Modelos Lineales

Análisis de la Covarianza

Fernando Massa; Bruno Bellagamba

13 de junio 2024

FCEA FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN
IESTA INSTITUTO DE ESTADÍSTICA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Temario

Objetivos

- Hoy plantearemos al análisis de la covarianza (ANCOVA).
- Veremos como sirve de nexo entre los modelos de ANOVA y de regresión.
- Veremos como esto nos sirve para obtener uns perspectiva jerárquica del modelo lineal general.
- Plantearemos la interacción entre una variable cuantitativa y una cualitativa y que interpretación tiene.

Introducción

Asuma que como estadístico se le plantea la siguiente situación:

Introducción

Asuma que como estadístico se le plantea la siguiente situación:

> *Se quiere saber cuál de los 4 profesores del curso de cálculo II logra mejores resultados.*

Asuma que como estadístico se le plantea la siguiente situación:

> *Se quiere saber cuál de los 4 profesores del curso de cálculo II logra mejores resultados. Se reparten los 100 alumnos de una cierta generación en 4 grupos de 25 alumnos, uno para cada profesor.*

Asuma que como estadístico se le plantea la siguiente situación:

> *Se quiere saber cuál de los 4 profesores del curso de cálculo II logra mejores resultados. Se reparten los 100 alumnos de una cierta generación en 4 grupos de 25 alumnos, uno para cada profesor.*
>
> *Se decide comparar los puntajes correspondientes a la evaluación final del curso mediante el modelo de ANOVA.*
>
> *Posteriormente, si se detecta un efecto significativo, las comparaciones múltiples podrán determinar qué profesor presentó el mejor desempeño.*

Introducción

Asuma que como estadístico se le plantea la siguiente situación:

> *Se quiere saber cuál de los 4 profesores del curso de cálculo II logra mejores resultados. Se reparten los 100 alumnos de una cierta generación en 4 grupos de 25 alumnos, uno para cada profesor.*
>
> *Se decide comparar los puntajes correspondientes a la evaluación final del curso mediante el modelo de ANOVA.*
>
> *Posteriormente, si se detecta un efecto significativo, las comparaciones múltiples podrán determinar qué profesor presentó el mejor desempeño.*

A simple vista, la situación no presenta grandes problemas y el método de análisis también parece correcto, pero...

Asuma que como estadístico se le plantea la siguiente situación:

> *Se quiere saber cuál de los 4 profesores del curso de cálculo II logra mejores resultados. Se reparten los 100 alumnos de una cierta generación en 4 grupos de 25 alumnos, uno para cada profesor.*
>
> *Se decide comparar los puntajes correspondientes a la evaluación final del curso mediante el modelo de ANOVA.*
>
> *Posteriormente, si se detecta un efecto significativo, las comparaciones múltiples podrán determinar qué profesor presentó el mejor desempeño.*

A simple vista, la situación no presenta grandes problemas y el método de análisis también parece correcto, pero... ¿Y si alguno de los grupos tiene los alumnos más *capaces*?¿No se vería beneficiado el profesor de dicho grupo?¿No se sesgarían los análisis?

Introducción

El propósito principal (pero no el único) del modelo de ANCOVA es el de realizar el clásico análisis de la varianza **ajustado** los resultados por la presencia de una variable cuantitativa.

Introducción

El propósito principal (pero no el único) del modelo de ANCOVA es el de realizar el clásico análisis de la varianza **ajustado** los resultados por la presencia de una variable cuantitativa. En este ejemplo podría ser el puntaje obtenido por los alumnos en el curso de cálculo I.

Introducción

El propósito principal (pero no el único) del modelo de ANCOVA es el de realizar el clásico análisis de la varianza **ajustado** los resultados por la presencia de una variable cuantitativa. En este ejemplo podría ser el puntaje obtenido por los alumnos en el curso de cálculo I.

La bibliografía suele referirse a estas variables explicatias como *covariables*, de ahí el nombre del método.

Introducción

El propósito principal (pero no el único) del modelo de ANCOVA es el de realizar el clásico análisis de la varianza **ajustado** los resultados por la presencia de una variable cuantitativa. En este ejemplo podría ser el puntaje obtenido por los alumnos en el curso de cálculo I.
La bibliografía suele referirse a estas variables explicatias como *covariables*, de ahí el nombre del método.

De incluir una variable de índole cuantitativa en el modelo de ANOVA (por ejemplo, a una vía), la ecuación pasa a ser:

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

Introducción

El propósito principal (pero no el único) del modelo de ANCOVA es el de realizar el clásico análisis de la varianza **ajustado** los resultados por la presencia de una variable cuantitativa. En este ejemplo podría ser el puntaje obtenido por los alumnos en el curso de cálculo I.
La bibliografía suele referirse a estas variables explicatias como *covariables*, de ahí el nombre del método.

