Universidad Técnica Federico Santa María
Departamento de Informática
Análisis Inteligente de Datos

Capítulo 2:
Métodos Estadísticos
II semestre de 2010

Héctor Allende: hallende@inf.utfsm.cl

Modelos Regresión (Metamodelo)

- El análisis de regresión consiste en determinar cómo una variable *y*, está relacionada (asociada , correlacionada o ligada) con una o más variables explicativas $x_1, \ldots, x_n$

y	Respuesta Variable dependiente Salida
x_i	Regresores Variables explicativas Variable independiente Entrada

Profesor H.Allende 2

Modelo Regresión lineal Simple

Modelo de Regresión ≡ Modelo Explicativo Estático

$$y_{ij} = \beta_0 + \beta_1 x_i + u_{ij} \quad \text{(Hipótesis Estructural)}$$

Supuestos básicos

y_{ij}, u_{ij} : variables aleatorias dependiente ; β_0, β_1: Parámetros y
x_i : variable explicativa deterministica (estocástica).

Supuestos distribucionales usuales.

1. E $[u_{ij}]$=0.
2. Var $[u_{ij}]=\sigma^2$, Cte ; **Perturbación independiente de *x*.**
3. $u_{ij} \sim N(0,\sigma^2)$.
4. E $[u_{ij}\, u_{kh}]=0$, $\forall$ (i,j)≠(k,h) $\Rightarrow$

Profesor H.Allende 3

Modelos Regresión Simple

1- E[y_{ij}/x_i]= $\beta_0 + \beta_1 x_i$ **2- Var** $[y_{ij}/x_i]=\sigma^2$.
***3 - f (y_{ij}/x_i)* es normal** **4- Las obs. son independientes entre si**

1.2 Estimación de parámetros.

$$y_{ij} = \beta_0 + \beta_1 x_i + u_{ij} \Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Leftrightarrow E[y_i / x_i] = \beta_0 + \beta_1 x_i$$

Residuo (e_{ij})= Valor observado (y_{ij})– Valor previsto $(\hat{y}_i)$.

1.2.1 Método de Máxima Verosimilitud. (función de Verosimilitud).

$$\ell(\beta_0, \beta_1, \sigma^2, y_{ij}) = \frac{1}{\sqrt{2\pi}\,\sigma} Exp\left[-\frac{1}{2\sigma^2}(y_{ij} - \beta_0 - \beta_1 x_i)^2 \right]$$

$$\Rightarrow L(\beta_0, \beta_1, \sigma^2) = \log\left\{ \prod_{i=1}^{n} \ell(\beta_0, \beta_1, \sigma^2, y_{ij}) \right\}$$

Profesor H.Allende 4

Modelos Regresión Simple

Derivando $L(\,)$ con respecto a los parámetros :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{Cov(x,y)}{S_x^2} \qquad \text{y} \qquad \hat{\sigma}^2 = \frac{\sum\sum e_{ij}^2}{n}$$

Métodos: Mínimos Cuadrados y Máxima verosimilitud.

$$Maximizar\,L(\beta_0, \beta_1, \sigma^2) \Leftrightarrow \underset{(\beta_0,\beta_1)}{Min} \sum\sum (y_{ij} - \beta_0 - \beta_1 x_i)^2$$

$$M = \sum\sum (y_{ij} - \beta_0 - \beta_1 x_i)^2 = \sum\sum e_{ij}^2$$

$$\Rightarrow \hat{S}_R^2 = \frac{\sum\sum e_{ij}^2}{n-2}$$

Profesor H.Allende 5

Distribución de los Parámetros

$\hat{\beta}_0$ y $\hat{\beta}_1$ son vari ables aleatorias

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \approx N(\beta_0, V(\hat{\beta}_0))$$

$$\hat{\beta}_1 = \frac{Cov(x,y)}{S_x^2} \approx N(\beta_1, V(\hat{\beta}_1))$$

Profesor H.Allende 6

Propiedades de los estimadores.

Propiedades de la ley de probabilidades de $\hat{\beta}_1$

$$E[\hat{\beta}_1] = \beta_1$$

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{nS_x^2}$$

$$\hat{\beta}_1 \approx N\left(\beta_1 \; ; \; \frac{\sigma^2}{nS_x^2}\right)$$

Intervalos de confianza

$$IC_{\gamma=(1-\alpha)} = \left[\hat{\beta}_1 \pm t_{1-\alpha/2} \frac{\hat{S}_R}{\sqrt{n}S_x}\right]$$

Propiedades de los estimadores.

Propiedades de la ley de probabilidades de $\hat{\beta}_0$

$$E[\hat{\beta}_0] = \beta_0$$

$$Var[\hat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)$$

$$\hat{\beta}_0 \approx N\left(\beta_0 \; ; \; \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{nS_R^2}\right)\right)$$

Intervalos de confianza

$$IC_{\gamma=(1-\alpha)} = \left[\hat{\beta}_0 \pm t_{1-\alpha/2} \frac{\hat{S}_R}{\sqrt{n}} \sqrt{1 + \frac{\bar{x}^2}{S_x^2}}\right]$$

Propiedades de los estimadores.

Propiedades de la ley de probabilidades $\hat{S}_R^2$

$$E[\hat{S}_R^2] = \sigma^2$$

$$Var[\hat{S}_R^2] = \frac{2\sigma^4}{n-2}$$

$$\frac{\sum \sum e_{ij}^2}{\sigma^2} \approx \chi^2_{(n-2)}$$

Intervalos de confianza para σ^2

$$IC(\sigma^2)_{\gamma=(1-\alpha)} = \left[\frac{(n-2)\hat{S}_R^2}{\chi^2_{1-\alpha/2}} \; ; \; \frac{(n-2)\hat{S}_R^2}{\chi^2_{\alpha/2}}\right]$$

Contraste de regresión.

