Quinta Clase de Análisis de Datos

Prof: Boris Panes
Universidad Del Desarrollo

Septiembre 28, 2024

Durante la clase tendremos la oportunidad de conversar sobre el avance de su proyecto T2

¿Por qué o cuando es necesario o recomendable incluir nuevas variables en el análisis?

¿Cómo identificamos e incluimos estas nuevas variables en el análisis del problema?

Resumen del caso de estudio: *School System in California*

Estamos a fines de los 90 en California, USA. El equipo del departamento de educación esta buscando ideas para mejorar el rendimiento de los alumnos en el estado. Se cuenta con abundantes datos relacionados con los estudiantes, profesores, establecimientos y notas en exámenes estandarizados

En términos generales es posible **hipotetizar** que el rendimiento de los alumnos debería estar correlacionado con la calidad de la enseñanza impartida por los colegios. Esta calidad debería aumentar mientras mas personalizado es el trato con los estudiantes. Para aumentar el nivel de dedicación inicialmente se plantea **reducir el numero de estudiantes por profesor.**

Extracto de la tabla de datos disponible

```
In [5]: df[["district","enrl_tot","teachers","str","testscr","el_pct","computer"]]
Out[5]:
```

	district	enrl_tot	teachers	str	testscr	el_pct	computer
0	Sunol Glen Unified	195	10.900000	17.889910	690.799988	0.000000	67
1	Manzanita Elementary	240	11.150000	21.524664	661.200012	4.583333	101
2	Thermalito Union Elementary	1550	82.900002	18.697226	643.599976	30.000002	169
3	Golden Feather Union Elementary	243	14.000000	17.357143	647.700012	0.000000	85
4	Palermo Union Elementary	1335	71.500000	18.671329	640.849976	13.857677	171
...	...	...	...	...	...	...	...

Resumen del caso de estudio: *School System in California*

En términos prácticos se plantea que debería existir una **relación entre las notas de los exámenes y la cantidad de estudiantes por profesor**. Se entiende que deben haber algunos otros factores que son relevantes para esta evaluación. **Considerando esta situación se decide** comenzar por **un análisis simple pero bien definido** basado en un modelo de **regresión lineal**

Básicamente, se plantea que la **relación entre notas y numero de alumnos por profesor** sigue una **tendencia lineal**, mientras mas pequeños los cursos mejor es el rendimiento. Los resultados de este primer análisis permiten comprender que **la tendencia existe pero no es definitiva**. El paso siguiente es **considerar otras variables disponibles en el problema.**

Las métricas estadísticas de las variables disponibles con respecto a la variable dependiente y las independientes entregan información cuantitativa que nos permite entender cuantas y que variables deberíamos incluir en la extensión de la regresión lineal. La correcta identificación y tratamiento de estas variables permitirán aumentar el nivel de precisión de los resultados de los modelos.

Regresión Lineal Total

Regresión lineal simple

Presentación de datos
Ecuación de la recta mas correcciones (Regresión Lineal Simple)
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

Sesgo de variable omitida
Ecuación del hiperplano más correcciones (Regresión Lineal Múltiple)
Factores y variables categóricas
Multicolinealidad, factores de confusión e interacciones

Contenidos

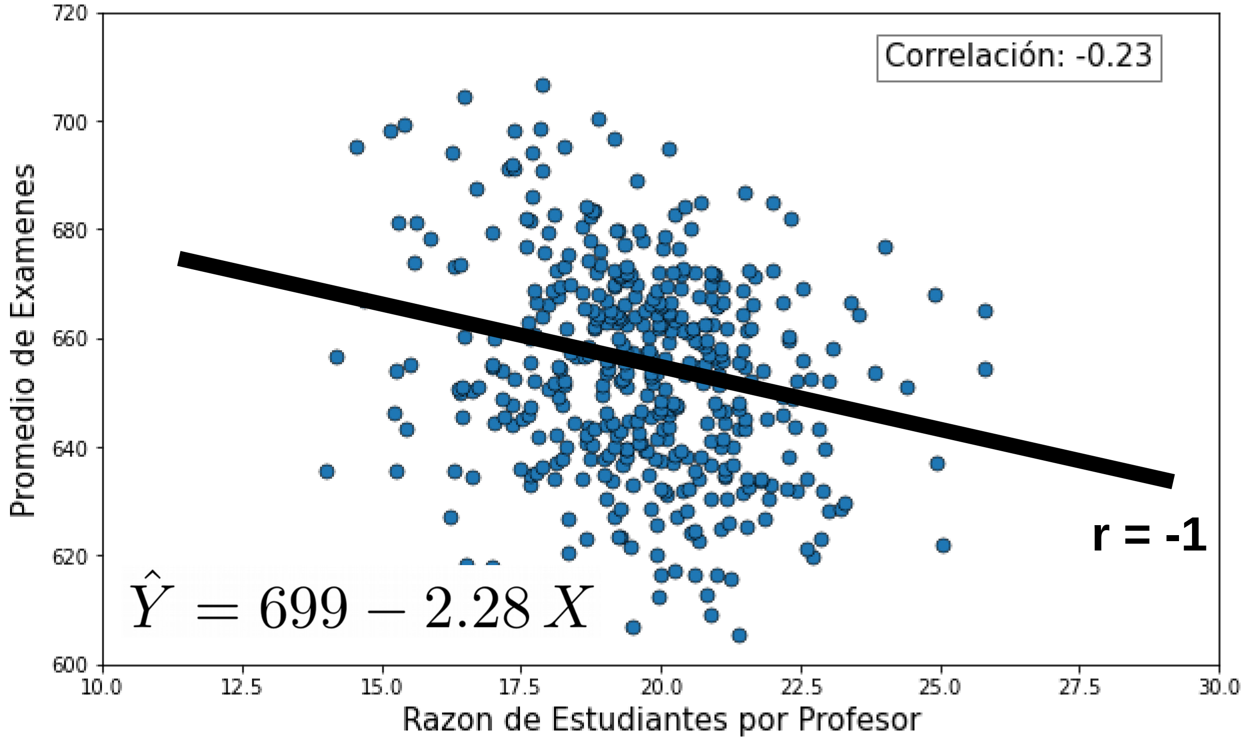
Discusión de elementos relacionados con la Regresión Lineal

