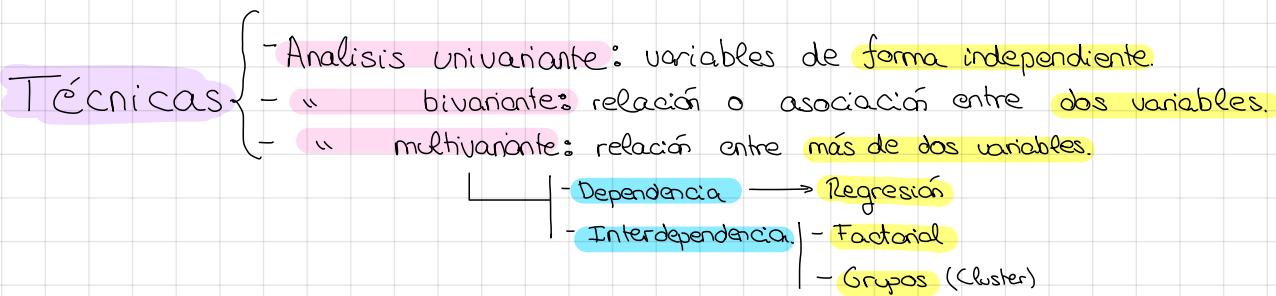


Tema 1: Análisis de dependencia.



MODELO LINEAL

- Y: dependiente / regresando
- X: independiente / regresor.
- E: perturbación aleatoria o componente estocástica. { No la puedo controlar.

Relación verdadera }
$$Y = \beta_1 + \beta_2 X + E$$

Recta estimada }
$$\hat{Y} = b_1 + b_2 X$$

Residuos $\rightarrow e = Y - \hat{Y}$

↑ ↑ ↑ Redes Estimados

MODELO LINEAL SIMPLE

$$y_{nxi} = X_{nxi} \beta_{2xi} + e_{nxi}$$

MODELO LINEAL GENERAL

$$y_{nxi} = X_{nxi} \beta_{kxi} + e_{nxi}$$

ESTIMACIÓN POR MÍNIMOS CUADRADOS. \rightarrow Calculos para hallar la mejor estimación de $\hat{\beta}$ para el cual E es el mínimo posible.

DEMOSTRACIÓN EXAMEN

$$\begin{aligned} \sum e_i^2 &= e'e = (y - \hat{y})'(y - \hat{y}) = (y - x\hat{\beta})'(y - x\hat{\beta}) = (y' - \hat{\beta}'x')(y - x\hat{\beta}) = \\ &= y'y - y' \cdot (x\hat{\beta}) - (\hat{\beta}'x') \cdot y + (\hat{\beta}'x') \cdot (x\hat{\beta}) \\ &\quad \underbrace{- 2\hat{\beta}'x'y}_{\text{(porque va a quedar una matriz del tipo } 1 \times 1\text{)}} \end{aligned}$$

poner los términos tener sentido en el orden.

$$\left\{ \begin{array}{l} y = x\beta + e \\ \hat{y} = x\hat{\beta} \end{array} \right.$$

$= y'y - 2\hat{\beta}'x'y + \hat{\beta}'x' \cdot x\hat{\beta} \rightarrow$ Queremos buscar un min. Calcularemos las derivadas.

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}} = 0 - 2x'y + 2x'x\hat{\beta} = 0$$

Despejamos: $2x'x\hat{\beta} = 2x'y \rightarrow (x'x)\hat{\beta} = x'y$

ESTIMADOR (OLS) MCO

$$\hat{\beta} = (x'x)^{-1} x'y$$

Tema 2

PROPIEDADES DE LOS ESTIMADORES.

ESTIMADORES INSENGADOS



DEMOSTRACIÓN EXAMEN

$$\begin{aligned} E(\hat{\beta}) &= E((x'x)^{-1} \cdot x'y) = E\left((x'x)^{-1} x' (x\beta + \varepsilon)\right) = \\ &= E\left(\beta + (x'x)^{-1} x' \cdot \varepsilon\right) = \beta + E((x'x)^{-1} x' \cdot \varepsilon) = \beta + (x'x)^{-1} x' \cdot 0 = \boxed{\beta} \end{aligned}$$

1 YA NO SE PONE LA E
cte El valor esperado de las perturbaciones es 0.

Todo esto se tiene que multiplicar en cada miembro

Propiedades bajo normalidad

* Si no hay normalidad, no se puede aplicar inferencia estadística.

