USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

El modelo de análisis de la covarianza no paramétrico

Antón Quintela Ferreiro

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

El modelo de análisis de la covarianza no paramétrico

Antón Quintela Ferreiro

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística e Investigación Operativa
Título: El modelo de análisis de la covarianza no paramétrico
Breve descripción del contenido:
El modelo de análisis de la covarianza (ANCOVA), dentro del marco del modelo lineal general, se trata de un modelo de regresión lineal que en su caso más sencillo combina dos covariables de distinta naturaleza, continua y categórica, véase [Faraway, 2004]. El modelo ANCOVA es útil para analizar si la relación supuestamente lineal entre dos variables continuas (respuesta y explicativa) se ve afectada por la presencia de la variable categórica. Esta suposición de linealidad en la relación puede ser excesivamente fuerte en contextos aplicados donde se requiere una aproximación más flexible. Dicha aproximación (llamada, ANCOVA no paramétrico) fue introducida por Young y Bowman (1995). El objetivo del trabajo es explorar la potencialidad del modelo ANCOVA no paramétrico, formulado a través de estimadores no paramétricos de la función de regresión, y evaluar su comportamiento mediante un estudio de simulación ilustrando su aplicabilidad sobre un conjunto de datos reales.

Índice

Resumen

En este trabajo se llevará a cabo un estudio detallado del modelo del análisis de la covarianza (ANCOVA) no paramétrico, es decir, una presentación en detalle del modelo, su estimación y los contrastes asociados. Además, se explicará el porqué de la introducción de este modelo, y qué ventajas presenta respecto a la formulación del ANCOVA dentro del marco del modelo lineal general. Por último, cabe destacar que, tanto para ilustrar las diferencias entre el modelo no paramétrico y el lineal, como para entender con mayor facilidad el funcionamiento del modelo no paramétrico, se utilizará un conjunto de datos relativo al precio de distintos modelos de coches.

Abstract

In this work, a detailed study of the non-parametric ANCOVA model will be carried out, that is, a detailed presentation of the model, the estimation procedure and its associated hypothesis testing tools. In addition, the reason behind the introduction of this model will be explained, jointly with the advantages it presents with respect to the formulation of ANCOVA within the framework of the general linear model. Finally, it should be noted that, both to illustrate the differences between the non-parametric and the linear model, and to understand more easily the performance of the non-parametric model, a dataset related to the price of different car models will be used.

Capítulo 1

Introducción

En este capítulo se introducirá el modelo lineal general, para así poder entender mejor cómo encaja el ANCOVA lineal en este modelo y qué limitaciones trae consigo este mismo hecho. A continuación, se desarrollará en detalle el ANCOVA lineal y se expondrán los inconvenientes que presenta el modelo y que motivan la aparición del ANCOVA no paramétrico.

En consecuencia, este capítulo se dividirá en dos secciones: en la primera sección, se introducirá el modelo lineal general y se verá cómo encaja el ANCOVA en este modelo. En la segunda sección se estudiará en detalle el ANCOVA lineal y se explicitarán los procedimientos de estimación y sus contrastes asociados. En esta sección, con ayuda de los datos de [Manish Kumar, 2019], se procederá a explicar cuáles son sus principales limitaciones, dando así una motivación para introducir el modelo no paramétrico, pero a la vez señalando las situaciones en las que el modelo es apropiado.

1.1. Introducción al modelo lineal general

Sea Y una variable respuesta y $X_1, \ldots, X_{p-1}$ un conjunto de variables explicativas. Suponiendo que la relación entre la respuesta y las explicativas es lineal con respecto a unos parámetros, es posible formular el modelo lineal múltiple tal y como se plantea en [Faraway, 2004], el cual se expresa de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon, \tag{1.1}$$

donde $\varepsilon \sim N\left(0, \sigma^2\right)$ es el error y $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son los parámetros. Por simplicidad, se planteará el modelo (1.1) para diseño fijo, en el cual, dada una muestra de tamaño n, esta puede expresarse como:

$$Y_j = \beta_0 + \beta_1 x_{j,1} + \cdots + \beta_{p-1} x_{j,p-1} + \varepsilon_j, \qquad j = 1, \ldots, n,$$

donde Y_j es la respuesta del individuo j-ésimo, $x_{j,l}\ \forall\ l \in \{1, \ldots, p-1\}$ son los valores del individuo j-ésimo en las variables explicativas y ε_j es el error del individuo j-ésimo. Estos errores se suponen independientes, normales y homocedásticos, quedando estas dos últimas propiedades reflejadas en la siguiente expresión: $\varepsilon_j \sim N\left(0, \sigma^2\right)$ para $j = 1, \ldots, n$.

