Estadística Bayesiana

Clase 11: Modelos de análsis de varianza de un factor

Isabel Cristina Ramírez Guevara

Escuela de Estadística
Universidad Nacional de Colombia, Sede Medellín

Medellín, 16 de septiembre de 2020

Modelo de análisis de varianza

El análisis de varianza es un método estadístico para determinar sí diversos conjuntos de muestras aleatorias de una determinada variable proceden de la misma población.

Modelo de análisis de varianza

El análisis de varianza es un método estadístico para determinar sí diversos conjuntos de muestras aleatorias de una determinada variable proceden de la misma población.
En general, cada conjunto muestral se caracteriza por estar afectado por un tratamiento específico, que puede influir en los valores que tome la variable de estudio (continua).

Modelo de análisis de varianza

El análisis de varianza es un método estadístico para determinar sí diversos conjuntos de muestras aleatorias de una determinada variable proceden de la misma población.
En general, cada conjunto muestral se caracteriza por estar afectado por un tratamiento específico, que puede influir en los valores que tome la variable de estudio (continua).
Estos métodos nacieron en la agricultura, se tenía como objetivo establecer si los diferentes niveles de fertilizante (tratamientos) generaban un cambio en las cosecha.

Modelo de análisis de varianza

El análisis de varianza es un método estadístico para determinar sí diversos conjuntos de muestras aleatorias de una determinada variable proceden de la misma población.
En general, cada conjunto muestral se caracteriza por estar afectado por un tratamiento específico, que puede influir en los valores que tome la variable de estudio (continua).
Estos métodos nacieron en la agricultura, se tenía como objetivo establecer si los diferentes niveles de fertilizante (tratamientos) generaban un cambio en las cosecha.
Se denomina factor a la variable que supuestamente tiene influencia sobre la variable estudiada a la que se denomina dependiente. En el ejemplo anterior el factor es el fertilizante y la variable dependiente la cosecha. En el análisis de varianza, el factor se introduce de forma discreta.

Modelo de análisis de varianza - un factor

Suponga que se tiene una variable categórica A (factor) con a niveles y una variable de respuesta continua Y. Cuando se tiene que el nivel de la variable categórica A influencia la media de la variable continua Y, es equivalente a definir diferente medias de Y para cada categoría de A. Por lo tanto, si se supone un distribución normal para la variable de respuesta Y, se tiene el modelo:

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, \cdots, a$$

$$j = 1, \cdots, n$$

donde μ_i es la media de Y para la i-ésima categoría de A y $\varepsilon_{ij} \sim \text{N}(0, \sigma^2)$.

Modelo de análisis de varianza - un factor

Si $\mu_i = \mu + \alpha_i$, este modelo se puede escribir como:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \cdots, a$$

donde μ media global de la respuesta Y y α_i es el efecto del i-ésimo nivel del factor A.

Esta parametrización se utiliza por dos razones:

Modelo de análisis de varianza - un factor

Si $\mu_i = \mu + \alpha_i$, este modelo se puede escribir como:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \cdots, a$$

donde μ media global de la respuesta Y y α_i es el efecto del i-ésimo nivel del factor A.

Esta parametrización se utiliza por dos razones:

- Separar el efecto de la media global del efecto de la variable categórica A o factor.

Modelo de análisis de varianza - un factor

Si $\mu_i = \mu + \alpha_i$, este modelo se puede escribir como:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \cdots, a$$

donde μ media global de la respuesta Y y α_i es el efecto del i-ésimo nivel del factor A.

Esta parametrización se utiliza por dos razones:

- Separar el efecto de la media global del efecto de la variable categórica A o factor.
- Permite una generalización del modelo ANOVA cuando se tienen más factores en el modelo.

Modelo de análisis de varianza - un factor

La verosimilitud de este modelo es:

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

Para tener una solución única del sistema se requiere restricciones. Dos de las restricciones que se utilizan con más frecuencia son:

Modelo de análisis de varianza - un factor

La verosimilitud de este modelo es:

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

Para tener una solución única del sistema se requiere restricciones. Dos de las restricciones que se utilizan con más frecuencia son:

- Utiliza un nivel del factor A como nivel de referencia, lo que implica que $\alpha_r = 0$ siendo r el nivel de referencia.

