CAP 9: ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES

El procedimiento ANOVA para muestras independientes corresponde al análisis de varianza de una vía, pues solo considera una única variable independiente (de tipo categórica, un factor) cuyos niveles definen los grupos que se están comparando.

Por ejemplo:

H0: El tiempo de ejecución promedio para instancias de tamaño E es igual para los tres algoritmos.
HA: El tiempo de ejecución promedio para instancias de tamaño E es diferente para al menos un algoritmo.

H0: Las varianzas de las k muestras son iguales.
HA: Al menos una de las muestras tiene varianza diferente a alguna de las demás.

CONDICIONES PARA USAR ANOVA DE UNA VÍA PARA MUESTRAS INDEPENDIENTES

Las condiciones que se deben verificar son:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las k muestras son obtenidas de manera aleatoria e independiente desde la(s) población(es) de origen.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. Las k muestras tienen varianzas aproximadamente iguales (homocedasticidad).

ANÁLISIS POST-HOC

- Correcciones de Bonferroni (conservador) y Holm.
- Prueba HSD de Tukey.
- Prueba de comparación de Scheffé (conservador).

CAP 10: ANOVA DE UNA VÍA PARA MUESTRAS CORRELACIONADAS

El procedimiento ANOVA para muestras correlacionadas corresponde al análisis de varianza de una vía, pues solo considera una única variable independiente (de tipo categórica, un factor) cuyos niveles definen los grupos que se están comparando.

En este caso podemos distinguir entre dos escenarios:

- Diseño con medidas repetidas: A cada sujeto se le toman medidas en las diferentes condiciones, por ejemplo, registrar los tiempos de ejecución para una misma instancia de un problema con k algoritmos diferentes.

- Diseño con bloques aleatorios: Cada bloque contiene diferentes sujetos agrupados según una determinada característica, por ejemplo, registrar tiempos de ejecución usando instancias de grafos diferentes, pero similares (como que tengan el mismo número de vértices y aristas), para los k algoritmos.

Por ejemplo:

H0: El tiempo de ejecución promedio es igual para los cuatro algoritmos. **HA**: El tiempo de ejecución promedio es diferente para al menos un algoritmo.

CONDICIONES PARA USAR ANOVA DE UNA VÍA PARA MUESTRAS CORRELACIONADAS

Las condiciones que se deben verificar son:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las mediciones son independientes al interior de cada grupo.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. La matriz de varianzas-covarianzas es esférica. Como explica Horn, esta condición establece que las varianzas entre los diferentes niveles de las medidas repetidas deben ser iguales.

CAP 11: INFERENCIA NO PARAMÉTRICA CON MEDIANAS

PRUEBAS PARA UNA O DOS MUESTRAS

Prueba de suma de rangos de Wilcoxon (U de Mann-Whitney o U de Whitney- Mann)

Esta prueba es una alternativa no paramétrica a la prueba t de Student con muestras independientes.

Las condiciones que se deben verificar son:

1. Las observaciones de ambas muestras son independientes.
2. La escala de medición empleada, debe ser a lo menos ordinal, de modo que tenga sentido hablar de relaciones de orden (igual que, menor que, mayor o igual que).

Por ejemplo:

H0: No hay diferencia en la usabilidad de ambas interfaces (se distribuyen de igual forma). **HA**: Sí hay diferencia en la usabilidad de ambas interfaces (distribuciones distintas).

Prueba de rangos con signo de Wilcoxon

Esta prueba es parecida a la anterior (prueba de suma de rangos de Wilcoxon). Sin embargo, en este caso es la alternativa no paramétrica a la prueba t de Student con muestras pareadas.

Las condiciones que se deben verificar son:

1. Los pares de observaciones son independientes.
2. La escala de medición empleada para las observaciones es intrínsecamente continua.
3. La escala de medición empleada para ambas muestras debe ser a lo menos ordinal.

Por ejemplo:

H0: Las mismas personas no perciben diferencia en la usabilidad de ambas interfaces.
HA: Las mismas personas consideran que la interfaz A tiene mejor usabilidad que la interfaz B.