Por último, dada una muestra de observaciones de tamaño n, el modelo (1.1) se puede expresar en forma matricial, pasando a llamarse modelo lineal general, el cual se puede expresar como sigue:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{1.2}$$

equivalentemente, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, siendo $\mathbf{Y}$ el vector de la variable respuesta, $\mathbf{X} \in \mathcal{M}_{n\times p}(\mathbb{R})$ la matriz de diseño donde cada columna, excepto la primera, representa los valores de todas las observaciones para una variable y las filas son el valor de todas las variables para un individuo. Además, $\boldsymbol{\beta}$ es el vector de parámetros y $\boldsymbol{\varepsilon} \in N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I_n}\right)$. En la sección dedicada al ANCOVA lineal, tras introducir este modelo, se presentará con la notación matricial, de forma que será más sencillo entender como, efectivamente, el ANCOVA lineal es un caso particular del modelo lineal general. Por último, dado que en el modelo lineal general el único efecto que se tiene en cuenta sobre la variable respuesta Y, es el conjunto de variables explicativas $X_1, \ldots, X_{p-1}$, sería conveniente comprobar que este efecto es suficientemente significativo. Para ello, se desarrolla el test de no efecto que se detallará en la siguiente sección.

1.2. Test de no efecto

Para averiguar si, efectivamente, el efecto de la variable continua sobre la respuesta es significativo, el contraste planteado es el siguiente:

$$\begin{cases} H_0 : \boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \\ H_a : \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}. \end{cases} \tag{1.3}$$

En este contraste, en la hipótesis nula se especifica un modelo donde la mejor predicción de la variable respuesta es su media μ. Por tanto, en el contraste (1.3) se confrontan un modelo

lineal múltiple frente a un modelo que consiste en hacer la media de las respuestas. Para llevar a cabo este contraste se empleará el denominado test F, basado en el siguiente estadístico:

$$\frac{(RSS_0 - RSS)\,/\,(p-1)}{RSS/\,(n-p)} \in F_{p-1,n-p},$$

donde $F_{k,l}$ es la F de Snédecor de k y l grados de libertad y:

$$RSS = \sum_{i=1}^{n}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \ldots + \hat{\beta}_{p-1}X_{i,p-1}\right)\right]^2 \qquad RSS_0 = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2,$$

siendo $\bar{Y}$ la media muestral de los $Y_1, \ldots, Y_n$. RSS y RSS_0 representan la suma de los residuos al cuadrado a las que daría lugar el modelo lineal general y el ajuste por la media de los datos, respectivamente. Ambas representan la variabilidad de las respuestas en sus respectivos modelos, y el estadístico de contraste compara la diferencia entre RSS_0 y RSS con RSS, ajustando los grados de libertad. En otras palabras, justifica si la reducción del error que se produce al emplear el modelo lineal general, en lugar de únicamente la media de las respuestas, compensa el aumento del número de parámetros, es por ello que tiene en cuenta los grados de libertad.

1.3. ANCOVA lineal

En esta sección se plantea un modelo de regresión en el cual se pretende explicar una variable respuesta continua Y mediante dos variables explicativas: X continua y X_g categórica, siendo esta última la responsable de la división de la muestra en grupos para los que todas las observaciones tienen el mismo valor en X_g. Es importante recalcar que el modelo ANCOVA está enmarcado dentro del modelo lineal general, por lo que ha de satisfacer las hipótesis de normalidad, homogeneidad de varianzas, linealidad e independencia tanto dentro de los grupos como entre ellos. A continuación, y de forma análoga a como se describe en [Maxwell y Kelley, 1990], se distinguirán dos modelos en función de que efecto tenga X_g sobre Y, o lo que es lo mismo, sobre la recta de regresión. Si el efecto del grupo solo afecta a la ordenada en el origen y no a la pendiente de la recta se plantea un modelo sin interacción, pero si por el contrario, el grupo afecta a ambos elementos (pendiente y ordenada en el origen), como el cambio de grupo puede potenciar o atenuar el efecto de la explicativa continua X sobre la respuesta (modifica su pendiente), se plantea un modelo con interacción.