Modelo de análisis de varianza - un factor

La verosimilitud de este modelo es:

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$$

Para tener una solución única del sistema se requiere restricciones. Dos de las restricciones que se utilizan con más frecuencia son:

- Utiliza un nivel del factor A como nivel de referencia, lo que implica que $\alpha_r = 0$ siendo r el nivel de referencia.
- La suma de los efectos es igual a cero, por lo tanto:

$$\sum_{i=1}^{a} \alpha_i = 0$$

lo cual es equivalente a:

$$\alpha_1 = -\sum_{i=2}^{a} \alpha_i$$

Definiendo las distribuciones a priori

Los parámetros a estimar son μ, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_a)$ y σ^2. Se supone que todos los parámetros son independientes, por lo tanto:

$$p(\mu, \boldsymbol{\alpha}, \tau) = p(\mu) \prod_{i=1}^{a} p(\alpha_i) p(\tau)$$

Definiendo las distribuciones a priori

Los parámetros a estimar son μ, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_a)$ y σ^2. Se supone que todos los parámetros son independientes, por lo tanto:

$$p(\mu, \boldsymbol{\alpha}, \tau) = p(\mu) \prod_{i=1}^{a} p(\alpha_i) p(\tau)$$

$$\mu \sim \mathsf{N}(\theta_\mu, \sigma^2_\mu)$$

$$\alpha_i \sim \mathsf{N}(\mu_{\alpha_i}, \sigma^2_{\alpha_i}) \quad \text{para} \quad i = 1, \cdots, a$$

$$\tau \sim \mathsf{Gamma}(\lambda, \beta)$$

donde $\tau = \frac{1}{\sigma^2}$ es la precisión. Utilizar una distribución a priori Gamma para la precisión τ corresponde a una distribución a priori Gamma-inversa para σ^2.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\alpha}$ y μ, θ_μ =0 y μ_{α_j} =0.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\alpha}$ y μ, θ_μ =0 y μ_{α_j} =0. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que los diferentes niveles del factor A no tiene efecto en Y.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\alpha}$ y μ, θ_μ =0 y μ_{α_j} =0. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que los diferentes niveles del factor A no tiene efecto en Y. Las varianzas a priori $\sigma^2_{\beta_j}$ y σ^2_μ se establecen igual a un valor grande para representar una alta incertidumbre o ignorancia a priori.

Definiendo las distribuciones a priori

Cuando no hay información disponible para definir las distribuciones a priori usualmente se selecciona como media a priori para los parámetros $\boldsymbol{\alpha}$ y μ, θ_μ =0 y μ_{α_j} =0. Dicho valor centra nuestro conocimiento a priori alrededor de cero, lo cual corresponde al supuesto que los diferentes niveles del factor A no tiene efecto en Y. Las varianzas a priori $\sigma^2_{\beta_j}$ y σ^2_μ se establecen igual a un valor grande para representar una alta incertidumbre o ignorancia a priori. De la misma manera, para τ se utilizan valores pequeños para los parámetros de la distribución a priori, haciendo que la distribución sea no informativa. En la práctica se hace $\lambda = \beta$ y se definen valores como: 1, 0.1, 0.001.

Verificar los supuestos

Se debe verificar los siguientes supuestos:

- **La varianza es constante de los errores.** Se evalúa gráficamente estos supuestos mediante los gráficos de residuos vs. valores ajustados y residuos vs. niveles del factor A, se espera que los residuos se distribuyen al azar alrededor del valor cero.

Verificar los supuestos

Se debe verificar los siguientes supuestos:

- **La varianza es constante de los errores.** Se evalúa gráficamente estos supuestos mediante los gráficos de residuos vs. valores ajustados y residuos vs. niveles del factor A, se espera que los residuos se distribuyen al azar alrededor del valor cero.
- **Normalidad para los errores del modelo.**. La normalidad se chequea a través del gráfico de probabilidad normal construido con los residuos del ajuste, y se espera que la nube de puntos caiga sobre la recta de probabilidad normal, mostrando una asociación lineal entre los cuantiles muestrales de los residuales vs. los cuantiles teóricos estimados bajo supuesto de normalidad.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando boxplot.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando boxplot.
3. Seleccionar la forma más apropiada para el modelo: modelo anova a considerar.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando boxplot.
3. Seleccionar la forma más apropiada para el modelo: modelo anova a considerar.
4. Estimar los parámetros.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando boxplot.
3. Seleccionar la forma más apropiada para el modelo: modelo anova a considerar.
4. Estimar los parámetros.
5. Evaluar el modelo: análisis de los residuales para evaluar supuestos, gráficos de probabilidad normal, interpretar estimaciones de parámetros que resulten de interés.

Análisis de varianza (ANOVA)

El trabajo de análisis contempla una serie de tareas que pueden resumirse en las siguientes:

1. Comprensión del problema.
2. Desarrollar un análisis preliminar: Análisis descriptivos de los datos por ejemplo usando boxplot.
3. Seleccionar la forma más apropiada para el modelo: modelo anova a considerar.
4. Estimar los parámetros.
5. Evaluar el modelo: análisis de los residuales para evaluar supuestos, gráficos de probabilidad normal, interpretar estimaciones de parámetros que resulten de interés.
6. Reportar los resultados.

Ejemplo

El director de un colegio desea contratar un nuevo profesor de matemáticas. Por esta razón realiza un pequeño estudio. Un grupo de 26 estudiantes fueron divididos aleatoriamente en cuatro grupos. En todos los grupos se enseñaron los mismos tópicos de matemáticas por dos horas diarias en una semana. Luego se les realiza la misma prueba a todos los estudiantes. Se obtienen los siguientes resultados en la prueba: