s



**Análisis exploratorio de datos univariados parte 1​**

[Análisis exploratorio de datos univariados parte 1​ | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/8rQiG/analisis-exploratorio-de-datos-univariados-parte-1)

La estadística es la ciencia del aprendizaje a partir de los datos.

Comprende el diseño del estudio o del experimento que

sea relevante para satisfacer nuestros intereses,

la recolección de los datos,

el modelamiento y el análisis con el propósito de crear

nuevo conocimiento y soportar la toma de decisiones en la organización,

tomando en consideración que frecuentemente la información

a nuestro alcance es tanto limitada como variable.

El proceso de aprendizaje

a partir de los datos semeja estrechamente el método científico,

un conjunto de principios y procedimientos

que emplean los científicos en la búsqueda de nuevo conocimiento

y se desarrolla en cuatro pasos: definición del problema,

recolección de los datos,

síntesis de los datos,

análisis de los datos,

interpretación del análisis y comunicación de los resultados.

El presente curso se ocupa parcialmente y a un nivel

introductorio de los últimos dos pasos del proceso descrito.

La relación entre ellos y los temas tratados en el curso es la siguiente.

Mientras que el análisis exploratorio provee la síntesis requerida en los datos,

los métodos de regresión y clasificación lineales y los

modelos lineales generalizados sustentan el análisis.

Se asumirá en consecuencia,

que el problema está claramente definido y que los datos han sido recolectados

apropiadamente a fin de ofrecer a través de ellos una solución al problema identificado.

De lo contrario, la interpretación de los resultados del análisis puede conducir

a conclusiones equivocadas en la medida en que el análisis se

desarrolla a partir de un conjunto de datos que no representa el problema,

o que a pesar de representarlo,

resulte incompleto o contenga información impropia.

El análisis exploratorio de datos puede definirse como el arte de observar uno o más

conjuntos de datos en un esfuerzo por comprender la estructura que subyace a ellos.

En la presente sesión desarrollaremos las

herramientas del análisis exploratorio para una sola variable,

denominado "análisis univariado", mientras que en la

siguiente sesión nos ocuparemos del análisis exploratorio bivariado.

Independientemente de si el conjunto de mediciones

a nuestro alcance es la población de interés como un todo

o una muestra convenientemente seleccionada y diseñada para inferir sobre ella,

un primer paso en la formulación de conclusiones acerca de la

población es la descripción precisa de los datos.

Nuestra tarea inicialmente consiste en organizar,

sintetizar y describir los datos.

Esto es, en darle sentido a los datos,

reduciendo un conjunto grande de mediciones

a unas pocas medidas descriptivas que caractericen las mediciones originales.

Las dos herramientas principales para describir o

explorar un conjunto de mediciones son

los métodos gráficos y los métodos de descripción numérica.

En esta presentación haremos caso omiso de los

primeros para concentrarnos en los segundos,

los métodos de descripción numérica.

Las medidas de descripción numérica son comúnmente empleadas para construir una

imagen clara sobre las características de un conjunto de mediciones,

algo difícil de lograr solo a partir de instrumentos gráficos.

Las medidas descriptivas de mayor uso son las medidas de

tendencia central y las medidas de variabilidad.

A través de ellas se busca localizar adecuadamente el

centro de la distribución de los datos para luego establecer

su dispersión alrededor del centro identificado.

Aunque su naturaleza es diferente,

en lo sucesivo no se hará una distinción entre las medidas descriptivas de una población;

denominadas "parámetros"; y las medidas descriptivas de una muestra;

denominadas "estadísticas".

En lugar de pretender calcular los parámetros de una población,

las estadísticas se utilizan en la práctica

para estimar los parámetros de interés asociados.

Entre las medidas de tendencia central de un conjunto de mediciones,

se destacan la moda,

la mediana y la media,

que definiremos a continuación.

La moda de un conjunto de mediciones se define como

aquella medición que ocurre o se observa con mayor frecuencia.

Es aplicable como medida de tendencia central tanto

para mediciones cuantitativas como cualitativas.

Debe notarse que algunas distribuciones tienen más de una moda.

De esta forma, podemos hablar de distribuciones bimodales,

trimodales, y así sucesivamente.

La mediana de un conjunto de mediciones se define de manera diferente,

según si el número de mediciones en el conjunto es par o impar.

Para un número impar de mediciones,

la mediana se define como el valor central de ese conjunto,

una vez las mediciones se encuentren ordenadas de menor a mayor.

Si ese número es par,

la mediana es el promedio de los dos valores centrales.

En cualquier caso, la mediana divide

el conjunto ordenado en dos grupos que contienen igual número de mediciones.

Es aplicable como medida de tendencia central

exclusivamente para mediciones cuantitativas.

A diferencia de la moda, la mediana es única a un conjunto de datos.

La media, o media aritmética,

de un conjunto de n mediciones que denotaremos por y1, y2 hasta yn,

se define como aquel valor que resulta de sumar las

mediciones en el conjunto y luego dividir

este resultado por el número total de mediciones.

y barra es igual a la suma de los y\_i, dividida por n.

Es, en efecto, el punto de balance del conjunto de datos.

La media de una muestra, convenientemente seleccionada,

de una población de interés,

se utiliza frecuentemente para estimar la media desconocida de esa población.

Es aplicable como medida de tendencia central

exclusivamente para mediciones cuantitativas.

También como la mediana, la media es única a un conjunto de datos.

Pero a diferencia de la moda y la mediana,

la media está altamente influida por los valores

extremos en el conjunto denominados "outliers".

En la práctica no estamos limitados

al uso de una sola de las medidas anteriormente descritas.

En ocasiones conviene usar más de una de ellas para ofrecer una

descripción adecuada de la tendencia central en los datos.

La relación entre estas medidas depende de la asimetría de los datos.

Si la distribución de los datos es simétrica y en forma de campana,

moda, mediana y media serán iguales.

Pero si la distribución es asimétrica,

con un solo pico y una larga cola que se extiende en una dirección,

la media mu se desplaza en el sentido de la cola,

mientras que la mediana Md ocupa un lugar intermedio

entre la moda Mo y la media.

**Análisis exploratorio de datos univariados parte 2**

Guardar nota

[Análisis exploratorio de datos univariados parte 2 | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/sFmvE/analisis-exploratorio-de-datos-univariados-parte-2)Las medidas de tendencia central no ofrecen una

imagen completa de la distribución de un conjunto de mediciones.

Además de localizar el centro de la distribución,

es deseable contar con alguna medida de la

variabilidad o dispersión presente en los datos.

Consideraremos a continuación las definiciones de rango, rango intercuartil, varianza,

desviación estándar y coeficiente de variación,

como las medidas de dispersión más comunes de un conjunto de datos.

En todos los casos, valores altos de estas medidas

estarán asociados con una mayor dispersión en los datos.

El rango de un conjunto de mediciones,

la medida de variabilidad más simple,

pero quizá menos útil,

se define como la diferencia entre los valores máximo y mínimo del conjunto.

El rango es sensible a los valores extremos y

no permite discriminar entre conjuntos de datos con el mismo rango,

pero que presentan diferentes grados de concentración.