1.3.1. ANCOVA sin interacción

A continuación, se introducirá la notación que se utilizará durante el resto del trabajo para denotar a los individuos en muestras divididas en grupos. Sea $i \in \{1, 2, \ldots, I\}$ el grupo de la variable categórica al que pertenece una observación, con I siendo el número total de categorías

de X_g. Sea $j \in \{1, 2, \ldots, n_i\}$ el índice de una observación dentro del grupo i-ésimo, donde n_i es el número total de observaciones de dicho grupo. Entonces es posible expresar Y_{ij}, la respuesta de la observación j-ésima del grupo i-ésimo como:

$$\begin{aligned} Y_{1j} &= \mu + \gamma X_{1j} + \varepsilon_{1j}, && j = 1, \ldots, n_1, \\ Y_{ij} &= \mu + \alpha_i + \gamma X_{ij} + \varepsilon_{ij}, && i = 2, \ldots, I,\ j = 1, \ldots, n_i, \end{aligned} \tag{1.4}$$

donde μ es el intercepto correspondiente a la regresión simple de Y frente a X en el primer grupo (grupo de referencia); α_i es la diferencia entre el intercepto asociado a la regresión simple de Y frente a X en el grupo i-ésimo y μ; y γ es la pendiente de la regresión simple de Y frente a X en cualquier grupo, pues se supone paralelismo de las rectas. Por último, los $\varepsilon_{ij} \in N(0, \sigma^2)$ son los errores, que se supondrán independientes tanto dentro del grupo como entre ellos. A continuación, con el objetivo de ver como encaja el ANCOVA lineal sin interacción en el modelo (1.2) se tomará de nuevo una muestra de tamaño n bajo diseño fijo y se procederá a exponerlo en forma matricial:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 & x_{1,1} \\ \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 & x_{1,n_1} \\ 1 & 1 & 0 & \cdots & 0 & x_{2,1} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & x_{2,n_2} \\ 1 & 0 & \ddots & 0 & 0 & x_{I,1} \\ \vdots & \vdots & \ddots & \ddots & 0 & \vdots \\ \vdots & \vdots & \cdots & 0 & 1 & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & 1 & x_{I,n_I} \end{pmatrix} \cdot \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_I \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{I1} \\ \vdots \\ \varepsilon_{In_I} \end{pmatrix}. \tag{1.5}$$

Una vez expuesto el modelo teórico (1.4) hay que explicitar cómo son los estimadores de μ, α_i y γ, denotados respectivamente por, $\hat{\mu}$, $\hat{\alpha}_i$, y $\hat{\gamma}_i$. Como se ha descrito anteriormente, γ es la pendiente de la regresión de Y sobre X para cualquier grupo, o lo que es lo mismo, es la pendiente de la regresión de Y sobre X tras quitar el efecto del grupo. La estimación de Y puede llevarse a cabo mediante una regresión particionada, obteniéndose el siguiente estimador:

$$\hat{\gamma} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\bullet}\right)\left(X_{ij} - \bar{X}_{i\bullet}\right)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i\bullet}\right)^2},$$

donde $\bar{Y}_{i\bullet}$ representa el promedio de las respuestas asociadas al i-ésimo grupo y $\bar{X}_{i\bullet}$ representa el promedio de los valores de la explicativa continua asociados al i-ésimo grupo. En función de

este estimador se obtienen los otros estimadores:

$$\hat{\mu} = \bar{Y}_{\bullet\bullet} - \hat{\gamma}\bar{X}_{\bullet\bullet}$$

$$\widehat{\alpha_i} = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet} - \hat{\gamma}\left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right),$$

donde $\bar{Y}_{\bullet\bullet}$ y $\bar{X}_{\bullet\bullet}$ representan, respectivamente, el promedio de respuestas y el promedio de valores de la variable explicativa.

