Tema 4
Técnicas de regresión
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

Objetivos

- Técnicas de regresión
- Regresión simple (mínimos cuadrados)
- Estimación de los parámetros. Minimización de errores
- Atribución
- Predicción

Utilidad

- Predicción de nuevos datos
- Creación de modelos
- Evaluación de la significatividad de las variables de predicción
- Base de técnicas de la estadística clásica

Modelos

- Representación matemática que pretendre aproximarse a la realidad y hacer predicciones

Regresión simple

- Variable dependiente Y formulada a partir de variable independiente X
- Error distribuido de manera aleatoria
- β_0, β_1 parámetros de la regresión

$$Y = \beta_0 + \beta_1 X.$$

- Para las observaciones de las poblaciones:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Para cada observación de una muestra de las poblaciones:

$$y_i = \hat{b}_0 + \hat{b}_1 x_i + \varepsilon_i$$

- Valor del ajuste:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i.$$

El error

- Es la diferencia entre el valor real y el ajustado de la observación:

$$\varepsilon_i = \hat{y}_i - y_i.$$

Supuestos a comprobar en los modelos de regresión simple:

- El modelo es lineal
- Normalidad y homocedasticidad:
  Los errores se distribuyen como $N(0, \sigma)$, con σ constante
- Los errores son variables aleatorias independientes

Si algún supuesto no se cumple, puede faltar alguna variable o puede ser necesario aplicar otro tipo de regresión

Coeficientes de regresión:

- β_0, el intercepto: $\beta_0 = E(y|x = 0)$
- β_1, pendiente de la recta: $\beta_1 = E(y|x + 1) - E(y|x)$

Varianza del modelo:

- σ^2, se determina a través de los errores del modelo

Obtener causalidad:

- Se necesitan datos experimentales
- Se necesita relación entre las variables fuera de la estadística
- En muchos casos solamente podemos probar asociación

Mínimos cuadrados:

- Buscamos $q(\hat{b}_0, \hat{b}_1)$ la mínima diferencia cuadrática entre cada $\hat{y}_i$ e y_i:

$$q(\hat{b}_0, \hat{b}_1) = \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2$$

- Obtenemos estimadores de los estimadores $\hat{b}_0, \hat{b}_1$:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

- Con estos estimadores queda definida la recta de regresión
- $\sum_{i=1}^{n} e_i = 0$. Por tanto, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ y $\bar{y} = \bar{\hat{y}}$
- La recta de regresión pasa por el punto $(\bar{x}, \bar{y})$

Estimación de la varianza σ^2

- Podemos estimar la varianza con s^2:

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- $\sum_{i=1}^{n} e_i^2$ es llamado SSE (suma de cuadrados residual)
- Si aplicamos el método de máxima verosimilitud (MLE) en lugar del de mínimos cuadrados, la estimación de los coeficientes es la misma pero la de σ^2 es $\hat{\sigma}^2 = SSE/n$

- Es X significativa en la predicción de Y?
- Calculamos el intervalo de confianza de la estimación de β_1
- Si no contiene el 0 entonces es significativa
- El intervalo:

$$\beta_1 \pm t_{1-\alpha/2}(n-2)s_{\beta_1},$$

  donde $s^2_{\beta_1} = s^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $t_{1-\alpha/2}(n-2)$ es la densidad de la t de student con $n-2$ grados de libertad. Para n grande y $\alpha = 0{,}05$ podemos tomarlo como $1{,}96$
- En caso de tener el p-valor, que este sea menor que α equivale a que X es significativa

- Para hacer predicción puntual:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

- Para calcular el intervalo de confianza de una predicción:

$$\hat{y} \pm t_{1-\alpha/2}(n-2)s_{\text{pred}}$$

donde

$$s_{\text{pred}} = s^2 \left(1 + \frac{1}{n} + \frac{x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Descomponemos la varianza

- SST=SSE+SSR
- $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- SST, suma de cuadrados total: mide la variabilidad de la muestra de Y respecto a la media muestral
- SSE, suma de cuadrados del error: mide la variabilidad de los datos de Y respecto a los valores ajustados
- SSR, suma de cuadrados de la regresión: mide la variabilidad de los valores ajustados respecto a la media muestral

Casos extremos:

- $y_i = \hat{y}_i$: entonces SSE$= 0$ y por tanto SST$=$SSR. Todos los residuos son cero, situación ideal
- $\hat{y}_i = \bar{y}$: entonces SST$=$SSE; es el peor escenario de un modelo de regresión, tendríamos $\beta_1 = 0$ y $\beta_0 = \bar{y}$
- Normalmente encontramos situaciones intermedias

Utilidad:

- Realización de la tabla del análisis de la varianza
- Obtención de medidas de la bonda del ajuste del modelo

El coeficiente de determinación R^2

- Representa la proporción de la varianza de Y explicada por el modelo
- Es una de las medidas de la bondad del ajuste más usadas
- $R^2 =$SSR$/$SST$= 1-$SSE$/$SST
- Es adimensional y $0 \leq R^2 \leq 1$
- En los casos extremos anteriores es cuando tenemos $R^2 \in \{0, 1\}$

Análisis de los residuos:

- Aleatoriedad
- Normalidad
- Homogeneidad de la varianza
- Outliers

Si algún supuesto no se cumple o R^2 es bajo, quizás el modelo no es adecuado

Alternativas:

- Transformación de las variables
- Usar otro modelo