El rango intercuartil de un conjunto de mediciones,

ordenados de menor a mayor según su magnitud,

se define como la diferencia entre los cuartiles superior e inferior de la distribución,

también llamado los percentiles 75 y 25,

donde el percentil P es aquel valor en el conjunto que separa a su izquierda

a lo sumo el P por ciento de los datos y

a su derecha a lo sumo el 100 menos P por ciento de los datos.

La mediana es tanto el cuartil intermedio como el percentil 50 de la distribución.

El rango intercuartil ignora los valores extremos,

pero puede ser útil al momento de comparar la variabilidad de 2 o más conjuntos de datos.

La varianza de un conjunto de mediciones se define como la suma de

los cuadrados de las desviaciones de cada una de ellas,

con respecto a la media y barra del conjunto,

dividida luego por n menos 1,

suma de los cuadrados de yi menos y barra,

dividida por n menos 1.

La varianza permite comparar la variabilidad de 2 conjuntos de datos,

como también interpretar la variabilidad de un solo conjunto.

No obstante, es sensible a la escala de medición de los datos.

La desviación estándar s de un conjunto de

mediciones se define como la raíz cuadrada positiva de la varianza.

A diferencia de la varianza,

la desviación estándar se expresa en las mismas

unidades de cada una de las mediciones en el conjunto.

Tanto la varianza como la desviación estándar son

altamente sensibles a los valores extremos.

Las desviaciones de estos valores con respecto a la media dominan estas medidas,

lo que hace que en estas circunstancias ellas no

representen adecuadamente la dispersión presente en los datos.

Buscando interpretar la desviación estándar de un conjunto de mediciones,

la siguiente regla es útil: para aquellas

distribuciones en forma de campana es posible afirmar que el

intervalo y barra más o menos s

comprende aproximadamente el 68 por ciento de las mediciones en el conjunto,

el intervalo y barra más o menos 2s aproximadamente el 95 por ciento y el

intervalo y barra más o menos 3s aproximadamente el 99,7 por ciento.

Finalmente, el coeficiente de variación de un conjunto de mediciones,

una medida libre de unidades,

se define como el cociente entre la desviación estándar y

el valor absoluto de la media: CV es igual a Sigma dividida por el valor absoluto de Mu.

Es, en esencia, la desviación estándar expresada en múltiplos de la media,

aislando de esta forma el impacto de la escala sobre la medida de dispersión.

Por tal razón, permite comparar la variabilidad de 2

conjuntos de mediciones que tienen unidades distintas.

Un recurso gráfico que combina algunas de las

medidas anteriores es el denominado diagrama de caja.

Se construye dibujando una caja entre los cuartiles inferior Q\_1 y superior Q\_3,

que la atraviesa una línea en el lugar ocupado por la mediana Q\_2.

Luego, una nueva línea recta perpendicular a las anteriores conecta

la caja con el valor mínimo del conjunto y otra más con el valor máximo.

Una mirada rápida al diagrama permite reconocer las características de una distribución,

los valores extremos, la tendencia central mediana,

la dispersión, rango intercuartil y la asimetría.

A través del diagrama, también es posible comparar diferentes

conjuntos de datos en relación con las medidas anteriores,

su centro, dispersión y simetría.

**Análisis exploratorio de datos bivariados**

Guardar nota

[Análisis exploratorio de datos bivariados | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/lxnVn/analisis-exploratorio-de-datos-bivariados)

En la sesión anterior hemos hecho referencia

a los métodos gráficos y a los métodos de descripción numérica como

instrumentos apropiados para la síntesis de datos relacionados con una sola variable.

Frecuentemente más de una variable está involucrada en el estudio de los fenómenos de

interés y podríamos estar interesados

ahora en explorar la relación que pueda existir entre ellas.

La presente sesión se ocupará de las herramientas para la

síntesis de los datos disponibles sobre dos variables.

Consideremos primero el problema

de sintetizar los datos relativos a dos variables cualitativas.

Una opción consiste en construir clasificaciones cruzadas de los

datos para formar así lo que se denomina una tabla de contingencia.

Las filas de la tabla estarán asociadas con las categorías de la primera variable,

y las columnas con las categorías de la segunda variable.

En las celdas de la tabla se registra el número de observaciones n

y j que pertenecen simultáneamente

a cada categoría de la primera y cada categoría de la segunda variable.

Una forma de reconocer la existencia de alguna relación entre las dos variables,

consiste en examinar a lo largo de cada fila el porcentaje de las observaciones

que se clasifican en cada una de sus celdas

frente al total de su respectiva columna, n punto J.

Una diferencia marcada entre esos porcentajes sugiere la

presencia de una relación entre las dos variables.

Un ejercicio similar podría realizarse con el mismo

propósito a lo largo de cada columna de la tabla.

Reproduce el video desde :1:53 y sigue la transcripción1:53

Por su parte, la relación entre dos variables cuantitativas se puede sintetizar

a través de un recurso gráfico conocido como diagrama de dispersión.

Mientras que los ejes del diagrama están asociados con cada una de las variables,

cada punto en el diagrama corresponde a una observación

particular de la primera y el respectivo valor de la segunda variable.

La curva que se ha ajustado a las observaciones en

el diagrama representa una síntesis de la relación entre las dos variables.

Revela la forma general y el sentido de la

relación entre ellas y permite predecir el valor que

asumiría la segunda variable para un valor particular

de la primera no especificado en el conjunto de datos.

La proximidad de los puntos a la curva

muestra la fortaleza de la asociación entre las dos variables,

algo ciertamente difícil de precisar por medio de solo una gráfica.

Más adelante se desarrollarán algunos métodos para construir la

línea que proporcione el mejor ajuste a los datos disponibles.

Una medida numérica libre de unidades que describe con precisión

la fortaleza de la relación entre dos variables

cuantitativas es el coeficiente de correlación.

Este coeficiente, denotado por r,

mide el grado de asociación lineal y no otra,

que pueda existir entre dos variables cuantitativas

y se calcula como r es igual a la suma de xy menos x barra por y y

menos y barra dividida por la raíz cuadrada de la suma de los cuadrados

de xy menos x barra por la suma de los cuadrados de yi menos y barra.

Dados n pares de observaciones,

x y yi de las variables x y y.

Conviene hacer algunas precisiones sobre el coeficiente de correlación.

r toma valores entre menos 1 y 1.

Un valor de r próximo a más o menos 1

indica que los puntos del diagrama de dispersión están próximos a una línea recta.

Un valor positivo de r indica una asociación positiva entre las dos variables.

Cuando una de las dos variables crece,

la otra también crecerá.

Un valor de r cercano a 0 indica que no hay relación entre las dos variables,

o que entre ellas existe una relación, pero no lineal.

Un valor alto de r positivo o negativo no

permite concluir que una de las variables sea causa de la otra.

Otros factores podrían explicar esas variaciones.

Debido a que en el cálculo de r,

las dos variables se han estandarizado,

r es invariante ante cambios en la escala de cualquiera de las variables.

Idioma de la transcripción: Español

​

**Regresión lineal y correlación - parte 1**

Guardar nota

[Regresión lineal y correlación - parte 1 | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/Nefo6/regresion-lineal-y-correlacion-parte-1)

El análisis de regresión es una de las técnicas más ampliamente utilizadas en el

campo de la estadística.