- Estimadores de β siguen una ley $\rightarrow N(\beta, \sigma^2 (x'x)^{-1})$
- Estimadores de y siguen una ley $\rightarrow N(y, \sigma^2 H)$
- Residuos de la regresión OLS siguen una ley $\rightarrow N(0, \sigma^2 M)$

En caso de querer hacer un cor pero hay una variable categórica y no te lo permite:

datos <- select (basedatos, columnas, que, quiero)

Si se quiere hacer una base de datos nueva:

datos <- data.frame (columnas, que, queremos)

Tema 3, 4 Suposiciones del modelo de regresión lineal.

Se debe cumplir:

DIAGNOSIS

FUNCIONALIDAD

- Linealidad: (Además del ggplot, se debe hacer el gráfico de los residuos frente a los valores estimados.)

SOLUCIONES A LA NO LINEALIDAD

- Transformar las variables (logaritmos)
- Introducir potencias en las variables indepes (x^2, x^3, x^4)
- Eliminar datos atípicos.
- Introducir variables adicionales.

PARÁMETROS

- En todas las observaciones los valores deben ser iguales. (ctes)

LOS ERRORES

$$\varepsilon \rightarrow N(0, \sigma^2 I)$$

- Si se observan los errores:
- Esperanza nula: El valor esperado es 0 $E(\varepsilon) = 0$
 - Homocedasticidad: Varianza de los errores es cte. $\text{Var}(\varepsilon_i) = \sigma^2$ (Breuch Pagan)
 - Independencia: perturbaciones indepes $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ (Durbin Watson) (NO AUTOCORRELACIÓN)
 - Normalidad: siguen normalidad los errores (qq-plot y pp-plot) ó (Shapiro-wilks)

REGRESORES (x: indepe)

- Exogeneidad: son fijos
- Mensurabilidad: se miden sin error.
- No colinealidad: No son dependientes entre ellos.
 - Si hay varias x hay que comprobar la multicolinealidad (Si hay relación entre las indepes (corplot))
- car::vif (modelo 1)
Genera mucha dispersión en los β estimados.
- Identificabilidad: El número de x es menor o igual que observaciones.



→ Cosas que hay que comprobar en los test de R

Tema 5

Variables ficticias.

Hay que hacer un excel en el ge:

- Dummy Aditiva (DA):

Se le asocia 1 a una variable y 0 a otra.

- Dummy Multiplicativa (DM):

(DA · X) se multiplican los valores de DA por la X.

ANÁLISIS ESTADÍSTICO

- Estudio de la particularidad de una observación.

- Valores atípicos: 1 en el punto y 0 en el resto

- Cambio estructural:

- Comprobar si hay cambios en un periodo de tiempo.

- Estacionalidad:

- Para cada estación.

- Variables categóricas - k:

Si hay K valores, hay que construir K DA.

Si hay 2 valores sólo necesitamos una DA.

ESTIMACIÓN POR PASOS

Forward: selecciona hacia delante aquella variable que mejore el modelo.

Backward: Se inicia con todas las variables disponibles. Se prueba a eliminar una variable y si mejora el modelo queda excluida.

Doble o mixto (both): Mix de las otras dos.

CRITERIO DE INFORMACIÓN DE AKAIKE

Cuanto menor sea el AIC mejor será el modelo.