De incluir una variable de índole cuantitativa en el modelo de ANOVA (por ejemplo, a una vía), la ecuación pasa a ser:

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

Es posible pensar que el modelo de ANCOVA está a medio camino entre el análisis de regresión y el análisis de la varianza.

Razones para realizar este análisis:

- Ganar precisión en las estimaciones al reducir la varianza del error (σ^2).

Razones para realizar este análisis:

- Ganar precisión en las estimaciones al reducir la varianza del error (σ^2).
- Del item anterior se desprende que si la disminución en σ^2 compensa y supera la pérdida del grado de libertad usado como moneda de cambio, **aumenta** la potencia de la prueba de comparación de medias de tratamientos.

Razones para realizar este análisis:

- Ganar precisión en las estimaciones al reducir la varianza del error (σ^2).
- Del item anterior se desprende que si la disminución en σ^2 compensa y supera la pérdida del grado de libertad usado como moneda de cambio, **aumenta** la potencia de la prueba de comparación de medias de tratamientos.
- Captar información de otros factores que inciden sobre Y pero que no pueden ser captados por el investigador.

ANCOVA

Razones para realizar este análisis:

- Ganar precisión en las estimaciones al reducir la varianza del error (σ^2).
- Del item anterior se desprende que si la disminución en σ^2 compensa y supera la pérdida del grado de libertad usado como moneda de cambio, **aumenta** la potencia de la prueba de comparación de medias de tratamientos.
- Captar información de otros factores que inciden sobre Y pero que no pueden ser captados por el investigador.
- Ajustar los valores de la variable de respuesta Y en situaciones donde el diseño experimental no permita asignar aleatoriamente los sujetos a cada tratamiento. De esta manera, se logra evitar posibles **sesgos**.

Supuestos

Existen 3 supuestos importantes a tener en cuenta:

Supuestos

Existen 3 supuestos importantes a tener en cuenta:

- **La variable dependiente Y está linealmente asociada a x.**

  Si este supuesto se cumple, parte de la variabilidad de Y puede ser *removida* del análisis, reduciendo la varianza de los errores. La forma de chequear este supuesto es mediante la prueba $H_0)\beta = 0$,

Supuestos

Existen 3 supuestos importantes a tener en cuenta:

- **La variable dependiente Y está linealmente asociada a x.**

  Si este supuesto se cumple, parte de la variabilidad de Y puede ser *removida* del análisis, reduciendo la varianza de los errores. La forma de chequear este supuesto es mediante la prueba $H_0)\beta = 0$,

- **La pendiente de x es común a todos los tratamientos**.

  Esto se desprende del supuesto anterior y es posible chequearlo mediante la prueba $H_0)\beta_1 = \beta_2 = \ldots = \beta_J$

Supuestos

Existen 3 supuestos importantes a tener en cuenta:

- **La variable dependiente Y está linealmente asociada a x.**

  Si este supuesto se cumple, parte de la variabilidad de Y puede ser *removida* del análisis, reduciendo la varianza de los errores. La forma de chequear este supuesto es mediante la prueba $H_0) \beta = 0$,

- **La pendiente de x es común a todos los tratamientos**.

  Esto se desprende del supuesto anterior y es posible chequearlo mediante la prueba $H_0) \beta_1 = \beta_2 = \ldots = \beta_J$

- **La covariable NO afecta el promedio de Y entre los tratamientos**.

  La violación de este supuesto jugaría en contra de los inereses del análisis. Para testear este supuesto, es posible llevar a cabo un ANOVA sobre la covariable x antes del ANCOVA.

Caso general

En el caso anterior solo se incluía una covariable en el modelo a una vía pero puede generalizarse, incluyendo más de una.

$$y_{ij} = \mu + \alpha_j + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + \varepsilon_{ij} \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

Caso general

En el caso anterior solo se incluía una covariable en el modelo a una vía pero puede generalizarse, incluyendo más de una.

$$y_{ij} = \mu + \alpha_j + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + \varepsilon_{ij} \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

E incluso es posible incluir p covariables al modelo a dos vías.

$$y_{ijk} = \mu + \alpha_j + \gamma_k + \delta_{jk} + \beta_1 x_{1ijk} + \ldots + \beta_p x_{pijk} + \varepsilon_{ij} \quad i = 1, \ldots, n_{jk} \quad j = 1, \ldots, J \quad k = 1, \ldots, K$$

Caso general

En el caso anterior solo se incluía una covariable en el modelo a una vía pero puede generalizarse, incluyendo más de una.

$$y_{ij} = \mu + \alpha_j + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + \varepsilon_{ij} \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

E incluso es posible incluir p covariables al modelo a dos vías.

$$y_{ijk} = \mu + \alpha_j + \gamma_k + \delta_{jk} + \beta_1 x_{1ijk} + \ldots + \beta_p x_{pijk} + \varepsilon_{ij} \quad i = 1, \ldots, n_{jk} \quad j = 1, \ldots, J \quad k = 1, \ldots, K$$

De manera general, expresaremos el modelo de análisis de covarianza en forma matricial a través de la ecuación:

$$Y = Z\alpha + X\beta + \varepsilon$$

- X contiene las covariables y β los parámetros asociados a las mismas
- Z contiene las dummies asociadas de los factores del modelo y α contiene los efectos de cada factor.