Prueba de hipótesis. $H_0 : \beta_1 = 0$ v/s $H_1 : \beta_1 \neq 0$

Sea Variación Total $= VT = \sum_i \sum_j \left(y_{ij} - \bar{y}\right)^2$

Variación no Explicada $= VNE = \sum \sum \left(y_{ij} - \hat{y}_i\right)^2$

Variación Explicada $= VE = \sum n_i \left(\hat{y}_i - \bar{y}\right)^2$

Estadístico

$$F_0 = \frac{\hat{\beta}_1^2 nS_x^2}{\hat{S}_R^2} = \frac{VE}{\hat{S}_R^2} \approx F_{(1;n-2)}$$

Para la región crítica P($F_{(1,n-2)}$ ≤ C)=1-α, se rechaza H_0 para F_0>C.

Predicción de las medias condicionales.

$$\hat{y}_h = \bar{y} + \hat{\beta}_1(x_h - \bar{x})$$

$$\Rightarrow E[\hat{y}_h] = \beta_0 + \beta_1 x_h = m_h \quad \left(\hat{E}[y/x_h]\right)$$

$$Var[\hat{y}_h] = \left\{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum n_i (x_i - \bar{x})^2}\right\}\sigma^2 = v_{hh}\sigma^2$$

Intervalo de confianza para las medias. $IC_\gamma(m_h) = \left[\hat{y}_h \pm t_{\alpha/2}\hat{S}_R\sqrt{v_{hh}}\right]$

Se desea prever el valor de y para $x = x_h$. Intuitivamente

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

Corresponde a la Estimaciones de las medias condicionales.

Criterio de predicción. Error cuadrático medio mínimo

$$E\left[(y - \tilde{y}/x)^2\right] = Var(y/x) + \left(E[y/x] - \tilde{y}\right)^2$$

Intervalo de Confianza

Intervalo de Confianza para las observaciones $IC_\gamma(y_h) = [\hat{y}_h \pm t_{\alpha/2}\hat{S}_R\sqrt{1 + v_{hh}}]$

Coeficiente de Correlación.

Coeficiente de determinación: $R^2 = \frac{VE}{VT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

Coeficiente de correlación: r $\quad \sqrt{R^2} = r = \frac{Cov(x, y)}{S_x S_y}$

Relación entre $\hat{S}_R^2$ y r : $\quad r = \left\{1 - \frac{(n-2)\hat{S}_R^2}{nS_y^2}\right\}^{1/2}$

Contraste de linealidad.

Hipótesis.

$$H_0 : E[y_{ij}/x_i] = \beta_0 + \beta_1 x_i \quad \text{v/s} \quad H_1 : E[y_{ij}/x_i] \neq \beta_0 + \beta_1 x_i$$

$$\text{Varianza entre las medias y las rectas: } \hat{S}_{12}^2 = \frac{\sum n_i (\bar{y}_i - \hat{y}_i)^2}{d-2}$$

$$\text{Varianza de la perturbación sin la linealidad } \hat{S}_2^2 = \frac{\sum\sum (y_{ij} - \bar{y}_i)^2}{n-d}$$

Estadístico de Prueba: $$F_0 = \frac{\hat{S}_{12}^2}{\hat{S}_2^2} \approx F_{(d-2,n-d)}$$

Región crítica P ($F_{(d-2,n-d)} \leq C$)=1-α, se rechaza H_0 para F_0>C.

Profesor H.Allende

13

Análisis de residuos.

Tiene por objeto contrastar a posteriori las hipótesis de linealidad del modelo. Es especialmente importante cuando se tiene un solo valor de la variable y para cada valor de la variable de control x.

El análisis de los residuos se utiliza para verificar:

Si su distribución es aproximadamente normal.
Si su variabilidad es constante y no depende de x.
Si presentan evidencia de una relación no lineal.
Si existen observaciones atípicas o hetereogéneas.

Profesor H.Allende

14

Regresión Lineal

- Modelo Lineal General:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + u_j \qquad j = 1..n$$

$$u_j \sim N(0, \sigma^2) \quad \text{i.i.d}$$

- Valor esperado:

$$E[y_j / x_1, x_2..x_k] = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ij}$$

Profesor H.Allende

15

REGRESION GENERAL

La variable de respuesta "y" depende de muchas variables $x_1, x_2, \ldots, x_n$, aunque algunas de estas son no observables.
El modelo de regresión pretende develar efecto de las variables explicativas más importantes y representa las restantes mediante una v.a. la perturbación.
Es decir: $y = f(x_1, x_2, \ldots, x_k) + \underbrace{g(x_{k+1}, \ldots, x_\infty)}_{\mu}$

Suponga que en el rango de interés, la función f admite una aproximación lineal: $f(x_1, x_2, \ldots, x_k) = \sum_{j=0}^{k} \beta_j x_j$, donde $x_0 = 1$.

En tal caso $y = \sum_{j=0}^{k} \beta_j x_j + \mu$.

Ejemplo: Modelo para predecir la valor de las propiedades en Viña del Mar de sus características físicas , geográficas etc.

Profesor H.Allende

16

REGRESION GENERAL

Se hacen las siguientes hipótesis sobre la distribución de las variables:
-Para cada conjunto fijo de las x, la distribución de y es normal

$$E[y / x_1, \ldots, x_k] = \sum_{j=0}^{k} \beta_j x_j \qquad Var[y/x_1, \ldots, x_k] = \sigma^2 = cte.$$

Las variables y_i son independientes entre si.
-El nº de variables explicativas es menor que el nº de observaciones.
-Las x's son realmente distintas y no existen entre ellas relaciones lineales exactas.

Luego $y_i = \beta_0 + \sum_{j=0}^{k} \beta_j x_{ji} + \mu_i .$

Donde cada coeficiente β_j mide el efecto marginal sobre la respuesta de un aumento unitario en x_j.