- Evaluación del ajuste obtenido por el modelo lineal
  - ESR y Coeficiente de determinación R^2
- Supuestos de MCO
  - Distribución de los estimadores
  - Test de hipótesis

Regresión Lineal Múltiple

- Sesgo de variable omitida
- Extensión de la Regresión Lineal Simple
- Evaluaciones
- Aspectos emergentes en problemas multivariables
  - Heterocedasticidad y homocedasticidad
  - Multicolinealidad entre potenciales regresores
  - Variables categóricas

Gráfico de dispersión y recta de tendencia



En este gráficos tenemos

puntos en azul con r = -0.23

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

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

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

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

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 Estándar 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

También podemos notar que **ESR debería ser similar a la desviación estándar de Y** siempre y cuando las predicciones Y_hat se encuentren cerca del promedio

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

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

$$\longrightarrow \qquad \longrightarrow \qquad R^2 = \frac{SE}{ST}$$

$$e_i = Y_i - \hat{Y}_i \qquad ST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Cuando b_1 es cero, tenemos que $b_0 = \langle Y \rangle$ y por lo tanto las predicciones de la regresión lineal están dadas por un solo valor igual a $\langle Y \rangle$. 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)

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

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 punto a punto, pero quizás si en promedio (esto depende de la distribución de los errores)**.

El valor de ESR es útil para comparar entre modelos

Notar las unidades de ESR y compararlos con la desviación estándar de los valores de Y (desviación estándar de la notas de los tests es 19.1)

Además también es posible usar ESR y R^2 para comparar entre diferentes alternativas a MCO para calcular los estimadores, **por ejemplo usando mas variables independientes**

Supuestos de MCO y consecuencias

Supuestos de MCO: significancia de la tendencia observada

Bajo los siguientes supuestos es justificado plantear que los valores obtenidos de b_0 y b_1 a partir del método de mínimos cuadrados converjan a los valores reales de la regresión lineal que genera los datos observados

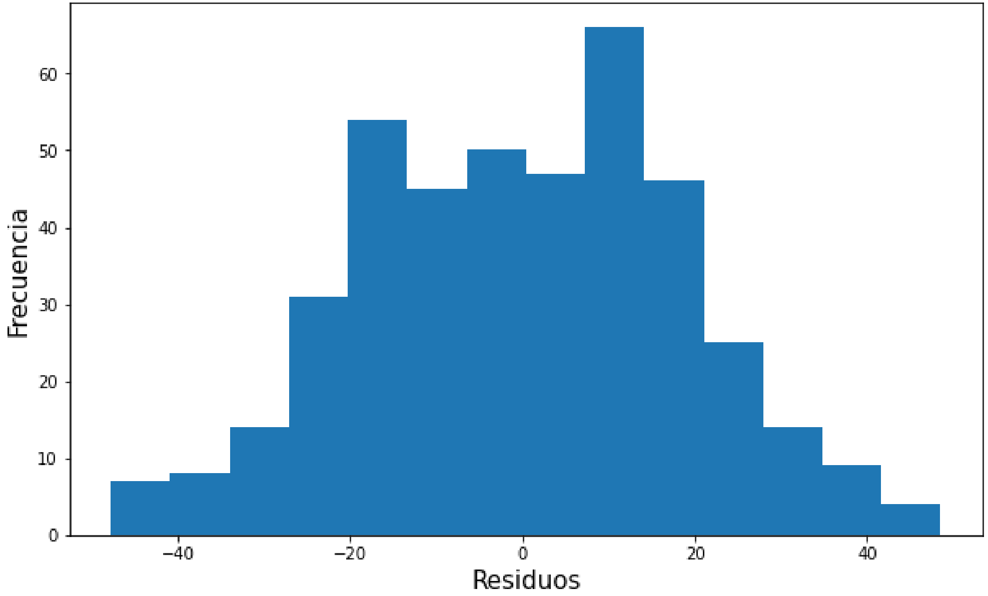
1- **La media de los errores e_i condicionado sobre X_i es igual a cero**. En la práctica esto quiere decir que se espera que los errores distribuyan normalmente en torno al valor Y predicho por la regresión lineal. X y e son incorrelacionados. Presencia de factores externos. **Este supuesto implica que los estimadores son insesgados**