A través de él se busca modelar la relación funcional que pueda existir entre

una variable de respuesta, la variable dependiente del modelo,

y una o más variables explicativas, las variables independientes.

Reproduce el video desde ::34 y sigue la transcripción0:34

El modelo de regresión puede satisfacer en la práctica cualquiera de dos propósitos.

Explicación, permite establecer cuáles de las variables explicativas tienen algún

efecto sobre la variable de respuesta.

Predicción, permite estimar el impacto sobre la variable de respuesta debido

a cambios especificados en las variables explicativas.

Reproduce el video desde ::55 y sigue la transcripción0:55

Para que tales propósitos adquieran sentido, debe existir algún tipo de

asociación entre la variable de respuesta y las variables explicativas.

Reproduce el video desde :1:5 y sigue la transcripción1:05

En la presente sesión, consideraremos el modelo de regresión lineal simple.

Por medio del cual,

la variable dependiente y se relaciona con una sola variable independiente x,

a través de una función lineal de la forma y = beta0 + beta1x + épsilon.

Donde beta0 es el término constante, el intercepto verdadero,

y representa el valor esperado de y cuando x = 0.

Beta1 es el coeficiente asociado con x, la pendiente verdadera,

y representa el cambio esperado en y debido a un incremento unitario en x.

Y épsilon es una variable aleatoria denominada el término de error,

que representa las desviaciones inevitables de las observaciones frente

a la línea de regresión.

Incorpora los efectos sobre la variable dependiente de todos aquellos factores

ignorados o desconocidos no incluidos en el modelo.

Reproduce el video desde :1:53 y sigue la transcripción1:53

Y, como función de épsilon,

es entonces una variable aleatoria que comprende dos partes.

Un componente estocástico, y uno no-estocástico.

Reproduce el video desde :2:3 y sigue la transcripción2:03

En efecto, los valores de la variable independiente generalmente se consideran

predeterminados, de tal manera que la única fuente de aleatoriedad en el modelo

es el término de error.

Reproduce el video desde :2:14 y sigue la transcripción2:14

La idea básica detrás del análisis de regresión consiste en producir

estimaciones de los parámetros desconocidos beta0 y beta1,

a partir de los datos disponibles, para satisfacer los propósitos you mencionados.

Reproduce el video desde :2:28 y sigue la transcripción2:28

El modelo estimado, denotado por, y-gorro = beta0-gorro + beta1-gorro x.

Representa el lugar geométrico de los valores estimados de la

variable dependiente, y su construcción será discutida más adelante.

Reproduce el video desde :2:44 y sigue la transcripción2:44

El primer supuesto que hace el modelo de regresión,

es que la relación entre las variables dependiente e independiente es lineal.

Reproduce el video desde :2:51 y sigue la transcripción2:51

De acuerdo con este supuesto,

la pendiente de la ecuación se mantiene invariante ante cambios en x.

Otros supuestos se refieren al término de error épsilon,

y aplicado a la i-ésima observación, y son los siguientes.

Los errores tienen media cero.

Reproduce el video desde :3:7 y sigue la transcripción3:07

Los errores tienen la misma varianza sigma cuadrado.

Reproduce el video desde :3:11 y sigue la transcripción3:11

Los errores son independientes entre sí.

Los errores tienen una distribución normal.

Reproduce el video desde :3:18 y sigue la transcripción3:18

Como consecuencia de lo anterior, los valores y(i) de la variable dependiente,

también tienen una distribución normal,

cuyos valores medios se localizan en la línea de regresión.

Y cuyas varianzas se mantienen constantes para cualquier valor de la variable

independiente.

Reproduce el video desde :3:34 y sigue la transcripción3:34

Antes de poner en práctica el modelo,

conviene verificar que todos los supuestos anteriores se satisfacen.

Un punto de partida es la construcción de un diagrama de dispersión,

con los pares de observaciones xi, yi.

Reproduce el video desde :3:48 y sigue la transcripción3:48

Examine si los puntos se localizan alrededor de una línea recta,

o si responden a algún otro patrón.

También, determine si hay observaciones atípicas que se apartan del patrón general

de los datos, y que deberían ser removidas.

Reproduce el video desde :4:2 y sigue la transcripción4:02

Muchas relaciones entre variables no son lineales.

Si así lo revela el diagrama de dispersión,

podría intentarse una transformación de cualquiera de las dos variables, a fin

de producir una relación aproximadamente lineal entre las variables transformadas.

Reproduce el video desde :4:18 y sigue la transcripción4:18

Entre las transformaciones más comunes se encuentran,

la raíz cuadrada del logaritmo natural, y el valor recíproco de una variable.

Como se ha mencionado, el intercepto beta0,

y la pendiente beta1 en el modelo de regresión,

son parámetros desconocidos que ahora deben estimarse a partir de los datos.

Se pretende encontrar aquella línea de regresión que produzca el mejor ajuste

a los datos, es decir, que haga mínimas las discrepancias entre cada

observación xi, yi, y la línea estimada.

El método de estimación más común se denomina el método de mínimos cuadrados.

Procedimiento que, al momento de estimar los parámetros beta0 y beta1,

busca minimizar la suma de los cuadrados de los errores en los que se incurriría al

estimar cada y(i) a través del modelo.

Suma de los cuadrados de y(i)- yi-gorro,

es igual a la suma de los cuadrados de yi- (beta0-gorro + beta1-gorro xi).

Reproduce el video desde :5:16 y sigue la transcripción5:16

Después de tomar las derivadas parciales de la anterior suma de cuadrados con

respecto a cada uno de los parámetros, e igualarlas a cero.

Y una vez verificadas las condiciones de segundo orden para un mínimo,

se obtienen los siguientes estimadores de beta0 y beta1.

Beta1-gorro = Sxy sobre Sxx.

Beta0-gorro = y-barra- beta1x-barra.

Donde Sxy es igual a la suma de, xi- x-barra x yi- y-barra.

Y Sxx es igual a suma de los cuadrados de, xi- x-barra.

[MUSIC]

**Regresión lineal y correlación - parte 2**

[Regresión lineal y correlación - parte 2 | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/JjZPx/regresion-lineal-y-correlacion-parte-2)

La precisión lograda en la estimación de los parámetros del modelo puede

establecerse a través de los errores estándar de los estimadores definidos con.

Sigma beta1-gorro = sigma épsilon / la raíz cuadrada de Sxx.

Sigma beta0-gorro = sigma épsilon x la raíz cuadrada de 1 sobre n +

x-barra al cuadrado sobre Sxx.

Donde sigma épsilon al cuadrado es la varianza del término de error,

que también es la varianza de y(i).

Parámetro que suele estimarse a través de la denominada varianza residual,

definida como S-épsilon al cuadrado = la suma de los cuadrados de y(i)-

y(i)-gorro / n- 2 = a la suma de cuadrados de los residuos / n- 2.

La diferencia yi- yi-gorro, que hace parte de la expresión anterior,

se conoce como el residuo asociado con la i-ésima observación,

y representa el error incurrido al estimar yi por medio de yi-gorro.

Reproduce el video desde :1:13 y sigue la transcripción1:13

A partir de los resultados anteriores es posible calcular un intervalo de confianza

para cada uno de los parámetros.