μ_i: perturbación aleatoria ; $\mu_i \sim N[0, \sigma^2]$, $\forall$ i=1,...,n.
Var [μ_i]= σ^2 = cte, $\forall$ i=1,...,n. ; E[$\mu_i \mu_j$]=0, si i$\neq$j

Profesor H.Allende

17

Estimación de Parámetros

Sea $y_i = \sum_{j=0}^{k} \beta_j x_{ji} + \mu_i$ j=1,...n ; $x_0 = 1$ y sea $Q = \sum_{i=1}^{n} (y_i - \sum_{j=0}^{k} \beta_j x_{ji})^2$

Bajo el supuesto de normalidad de la variable aleatoria y se sabe que

$$\underset{\underline{\beta}}{Min}\ Q \Leftrightarrow \underset{\underline{\beta}}{Max}\ \ln \prod_{i=1}^{n} f(y_i, \underline{\beta}, \sigma^2)$$

Derivando con respecto a β_0 y a β_j, se obtiene las siguientes ecuaciones

notación matricial: $X^t Y = X^t X \underline{\hat{\beta}}$

$$\text{donde } Y^t = (y_1, \ldots, y_n) \wedge X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}$$

Como por hipótesis X'X no es singular se tiene que $\hat{\beta} = (X^t X)^{-1} X^t Y$

Profesor H.Allende

18

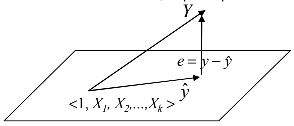
Interpretación geométrica.

Considere los vectores de R^n ; $1, X_1, X_2,...,X_k$ que forman las columnas de la matriz de diseño X. El objetivo de la estimación es determinar $\hat{y}$, como CL de X

$$\hat{y} = \beta_0 1 + \beta_1 X_1 + \ldots + \beta_k X_k$$

$\hat{y}$ está contenido en el subespacio generado por los vectores $<1, X_1, X_2,...,X_k>$

El criterio de mínimos cuadrados, impone que el norma del vector $\|e\| = \|y - \hat{y}\|$ sea mínima.



Del teorema de proyección se tiene que: $\hat{e}$ es $\perp$ $1, X_1, \ldots, X_k$ e $\hat{y}$

Es decir $1^t e = X_1{}^t e = \ldots = X_k{}^t e = \hat{y}^t e = 0$

$$X^t e = X^t\left(Y - \hat{Y}\right) = X^t\left(Y - X\hat{\beta}\right) = 0 \Rightarrow X^t Y = X^t X \underline{\hat{\beta}}$$

Profesor H.Allende

19

Interpretación geométrica.

Por lo tanto

$$\hat{Y} = \hat{Y} + e$$

$$\hat{Y} = X\underline{\hat{\beta}} = \underbrace{X\left(X^t X\right)^{-1} X^t}_{V} Y$$

$$\hat{Y} = V \mathrm{Y}$$

Siendo V la matriz de proyeción (simétrica e idempotente).

$V^t = V$ y $V^2 = V$

Esta matriz juega un rol importante en la etapa de diagnóstico.

$$e = Y - \hat{Y} = Y - V\mathrm{Y} = \left(I - V\right)\mathrm{Y}$$

Profesor H.Allende

20

EJEMPLO 1

Ejemplo:

Los siguientes datos muestran el indicador global de desarrollo regional y, en términos del número de automóviles por mil habitantes (x_1) y el número de teléfonos por mil habitantes (x_2) en ocho de las 13 regiones del país.

Profesor H.Allende

21

Región	Indicador y	Automóviles x_1	Teléfonos x_2
I	64	58	111
II	78	84	131
III	83	78	158
IV	88	81	147
V	89	82	121
VI	99	102	165
VII	101	85	174
VII	102	102	169

$$\underline{Y} = \beta_0 1 + \beta_1 \underline{X}_1 + \beta_2 \underline{X}_2 + \underline{u} \qquad \underline{Y} = X\underline{\beta} + \underline{u} \qquad \text{con} \qquad X = \begin{pmatrix} 1 & 58 & 111 \\ 1 & 84 & 131 \\ 1 & 78 & 158 \\ 1 & 81 & 147 \\ 1 & 82 & 121 \\ 1 & 102 & 165 \\ 1 & 85 & 174 \\ 1 & 102 & 169 \end{pmatrix}$$

$$X^t X = \begin{pmatrix} 8 & 672 & 1176 \\ 672 & 57822 & 100453 \\ 1176 & 100453 & 176785 \end{pmatrix} \qquad X^t Y = \begin{pmatrix} 704 \\ 60251 \\ 105288 \end{pmatrix}$$

Resolviendo la ecuación matricial $X^t Y = X^t X \underline{\beta}$ se obtiene:

$$\underline{\beta}^t = (9{,}05\ ;\ \ 0{,}52\ ;\ \ 0{,}24) = (\beta_0\ ,\ \ \beta_1\ ,\ \ \beta_2)$$

Profesor H.Allende

22

Conclusiones

Conclusiones.

1. Cualquier coeficiente de regresión estimado $\hat{\beta}_i$; puede interpretarse como la pendiente de la recta de regresión de los residuos de una regresión y respecto a todas las otras variables (parte de y no explicada por el resto de las x) con la contribución diferencial de x_i.
2. El coeficiente de regresión $\hat{\beta}_i$; tiene que interpretarse como el efecto diferencial de la variable x_i, eliminando los efectos de las otras variables explicativas.
3. El efecto sobre los coeficientes de regresión de las variables relevantes para explicar y, es distinto cuando las variables incluídas son independientes de las excluídas que cuando no lo son: en el primer caso no afectarán a los coeficientes $\hat{\beta}_i$, pero en el segundo pueden distorsionarlos apreciablemente.

Profesor H.Allende

23

Propiedades de los estimadores

2.3.1 Esperanza.

Sea: $C = \left(X^t X\right)^{-1} X^t$

Se puede demostrar que: $\underline{\hat{\beta}} = \beta + Cu$

Luego, $E\left(\underline{\hat{\beta}}\right) = E\left(\beta + Cu\right) = \beta + CE\left(u\right) = \beta.$

2.3.1 Covarianzas.

Sea $\hat{\beta} - \beta = Cu$

Se puede demostrar que: $\Sigma = E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^t\right] = \sigma^2\left(X^t X\right)^{-1}$

Llamando q_{ij} a los elementos de la matriz Σ, se concluye que: $\hat{\beta}_j \sim N\left(\beta_j ; \sigma^2 q_{jj}\right)$

La matriz $(X^t X)^{-1}$ en general no es diagonal, por lo tanto, su inversa tampoco lo será y los coeficientes $\hat{\beta}$ no serán independientes al no tener covarianzas nulas.