2- **Los valores de X_i e Y_i son independientes e idénticamente distribuidos (i.i.d).** En la práctica esto significa que la muestra de datos es una representación aleatoria simple de los datos. Es una condición sobre la metodología de toma de datos. **Este supuesto se aplica para obtener la varianza muestral de los estimadores**

3- **Los datos atípicos son improbables.** Esto quiere decir que la distribución de probabilidad de las variables X_i e Y_i tiene una curtosis finita (decrece rápidamente para valores alejados de la media). **Aproximación utilizando durante la obtención de la distribución de los estimadores de la regresión**

Supuestos de MCO: errores aleatorios

Histograma de residuos muestra que la media global de estos es nula. Esto es un indicio que la media condicionada para todo X también lo es



¿Cómo podríamos calcular la distribución de u dado un X en particular, tal como indica el supuesto original?

En términos generales podemos notar que las evaluaciones del modelo de regresión lineal muestran una consistencia aceptable con los datos. Además podemos ver que estos son consistentes con los supuestos. En este caso la recta aproxima muy de cerca la tendencia de los valores de <Y> condicionados sobre X **(discutir observaciones futuras en promedio)**

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.

$\bar{X}$

$\hat{\beta}_1$

$\longrightarrow$ **Remuestreo o Tratamiento aproximado: TLC** $\longrightarrow$

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

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

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

Distribución muestral de los estimadores MCO

Asumiendo que se cumplen los tres supuestos de MCO se puede obtener que los estimadores son insesgados y consistentes

Insesgado: la esperanza de los coeficientes es igual al valor real para n grande

$$\mathrm{E}(\hat{\beta}_1) = \beta_1$$

Consistente: la desviación estándar del estimador se aproxima a cero con n grande

$$\sigma_{\hat{\beta}_1} \to 0 \text{ para } n \to \infty$$

Utilizando las expresiones para la dispersión de los estimadores, más las métricas de evaluación podemos reportar el resultado de la regresión de la siguiente forma

$$\widehat{CalificaciónExamen} = \underset{(10{,}4)}{698{,}9} - \underset{(0{,}52)}{2{,}28} \times REM,\ R^2 = 0{,}051,\ ESR = 18{,}6.$$

Test de hipótesis

Comparación entre modelos con distintos coeficientes utilizando la información generada en el capitulo sobre distribución muestral de los estimadores

Test estadístico AB: promedio de tiempo de sesión comparado entre las paginas A y B