Estimación

De aquí en adelante asumiremos que los factores asociados a la parte *"ANOVA"* del modelo son especificados bajo la restricción donde se anula su primer nivel.

Estimación

De aquí en adelante asumiremos que los factores asociados a la parte *"ANOVA"* del modelo son especificados bajo la restricción donde se anula su primer nivel.

Podemos reexpresar el modelo en la forma:

$$\begin{aligned} Y &= Z\alpha + X\beta + \varepsilon \\ &= (Z, X)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \varepsilon \\ &= U\theta + \varepsilon \end{aligned}$$

Estimación

De aquí en adelante asumiremos que los factores asociados a la parte *"ANOVA"* del modelo son especificados bajo la restricción donde se anula su primer nivel.

Podemos reexpresar el modelo en la forma:

$$\begin{aligned} Y &= Z\alpha + X\beta + \varepsilon \\ &= (Z, X)\begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \varepsilon \\ &= U\theta + \varepsilon \end{aligned}$$

La estimación por mínimos cuadrados de θ surgiría de la ecuación:

$$\begin{aligned} U'U\theta &= U'Y \\ \begin{pmatrix}Z' \\ X'\end{pmatrix}(Z, X)\begin{pmatrix}\hat{\alpha} \\ \hat{\beta}\end{pmatrix} &= \begin{pmatrix}Z' \\ X'\end{pmatrix}Y \\ \begin{pmatrix} Z'Z & Z'X \\ X'Z & X'X \end{pmatrix}\begin{pmatrix}\hat{\alpha} \\ \hat{\beta}\end{pmatrix} &= \begin{pmatrix}Z'Y \\ X'Y\end{pmatrix} \end{aligned}$$

Desarrollando la ecuación anterior :

Desarrollando la ecuación anterior :

$$\begin{aligned} Z'Z\hat{\alpha} + Z'X\hat{\beta} &= Z'Y \\ X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Desarrollando la ecuación anterior :

$$\begin{aligned} Z'Z\hat{\alpha} + Z'X\hat{\beta} &= Z'Y \\ X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Premultiplicando la primera ecuación por $\left(Z'Z\right)^{-1}$ se obtiene que:

$$\hat{\alpha} = \left(Z'Z\right)^{-1} Z'Y$$

Desarrollando la ecuación anterior :

$$\begin{aligned} Z'Z\hat{\alpha} + Z'X\hat{\beta} &= Z'Y \\ X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Premultiplicando la primera ecuación por $\left(Z'Z\right)^{-1}$ se obtiene que:

$$\hat{\alpha} = \left(Z'Z\right)^{-1} Z'Y - \left(Z'Z\right)^{-1} Z'X\hat{\beta}$$

Desarrollando la ecuación anterior :

$$\begin{aligned} Z'Z\hat{\alpha} + Z'X\hat{\beta} &= Z'Y \\ X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Premultiplicando la primera ecuación por $\left(Z'Z\right)^{-1}$ se obtiene que:

$$\begin{aligned} \hat{\alpha} &= \left(Z'Z\right)^{-1} Z'Y - \left(Z'Z\right)^{-1} Z'X\hat{\beta} \\ \hat{\alpha} &= \hat{\alpha}_0 - \left(Z'Z\right)^{-1} Z'X\hat{\beta} \end{aligned}$$

Siendo $\hat{\alpha}_0$ el estimador que se obtendría en el modelo de ANOVA.

Desarrollando la ecuación anterior :

$$\begin{aligned} Z'Z\hat{\alpha} + Z'X\hat{\beta} &= Z'Y \\ X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Premultiplicando la primera ecuación por $\left(Z'Z\right)^{-1}$ se obtiene que:

$$\begin{aligned} \hat{\alpha} &= \left(Z'Z\right)^{-1} Z'Y - \left(Z'Z\right)^{-1} Z'X\hat{\beta} \\ \hat{\alpha} &= \hat{\alpha}_0 - \left(Z'Z\right)^{-1} Z'X\hat{\beta} \end{aligned}$$

Siendo $\hat{\alpha}_0$ el estimador que se obtendría en el modelo de ANOVA. De esta forma, se puede observar que, salvo que $Z'X = 0$, se están corrigiendo los efectos del modelo ANOVA por la presencia de las covariables presentes en X.