Profesor H.Allende

24

El Teorema de Gauss-Markov

El teorema de Gauss-Markov : Fundamento teórico principal del método de mínimos cuadrados en modelos lineales y establece que si las siguientes hipótesis son ciertas:

a) Todos los valores de la variable aleatoria dependiente están generados por el modelo lineal: $Y = X\underline{\beta} + U$

b) Las perturbaciones u_i son no correlacionadas.

c) Todas las perturbaciones tienen la misma varianza.

d) Las perturbaciones son independientes de las v. a. X_i

e) Las variables X_i se obtienen sin errores de medida.

f) Se desea encontrar el estimador *(óptimo ECM)*, dentro de la clase de estimadores insesgados (centrados), que sean funciones lineales de Y. *El* estimador óptimo insesgado tendrá varianza mínima.

Entonces: Gauss-Markov aseguran que los estimadores mínimo cuadráticos son "óptimos" en el sentido restringido dado por f) - g), independiente de la distribución de U.

Estimación de la Varianza.

El modelo de regresión múltiple quedará especificado al estimar β y la varianza σ² de la perturbación $e = Y - \hat{Y} = (I - V)Y$

V es una matriz idempotente, luego $(I\text{-}V)$ también lo es. $\hat{\sigma}^2 = \frac{1}{n} e^t e = U^t (I - V) U$

La expresión es una forma cuadrática de variables aleatorias normales N(0,σ²) e independientes. Luego, $\frac{1}{\sigma^2} e^t e \sim \chi^2_{rang(I-V),gl}$.

Como (I-V) proyecta a Y sobre el complemento ortogonal al espacio definido por X, tendrá rango n-k-1. $\Rightarrow \frac{1}{\sigma^2} e^t e \sim \chi^2_{(n-k-1)}$.

Finalmente, el estimador insesgado para σ², llamado varianza residual es $\hat{S}_R^2$:

$$\hat{S}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}$$

Intervalos de Confianza y Pruebas de Hipótesis

Intervalos de confianza

Si se verifica que $\hat{\beta}_i$ y $\hat{S}_R^2$ son independientes, entonces $\frac{\hat{\beta}_i - \beta_i}{\hat{S}_R \sqrt{q_{ii}}} \sim t_{(n-k-1),gl}$.

Luego, un intervalo de confianza para β_i de nivel γ =1-α

$$IC_\gamma = \left[\hat{\beta}_i \pm t_{n-k-1(\alpha/2)} \hat{S}_R \sqrt{q_{ii}}\right]$$

Pruebas o contrastes.

Se desea contrastar que la vriable aleatoria $\hat{\beta}_i$ tiene media β_i^* . El test se realiza basado en la estadística t : siendo $E[\hat{\beta}_i] = \beta_i^*$.

$$t = \left(\frac{\hat{\beta}_i - \beta_i^*}{\hat{S}_R \sqrt{q_{ii}}}\right) \approx t_{(n-k-1)gl}$$

Intervalos de Confianza y Pruebas de Hipótesis

Una prueba importante es $H_0 : \beta_i^* = 0.$ Bajo H_0 $\left(t_0 = \frac{\hat{\beta}_i}{\hat{S}_R \sqrt{q_{ii}}}\right) \sim t_{(n-k-1),gl}$.

Rechazandosé H_0 para t_0 > c (valor crítico).

Regiones de confianza para conjuntos de coeficientes.

Como los coeficientes $\hat{\beta}_i$ son dependientes, Los intervalos de confianza individuales pueden dar una imagen errónea de sus valores conjuntos.

Sea $F = \frac{(\hat{\beta} - \underline{\beta})^t (X^t X)(\hat{\beta} - \underline{\beta})}{(k+1)\hat{S}_R^2} \sim F_{(k+1, n-k-1),gl}$.

Luego, la región de confianza de nivel (1-α) se obtiene calculando un valor crítico de la tabla F: $P\left(F_{(k+1,n-k-1)} \geq F_c\right) = \alpha$

Entonces, el elipsoide confidencial contendrá aquellos valores β tales que:

$$\left(\hat{\underline{\beta}} - \underline{\beta}\right)^t \left(X^t X\right)\left(\hat{\underline{\beta}} - \underline{\beta}\right) \leq F_c \hat{S}_R^2 (k+1)$$

Intervalos de confianza para la varianza

Intervalos de confianza para la varianza.

Un intervalo de confianza de nivel γ =1-α para σ² es:

$$IC_\gamma = \left[\frac{(n-k-1)\hat{S}_R^2}{\chi^2_{(1-\alpha/2)(n-k-1)gl}} ; \frac{(n-k-1)\hat{S}_R^2}{\chi^2_{\alpha/2(n-k-1)gl}}\right]$$

Para intervalos de confianza de una cola:

$$IC_\gamma = \left[0 ; \frac{(n-k-1)\hat{S}_R^2}{\chi^2_{(\alpha, n-k-1)}}\right]$$

Contraste de regresión

El contraste de regresión para coeficientes individuales.

$H_0 : \beta_h = 0$ v/s $H_1 : \beta_h \neq 0,$ Estadística $t = \left[(\beta_h) / (\hat{S}_R \sqrt{q_{hh}})\right]$ ~ $t_{(n-k-1)gl}$.

Usando ANDEVA.

VE(k): Variación explicada por el modelo completo.

VE(k-1): Variación explicada por el modelo sin la variable x_h.

Δ VE = VE(k)-VE(k-1)= (β_h / q_{hh})

Si β_h=0, ΔVE depende solo del error experimental.