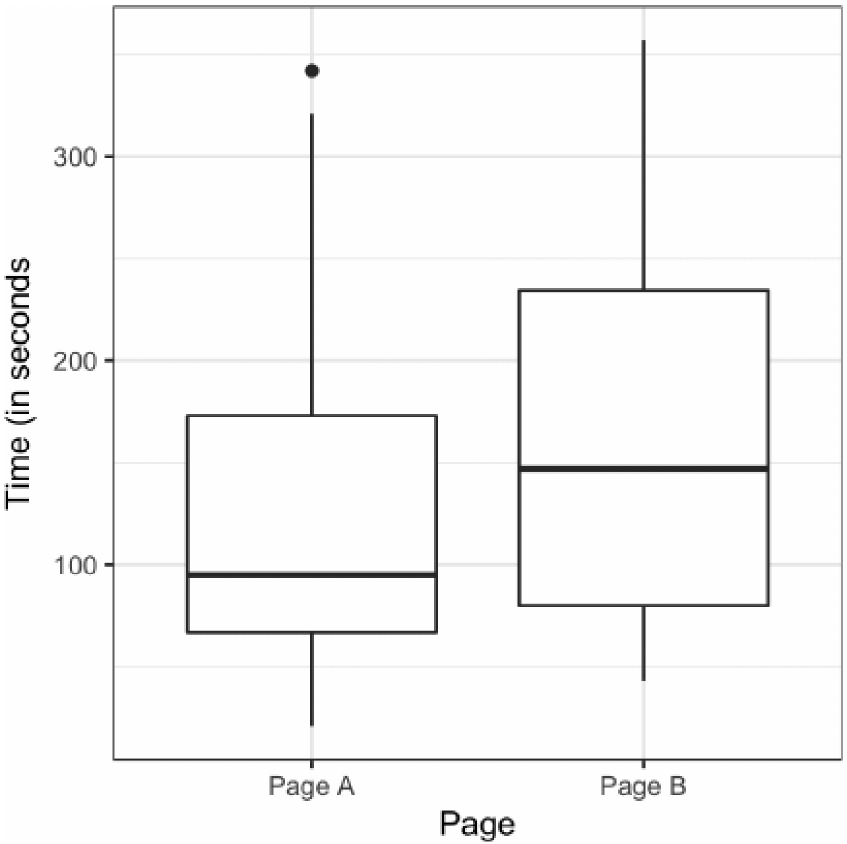


Figure 3-3. Session times for web pages A and B

Numero total de sesiones (A+B) = 36
Sesiones en A = 21
Sesiones en B = 15

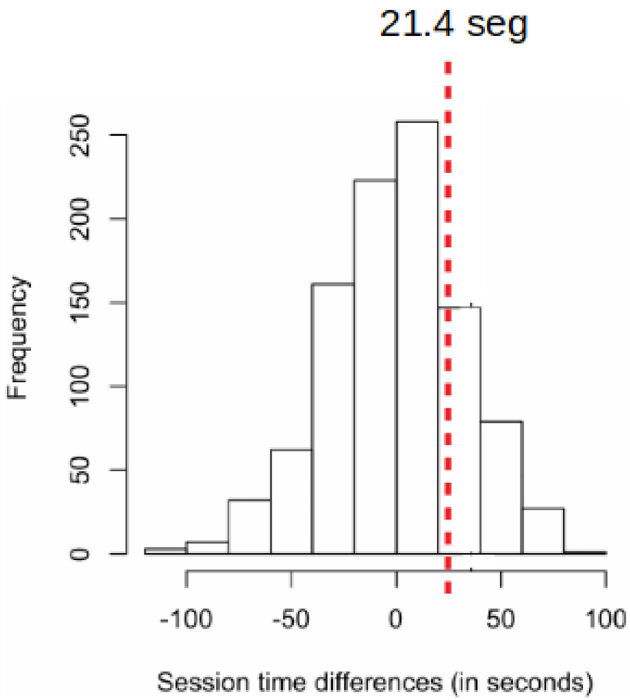


Figure 3-4. Frequency distribution for session time differences between pages A and B

Diferencia entre promedios = 21.4 seg

Test estadístico aproximado

A partir del proceso de muestreo es posible obtener empíricamente las distribuciones de las métricas estadísticas.

Por otro lado es posible verificar que ciertos aspectos de este tipo de distribuciones pueden ser obtenidas de manera analítica utilizando por ejemplo la ley de los grandes números

Dado ciertas condiciones sobre las variables X_i es posible anticipar el valor de la media y la desviación estándar de la métrica asociada a la media <X>, estos valores resultan ser

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

A partir de estas expresiones **es posible extender la discusión hacia los test de hipótesis** entre las diferencias de medias, como en el test-AB. En este caso obtenemos que para un test-AB la significancia de la hipótesis alternativa con respecto a la hipótesis nula, dada por una diferencia nula esta dada por

$$t_{obs} \simeq \frac{d_{AB}}{\sigma_X/\sqrt{n}} \longrightarrow \text{p-value} = P(t > t_{obs})$$

usando una distribución-t

Este término proviene de la distribucion de d_{AB}

Test estadístico aproximado

El error muestral para la diferencia de medias es una medida de la variabilidad esperada en la diferencia debido al muestreo. Se calcula combinando los errores muéstrales de cada grupo. La fórmula para el error muestral de la diferencia de dos medias es:

$$\sigma_X/\sqrt{n} \approx \sqrt{\sigma^2_{X_A}/n_A + \sigma^2_{X_B}/n_B}$$

La distribución de base para el test de hipótesis se puede construir empíricamente bajo la hipótesis nula usando un muestreo por permutación aleatoria entre los grupos. Por otro lado las expresiones anteriores son construidas considerando los datos observados para ambos grupos.

El punto clave es que ambas expresiones para el error muestral coinciden y por lo tanto es posible usar el test estadístico mostrado en la lamina anterior para calcular el p-value

Test estadístico de los coeficientes

A partir de la obtención de una distribución muestral de los estimadores b_0 y b_1 es posible definir un test estadístico basado en esta distribución

H_0 = La hipótesis nula asume que la diferencia entre dos estimadores alternativos es nula
H_1 = La hipótesis alternativa asume que la diferencia es distinta de cero **(hipótesis bilateral)**

El test estadístico se define a partir de la diferencia entre el valor obtenido del coeficiente y el valor esperado del coeficiente **(normalmente una regresión con pendiente nula)** dividido por el error estándar de la distribución del estimador. Para una muestra grande de eventos el test estadístico sigue una **distribución normal estándar**

$$t_{\text{obs}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sigma_{\hat{\beta}_1}} \longrightarrow \text{p-value} = P(|t| > |t_{\text{obs}}|)$$