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$X'Z\hat{\alpha} + X'X\hat{\beta} = X'Y$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned} X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\ X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \end{aligned}$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned} X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\ X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\ X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \end{aligned}$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned}
X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\
X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y + X'\left[I - Z\left(Z'Z\right)^{-1}Z'\right]X\hat{\beta} &= X'Y
\end{aligned}$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned}
X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\
X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y + X'\left[I - Z\left(Z'Z\right)^{-1}Z'\right]X\hat{\beta} &= X'Y
\end{aligned}$$

Si definimos la matriz de proyección $P_Z = Z\left(Z'Z\right)^{-1}Z'$, entonces:

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned}
X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\
X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y + X'\left[I - Z\left(Z'Z\right)^{-1}Z'\right]X\hat{\beta} &= X'Y
\end{aligned}$$

Si definimos la matriz de proyección $P_Z = Z\left(Z'Z\right)^{-1}Z'$, entonces:

$$X'P_ZY + X'(I - P_Z)X\hat{\beta} = X'Y$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned}
X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\
X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y + X'\left[I - Z\left(Z'Z\right)^{-1}Z'\right]X\hat{\beta} &= X'Y
\end{aligned}$$

Si definimos la matriz de proyección $P_Z = Z\left(Z'Z\right)^{-1}Z'$, entonces:

$$\begin{aligned}
X'P_ZY + X'(I - P_Z)X\hat{\beta} &= X'Y \\
X'(I - P_Z)X\hat{\beta} &= X'[I - P_Z]Y
\end{aligned}$$

Sustituyendo la expresión de $\hat{\alpha}$ en la segunda ecuación se obtiene:

$$\begin{aligned}
X'Z\hat{\alpha} + X'X\hat{\beta} &= X'Y \\
X'Z\left[\left(Z'Z\right)^{-1}Z'Y - \left(Z'Z\right)^{-1}Z'X\hat{\beta}\right] + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y - X'Z\left(Z'Z\right)^{-1}Z'X\hat{\beta} + X'X\hat{\beta} &= X'Y \\
X'Z\left(Z'Z\right)^{-1}Z'Y + X'\left[I - Z\left(Z'Z\right)^{-1}Z'\right]X\hat{\beta} &= X'Y
\end{aligned}$$

Si definimos la matriz de proyección $P_Z = Z\left(Z'Z\right)^{-1}Z'$, entonces:

$$\begin{aligned}
X'P_ZY + X'(I - P_Z)X\hat{\beta} &= X'Y \\
X'(I - P_Z)X\hat{\beta} &= X'[I - P_Z]Y \\
\hat{\beta} &= \left[X'(I - P_Z)X\right]^{-1}X'[I - P_Z]Y
\end{aligned}$$

En un contexto de diseño de experimentos, el principal objetivo inferencial recae sobre hipótesis que refieren a los *efectos*. Recordando que $\theta = (\alpha, \beta)$, las pruebas de hipótesis lineales a realizar serían de la forma:

$$
\begin{aligned}
H_0) \quad & C\theta = \mathbf{0} \\
H_0) \quad & (C_1, \mathbf{0}) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{0} \\
H_0) \quad & C_1 \alpha = \mathbf{0}
\end{aligned}
$$

En un contexto de diseño de experimentos, el principal objetivo inferencial recae sobre hipótesis que refieren a los *efectos*. Recordando que $\theta = (\alpha, \beta)$, las pruebas de hipótesis lineales a realizar serían de la forma:

$$
\begin{aligned}
H_0) \quad & C\theta = \mathbf{0} \\
H_0) \quad & (C_1, \mathbf{0}) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{0} \\
H_0) \quad & C_1 \alpha = \mathbf{0}
\end{aligned}
$$

No obstante, también puede ser de interés realizar pruebas de hipótesis sobre los coeficientes asociados a las covariables.

$$
\begin{aligned}
H_0) \quad & C\theta = \mathbf{0} \\
H_0) \quad & (\mathbf{0}, C_2) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{0} \\
H_0) \quad & C_1 \beta = \mathbf{0}
\end{aligned}
$$

A modo de ejemplo consideraremos el caso del modelo a una vía incluyendo una variable explicativa.

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

A modo de ejemplo consideraremos el caso del modelo a una vía incluyendo una variable explicativa.

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

Al momento de poner a prueba la hipótesis de que el promedio de Y varía entre los niveles del factor A (ajustando por el efecto de x), la prueba a realizar es:

$$H_0) \quad \alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$$
$$H_1) \quad \text{no } H_0$$

A modo de ejemplo consideraremos el caso del modelo a una vía incluyendo una variable explicativa.