Luego, una estadística $F = \frac{\Delta VE}{\hat{S}_R^2} \approx F_{(1,n-k-1)}$

Contraste de regresión

El contraste de regresión para grupos de coeficientes.

$$H_0: \beta_1 = \ldots = \beta_k = 0 \quad \text{v/s} \quad H_1: \text{algún } \beta_i \neq 0, \quad i = \overline{1,k}$$

Sea $\underline{\beta}^*$ el vector de coeficientes que no incluye a la componente β_0

$$F = \frac{\hat{\underline{\beta}}^{*t}\left(X^{*t}X^*\right)^{-1}\hat{\underline{\beta}}^*}{k\hat{S}_R^2} \approx F_{(k;n-k-1)}$$

Descomposición de la varianza.

Pitágoras Generalizado : $\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$

Tabla de ANDEVA.

Fuente	Suma de Cuadrados	g.l	Varianza	Contraste
VE	$\sum\left(\hat{y}_i - \overline{y}\right)^2$	k	$\hat{S}_e^2$	
VNE	$\sum\left(y_i - \hat{y}_i\right)^2$	n-k-1	$\hat{S}_R^2$	$F = \frac{\hat{S}_e^2}{\hat{S}_R^2}$
VT	$\sum\left(y_i - \overline{y}\right)^2$	n-1	$\hat{S}_y^2$	

Profesor H.Allende

31

Correlación en Regresión Múltiple

El contraste de regresión establece que la VE es significativamente mayor que VNE. Bajo H_0, $F_{(k,n-k-1)gl} \sim F = \frac{VE}{k\hat{S}_R^2}$

El coeficiente de determinación.

Es una medida descriptiva global del ajuste de un modelo:
Al valor R se le denomina coeficiente de correlación múltiple.

$$R^2 = \frac{VE}{VT} = \frac{\sum\left(\hat{y}_i - \overline{y}\right)^2}{\sum\left(y_i - \overline{y}\right)^2}$$

Observaciones.

1. Desde un punto de vista estricta la correlación se define solo para v.a., al ser X variables fijas el nombre no es totalmente correcto.
2. R^2 aumenta cuando k aumenta.
3. R^2 es muy sensible con respecto a la formulación del modelo y a la elección de la variable dependiente "y".

Profesor H.Allende

32

El coeficiente de determinación corregido.

Para evitar que R^2 aumente cuando k aumenta, se define un R^2-corregido como:

$$\overline{R}^2 = 1 - \left\{\frac{\sum e_i^2/(n-k-1)}{\sum\left(y_i - \overline{y}\right)/n-1}\right\}$$

Donde se verifica: 1) $\overline{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$

2) $\hat{S}_R^2 = \hat{S}_y^2\left(1 - \overline{R}^2\right)$.

R^2 y el Test de F Regresión.

Una forma alternativa para contrastar la hipótesis de que todos los coeficiente de regresión son cero es: $F = \frac{VE}{k} \Big/ \frac{VNE}{n-k-1} = \frac{VE}{VNE}\left(\frac{n-k-1}{k}\right)$

Mientras $1 - R^2 = \frac{VNE}{VT}$

Luego, el contraste F de regresión puede escribirse: $F_{(k,n-k-1)} = \left(\frac{R^2}{1-R^2}\right)\left(\frac{n-k-1}{k}\right)$

Profesor H.Allende

33

Correlación Parcial Múltiple

Correlación Parcial.

Dado un conjunto de variables $\left(x_1, \ldots, x_p\right)$, el coeficiente de correlación parcial entre dos de ellas, algún x_i y x_j, es una medida adimensional de su relación lineal, cuando se eliminan de ambas los efectos debidos al resto de las variables.

Definición:

Consideremos k regresores $\left(x_1, \ldots, x_k\right)$. Entonces el coeficiente de correlación parcial entre x_1 y x_2 se define como el coeficiente de correlación Lineal de Pearson entre x_1, x_2.

$$r_{12\cdot34\ldots k} = \text{cov}\left(e_{1\cdot34\ldots k}, e_{2\cdot34\ldots k}\right)/\hat{S}^2_{e_{2\cdot34\ldots k}}$$

Es decir $r_{12\cdot34\ldots k}$ es el coeficiente del modelo $e_{1\cdot34\ldots k} = \left(r_{12\cdot34\ldots k}\right) * e_{2\cdot34\ldots k} + u$

Donde $e_{1\cdot34\ldots k}$ y $e_{2\cdot34\ldots k}$ son los residuos de la regresión múltiple de x_1 y x_2 con respecto al resto de las variables de control $\left(x_3, x_4, \ldots, x_k\right)$.

$$x_1 = \beta_0 + \sum_{j=3}^{k}\beta_j x_j + e_{1\cdot34\ldots k}$$

$$x_2 = \tilde{\beta}_0 + \sum_{j=3}^{k}\beta_j x_j + e_{2\cdot34\ldots k}$$

Profesor H.Allende

34

Al estar los residuos depurados de los efectos de las restantes variables, el $r_{12\cdot34\ldots k}$ representa la relación entre x_1 y x_2 que no pueden explicarse por las variables restantes.

El coeficiente de correlación parcial entre la variables de respuesta y un regresor x_i (notación: $r_{y_i\cdot R}$) se obtiene fácilmente a partir de la estadística "t" ; $t_i = \hat{\beta}_i/\hat{\sigma}\left(\beta_i\right)$

Entonces $r^2_{y_i\cdot R} = \frac{t_i^2}{t_i^2 + n - k - 1}$

Regresión con variables ortogonales

Es un caso especial de regresión múltiple donde todas las variables explicativas satisfacen

$$\sum_{i=1}^{h}\left(x_{ij} - \overline{x}_j\right)\left(x_{ih} - \overline{x}_h\right) = 0, \quad \forall j,h$$

$$\Rightarrow X'X = Diagonal\left(\sum\left(x_{i1} - \overline{x}_1\right)^2, \ldots, \sum\left(x_{ik} - \overline{x}_k\right)^2\right)$$

$$\Rightarrow \hat{\beta}_j = \frac{\sum\left(x_{ij} - \overline{x}_j\right)\left(y_i - \overline{y}\right)}{\sum\left(x_{ij} - \overline{x}_j\right)^2} = \frac{\sum\tilde{x}_j\tilde{y}_j}{\sum\tilde{x}_j^2}$$

Profesor H.Allende

35

Predicción

Predicción del valor medio.