Sus límites se definen de tal manera que con una alta certeza comprendan el valor

desconocido de esos parámetros.

Mientras que el intervalo de confianza para beta0 no reviste un especial interés,

el intervalo de confianza para beta1 hará una contribución al momento de

caracterizar la relación entre las variables del modelo.

Y se calcula como, beta1-gorro- t1- alfa-medios x

S-épsilon x la raíz cuadrada de 1 sobre Sxx en el límite inferior.

Beta1-gorro + t1- alfa-medios x S-épsilon x la raíz cuadrada de

1 sobre Sxx en el límite superior.

Donde alfa es el nivel de riesgo asumido en la inferencia, y t es el cuantil 1

-alfa-medios de la distribución t, con (n -2) grados de libertad.

[MUSIC]

Para establecer la significancia de la variable independiente en el modelo

se contrastan las hipótesis H0:beta1 = 0 versus Ha:beta1 diferente de 0.

La prueba indica rechazar la hipótesis nula si el valor de la estadística

t en valor absoluto es mayor que t1- alfa-medios n -2.

Donde la estadística t = beta1-gorro- 0 /

S-épsilon x la raíz cuadrada de 1 sobre Sxx.

El rechazo de H0 nos lleva a concluir que x, la variable independiente, ciertamente

contribuye a explicar la variación observada en y, la variable dependiente.

Esto es, que el modelo univariado es significativo.

Llega el momento de establecer si el modelo lineal,

y no otro que incorpore, por ejemplo, algunas potencias de x,

constituye una representación apropiada para la relación que existe entre y y x.

Los gráficos son siempre un buen punto de partida para determinar una eventual

falta de ajuste del modelo a los datos.

Un diagrama de dispersión de y frente a x.

Y luego un gráfico de los residuos frente a los valores proyectados de y,

pueden revelar cualquiera de los siguientes problemas.

Reproduce el video desde :3:21 y sigue la transcripción3:21

La presencia de outliers, observaciones atípicas, o de observaciones erróneas.

Reproduce el video desde :3:27 y sigue la transcripción3:27

Alguna violación a los supuestos del modelo.

Linealidad entre y y x, independencia, normalidad,

y varianza constante del término de error.

Reproduce el video desde :3:38 y sigue la transcripción3:38

Mientras que la primera gráfica indica que ninguno de estos problemas está presente,

la segunda sugiere que un modelo polinomial de orden superior podría

resultar más apropiado.

Reproduce el video desde :3:50 y sigue la transcripción3:50

Por otra parte, una gráfica de los residuos frente a los valores

predeterminados de x, permite verificar el supuesto de varianza constante que se

impone sobre el término de error.

Reproduce el video desde :4:1 y sigue la transcripción4:01

La primera gráfica indica una varianza homogénea.

Y la segunda muestra una varianza creciente en el término de error a medida

que x aumenta.

Reproduce el video desde :4:11 y sigue la transcripción4:11

Más adelante estaremos en capacidad de verificar

los supuestos de independencia y normalidad a los que he hecho referencia.

Reproduce el video desde :4:19 y sigue la transcripción4:19

Tan pronto se ha logrado especificar el modelo estimado de regresión,

es necesario establecer su capacidad predictiva de los valores de y.

Reproduce el video desde :4:28 y sigue la transcripción4:28

Una forma de hacerlo consiste, en interpretar la desviación estándar

residual alcanzada en el contexto del problema.

Otra, quizá más general,

se basa en el concepto de correlación que desarrollaremos a continuación.

Reproduce el video desde :4:41 y sigue la transcripción4:41

Considere dos métodos de predicción de los valores de y, que se comparan a través de

la suma de los cuadrados de los errores de predicción en los que se incurre.

Reproduce el video desde :4:52 y sigue la transcripción4:52

El primero de ellos utiliza el modelo de regresión,

y su suma de cuadrados de los errores es.

Suma de cuadrados de los residuos igual a la suma de los cuadrados de y(i)-

y(i)-gorro.

Reproduce el video desde :5:4 y sigue la transcripción5:04

El segundo ignora el modelo,

y emplea como predictor la media aritmética de los valores de y.

Caso en el cual, la suma de cuadrados de los errores es.

Suma de cuadrados total igual a la suma de los cuadrados de y(i)- y-barra.

Reproduce el video desde :5:19 y sigue la transcripción5:19

En consecuencia, la reducción proporcional en el error que se logra al utilizar

como predictor el modelo de regresión en lugar de la media aritmética es,

ryx al cuadrado = la suma de cuadrados total- la suma de cuadrados de los

residuos / la suma de cuadrados total.

Reproduce el video desde :5:36 y sigue la transcripción5:36

Cociente que recibe el nombre de coeficiente de determinación,

y se denota por r cuadrado.

Este coeficiente tomará siempre valores en el intervalo (0, 1).

Valores de r cuadrado próximos a 1 indican que el modelo ha logrado un buen

ajuste a los datos disponibles.

Puede demostrarse que r cuadrado es, en efecto,

el cuadrado del coeficiente de correlación de x y y.

Reproduce el video desde :5:58 y sigue la transcripción5:58

De esta forma, una correlación de -0.3 indica que el modelo de regresión solo

logra un 9% de reducción en el error de predicción frente a la media aritmética.

La relación, suma de cuadrados total igual a la suma de cuadrados de

los residuos más la suma de cuadrados de la regresión,

conocida como la ecuación fundamental del análisis de varianza.

Permite interpretar la suma de cuadrados de la regresión como aquella parte de la

variación total en la variable dependiente,

que logra explicar el modelo de regresión.

Suma de cuadrados de la regresión = la suma de los cuadrados de yi-gorro-

y-barra, e igual a ryx al cuadrado x la suma de cuadrados total.

Reproduce el video desde :6:38 y sigue la transcripción6:38

Una vez establecida la significancia del modelo, podremos emplearlo para realizar

pronósticos sobre el valor que subirá la variable dependiente,

para cualquier valor predeterminado de x.

Suponga que x1 hasta xn representan los valores de la variable independiente

empleados en la construcción del modelo.

Reproduce el video desde :6:57 y sigue la transcripción6:57

Nuestro pronóstico sobre el valor de y para un nuevo valor de x, que denotaremos

por xn + 1, se obtiene a partir del modelo estimado de la forma siguiente.

Yn+1-gorro = beta0-gorro + beta1-gorro xn+1.

Y su error estándar está dado por, sigma épsilon x la raíz

cuadrada de 1 sobre n + el cuadrado de xn + 1- x-barra / Sxx.

El así denominado intervalo de predicción para yn+1, que define límites de confianza

para el valor asumido por esa variable, se calcula como.

Yn+1-gorro- t1- alfa-medios x S-épsilon

x la raíz cuadrada de 1 + 1 sobre n + el cuadrado

de xn+1- x-barra / Sxx en el límite inferior.

