Supuestos y Validación del ANOVA

Clase 11

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Motivación

Desde la clase 1 estamos asumiendo que: $Y_{ij} \sim_{iid} Normal(\mu_i, \sigma^2)$ donde σ^2 es igual para todos los tratamientos.

¿Que sucede si esto NO es verdad? Es decir, que pasa si:

- La distribución de los datos NO es normal, sino que pertenece a otra familia.
- Los datos NO son independientes, sino que estas asociados entre ellos.
- La varianza de los datos σ^2 NO es constante, sino que cambia entre los tratamientos.

Outline

Normalidad

Asumir que los datos provienen de una distribución normal puede ser muy exigente!

¿Que propiedades tiene la normal?

- Es una distribución simétrica.
- Puede toma valores tanto positivos como negativos.
- Es una distribución continua.

Hay muchos contextos donde esto no pasa. Asumir normalidad en un modelo es en principio peligroso. Pero resulta que en nuestro contexto no lo es tanto. ¿Porqué?

Normalidad

En Proba 1 vieron uno de los resultados más importantes en estadística:

El teorema del límite central

Dada una muestra aleatoria $X_1, ..., X_n$ de una población con media μ y varianza σ^2, entonces se tiene que:

$$\bar{X} \sim AN\left(\mu, \frac{\sigma^2}{n}\right)$$

En palabras simples, la distribución de un promedios es aproximadamente normal, independiente de la distribución original de los datos.

Siempre hemos trabajado con los promedios (e.g $\bar{Y}_{i.}, \bar{Y}_{.j}, \bar{Y}_{..}$, etc) ya sea en las sumas de cuadrados o en los contrastes. Entonces, así los Y_{ij} NO sean normales, sus promedios muestrales eventualmente lo serán, y con eso es suficiente.

Outline

Independencia

La literatura asociada a probar el supuesto de independencia en una muestra es extremadamente limitada debido a todas las posibles estructuras de dependencia que pueden existir.

La dependencia entre los datos puede provenir de relaciones temporales, espaciales, grupos, etc. El objetivo no es ahondar en este tema, pero es importante que sepan de su existencia la hora de realizar un experimento.

El supuesto de independencia, en el contexto de diseño de experimentos, se trata de satisfacer en lo posible posible al momento del diseño, así como en toda su etapa de muestreo.

Independencia

Para esto se debe asegurar que los datos se recolecten de tal forma que la única influencia posible sobre la variable respuesta provenga de los factores de estudio y no otras fuentes. Entre eso, es necesario que las replicas del experimento no se afecten entre ellas (i.e repetir sujetos experimentales, instrumentos, etc)

Realmente es difícil asegurar todas estas condiciones en su totalidad, pero podemos determinarlas de ante mano y hacer algo al respeto. Por ejemplo utilizar bloques! Los bloques pueden ayudarles a incorporar la estructura de independencia que tengan en sus datos.

Outline

Homocedasticidad

Heterocedasticidad

Cuando la varianza NO es homogénea en los tratamientos (i.e $\sigma^2 = \sigma^2_{ij}$).

Si hay heterocedasticidad, todo el procedimiento de inferencia estadística se ve comprometido. Las pruebas estadíisticas NO son creíbles, y nuestras conclusiones tampoco serán acertadas, o al menos no al nivel de confianza que estamos afirmando.

Para identificar la heterocedasticidad, existen estrategias basadas en estadísticas descriptivas, gráficas (Box Plots) y pruebas estadísticas (Levene, Brown-Forsythe).

La corrección es por medio de una transformación que logre homogeneizar la varianza entre los tratamientos.

Outline

Motivación

Geoge Box

"Essentially, all models are wrong, but some are useful; the practical question is how wrong do they have to be to not be useful."

Nosotros tenemos un esquema mental de como funciona la realidad, pero no necesariamente significa que tenemos la razón:

¿Nuestros supuestos se cumplen? ¿Se incluyeron todos los factores en el diseño? ¿Podemos creer en el ANOVA?

La verdad es que no podemos probar si el modelo es la realidad, pero si podemos determinar si el modelo NO lo es, y así descartarlo.

Residuales

Para descartar nuestro modelo, necesitamos definir una medida que evidencie que tan bueno es su ajuste. En otras palabras, su error.

Habíamos visto que el SSE se asocia a lo que el modelo no explica, así que veamos como se construye:

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n} e_{ij}^2$$

Los términos e_{ij} se pueden ver como la contribución en el error asociado a la j-ésima observación del i-ésimo nivel del factor. Los definiremos como los residuales y serán nuestra medida.

Residuales

Los residuales e_{ij} se asocian a lo que el modelo no explica para esa observación particular. Luego, si existen patrones de cualquier tipo, significa que aún quedan cosas por explicar.

¿Que esperamos de los residuales?

- Los residuales deberían estar concentrados en el 0 y sin ninguna tendencia de sesgo. De lo contrario, habrían descaches evidentes.
- Los residuales deberían ser aleatorios con respecto a cualquier otra cantidad (i.e no tendencias de crecimiento, decrecieminto, etc). Si esto no sucede, quiere decir que hay información explicativa que NO se incluye en el modelo.

Si los residuales NO se comportan como queremos, quiere decir que el ANOVA no es adecuado y no es creíble.