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

Al momento de poner a prueba la hipótesis de que el promedio de Y varía entre los niveles del factor A (ajustando por el efecto de x), la prueba a realizar es:

$$H_0) \quad \alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$$
$$H_1) \quad \text{no } H_0$$

Bajo el cumplimiento de H_0 el modelo se *reduce* a:

$$y_{ij} = \mu^* + \beta x_{ij} + \varepsilon_{ij}$$

A modo de ejemplo consideraremos el caso del modelo a una vía incluyendo una variable explicativa.

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$

Al momento de poner a prueba la hipótesis de que el promedio de Y varía entre los niveles del factor A (ajustando por el efecto de x), la prueba a realizar es:

$$H_0) \quad \alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$$
$$H_1) \quad \text{no } H_0$$

Bajo el cumplimiento de H_0 el modelo se *reduce* a:

$$y_{ij} = \mu^* + \beta x_{ij} + \varepsilon_{ij}$$

Comparando la *SCRes* de ambos modelos se pude llevar a cabo la prueba mediente un estadístico F.

Las pruebas de hipótesis a realizar para comprobar los supuestos del ANCOVA (en el caso del modelo anterior) serían:

Las pruebas de hipótesis a realizar para comprobar los supuestos del ANCOVA (en el caso del modelo anterior) serían:

- **Relación lineal entre Y y X**

$$H_0) \quad \beta = 0$$
$$H_1) \quad \beta \neq 0$$

- **Igualdad de pendientes entre tratamientos**

$$H_0) \quad \beta_1 = \beta_2 = \ldots = \beta_J$$
$$H_1) \quad \text{no } H_0$$

Para realizar esta prueba es necesario ajustar un modelo *ampliado* donde se estime el modelo bajo H_1. Esto se debe a que el modelo ajustado en la práctica es el postulado bajo el cumplimiento de H_0.

Diagnóstico

De la misma manera que se hizo este recordatorio cuando se intrdujo el modelo de ANOVA, recordemos que la herramienta que se está empleando es el modelo lineal, por este motivo, para validar las inferencias es necesario realizar la etapa de diagnóstico.

Diagnóstico

De la misma manera que se hizo este recordatorio cuando se intrdujo el modelo de ANOVA, recordemos que la herramienta que se está empleando es el modelo lineal, por este motivo, para validar las inferencias es necesario realizar la etapa de diagnóstico.

- Linealildad (solo con x)
- Multicolinealidad (solo si se incluyen varias covariables)
- Homoscedasticidad
- Normalidad
- Observaciones atípicas

Diagnóstico

De la misma manera que se hizo este recordatorio cuando se intrdujo el modelo de ANOVA, recordemos que la herramienta que se está empleando es el modelo lineal, por este motivo, para validar las inferencias es necesario realizar la etapa de diagnóstico.

- Linealildad (solo con x)
- Multicolinealidad (solo si se incluyen varias covariables)
- Homoscedasticidad
- Normalidad
- Observaciones atípicas

Ante el no cumplimiento de alguno de estos supuestos, es posible llevar a cabo alguno de los procedimientos vistos en clases anteriores.

Ejemplo

Para poner en práctica el modelo de ANCOVA utilizaremos datos de un experimento donde se quisieron comparar 3 métodos de estudio. La forma de evaluarlos fue a través de la calificación final. Se usó como covariable la calificación en un curso previo.

Vayamos al  y veamos como ajustar estos modelos y como realizar la inferencia.

Volviendo a regresión

Como se mencionó anteriormente, este modelo es una especie de híbrido entre los modelos de regresión y de análisis de varianza.

Volviendo a regresión

Como se mencionó anteriormente, este modelo es una especie de híbrido entre los modelos de regresión y de análisis de varianza.

No obstante, de manera más global, todos ellos son casos particulares del *modelo lineal general*.

$$Y = X\beta + \varepsilon$$

Gracias a los conceptos adquiridos en estas últimas clases podemos pensar en una formulación *jerárquica* del modelo de *RLM* basada en el concepto de interacción.

Volviendo a regresión

Como se mencionó anteriormente, este modelo es una especie de híbrido entre los modelos de regresión y de análisis de varianza.

No obstante, de manera más global, todos ellos son casos particulares del *modelo lineal general*.

$$Y = X\beta + \varepsilon$$

Gracias a los conceptos adquiridos en estas últimas clases podemos pensar en una formulación *jerárquica* del modelo de *RLM* basada en el concepto de interacción.

En la próxima diapositiva realizaremos un ejemplo basado en el modelo *RLS* pero el razonamiento es válido para cualquier número de variables.

Representación jerárquica

El modelo de *RLS* se caracterizaba por la ecuación.

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

Representación jerárquica

El modelo de *RLS* se caracterizaba por la ecuación.