Yn+1-gorro + t1- alfa-medios x S-épsilon

x la raíz cuadrada de 1 + 1 sobre n + el cuadrado de

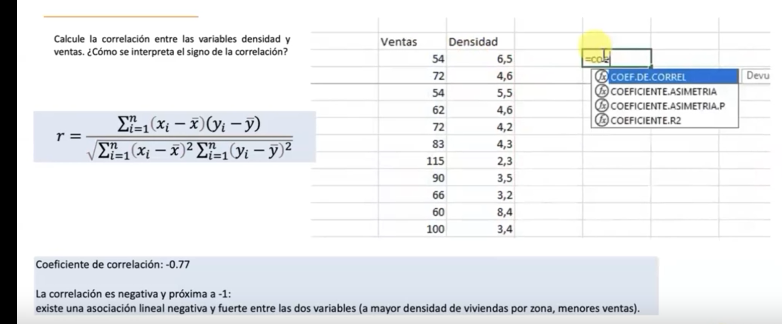
xn+1- x-barra / Sxx en el límite superior.

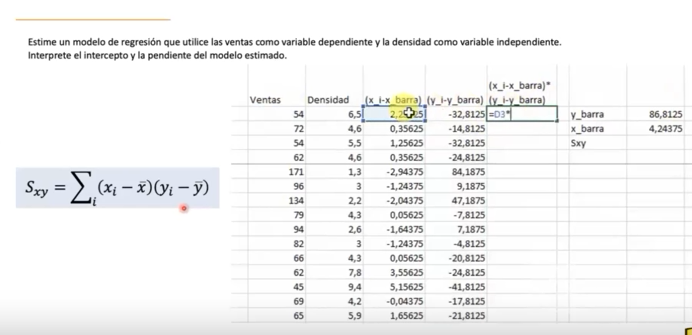
Donde t es el cuantil 1- alfa-medios de la distribución t,

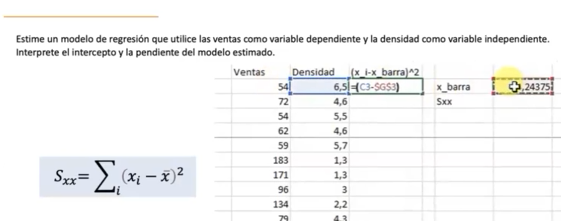
con (n -2) grados de libertad.

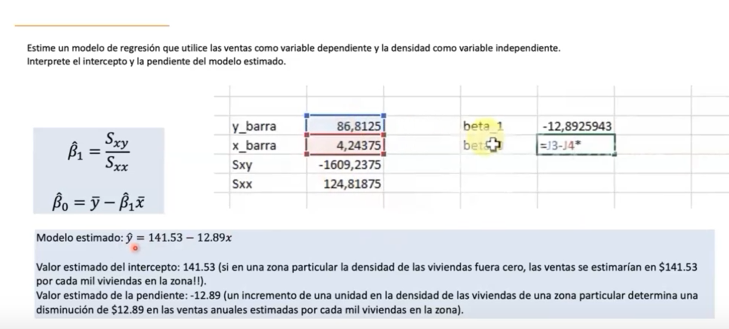
[MUSIC]