Residuales

¿Como evaluamos los residuales?

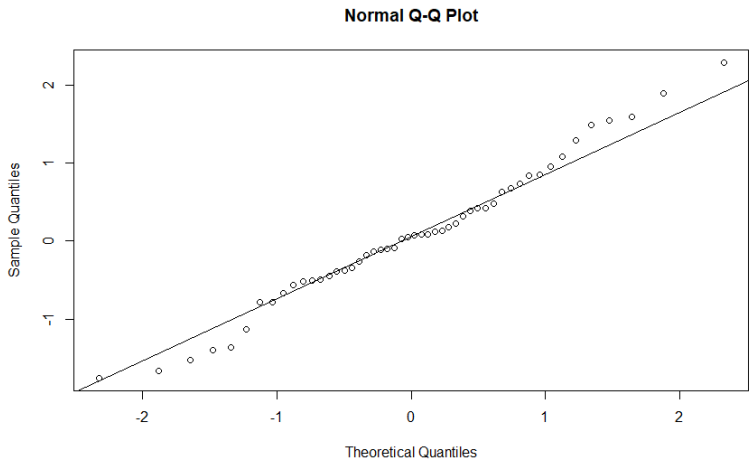
La estrategia es hacer gráficas de los RESIDUALES CONTRA TODO

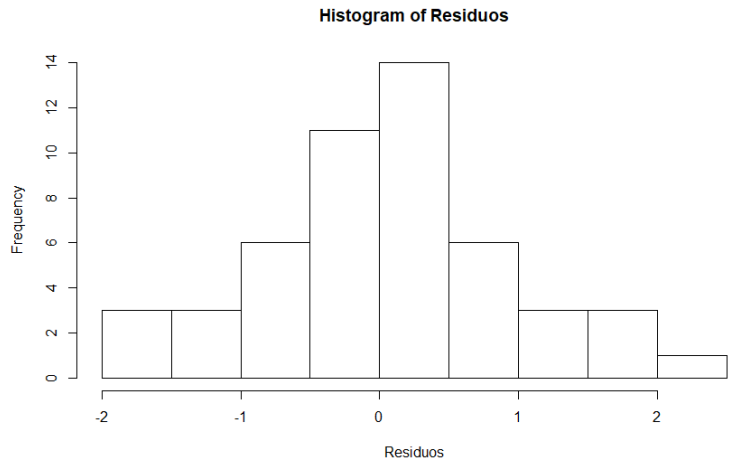
- Histogramas, qq plots (para verificar la distribución de los residuales)
- Gráficos de dispersión (para comprobar que no existen patrones: Residuales vs Tiempo, Residuales vs Y, Residuales vs $\hat{Y}$, Residuales vs Tratamientos, etc)
- Gráficos de cajas.

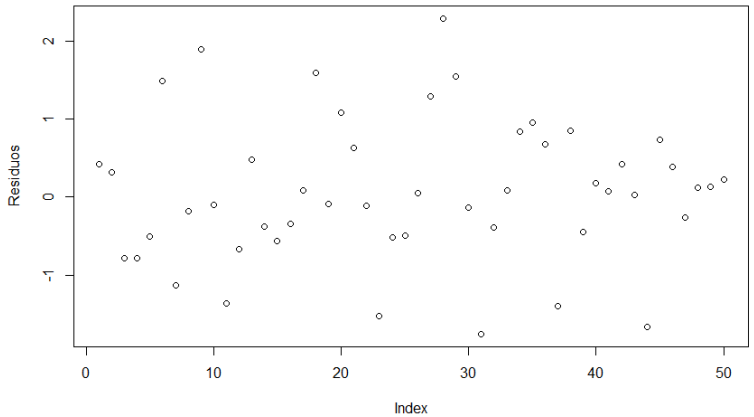
¿Que hacer si hay patrones?

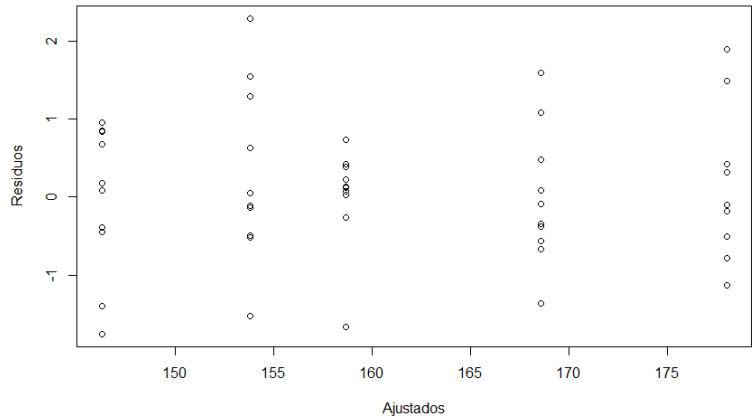
Los resultados del experimento simplemente no son creíbles. Hay que evaluar que falencias pudieron darse durante el experimento, para así corregirlo en uno futuro.

Ejemplos


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-2
-1
0
1
2


Histogram of Residuos
Frequency
Residuos
0
2
4
6
8
10
12
14


Residuos
Index
0
10
20
30
40
50


Residuos
Ajustados
150
155
160
165
170
175