A partir de este análisis podemos encontrar que, en el ejemplo de notas versus profesores por estudiante, la pendiente obtenida es distinta de 0 con una alta significancia estadística **($p < 1\%$)**. Es decir, **la alternativa de una regresión con pendiente nula es excluida.**

Existe una tendencia en los datos pero el error de las predicciones es considerable

Regresión Lineal Múltiple

Regresión Lineal Múltiple

Motivaciones para extender la discusión hacia una Regresión Lineal Múltiple

Pueden existir otros factores relevantes para explicar la variable objetivo

Evaluación del sesgo al escoger una variable X en particular (Sesgo de variable omitida)

Efectos propios de una Regresión Lineal Múltiple

Factores o variables categóricas que influyan en el calculo de una Regresión Lineal Simple aparecen una vez que extendemos el dominio de la variables predictivas

La multicolinealidad entre las variables independientes

Argumentos a partir del primer análisis de regresión lineal

En términos generales podemos notar dos situaciones que deberían llamar la atención del **análisis inicial del caso de estudio**

> Por un lado notamos que el valor de ESR es grande y R^2 es pequeño, por lo tanto es posible concluir que existen otras variables que pueden estar afectando las mediciones de manera considerable. Dado la distribucion de los u_i pareciera que estas variables introducen un ruido aleatorio a la medición que podría ser reducido considerando mediciones promedio para cada X_i
>
> Además, dado la naturaleza del problema y su complejidad intrínseca es razonable considerar que pueden existir otras variables explicativas de los resultados

Por lo tanto, dado que tenemos dos indicadores, uno cuantitativo y otro cualitativo, que nos hablan de la posible existencia de otros factores que podrían estar afectando las observaciones es razonable considerar estas variables explícitamente dentro del análisis

Lo anterior es claramente posible si disponemos de los datos asociados a estos posibles nuevos factores. Por ejemplo, las características **sociales de los estudiantes, profesores y colegios en cuestión**

Búsqueda de variables relevantes

Por lo tanto, el objetivo principal de extender el análisis hacia una regresión múltiple es lograr incorporar las variables omitidas, pero disponibles, en el ajuste de los datos. Esto permite por ejemplo preguntarse **cuanto afecta cada variable independiente cuando se mantienen fijas las demás**

6.1 Sesgo de variable omitida

Al haberse centrado únicamente en la ratio estudiantes-maestros, el análisis empírico de los Capítulos 4 y 5 ignoraba algunos factores que potencialmente podían ser determinantes importantes de las calificaciones en los exámenes, quedando recopilada su influencia en el término de error de la regresión. Estos factores omitidos incluyen las características de la escuela, tales como la calidad de los maestros y el uso del ordenador, y las características de los estudiantes, tales como el entorno familiar. Comenzamos por considerar una característica omitida de los estudiantes que resulta especialmente relevante en California debido a su gran población inmigrante: la prevalencia en el distrito escolar de estudiantes que por no ser hablantes nativos se encuentran todavía aprendiendo inglés.

→ testscr o la nota promedio es calculada considerando la nota de pruebas estandarizadas de lenguaje **(ingles)** y matemáticas

Sesgo de variable omitida: condiciones

Existencia de una variable relevante para la determinación de la variable dependiente:

En California existe una alta población hispánica, que se encuentra en general aprendiendo ingles. Es esperable que estos alumnos tengan un mas dificultad en las pruebas de lenguaje (ingles)

Existencia de una variable correlacionada con la variable independiente:

Las clases grandes asimismo tienen muchos estudiantes que aún están aprendiendo inglés

Consecuencia:

La regresión MCO de las calificaciones en las pruebas sobre la ratio estudiantes-maestros podría encontrar erróneamente una correlación y procurar un coeficiente estimado grande, cuando en realidad el verdadero efecto causal de la reducción del tamaño de las clases sobre las calificaciones obtenidas es pequeño, e incluso nulo.

Ambas condiciones deben ser cumplidas para que exista sesgo de variable omitida

TABLA 6.1 Diferencias en las calificaciones en los exámenes para los distritos escolares de California con alta y baja ratio de estudiantes por maestro, agrupados por porcentaje de estudiantes de inglés del distrito

	Ratio estudiantes-maestros < 20		Ratio estudiantes-maestros ≥ 20		Diferencia en la calificación en el examen, bajo REM vs alto REM	
	Promedio Calificación Examen	***n***	**Promedio Calificación Examen**	***n***	**Diferencia**	**Estadístico *t***
Todos los distritos	657,4	238	650,0	182	7,4	4,04
Porcentaje de estudiantes de inglés						
< 1,9 %	664,5	76	665,4	27	–0,9	–0,30
1,9-8,8 %	665,2	64	661,8	44	3,3	1,13
8,8-23,0 %	654,9	54	649,7	50	5,2	1,72
> 23,0 %	636,7	44	634,8	61	1,9	0,68