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

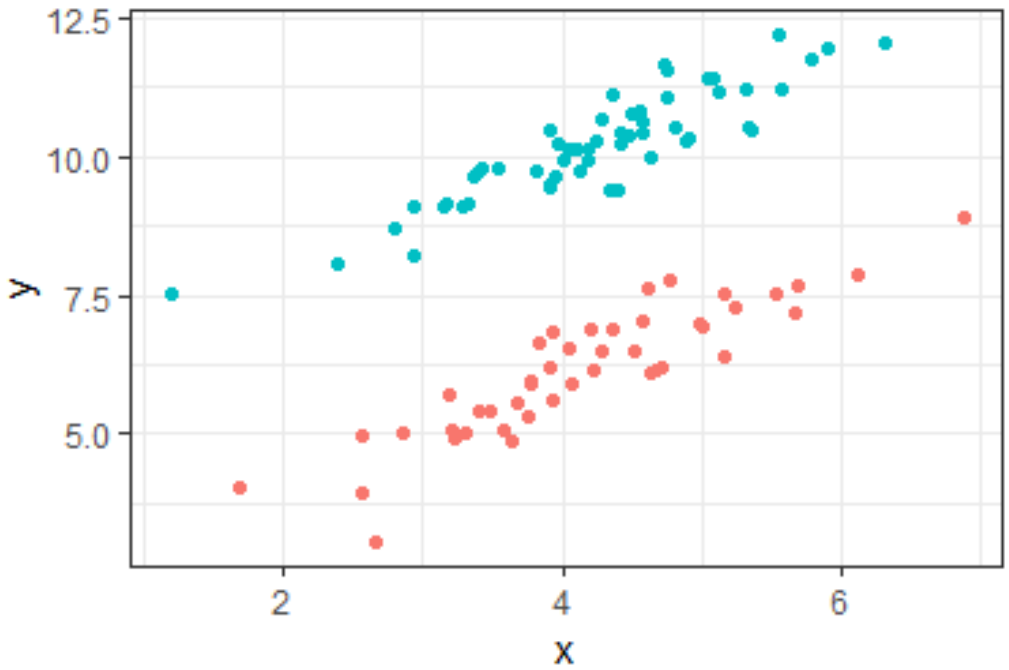
Este modelo es adecuado cuando la relación entre X y Y es lineal. Pero, ¿como adecuarlo a situaciones como esta?

Representación jerárquica

El modelo de *RLS* se caracterizaba por la ecuación.

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$$

Este modelo es adecuado cuando la relación entre X y Y es lineal. Pero, ¿como adecuarlo a situaciones como esta?



Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y.

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

$$y_i = \beta_{0j} + \beta_1 x_i + \varepsilon_i$$

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_i \end{aligned}$$

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_i \end{aligned}$$

Esta es la forma *jerárquica* ya que admite distintos niveles en la especificación de cada parámetro.

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_i \end{aligned}$$

Esta es la forma *jerárquica* ya que admite distintos niveles en la especificación de cada parámetro. En este caso, admitimos que la constante es una función lineal de una segunda variable explicativa, una *dummie* que diferencia los dos grupos de puntos.

Representación jerárquica

La observación clave del el gráfico anterior es que existe una *segunda* variable explicativa que altera la relación entre X e Y. En este caso, altera el valor de la constante. Parecen existir dos grupos, uno con un valor "bajo" de β_0 y otro con un valor "alto".

$$\begin{aligned} y_i &= \beta_{0j} + \beta_1 x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_i \end{aligned}$$

Esta es la forma *jerárquica* ya que admite distintos niveles en la especificación de cada parámetro. En este caso, admitimos que la constante es una función lineal de una segunda variable explicativa, una *dummie* que diferencia los dos grupos de puntos.