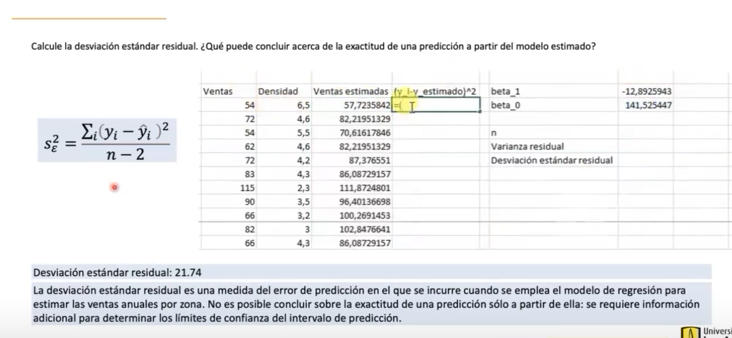
Sesión Sincrónica Semana 2

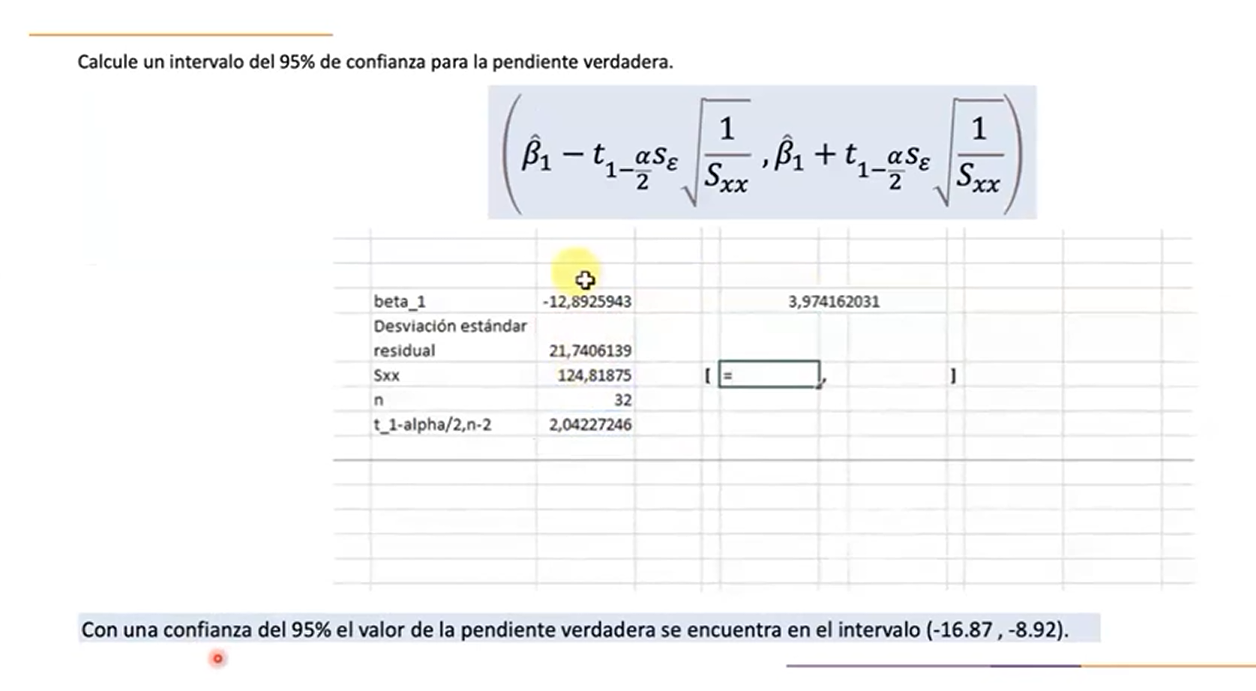




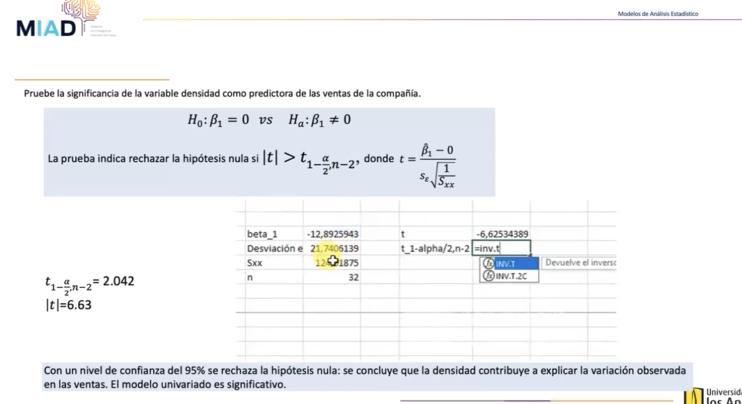


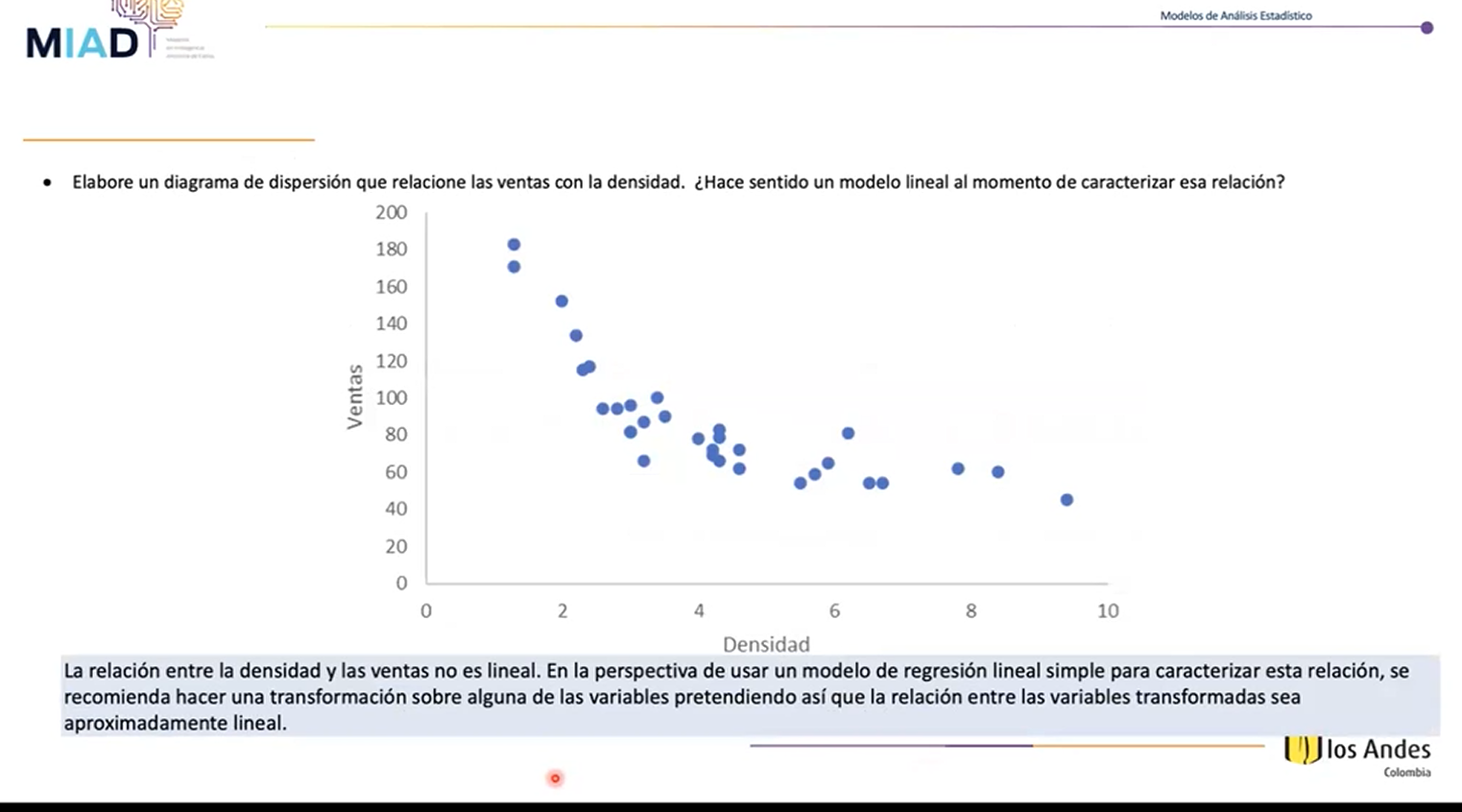


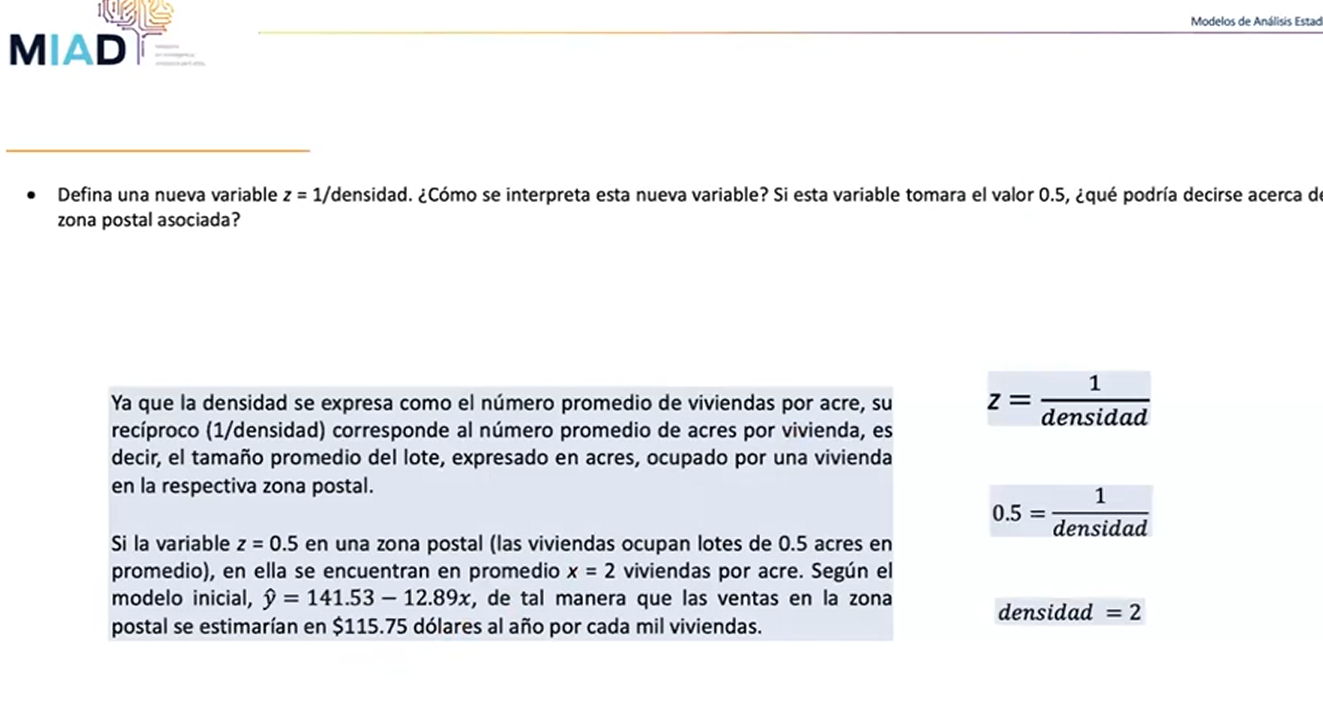


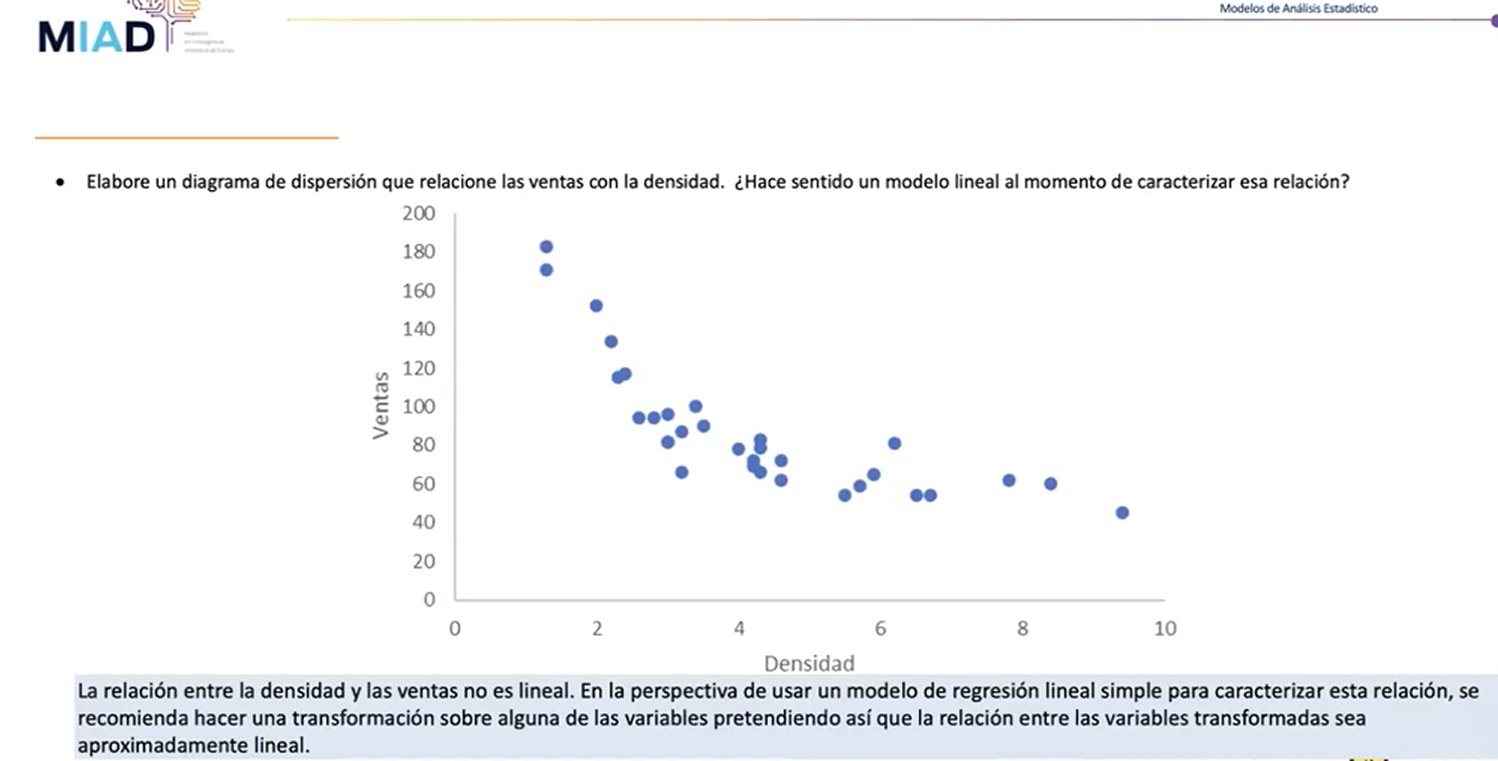


EN BUSCA DE ESTABLCER SI EL MODELO TIENE O NO UNA ALT CAPACIDAD PREDICTIVA SE EMPIEZA A DESARROLAR: Quiere decir cque con una confiana de 95% el valor de B1(Beta 1) s eencuentra dentro de los limite sindicados. (Esta explicación es del grafico de arriba)

Tiene mas sentido hacer este grafico antes de estimar mi modelo:







SEMANA 3

**Regresión múltiple y el modelo lineal general**

[Regresión múltiple y el modelo lineal general | Coursera](https://www.coursera.org/learn/modelos-de-analisis-estadistico/lecture/xaGd4/regresion-multiple-y-el-modelo-lineal-general)

En la sesión anterior se

presentó el modelo de regresión lineal simple que relaciona la

variable de respuesta con una sola variable explicativa cuantitativa.

Ahora el modelo de regresión se extiende,

permitiendo incorporar en él varias variables explicativas,

algunas de las cuales podrán ser categóricas.

El modelo de regresión lineal simple es muy restrictivo,

en la medida en que una sola variable explicativa puede

no dar cuenta de la variación observada en la variable dependiente.

A diferencia de él, el modelo de regresión múltiple relaciona la variable de

respuesta y con un conjunto de k variables explicativas cuantitativas,

a través de una función lineal de primer orden de la forma y igual a Beta\_0 más

Beta\_1 x\_1 más Beta\_2 x\_2 más Beta\_k x\_k más Épsilon.

Cada uno de los parámetros Beta

i para ir desde 1 hasta k representa el cambio esperado en y,

debido a un incremento unitario en x y cuando

se controlan los efectos de las demás variables independientes.