PRUEBA PARA MÁS DE DOS MUESTRAS

Prueba de Kruskal-Wallis

Si bien ANOVA es usualmente robusto ante desviaciones leves de las condiciones cuando las muestras son de igual tamaño, no ocurre lo mismo cuando los tamaños de las muestras difieren. En este caso, una alternativa es emplear la prueba de Kruskal-Wallis.

Las condiciones que se deben verificar son:

1. La variable independiente debe tener a lo menos dos niveles.
2. La escala de la variable dependiente debe ser, a lo menos, ordinal.
3. Las observaciones son independientes entre sí.

Por ejemplo:

H0: Todos los algoritmos son igual de eficientes (o, de manera similar, ningún algoritmo es menos ni más eficiente que los demás).
HA: Al menos uno de los algoritmos presenta una eficiencia diferente a al menos algún otro algoritmo.

Prueba de Friedman

Esta prueba es una alternativa del procedimiento ANOVA. Sin embargo, no es una extensión, puesto que no considera las diferencias relativas entre sujetos, y en consecuencia, el poder estadístico es bastante menor. Además, a veces no se puede comprobar que la escala de medición de la variable dependiente sea de intervalos iguales.

Las condiciones que se deben verificar son:

1. La variable independiente debe ser categórica y tener a lo menos tres niveles.
2. La escala de la variable dependiente debe ser, a lo menos, ordinal.
3. Los sujetos son una muestra aleatoria e independiente de la población.

Por ejemplo:

H0: Las interfaces tienen preferencias similares. **HA**: Al menos una interfaz obtiene una preferencia distinta a las demás.

CAP 12: REMUESTREO

BOOTSTRAPPING

Este método es términos generales, sigue los siguientes pasos:

1. Crear una gran cantidad B de nuevas muestras (cientos o miles) a partir de la muestra original. Cada muestra debe tener el mismo tamaño que la original y se construye mediante muestreo con reposición. Esto quiere decir que, al seleccionar un elemento de la muestra original, se devuelve a ella antes de tomar el siguiente, por lo que podría ser reelegido.
2. Calcular la distribución bootstrap y obtener el estadístico de interés para cada una de las muestras.
3. Usar la distribución bootstrap, la cual entrega información acerca de la forma, el centro y la variabilidad
4. de la distribución muestral del estadístico de interés.

Se divide en:

- Bootstrapping para una muestra.
- Bootstrapping para dos muestras independientes.
- Bootstrapping para dos muestras pareadas.

PRUEBAS DE PERMUTACIONES

En términos generales, las pruebas exactas de permutaciones para la diferencia entre dos grupos A y B (puede extenderse esta idea para más grupos) de tamaños n_AnA y n_B, respectivamente, sigue los siguientes pasos:

1. Calcular la diferencia entre el estadístico de interés observado para ambos grupos.
2. Juntar ambas muestras en una muestra combinada.

3. Obtener todas las permutaciones de la muestra combinada en que se pueden distribuir las observaciones en dos grupos de tamaños n_A y n_B.
4. Construir la distribución de las posibles diferencias, calculando la diferencia entre el estadístico de interés obtenido para ambos grupos en cada una de las permutaciones.
5. Calcular el valor p exacto, dado por la proporción de permutaciones en que el valor (absoluto, si es bilateral) de la diferencia calculada es menor/mayor o igual al valor (absoluto si es bilateral) de la diferencia observada.

En consecuencia, si la muestra es grande, suele tomarse una muestra aleatoria de las permutaciones posibles, procedimiento que suele denominarse simulación de Monte Carlo (no es insesgado).

En términos generales, el procedimiento a efectuar es:

1. Formular las hipótesis a contrastar (e identificar el estadístico de interés θ).
2. Crear una gran cantidad P de permutaciones (generalmente terminada en 9 para simplificar los cómputos) a partir de las muestras originales, usando muestreo sin reposición sobre la muestra combinada, y obtener el estadístico θ para cada una de las muestras.
3. Generar la distribución que el estadístico θ tendría si la hipótesis nula fuese cierta.
4. Determinar la probabilidad de encontrar un valor de θ al menos tan extremo como el observado en la distribución generada.