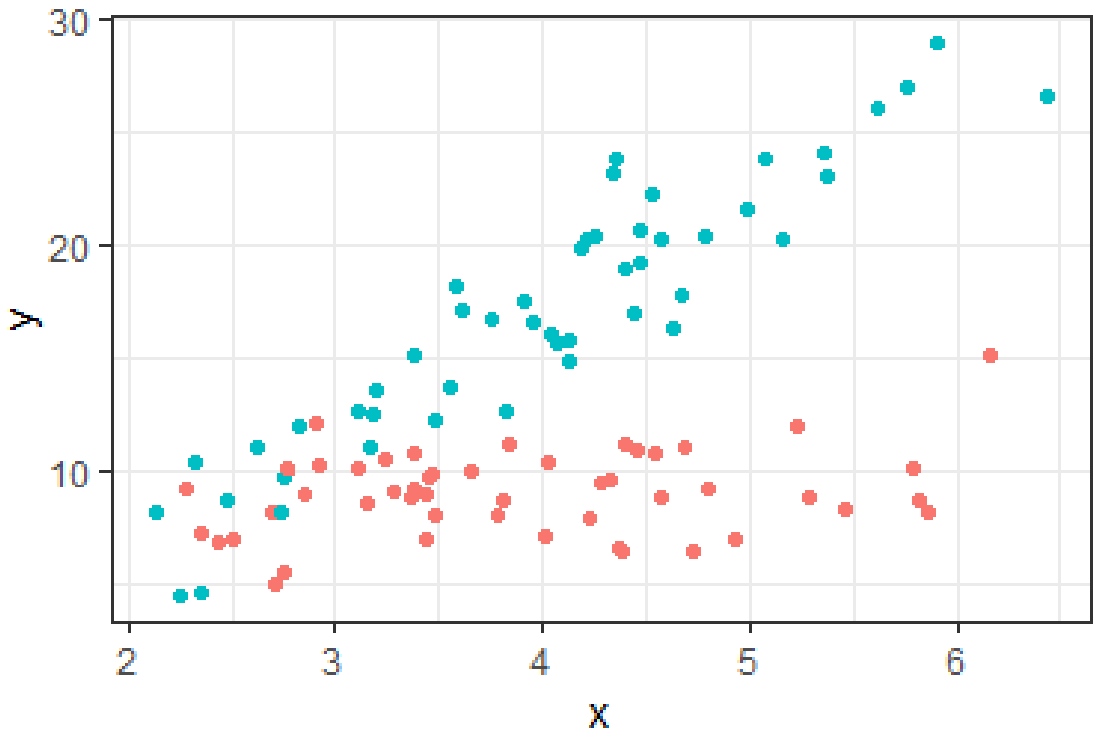
Si quisiéramos escribir todo el modelo en una única linea.

$$y_i = \beta_0 + \beta_{01} z_i + \beta_1 x_i + \varepsilon_i$$

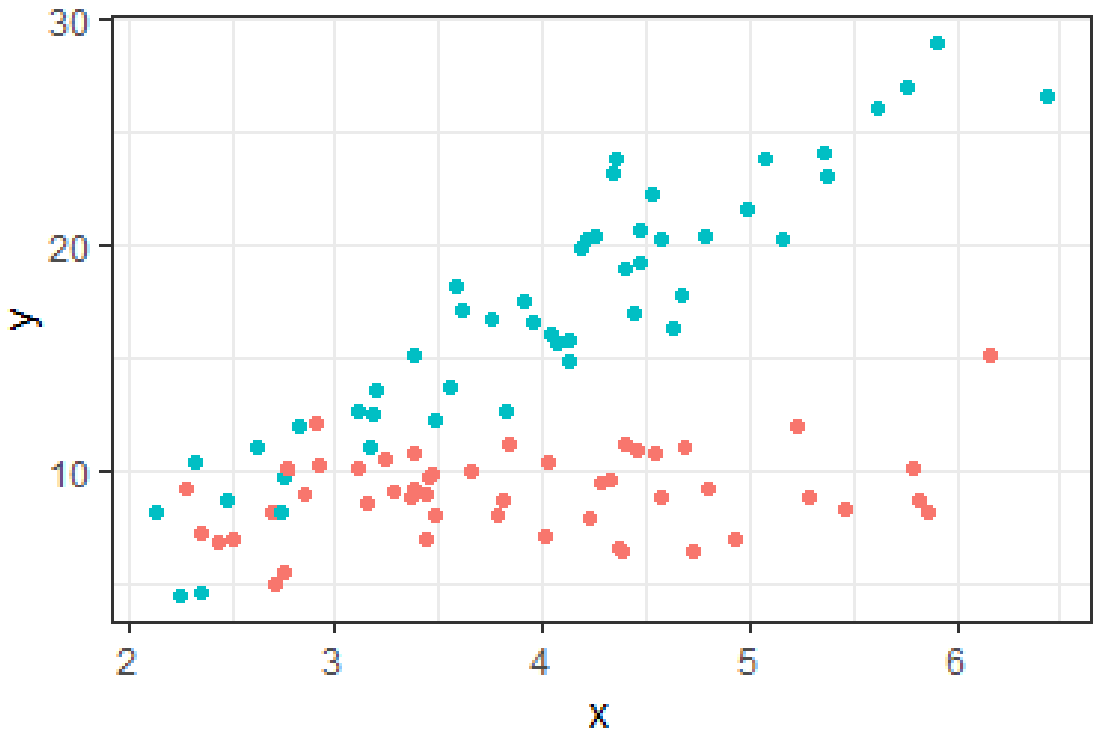
Lo cual no es más que el modelo de *ANCOVA*.

Representación jerárquica

Pero, ¿que tal esta otra situación?



Pero, ¿que tal esta otra situación?



¿Como debería expresarse el modelo en su formulación jearárquica para dar cuenta de esta inusual relación entre X e Y.?

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos.

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos. O que incluso las constantes dependen de los niveles de una variable cualitativa y las pendientes dependerían de los niveles de otra variable cualitativa.

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos. O que incluso las constantes dependen de los niveles de una variable cualitativa y las pendientes dependerían de los niveles de otra variable cualitativa.