Debido a que el cambio es constante y no depende del

valor asumido por cualquiera otra variable independiente,

los efectos de estas variables sobre la variable de respuesta se denominan aditivos.

De lo contrario, se dice que las variables independientes interactúan entre sí.

En el modelo anterior, cualquiera de las k variables

independientes podrá ser una función de otras variables en el conjunto,

tales como la potencia de alguna de ellas, por ejemplo,

x\_3 igual x\_1 al cuadrado;

un producto cruzado, por ejemplo,

x\_4 igual x\_1 por x\_2,

que permite modelar interacciones entre variables;

o un término no lineal, por ejemplo,

x\_5 igual logaritmo de x\_1,

aumentando así la capacidad del modelo para explicar

y. Aplicado a cada una de las observaciones en la muestra,

el modelo de regresión múltiple hace los siguientes

supuestos: los errores tienen media 0,

los errores tienen la misma varianza Sigma al cuadrado,

los errores son independientes entre sí,

los errores tienen una distribución normal.

El problema de modelar una situación experimental

no puede limitarse al uso de variables explicativas cuantitativas.

El modelo lineal general, cuya estructura asemeja el modelo de regresión múltiple,

permitirá estudiar la relación entre la variable de respuesta y un conjunto de variables

explicativas cuantitativas o cualitativas.

La palabra lineal en el modelo general se refiere a la

forma como los coeficientes Beta se incorporan al modelo,

no a la forma como las variables explicativas aparecen en él.

El modelo es lineal en los Betas y no necesariamente lineal en las variables

explicativas. Las técnicas que desarrollamos

a continuación para hacer inferencias sobre los coeficientes

del modelo de regresión múltiple corresponderán igualmente al modelo lineal general.