Regresión Lineal Múltiple

En términos prácticos la regresión lineal múltiple es una extensión de la regresión lineal simple, al menos en las siguientes características

	Regresión Lineal Simple	Regresión Lineal Múltiple
Columnas	X, Y	$X_1, X_2, ..., X_n, Y$
Ecuación de la “recta”	$\hat{Y} = \beta_0 + \beta_1 X$	$\hat{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$
Puntos de los datos	$Y_i = \hat{Y}_i + e_i$	$Y_i = \hat{Y}_i + e_i$
Coeficientes	β_0, β_1	$\beta_0, \beta_1, ..., \beta_n$
Optimización	Mínimos Cuadrados Ordinarios	Mínimos Cuadrados Ordinarios
Distribución Muestral	Progresivamente Normal	Progresivamente Normal

Regresión Lineal Múltiple

En términos prácticos la regresión lineal múltiple es una extensión de la regresión lineal simple, al menos en las siguientes características

Expresión para los coeficientes sigue siendo analítico

Regresión Lineal Simple

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{cov_{x,y}}{var_x}$$

Regresión Lineal Múltiple

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

Notación matricial, Capitulo 18 de S&W 2012

Tabla de datos

```
In [7]: df[["district","enrl_tot","teachers","str","testscr","el_pct"]]
Out[7]:
```

	district	enrl_tot	teachers	str	testscr	el_pct
0	Sunol Glen Unified	195	10.900000	17.889910	690.799988	0.000000
1	Manzanita Elementary	240	11.150000	21.524664	661.200012	4.583333
2	Thermalito Union Elementary	1550	82.900002	18.697226	643.599976	30.000002
3	Golden Feather Union Elementary	243	14.000000	17.357143	647.700012	0.000000
4	Palermo Union Elementary	1335	71.500000	18.671329	640.849976	13.857677
...	...	...	...	...	...	...
415	Las Lomitas Elementary	984	59.730000	16.474134	704.300049	5.995935
416	Los Altos Elementary	3724	208.479996	17.862625	706.750000	4.726101
417	Somis Union Elementary	441	20.150000	21.885857	645.000000	24.263039
418	Plumas Elementary	101	5.000000	20.200001	672.200012	2.970297
419	Wheatland Elementary	1778	93.400002	19.036402	655.750000	5.005624

420 rows × 6 columns

Para evaluar el sesgo de variable omitida es suficiente considerar al menos una variable extra y calcular variación en el valor de los coeficientes

Y = testscr
X_1 = str
X_2 = el_pct

Porcentaje de estudiantes aprendiendo ingles

Sesgo de variable omitida: evaluación

El cambio abrupto en los coeficientes es una señal de que existe un sesgo de variable omitida al considerar solo un regresor como estimador

Modelo 1: **testscr vs str**

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

Modelo 2: **testscr vs str + el_pct**

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

usando MCO

$$\hat{\beta}_0 = 699$$
$$\hat{\beta}_1 = -2.28$$

$\longrightarrow$

$$\hat{\beta}_0 = 686 \qquad \hat{\beta}_2 = -0.65$$
$$\hat{\beta}_1 = -1.18$$

Interpretación de los coeficientes b_1 y b_2 = variación de la variable Y dado una unidad de variación en alguna de las variables b dejando el resto fijas. **Mientras mayor es el sesgo de variable omitida mayor es el cambio en el valor del coeficiente**

Estimador de sesgo de variable omitida

El cambio de valor en el coeficiente es una señal de que el **valor estimado considerando una regresión lineal simple no es el más correcto que se puede obtener.** De hecho es posible mostrar que el valor de b_1 calculado a partir de MCO es sesgado cuando existe una correlación entre la variable X_1 y los errores e_i (incumplimiento de un requerimiento de MCO)

Si existen variables X_2 que afectan el valor de Y sus efectos deberían ser reflejados en el comportamiento de e_i. Por lo tanto si X_2 esta correlacionado con X_1 esto derivara en una correlación entre X_1 y e_i, generando un sesgo en el valor obtenido del coeficiente b_1

$$\text{Sesgo} = \beta_2 \cdot \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}$$

Para recuperar un valor insesgado de b_1 es necesario extender la dependencia de Y con respecto a las variables X_1 y ahora también X_2.

En resumen, ante la presencia de variables explicativas de Y es necesario considerar una expresión lo mas extendida posible para que los valores de los coeficientes sean insesgados

