12.30
		12.30 - 13.40

Publicación T1: Preparación de Datos

Publicación T2: Regresión Lineal

Publicación T3: Series Temporales

Más ejercicios con múltiples alternativas

Calendario y Evaluaciones

TRIM.	FECHA	HORA
TRIM.2	sábado, 24 de agosto de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 31 de agosto de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 7 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 14 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 28 de septiembre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 5 de octubre de 2024	11.20 - 12.30
		12.30 - 13.40
TRIM.2	sábado, 19 de octubre de 2024	11.20 - 12.30
		12.30 - 13.40

Más ejercicios con múltiples alternativas

Entrega T1: Limpieza y Estructura de Datos

Entrega T2: Regresión Lineal

Entrega T3: Series Temporales