Se divide en:

- Prueba de permutaciones para comparar una variable continua en dos muestras independientes.
- Prueba de permutaciones para comparar medias de más de dos muestras correlacionadas.

CAP 13: REGRESIÓN LINEAL (RLS)

La RLS asume que la relación entre dos variables, x e y puede ser modelada mediante una recta de la forma:

$$\acute{y} = \beta_0 + \beta_1 x$$

Donde:
$\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$ son los parámetros del modelo lineal.
$\boldsymbol{x}$ es la variable explicativa o predictor (variable independiente).
$\acute{y}$ es la variable de respuesta o de salida (variable dependiente).

* $\acute{y}$ es un estimador que podemos entender de la siguiente manera: dado un valor de x, el valor de γ es, en promedio, $\acute{y}$. En otras palabras, $\acute{y}$ corresponde al valor esperado de γ para un determinado valor de x. En la práctica, existe una diferencia entre el valor esperado $\acute{y}$ y el valor observado de γ. Esta diferencia se denomina residuo y se denota e. Así, se tiene:

$$\gamma = \acute{y} + e$$

CORRELACIÓN

Formalmente, podemos medir la fuerza de una relación lineal mediante la correlación (coeficiente de correlación de Pearson que toma un valor entre -1 y 1, si es mayor a 0 la relación es directa, si es menor es inversa).

REGRESIÓN LINEAL MEDIANTE MÍNIMOS CUADRADOS

Este método sirve para ajustar un modelo lineal, y es el más empleado ya que minimiza la suma de los cuadrados de los residuos.

Las condiciones que se deben verificar son:

1. Los datos deben presentar una relación lineal.
2. La distribución de los residuos debe ser cercana a la normal.
3. La variabilidad de los puntos en torno a la línea de mínimos cuadrados debe ser aproximadamente constante.
4. Las observaciones deben ser independientes entre sí. Esto significa que no se puede usar regresión lineal con series de tiempo.

Cuando contamos con más de una variable para construir una regresión lineal simple (RLS), lo más adecuado es que escojamos como predictor aquella variable que tenga la correlación más fuerte con la variable de respuesta.

REGRESIÓN LINEAL CON UN PREDICTOR CATEGÓRICO

* Solo se estudiará el cas de una variable dicotómica (así, siempre se cumple la condición de que los datos presentan una relación lineal).

Para usar una variable categórica con dos niveles, tenemos que convertirla a formato numérico, para lo cual creamos una nueva variable indicadora que toma los valores 0 y 1.

EVALUACIÓN DE UN MODELO DE RLS

Influencia de los valores atípicos

Los valores atípicos que se alejan horizontalmente del centro de la nube principal de puntos pueden, poten-cialmente, tener una gran influencia en el ajuste de la línea de regresión. Este fenómeno se conoce como apalancamiento (leverage en inglés), pues dichos puntos parecen tirar de la línea hacia ellos. Cuando un valor atípico ejerce efectivamente esta influencia, decimos que es un punto influyente.

Un buen método para identificar valores atípicos es usar los residuos (es decir, divididos por la estimación de su desviación estándar), pues esto nos permite establecer un rango fijo de valores aceptables y, en consecuencia, fijar un criterio para comparar residuos de distintos modelos.

Bondad de ajuste

Una medida muy útil que podemos usar para evaluar la bondad de ajuste de un modelo de regresión lineal con respecto a las observaciones es el coeficiente de determinación, cuyo valor varía entre 0 y 1, corresponde al porcentaje de la variabilidad de la respuesta que es explicado por el predictor.

Validación cruzada

Una estrategia frecuente para determinar si el modelo puede generalizarse es la validación cruzada, en la que el conjunto de datos se separa en dos fragmentos:

- Conjunto de entrenamiento: suele contener entre el 80 % y el 90 % de las observaciones (aunque es frecuente encontrar que solo contenga el 70 % de ellas), escogidas de manera aleatoria, y se emplea para ajustar la recta con el método de mínimos cuadrados.
- Conjunto de prueba: contiene el 10% a 30% restante de las instancias, y se usa para evaluar el modelo con datos nuevos.