Considera una muestra aleatoria de n observaciones de la

variable dependiente y sus valores asociados de las

k variables independientes que se relacionan a través del modelo de regresión lineal

múltiple de la manera siguiente: y igual a Beta\_0 más

Beta\_1 x\_i1 más Beta\_2 x\_i2 más Beta\_k x\_ik más Épsilon\_i,

para i igual 1,2 hasta n, n mayor que k,

donde x\_i1 hasta x\_ik son los valores predeterminados

de las variables independientes correspondientes a la observación y\_i.

Los estimadores de mínimos cuadrados de los parámetros

del modelo se obtienen siguiendo un procedimiento similar al ya ilustrado en

el modelo de regresión lineal simple y que en el presente caso nos lleva a resolver un

conjunto de ecuaciones simultáneas denominadas las ecuaciones normales del modelo.

En adición a los estimadores de los parámetros del modelo,

conviene contar con un estimador de la desviación estándar residual.

Como en el modelo de regresión lineal simple,

este estimador se obtiene tomando la raíz cuadrada del cociente

entre la suma de los cuadrados de los residuos y sus grados de libertad,

S\_Épsilon igual a la raíz cuadrada de la suma de cuadrados

de los residuos dividida por n menos (k más 1).

El coeficiente de determinación r cuadrado se define e

interpreta de la misma forma como en el modelo de regresión lineal simple.

Representa la proporción de la variación total observada en

la variable dependiente que logra explicar el modelo con k variables independientes.

R cuadrado igual a la suma de cuadrados total menos la suma de

cuadrados de los residuos dividida por la suma de cuadrados total.

Si todas las variables independientes no están correlacionadas entre sí,

r cuadrado es justamente la suma de los cuadrados de las correlaciones entre y y k de xi.

En caso contrario, y lo que es más usual,

se dice que estas variables presentan algún grado de multicolinealidad.

Cuando la correlación entre ellas es alta,

no resulta fácil establecer el valor predictivo de cada una

de las variables independientes en el modelo, como sería deseable.

Consideremos ahora la prueba de significancia global del modelo de regresión,

para lo cual se contrastan las hipótesis H\_0:

Beta\_1 igual a B\_2 igual a Beta\_K igual a 0;

H\_a: al menos uno de los Beta es diferente de 0.

Si la suma de cuadrados de la regresión es grande en

comparación con la suma de cuadrados de los residuos,

se concluye que las variables independientes en

su conjunto tienen algún valor predictivo.

En consecuencia, la hipótesis nula será

rechazada si el valor de la estadística F igual a la suma de

cuadrados de la regresión sobre k dividida por la suma de

cuadrados de los residuos sobre n menos paréntesis k más 1,

es mayor que el cuantil F\_1 menos Alfa.

La prueba anterior no permite identificar cuál o cuáles

de las variables independientes han contribuido a explicar y,

para lo cual se requiere el error estándar estimado de

cada uno de los estimadores de los coeficientes que para

Beta j gorro adquiere la forma S\_Beta j gorro

igual a S\_Épsilon por la raíz cuadrada de 1,

sobre la suma de los cuadrados de x\_ij menos x\_j barra por 1 menos R\_j al cuadrado,

donde x\_j barra es la media de las observaciones sobre la variable x\_j,

S\_Épsilon es la desviación estándar residual del modelo inicial

y R\_j al cuadrado es el coeficiente de

determinación de un modelo en el cual x\_j

se regresa sobre las demás variables independientes.

La fórmula anterior permite apreciar el efecto de la multicolinealidad sobre el modelo.

Si x\_j está altamente correlacionada

con una o más de las variables independientes restantes,

R\_j al cuadrado será próximo 1 y 1 menos R\_j al cuadrado cercano a 0,

de tal manera que el error estándar estimado de Beta\_j gorro

resultará grande y la estimación de Beta\_j altamente imprecisa.

El término 1 sobre 1 menos R\_j al cuadrado mide el incremento en la

varianza estimada de Beta\_j gorro debido a la presencia de multicolinealidad.

El intervalo de confianza para Beta\_j se construye a partir

del error estándar estimado de Beta\_j gorro de la

manera siguiente: Beta\_j gorro menos t\_1 menos

Alfa medios por S\_Beta j gorro en el límite inferior,

Beta\_j gorro más t\_1 menos Alfa medios por S Beta\_j gorro en el límite superior,

donde Alfa es de nuevo el nivel de riesgo asumido en la inferencia y t es el cuantil

1 menos Alfa medios de la distribución t

con n menos paréntesis k más 1 grados de libertad.

Para establecer la significancia individual de la variable x\_j en el modelo,

se contrastan las hipótesis: H\_0,

Beta\_j es igual a 0;

H\_a, Beta\_j es diferente de 0.

La prueba indica rechazar la hipótesis nula si

el valor de la estadística t en valor absoluto

es mayor que el cuantil t\_1 menos Alfa medios n menos paréntesis k más 1,

donde la estadística t es igual a Beta\_j gorro sobre S\_Beta j gorro.

El rechazo de H\_0 nos lleva a concluir que x j

tiene un valor predictivo único sobre la variable dependiente,

en adición a las contribuciones que en ese sentido

puedan hacer las demás variables independientes.

Finalmente, el pronóstico sobre el valor que asumirá la variable dependiente

para cualquier conjunto de valores predeterminados de x1 hasta xk se obtiene,

a semejanza del procedimiento descrito para una regresión simple,

sustituyendo estos últimos valores en el modelo estimado de regresión.

La semiamplitud del intervalo de predicción se calcula de la manera usual,

como el producto entre el factor t apropiado y el error estándar del pronóstico,

donde t es ahora el cuantil 1 menos Alfa medios de la distribución t con n

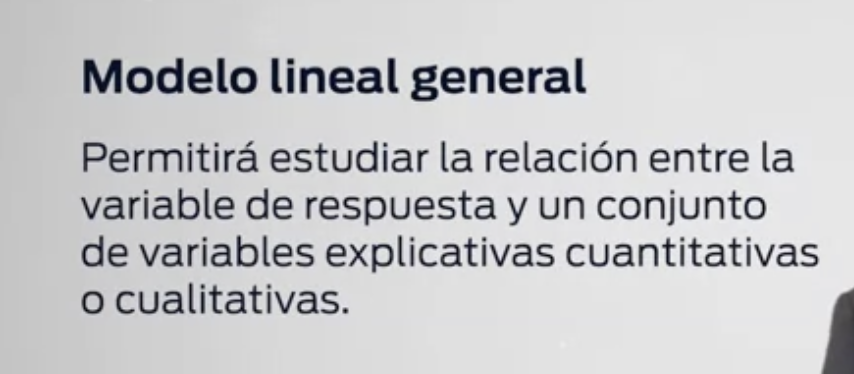
menos paréntesis k más 1 grados de libertad.

Supuestos del modelo de regresión lineal multiple:

A white paper with black text

Description automatically generated

El modelo lineal general es similar al modelo modelo lineal multiple. En el modelo lineal general, la palabra lineal no se refiere a que las variables esplicativas tenganque ser lienales, la palabra lienal hace referencia a como los betas aparecen en el modelo. Por lo tnato el modelo es lineal en los betas y no necesariamente lineal en las variables explicativas.



A text on a white background

Description automatically generated