La predicción del valor medio de la respuesta para ciertos valores concretos de las variables explicativas $x_h^t = \left(1, x_{1h}, \ldots, x_{kh}\right)$ será: $\hat{y}_h = x_h^t\,\hat{\underline{\beta}}$

$$E\left(\hat{y}_h\right) = x_h^t\,\beta = m_h$$

$$Var\left(\hat{y}_h\right) = \sigma^2 x_h^t\left(X^tX\right)^{-1}x_h = \sigma^2\nu_{hh} = \frac{\sigma^2}{\hat{n}_h}$$

Intervalo de confianza para m_h.

Un intervalo de confianza para m_h de nivel $\gamma = 1-\alpha$ es :

$$IC_\gamma = \left[\hat{y}_h \pm t_{\alpha/2}\hat{S}_R\sqrt{\nu_{hh}}\right]$$

Profesor H.Allende

36

Predicción de una observación.

Predicción de una observación.

La predicción de una observación y_h no observada se efectúa mediante mediante la media de la distribución condicionada, $\hat{y}_h$ dado x_h

$\hat{y}_h = x_h{}' \hat{\beta}$ $\qquad E(\hat{y}_h) = m_h$

Error cuadrático medio de la predicción.

$$E\left[(y_h - \hat{y}_h)^2\right] = \sigma^2(1 + v_{hh})$$

Intervalo de confianza para m_h.

Un intervalo de confianza para y_h de nivel $\gamma = 1-\alpha$ está dado por:

$$IC_\gamma = \left[\hat{y}_h \pm t_{\alpha/2} \hat{S}_R \sqrt{1 + v_{hh}}\right]$$

En la siguiente sección se describen los problemas principales que surgen al construír un modelo de regresión.

Profesor H.Allende 37

Diagnósis y validación de los modelos de regresión múltiple.

Hipótesis	Realidad
Las variables *X* toman valores distintos en la muestra.	Multicolinealidad: Las variables *X* toman valores semejantes en la muestra.
$E[y] = \beta' X$	Error de especificación, $E[y] \neq \beta' X$
La distribución de *u* es normal.	Falta de normalidad: *u* no es normal.
$Var[u] = cte.$ Homocedasticidad.	$Var[u] \neq cte.$ Heterocedasticidad.
u independientes entre si.	Autocorrelación: *u* dependientes.

Profesor H.Allende 38

Modelos Regresión No-Lineal

- Regresión No-Lineal:
  - Modelo *Allométrico*:

$$y_j = \beta_0 x_{1j}^{\beta_1} + \varepsilon_j \qquad j = 1..m$$

Usado para representar la relación existente entre el peso de una parte de una planta respecto a toda la planta

  - Modelo *Mitscherlich*:

$$y_j = \beta_0\left(1 - e^{-\beta_1(x_{1j} + \beta_2)}\right) + \varepsilon_j \qquad j = 1..m$$

y_j Rendimiento del cultivo
x_{1j} Cantidad de fertilizante utilizado
β_0 Límite superior del cultivo
β_2 Cantidad de fertilizante que ya hay en la tierra
β_1 Rapidez de crecimiento del cultivo

Relaciona el rendimiento de una cosecha con la cantidad de fertilizante utilizado

Profesor H.Allende 39

Modelos Lineales Generalizados

- Generalización:
  - La distribución de la variable de salida no necesariamente tiene que ser "Normal".
  - El valor esperado de la salida viene dado por:

$$g(E(y_j)) = \beta_0 \sum_{i=1}^{k} \beta_i x_{ij}$$

g(.) es una función monótona y diferenciable y se conoce como función de enlace.

Profesor H.Allende 40

Modelos Aditivos Generales

- Modelo:

$$g(E(y_j)) = \beta_0 \sum_{i=1}^{k} s_i(x_{ij})$$

donde s_i es una función arbitraria, usualmente suaves.

Ej: B-splines.

Profesor H.Allende 41

Análisis de Variancia (ANOVA)

- Método usado para identificar que parámetros son significativamente distintos de cero en el modelo lineal.
- La Técnica desarrollada por R . Fisher ha sido ampliamente usada en el análisis de experimentos en la agricultura y en la industria.

Profesor H.Allende 42

Ejemplo "Turnips for winter fodder"

- Ejemplo: "Nabos para el forraje invernal" Experimento para investigar el crecimiento de los nabos.

Treatments				Blocks			
Variety	Date	Density	Label	I	II	III	IV
Barkant	21-08-1990	1 Kg/ha	A	2.7	1.4	1.3	3.8
		2 Kg/ha	B	7.3	3.8	3.0	1.2
		4 Kg/ha	C	6.5	4.6	4.7	0.8
		8 Kg/ha	D	8.2	4.0	6.0	2.5
	28-08-1990	1 Kg/ha	E	4.4	0.4	6.5	3.1
		2 Kg/ha	F	2.6	7.1	7.0	3.2
		4 Kg/ha	G	24.0	14.9	14.6	2.6
		8 Kg/ha	H	12.2	18.9	15.6	9.9
Marco	21-08-1990	1 Kg/ha	J	1.2	1.3	1.5	1.0
		2 Kg/ha	K	2.2	2.0	2.1	2.5
		4 Kg/ha	L	2.2	6.2	5.7	0.6
		8 Kg/ha	M	4.0	2.8	10.8	3.1
	28-08-1990	1 Kg/ha	N	2.5	1.6	1.3	0.3
		2 Kg/ha	P	5.5	1.2	2.0	0.9
		4 Kg/ha	Q	4.7	13.2	9.0	2.9
		8 Kg/ha	R	14.9	13.3	9.3	3.6

Ejemplo "Turnips for winter fodder"

- Modelo lineal utilizado

$$y_j = \beta_0 + \beta_B x_{Bj} + \beta_C x_{Cj} + \ldots + \beta_R x_{Rj}$$
$$+ \beta_{II} x_{II,j} + \beta_{III} x_{III,j} + \beta_{IV} x_{IV,j} + \varepsilon_j \qquad j = 1..64$$

$$x_{ij} = \begin{cases} 1 & \text{si el punto j coincide con la bloque/tratamiento i} \\ 0 & \text{en todo otro caso} \end{cases}$$