Validación cruzada de k pliegues

Una buena manera de mejorar la estimación del error cuadrático medio es obtener más observaciones, de acuerdo al ya conocido teorema del límite central. Para esto, se puede usar una nueva manera de remuestreo: la validación cruzada de k pliegues.

La idea de fondo es la misma de la validación cruzada expuesta en el apartado anterior: usar un conjunto de entrenamiento para ajustar el modelo y otro de prueba para evaluarlo. Sin embargo, esta variante modifica este proceso a fin de obtener k estimaciones del error. Para ello se separa el conjunto de datos en k subconjuntos de igual tamaño y, como explica Amat Rodrigo (2016d), realizamos k estimaciones del error cuadrático medio de la siguiente manera:

1. Para cada uno de los k subconjuntos:
    a) Tomar uno de los k subconjuntos del conjunto de entrenamiento y reservarlo como conjunto de prueba.

b) Ajustar la recta de mínimos cuadrados usando para ello los k − 1 subconjuntos restantes.
c) Estimar el error cuadrático medio usando para ello el conjunto de prueba.

2. Estimar el error cuadrático medio del modelo, correspondiente a la media de los k errores cuadrados medios obtenidos en el paso 1.

Validación cruzada dejando uno fuera

Cuando la muestra disponible es pequeña, tema que reforzaremos en el capítulo siguiente, una buena alternativa es usar validación cruzada dejando uno fuera. El esquema es el mismo que para validación cruzada con k pliegues, pero ahora usaremos tantos pliegues como observaciones tenga el conjunto de entrenamiento.

CAP 14: REGRESIÓN LINEAL MÚLTIPLE (RLM)

La regresión lineal múltiple (RLM), correspondiente al caso de una única respuesta con múltiples predictores. Esta tiene la forma de:

$$\acute{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Donde:

Cada $\boldsymbol{x_i}$ es un predictor.
Cada $\boldsymbol{\beta_i}$ corresponde a un parámetro del modelo.
$\boldsymbol{k}$ es la cantidad de predictores.
ý es una estimación de la respuesta.

La RSM requiere verificar algunas condiciones:

1. La distribución de los residuos debe ser cercana a la normal.
2. La variabilidad de los residuos debe ser aproximadamente constante.
3. Los residuos deben ser independientes entre sí.
4. Cada variable se relaciona linealmente con la respuesta.

CONDICIONES PARA USAR RLM

Las condiciones que debemos cumplir para que un modelo de regresión lineal sea generalizable:

1. Las variables predictoras deben ser cuantitativas o dicotómicas (de ahí la necesidad de variables indi-cadoras para manejar más de dos niveles).
2. La variable de respuesta debe ser cuantitativa y continua, sin restricciones para su variabilidad.

1. Los predictores deben tener algún grado de variabilidad (su varianza no debe ser igual a cero). En otras palabras, no pueden ser constantes.
2. No debe existir multicolinealidad. Esto significa que no deben existir relaciones lineales fuertes entre dos o más predictores (coeficientes de correlación altos).
3. Los residuos deben ser homocedásticos (con varianzas similares) para cada nivel de los predictores.
4. Los residuos deben seguir una distribución cercana a la normal centrada en cero.
5. Los valores de la variable de respuesta son independientes entre sí.
6. Cada predictor se relaciona linealmente con la variable de respuesta.

Una opción adecuada cuando necesitamos saber cuáles predictores son estadísticamente significativos, es observar los valores p asociados a cada predictor. Habitualmente consideraremos significativos aquellos predictores para los cuales $p < 0,05$.

COMPARACIÓN DE MODELOS

En la sección anterior vimos que métricas como el AIC o el BIC nos pueden resultar útiles para comparar dos modelos de regresión lineal, considerando la noción general que un modelo es mejor mientras menor sea su valor de AIC (o BIC).