El modelo a estimar sería:

$$y_i = \beta_{0j} + \beta_{1k} x_i + \varepsilon_i$$

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos. O que incluso las constantes dependen de los niveles de una variable cualitativa y las pendientes dependerían de los niveles de otra variable cualitativa.

El modelo a estimar sería:

$$\begin{aligned} y_i &= \beta_{0j} + \beta_{1k} x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_{i1} \end{aligned}$$

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos. O que incluso las constantes dependen de los niveles de una variable cualitativa y las pendientes dependerían de los niveles de otra variable cualitativa.

El modelo a estimar sería:

$$\begin{aligned} y_i &= \beta_{0j} + \beta_{1k} x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_{i1} \\ \beta_{1k} &= \beta_1 + \beta_{11} z_{i1} \end{aligned}$$

Representación jerárquica

En este segundo caso podemos pensar que tanto la constante como la pendiente son específicas de ciertos grupos. O que incluso las constantes dependen de los niveles de una variable cualitativa y las pendientes dependerían de los niveles de otra variable cualitativa.

El modelo a estimar sería:

$$\begin{aligned} y_i &= \beta_{0j} + \beta_{1k} x_i + \varepsilon_i \\ \beta_{0j} &= \beta_0 + \beta_{01} z_{i1} \\ \beta_{1k} &= \beta_1 + \beta_{11} z_{i1} \end{aligned}$$

Al reunir todo en una sola ecuación:

$$y_i = \underbrace{\beta_0 + \beta_{01} z_{i1}}_{\beta_{0j}} + \underbrace{(\beta_1 + \beta_{11} z_{i2})}_{\beta_{1k}} x_i + \varepsilon_i$$

Interpretación Algunos puntos a destacar del modelo de la diapositiva anterior

- El parámetro β_{01} acompaña a la variable z_1, que es una variable *dummie* que refiere a cierto nivel de esta variable y adopta el valor cero, para el nivel de referencia. El valor de β_{01} se interpreta como la diferencia entre constantes para ambos grupos

Interpretación Algunos puntos a destacar del modelo de la diapositiva anterior

- El parámetro β_{01} acompaña a la variable z_1, que es una variable *dummie* que refiere a cierto nivel de esta variable y adopta el valor cero, para el nivel de referencia. El valor de β_{01} se interpreta como la diferencia entre constantes para ambos grupos
- En el momento en el que en la ecuación final surge un término como la multiplicación de dos variables, esto se trata de una **interacción**.

Interpretación Algunos puntos a destacar del modelo de la diapositiva anterior

- El parámetro β_{01} acompaña a la variable z_1, que es una variable *dummie* que refiere a cierto nivel de esta variable y adopta el valor cero, para el nivel de referencia. El valor de β_{01} se interpreta como la diferencia entre constantes para ambos grupos
- En el momento en el que en la ecuación final surge un término como la multiplicación de dos variables, esto se trata de una **interacción**.
- En este caso, la interacción entre x y z_2 va acompañada por el parámetro β_{11} el cuál se interpreta como la diferencia de pendientes entre los dos grupos generada por los grupos de la variable z_2 (el primero y el de referencia).

Interpretación Algunos puntos a destacar del modelo de la diapositiva anterior

- El parámetro β_{01} acompaña a la variable z_1, que es una variable *dummie* que refiere a cierto nivel de esta variable y adopta el valor cero, para el nivel de referencia. El valor de β_{01} se interpreta como la diferencia entre constantes para ambos grupos
- En el momento en el que en la ecuación final surge un término como la multiplicación de dos variables, esto se trata de una **interacción**.
- En este caso, la interacción entre x y z_2 va acompañada por el parámetro β_{11} el cuál se interpreta como la diferencia de pendientes entre los dos grupos generada por los grupos de la variable z_2 (el primero y el de referencia).
- Finalmente, β_0 y β_1 corresponden a la constante y pendiente de la ecuación de regresión en los grupos de referencia.

Ejemplo

Veamos un ejemplo donde se quiere obtener una expresión que permita predecir el precio de una propiedad en función del área de la misma y como la zona donde está emplazada puede influir sobre esta relación.

Vayamos al  a ver como ajustar estos modelos.

En la próxima clase

La próxima:

- Llevaremos a cabo un taller con los temas de estas últimas 3 o 4 clases.
- Veremos un ejemplo de ANOVA a 1 vía.
- Veremos otro ejemplo de ANOVA a 2 vías.
- Si da el tiempo, veremos un ejemplo de ANCOVA.

Bibliografía

- Carmona, Francesc (2003). *Modelos Lineales (notas de curso).* Departament d'Estadística.
- Farraway, Julian (2014). *Linear Models with R, second edition.* Chapman Hall/CRC.
- Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition.* John Wiley Sons, Inc.

¿Preguntas?

Muchas Gracias