Realización Regresión Lineal DEL MODELO

① Comprobar si hay relación lineal (la cual queremos estimar)

```

1 ## Caso Relación entre las ventas y la publicidad
2 Ventas <- ventas_publicidad$Ventas
3 Publicidad <- ventas_publicidad$Publicidad
4
5 ## Gráfico de dispersión
6 plot(Publicidad, Ventas)
7 ggplot(data = ventas_publicidad,
8         mapping = aes(x=Publicidad, y=Ventas)) + geom_point()
9
0 ## Regresión
1 regresion <- lm(Ventas ~ Publicidad, data = ventas_publicidad)
2 summary(regresion)
3 anova(regresion)
4
5 plot(Publicidad, Ventas)
6 abline(regresion)
    
```

1
son lo mismo pero
el segundo tiene α

2 Para estimarlo.

Gráfico de dispersión
con la recta estimada.

Comentar :
si se ajusta o no a una recta.
si tienen mucha dispersión.

Poner las
variables que
queremos
comprobar en
linealidad.

Depende de
si tengo
más o menos
de variables
independientes

② Estimamos el modelo lineal

Dep ~ Ind

Relación estimada:

$$\hat{V} = \beta_0 + \beta_1 \text{Pub}$$

↑ Tratar de que β_1 siempre
entre en los modelos

Con el modelo de
estimación explícita
en 87% de las ventas.

Cuento más próximo al
1 mejor es el modelo.

```

> regresion <- lm(Ventas ~ Publicidad, data = ventas_publicidad)
> summary(regresion)

Call:
lm(formula = Ventas ~ Publicidad, data = ventas_publicidad)

Residuals:
    Min      1Q  Median      3Q     Max 
-571.27 -102.69   83.69  153.92  286.66 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4012.7378   83.2563  48.20 <2e-16 ***
Publicidad   1.6558    0.1061  15.61 <2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 222.1 on 34 degrees of freedom
Multiple R-squared:  0.8776, Adjusted R-squared:  0.874 
F-statistic: 242.7 on 1 and 34 DF, p-value: < 2.2e-16

> anova(regresion)
Analysis of Variance Table

Response: Ventas
          Df Sum Sq Mean Sq F Value Pr(>F)    
Publicidad  1 12024435 12024435 43.69 < 2.2e-16 ***
Residuals  34 1673368  49338.5
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Por cada euro que
gastas en publicidad,
ganas 1,65 en ventas.

$$\hat{V} = 4012 + 1,65 \cdot \text{Pub}$$

ventas
independientes
de la
publicidad.

3 (Intercept) = cte $\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$

Publicidad = $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

Coincide con el t-test porque
solo hay una variable indep.,
si hubiera más de una no
coincidiría.

(como caer en la región de rechazo
me quedo con el valor estimado)

cae en la región de
rechazo.

2 Para comparar varios β
 H_0 : Todos los coeficientes
menores a la cte
son iguales a 0 (β_0 y β_1)
cte

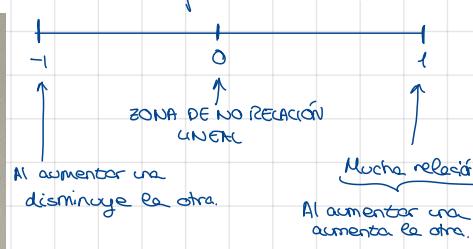
$H_0: \beta_1 = 0$ (Porque β_1 es la cte)
 $H_1: \beta_1 \neq 0$

Para comparar los coefs de correlación.
(los fijamos en la columna dependiente)

Correlaciones (el gráfico requiere llamar al paquete corrplot)

round(cor(medios_comunicación, method = "pearson"), 3)

corrplot(cor(medios_comunicación, method = "pearson"), type = "lower")



MEDIDAS DE BONPAD DE AJUSTE

VARIANZA RESIDUAL

- Un modelo será bueno cuando la dispersión de los residuos es pequeña

Desviación de los residuos.

1 Residual standard error: 222.1 on 34 degrees of freedom
Multiple R-squared: 0.8776, Adjusted R-squared: 0.874
F-statistic: 243.7 on 1 and 34 DF, p-value: < 2.2e-16

cuando hay más de un modelo para ver con cual nos quedamos

Comparar modelos con distintos números de variables.

Nos indica si esa nueva variable nos interesa meterla o no.

Utilizamos el \bar{R}^2 ajustado porque

{ - prox a 0 : explica poco de la variable dependiente
- prox a 1 : explica mucho de la variable dependiente