Evaluación del modelo multilineal

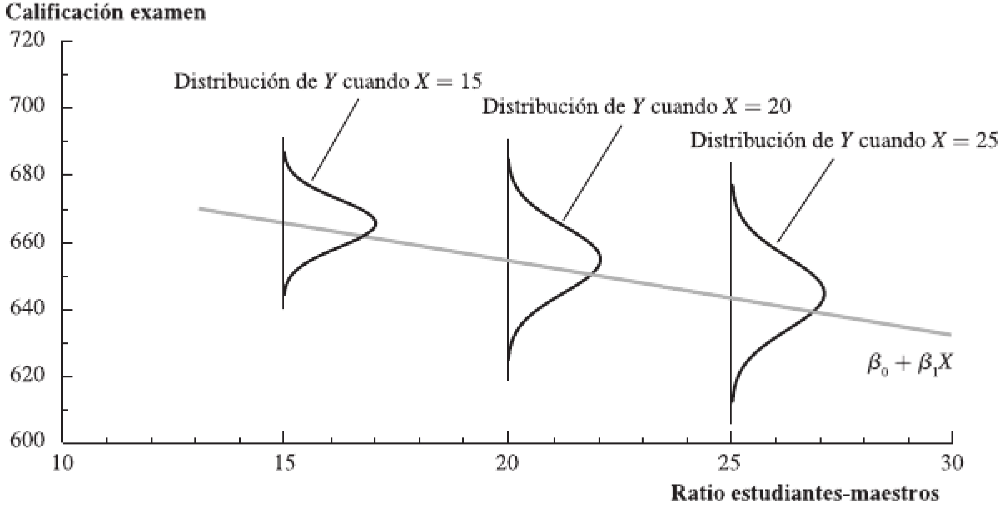
Las expresiones para ESR y R^2 son directamente utilizables en el caso de una regresión múltiple. En estas condiciones obtenemos que el R^2 de este hiperplano de regresión es $R^2=0.426$, el R^2 ajustado es $R^2=0.424$, y el error estándar de la regresión es ESR=14.5.

	RLS	RLM
R^2	0.05	0.42
ESR	18.6	14.5

Como se anunciaba desde un principio, las evaluaciones de la regresión lineal múltiple presentan mejores resultados que la regresión lineal simple, probablemente justificado por la estrategia de búsqueda de las variables extras

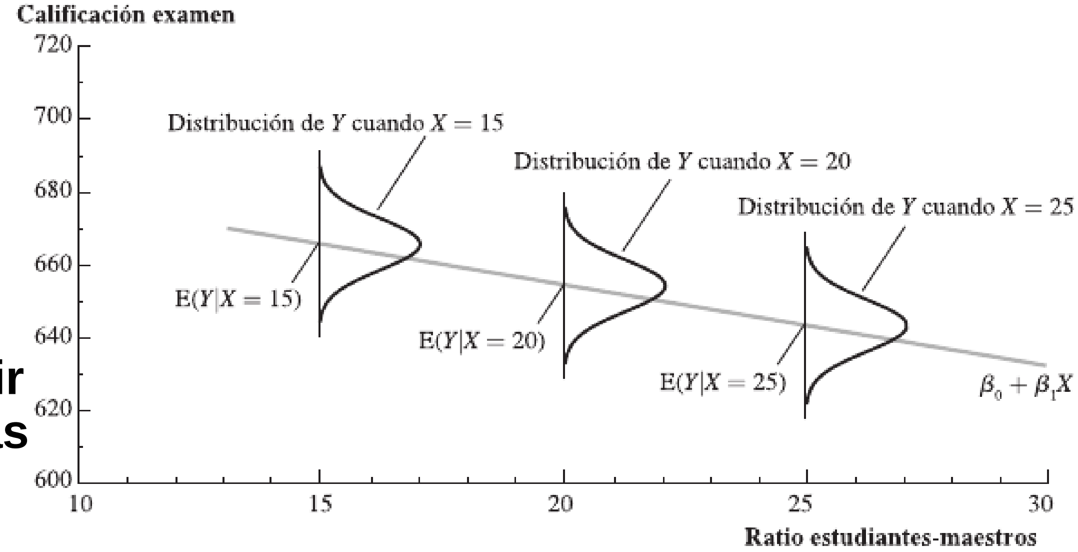
Tópicos extra

Heterocedasticidad



Estos términos definen el comportamiento de los errores en cuanto al valor de la varianza condicionada

Homocedasticidad



En términos generales y prácticos siempre es mejor suponer y utilizar las expresiones que consideran que el comportamiento de los errores es heterocedastico. Sin embargo **la opción de homocedasticidad puede venir como alternativa implícita en los programas estadísticos**

Supuestos de MCO: 3+Ausencia de Multicolinealidad

Cuarto supuesto de MCO: su cumplimiento es necesario para derivar expresiones de los coeficientes, sus **distribuciones muéstrales y test de hipótesis**

Problema y solución

La multicolinealidad entre dos vectores/columnas X_1 y X_2 se genera cuando estos son paralelos/proporcionales. En este escenario no es posible encontrar unívocamente los valores de los coeficientes y la interpretación de los resultados se vuelve inestable. **Debemos extraer esas variables**

Si $X_2 = \alpha X_1$

Efectivamente solo existe un regresor

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$Y = \beta_0 + (\beta_1 + \alpha\beta_2) X_1$$
$$Y = \beta_0 + \beta_1' X_1$$