SELECCIÓN DE PREDICTORES

La regresión jerárquica es un método que debemos considerar al momento de intentar probar una teoría y consiste en comenzar por incorporar en primer lugar aquellos predictores ya conocidos, en orden de importancia, en base a investigaciones previas. Una vez incor-porados todos los predictores ya conocidos, podemos incorporar otros nuevos si creemos que existen buenas y justificadas razones para ello.

Si, en lugar de probar una teoría, lo que queremos es explorar los datos, podemos usar otras estrategias:

- Selección hacia adelante.
- Eliminación hacia atrás.
- Regresión escalonada.
- Todos los subconjuntos.

EVALUACIÓN DE UN MODELO DE RLM

Identificación de valores con sobreinfluencia

- Residuo estandarizado.
- Valor predicho ajustado.
- Residuo estudiantizado.

- Diferencia en ajuste.
- Diferencia en betas.
- Distancia de Cook.
- Apalancamiento.
- Razón de covarianza.

Tamaño de la muestra

Una de las reglas más simplistas en verificar que se tengan al menos 10 o 15 observaciones por cada predictor.

CAP 15: REGRESIÓN LOGÍSTICA

La regresión logística es un modelo lineal generalizado, que admite una variable de respuesta cuyos residuos sigan una distribución diferente a la normal.

La regresión logística relaciona la distribución de la variable de respuesta con un modelo lineal usando como función de enlace la función logística estándar (*logit()*). Esta función describe una transición de cero a uno indicando la probabilidad de que ocurra algún evento.

Así, la regresión logística nos permite asociar la probabilidad de ocurrencia de un evento e a una combinación lineal de variables predictoras x1 , x2 , . . . , xn:

$$p(e) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

EVALUACIÓN DE UN CLASIFICADOR

Una forma de evaluar un modelo de clasificación de regresión logística, es de acuerdo a la cantidad de errores cometidos. Para ello, se construye una tabla de contingencia (matriz de confusión).

		Real		
		1 (+)	**0 (−)**	**Total**
Clasificación	**1 (+)**	VP	FP	$VP + FP$
	0 (−)	FN	VN	$FN + VN$
	Total	$VP + FN$	$FP + VN$	n

Donde:

- Verdades positivas (VP): Cantidad de instancias correctamente clasificadas como pertenecientes a la clase positiva.
- Falsos positivos (FP): Cantidad de instancias erróneamente clasificadas como pertenecientes a la clase positiva.

- Falsos negativos (FN): Cantidad de instancias erróneamente clasificadas como pertenecientes a la clase negativa.
- Verdaderos negativos (VN): Cantidad de instancias correctamente clasificadas como pertenecientes a la clase negativa.

Luego, se tiene que:

- Exactitud: Corresponde a la proporción de observaciones correctamente clasificadas.
- Sensibilidad: Indica cuán apto es el modelo para detectar aquellas observaciones pertenecientes a la clase positiva.
- Especificidad: Permite determinar cuán exacta es la asignación de elementos a la clase positiva.
- Precisión: También llamado valores predictivo positivo (VPP) indica la proporción de instancias clasificadas como positivas que realmente lo son.
- Valor predictivo negativo (VPN): Señala la proporción de instancias correctamente clasificadas como pertenecientes a la clase negativa.

La curva ROC indica que mientras más se aleje la curva de la diagonal, mayor es la precisión.

BONDAD DE AJUSTE DEL MODELO

El estadístico de log-verosimilitud, permite cuantificar la diferencia entre las probabilidades predichas y las observadas. Este indica que mientras menor sea su valor, mejor es el ajuste del modelo.

CONDICIONES PARA USAR REGRESIÓN LOGÍSTICA

Además de evaluar el desempeño del clasificador, es necesario verificar:

1. Debe existir una relación lineal entre los predictores y la respuesta transformada.
2. Los residuos deben ser independientes entre sí.

COMPARACIÓN DE MODELO

Podemos comparar modelos de regresión logística mediante la función anova().