- ¿Puede un cambio en el tratamiento cambiar el crecimiento de los nabos? ¿Existe alguna constante distinta de cero? → ANOVA

ANOVA

- Considerando el modelo de regresión lineal general:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + u_j \qquad j = 1..n$$

$$u_j \sim N(0, \sigma^2) \quad \text{i.i.d}$$

- Se estiman los parámetros a partir de los datos:

$$\hat{y}_j = \hat{\beta}_0 + \sum_{i=1}^{k} \hat{\beta}_i x_{ij}$$

ANOVA

- Sean los residuos: $e_j = Y_j - Y'_j$
- El tamaño de los residuos está relacionado con σ^2, y corresponde a la varianza de los u's.
- Estimando σ^2

$$S^2 = \frac{\sum_{j=1}^{m} (y_j - \hat{y}_j)^2}{m - (k+1)}$$

⇐ Suma al cuadrado de los residuos

⇐ Grados de libertad de los residuos

ANOVA

- Características de S^2 para comparar modelos:
  - Si el modelo ajustado es el adecuado, entonces S^2 es una buena estimación de σ^2
  - Si el modelo ajustado incluye términos redundantes, entonces S^2 es aún una buena estimación.
  - Si el modelo ajustado no incluye una o más entradas que deberían estar, entonces S^2 tendería a ser más grande que σ^2

ANOVA

- Sea Ω_1 un modelo lineal ajustado al conjunto de datos y sea $S_1^2 = \dfrac{RSS_1}{\upsilon_1}$ la estimación de la varianza.
- Sea Ω_0 un modelo lineal ajustado al conjunto de datos, pero con algunos parámetros en 0, y $S_0^2 = \dfrac{RSS_0}{\upsilon_0}$ es la estimación de la varianza.
  - Si S_0^2 es grande comparado con S_1^2, entonces Ω_0 es un modelo inadecuado.
  - Si son similares, entonces Ω_0 es un modelo satisfactorio.

Anova

- Sean
  - Suma de cuadrados extra $ESS = RSS_0 - RSS_1$
  - Grados extra de libertad $\upsilon_E = \upsilon_0 - \nu_1$
  - Estimación de la varianza $S_E^2 = \dfrac{ESS}{\upsilon_E}$
- Test de hipótesis:
  - Estadístico de test: $F = \dfrac{S_E^2}{S_1^2} = \dfrac{ESS/\upsilon_E}{RSS_1/\upsilon_1}$
  - F tiene una distribución F con parámetros ν_E y ν_1
  - Si F es aproximadamente 1 entonces Ω_0 es adecuado.
  - Si F es grande, entonces el modelo es inadecuado.

$$H_0: \ \Omega_0 \ \text{ es el modelo correcto}$$
$$H_1: \ \Omega_1 \ \text{ es el modelo correcto}$$

Profesor H.Allende

49

Ejemplo "Turnips for winter fodder"

- Modelo lineal utilizado

$$\Omega_0: \ y_j = \beta_0 + \beta_{II} x_{II,j} + \beta_{III} x_{III,j} + \beta_{IV} x_{IV,j} + \varepsilon_j \qquad j = 1..64$$

$$\Omega_1: \ y_j = \beta_0 + \beta_B x_{Bj} + \beta_C x_{Cj} + \ldots + \beta_R x_{Rj} + \beta_{II} x_{II,j} + \beta_{III} x_{III,j} + \beta_{IV} x_{IV,j} + \varepsilon_j \qquad j = 1..64$$

- Tablas ANOVA

	Df	Sum of Sq.	Mean Sq.	F. Value	Pr(F)
Block	3	163,73700	54,57891	2,27802	0,08867543
Residuals	60	1.437,53800	23,95897		

	Df	Sum of Sq.	Mean Sq.	F. Value	Pr(F)
Block	3	163,73700	54,57891	5,69043	0,00216381
Treat	15	1.005,92700	67,06182	6,99191	0,00000017
Residuals	45	431,61100	9,59135		

Profesor H.Allende

50

Ejemplo "Turnips for winter fodder"

- Donde la fila de residuos viene dado por:

$$RSS_0 = 1437.5 \qquad \upsilon_0 = 60 \qquad s_0^2 = \frac{1437.5}{60} = 23.96$$

$$RSS_1 = 431.6 \qquad \upsilon_0 = 45 \qquad s_0^2 = \frac{431.6}{45} = 9.59$$

  - Como las estimaciones de σ^2 son muy diferentes, entonces algunas de las entradas eliminadas de Ω_0 son necesarias
- La fila de tratamiento:

$$ESS = RSS_0 - RSS_1 = 1437.5 - 531.6 = 1005.9$$
$$\upsilon_E = \upsilon_0 - \upsilon_1 = 60 - 45 = 15$$
$$s_E^2 = \frac{1005.9}{15} = 67.06$$

Profesor H.Allende

51

Ejemplo "Turnips for winter fodder"

- Obteniendo el estadístico F:

$$\frac{s_E^2}{s_1^2} = \frac{67.06}{9.59} = 6.99$$

- Nivel de significancia se obtiene de la columna Pr(F)
- La fila "block" proviene al comparar el modelo Ω_0 con:

$$y_j = \beta_0 + \varepsilon_j \quad j = 1,2,..,64$$

Profesor H.Allende

52

Ejemplo "Turnips for winter fodder"

	Df	Sum of Sq.	Mean Sq.	F. Value	Pr(F)
Block	3	163,7367	54,5789	5,6904	0,0021638
Variety	1	83,9514	83,9514	8,7528	0,0049136
sowing	1	233,7077	233,7077	24,3665	0,0000114
density	3	470,3780	156,7927	16,3473	0,0000003
variety: sowing	1	36,4514	36,4514	3,8005	0,0574875
variety: density	3	8,6467	2,8822	0,3005	0,8248459
sowing: density	3	154,7930	51,5977	5,3796	0,0029884
variety; sowing: density	3	17,9992	5,9997	0,6256	6,6022439
Residuals	45	1.437,5380	23,9590		

Profesor H.Allende

53

Modelos Log-lineales

- Es una forma de investigar las relaciones entre variables categóricas.
- Es un tipo de modelo GLM:

$$Y_j \sim Pois(\mu_j)$$
$$\log(\mu_j) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_n x_{nj}$$

- Las asociaciones corresponden a los términos de interacción en el modelo.
  - Problema: Determinar qué parámetros β son nulos → ANOVA
  - Considerar una cantidad llamada ***desviación*** cuando se desea comparar dos modelos.

Profesor H.Allende

54

Ejemplo: "Breast cancer"

			State of Tumour			
			Minimal Inflammation		Greater Inflammation	
Centre	Age	Survived	Malignant Appearance	Benign Appearance	Malignant Appearance	Benign Appearance
Tokyo	Under 50	No	9	7	4	3
		Yes	26	68	25	9
	50-60	No	9	9	11	2
		Yes	20	46	18	5
	70 or over	No	2	3	1	0
		Yes	1	6	5	1
Boston	Under 50	No	6	7	6	0
		Yes	11	24	4	0
	50-60	No	8	20	3	2
		Yes	18	58	10	3
	70 or over	No	9	18	3	0
		Yes	15	26	1	1
Glamorgan	Under 50	No	16	7	3	0
		Yes	16	20	8	1
	50-60	No	14	12	3	0
		Yes	27	39	10	4
	70 or over	No	3	7	3	0
		Yes	12,0	11	4	1

Profesor H.Allende

55

Ejemplo: "Breast cancer"

	Df	Deviance	Resid. Df	Resid. Dev	Pr(F)
NULL			71		
centre	2	9,36190	69		
age	2	105,535	67		
survived	1	160,6009	66		
inflam	1	291,1986	65		
appear	1	7,5727	64		
centre:age	4	76,9628	60		
centre:survived	2	11,2698	58		
centre:inflam	2	23,2484	56		
centre:appear	2	13,3323	54		
age:survived	2	3,5257	52		
age:inflam	2	0,293	50		
age:appear	2	1,2082	48		
survived:inflam	1	0,9645	47		
survived:appear	1	9,6709	46		
inflam:appear	1	95,4381	45		

Profesor H.Allende

56

Construcción de modelos de regresión

- Una de las mayores dificultades en regresión múltiple es la construcción de un buen modelo, debido a que existen dos problemas

1. Se requiere incorporar la mayor cantidad de variables para alcanzar la mayor explicación de Y.
2. Se requiere tener la menor cantidad de variables para no aumentar la varianza.

Profesor H.Allende

57

Construcción de modelos de regresión

- Existen diversas maneras de seleccionar un buen modelo de regresión, una de ellas es crear todos los modelos posibles en base a las variables existentes y seleccionar el mejor a través de alguna medida de calidad.
- Otra manera es construir un modelo en forma dinámica, mediante:
  - Eliminación Progresiva.
  - Introducción Progresiva
  - Regresión paso a paso

Profesor H.Allende

58

Construcción de modelos de regresión

- Eliminación Progresiva
  1. El modelo se construye considerando todas las variables.
  2. Seleccionar la variable cuyo estadístico F sea mínimo de todos (X_j)
  3. Si el valor del estadístico es menor que un umbral de eliminación (F_{OUT}) reconstruir el modelo sin la variable e ir al paso 2.
  4. Fin del algoritmo.

Profesor H.Allende

59

Construcción de modelos de regresión

- Introducción Progresiva
  1. Selección de la variable más correlacionada con Y.
  2. Construcción del modelo de regresión simple $Y = \beta_0 + \beta_j X_j$.
  3. Verificar a través del estadístico F si Xj aporta significativamente al modelo.
  4. Selección de la variable con mayor correlación a Y dado el modelo existente, lo que equivale a seleccionar la variable con mayor estadístico F.
  5. Si el valor del estadístico F de la variable seleccionada es mayor a F_{IN}, incorporar la variable al modelo e ir a 4.
  6. Fin del algoritmo.

Profesor H.Allende

60

Construcción de modelos de regresión

- Regresión paso a paso.
  - Este método combina los dos algoritmos descritos anteriormente, para poder incorporar y eliminar variables a medida que construye el modelo.
  - Este método se basa en introducción progresiva, pero cuando se incorpora una nueva variable al modelo, se procede a eliminar las variables que ya pertenecían al modelo pero que ya no aportan en forma significativa a este.

Profesor H.Allende 61

Construcción de modelos de regresión

Indices de calidad para evaluar los modelos

Coef. de determinación múltiple: $R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$

Coef. de correlación corregido: $\overline{R}^2 = 1 - \frac{SS_E/(n-p)}{S_{yy}/(n-1)}$

Varianza residual: $MS_E = (1-\overline{R}^2)\frac{S_{yy}}{n-1}$

Estadístico Cp de Mallows: $C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p$

Criterio de Akaike: $AIC = n\ln(\hat{\sigma}^2) + n + 2p$

Profesor H.Allende 62

Ejemplo de Regresión Simple

t	0	1	2	3	4	5	6
V(t)	30	60	46	32	10	4	17
	20	40		26	14	8	
		20			12		
$\overline{V}$(t)	25	40	46	29	12	6	17

Sea $x_t = sen\ t$ $\quad y_t = \overline{V}(t)$

Luego $y(t) = a + b\,x_t + u_t$

$$\min_{a,b} Q(a,b) = \min_{a,b} \sum_t (y_t - a - bx_t)^2$$

Profesor H.Allende 63

$\hat{a} = \overline{y} - \hat{b}\overline{x} = 25{,}3$ $\quad \hat{b} = \frac{\mathbf{cov}(x,y)}{S_x^2} = 20$

$S_y^2 = 1276$ $\quad \sum (y_t - \hat{y}_t)^2 = 22{,}45$

% de Ajuste del Modelo =

$$1 - \frac{\sum \hat{e}_t^2}{S_y^2} = 0{,}98 * 100\% = 98\%$$

Profesor H.Allende 64