Valor indefinido de

$$(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$$

Variables independientes categóricas

Tratamiento de regresión lineal usando una variable binaria

Una variable categórica binaria se puede tratar como una variable numérica con dos valores. Los métodos de la regresión lineal funcionan de la misma forma para obtener los coeficientes correspondientes

Interpretación de los coeficientes como promedios o diferencias de promedios (S&W 2012)

Transformación de variables categóricas en variables binarias múltiples

Al pasar de variables categóricas a variables binarias podemos aplicar el mismo procedimiento mencionado anteriormente

Ingeniería de datos al servicio de la regresión lineal (B&B 2017)

Combinación entre variables continuas numéricas y variables binarias

Calculo del sesgo de variable omitida puede realizarse de la misma forma

Variables independientes categóricas

```
head(house[, 'PropertyType'])
Source: local data frame [6 x 1]

   PropertyType
         (fctr)
1     Multiplex
2 Single Family
3 Single Family
4 Single Family
5 Single Family
6     Townhouse
```

Transformación naive de las categorías en valores numéricos consecutivos, *en general*, puede introducir problemas de interpretación, puesto que la identificación es ambigua

Binarización: una vez que los elementos categóricos han sido binarizados es posible incluirlos en el análisis de regresión lineal

```
prop_type_dummies <- model.matrix(~PropertyType -1, data=house)
head(prop_type_dummies)
  PropertyTypeMultiplex PropertyTypeSingle Family PropertyTypeTownhouse
1                     1                         0                     0
2                     0                         1                     0
3                     0                         1                     0
4                     0                         1                     0
5                     0                         1                     0
6                     0                         0                     1
```

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

Entrega T2: Octubre 6, 2024

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

Entrega T1: Limpieza y Estructura de Datos

Entrega T2: Regresión Lineal

Entrega T3: Series Temporales

Más ejercicios con múltiples alternativas

Apéndice

Fuente de datos para regresión lineal

https://media.pearsoncmg.com/ph/bp/bp_stock_econometrics_4_cw/

Pearson

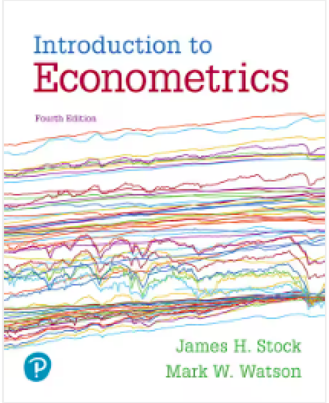
Introduction to Econometrics, 4th Edition

Stock • Watson

Help

COMPANION WEBSITE

Student Resources



Data Files

Download the data files you need for each chapter.

Additional Empirical Exercises

These are new exercises not found in the textbook.

STATA and EViews Tutorials

Learn to use these software programs.

Conducting a Regression Study Using Economic Data

Read basic guidelines for conducting an empirical analysis of economic data.

Ciencia de Datos en Las Noticias

Estudio detecta que el manejo deficiente de datos afecta la lucha contra la delincuencia

Por ejemplo, la falta de analistas de datos en Carabineros impide que la policía uniformada haga lecturas que se anticipen a las acciones criminales. También están al debe la Unidad de Análisis Financiero y la DGAC.

FABIAN LLANCA

"¿Inteligencia estatal para combatir la delincuencia? Brechas y desafíos en instituciones públicas" se llama el documento desarrolla-

El estudio plantea que la falta de científicos de datos impide que Carabineros haga análisis predictivos del comportamiento criminal.

gestión de datos en Carabineros, Sename, Directemar, Armada y DGAC; y fomentar la contratación de profesionales especializados con título técnico o universitario.

"Se requiere un cambio de paradigma en lo que respecta a la gestión y análisis de datos. Actualmente, la mayoría de los funcionarios de tecnología de la información en las instituciones ligadas a la seguridad realiza funciones de operador, soporte técnico o de desarrollador, cuando en verdad lo que se necesitan son más analistas de datos, administradores de bases de datos o científicos de datos para que el Estado pueda anticiparse y focalizar su acción contra la delincuencia", explica Alfonso España, cientista político e investigador de Horizontal.

¿Qué cosas no está haciendo Carabineros al no tener analistas de datos?

"Sin desmerecer su importante labor, la falta de analistas, administradores o científicos de datos, por ejemplo, no permite que Carabineros pueda realizar

Alcances de la regresión lineal

En el contexto de políticas publicas o económicas, podemos preguntarnos sobre los efectos de ciertas medidas sobre métricas especificas. Por ejemplo, podemos intentar medir el efecto que produce la reducción del número de alumnos en el rendimiento escolar.

En la práctica vamos a dedicar un tiempo considerable al estudio de este tema, lo cual nos permitirá analizar con cierto grado de profundidad una variedad de aspectos que van mas allá de la predicción, como lo son

- El cálculo de la incertidumbre asociada con una predicción
- Los efectos que pueden tener diferentes variables asociadas al proceso de modelamiento
- Extensión de 1 a N variables dependientes continuas

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