Con la introducción del modelo (1.4) surge la duda de cuando es razonable emplearlo. ¿Cómo se sabe si la variable categórica realmente tiene un efecto sobre la respuesta? ¿Y la continua? Para responder a estas preguntas surgen los siguientes contrastes:

Contraste del efecto de la variable continua

Para averiguar si, efectivamente, el efecto de la variable continua sobre la respuesta es significativo, el contraste planteado es el siguiente:

$$\begin{cases} H_0 : \gamma = 0, \\ H_a : \gamma \neq 0. \end{cases} \tag{1.6}$$

En este contraste, en la hipótesis nula se especifica un modelo donde tan solo la variable categórica tendría efecto sobre la respuesta. Este sería el caso de un modelo de análisis de la varianza (ANOVA), donde los parámetros a estimar serían las medias de los grupos. Por tanto, en el contraste (1.6) se confrontan un modelo ANOVA frente a un modelo ANCOVA. Para llevar a cabo este contraste se empleará el denominado test F, basado en el siguiente estadístico:

$$\frac{RSS_0 - RSS}{RSS/(n-I-1)} \in F_{1,n-I-1},$$

donde $F_{k,l}$ es la F de Snédecor de k y l grados de libertad y:

$$RSS = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu} - \widehat{\alpha_i} - \hat{\gamma}X_{ij})^2 \qquad RSS_0 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2.$$

Estas cantidades representan la suma de los residuos al cuadrado a las que daría lugar el ANCOVA sin interacción y el ANOVA, respectivamente. Ambas representan la variabilidad intragrupo para sus respectivos modelos, y el estadístico de contraste compara la diferencia entre RSS_0 y RSS con RSS, ajustando los grados de libertad. En otras palabras, justifica si la reducción del error que se produce al emplear el ANCOVA sin interacción en lugar del ANOVA compensa el aumento del número de parámetros, es por ello que tiene en cuenta los grados de libertad.

Contraste del efecto de la variable discreta

Análogamente, para averiguar si el efecto de la variable discreta sobre la respuesta es significativo, se plantea el siguiente contraste:

$$\begin{cases} H_0 : \alpha_i = 0 \ \forall \ i, \\ H_a : \exists \ j \text{ tal que } \alpha_j \neq 0. \end{cases} \tag{1.7}$$

Aquí, H_0 representa el caso en el que no hay efecto de la variable discreta (regresión lineal simple) y H_a el caso en que sí hay efecto de la variable discreta (ANCOVA sin interacción).

Para llevar a cabo el contraste (1.7) se empleará el denominado test F, basado en el siguiente estadístico:

$$\frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-I-1)} \in F_{I-1,n-I-1},$$

donde RSS es igual que en el contraste anterior, pues es la suma de residuos al cuadrado correspondiente al mismo modelo, sin embargo RSS_0 es ahora distinto ya que la hipótesis nula ha cambiado. Se expone su expresión a continuación:

$$RSS_0 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet} - \hat{\gamma}_0 (X_{ij} - \bar{X}_{\bullet\bullet}))^2 \qquad \text{con}$$

$$\hat{\gamma}_0 = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{\bullet\bullet}\right) \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{\bullet\bullet}\right)^2} \qquad \text{y} \qquad \hat{\mu}_0 = \bar{Y}_{\bullet\bullet} - \hat{\gamma}_0 \bar{X}_{\bullet\bullet}.$$

De forma análoga al test anterior, RSS_0 y RSS representan la suma de los residuos al cuadrado a las que daría lugar una regresión lineal simple y un ANCOVA sin interacción respectivamente. En este caso, RSS_0 es la suma de residuos al cuadrado asociada a un ajuste con una única recta, si no se tuviese en cuenta el efecto del grupo. En consecuencia, el test considera los grados de libertad para el cálculo del estadístico, pues este intenta corroborar que la reducción del error que se produce al emplear el ANCOVA sin interacción, en lugar de la regresión simple, compensa el aumento en el número de parámetros.

1.3.2. ANCOVA con interacción

Empleando la notación introducida en la sección anterior, se puede expresar Y_{ij}, la respuesta de la observación j-ésima del grupo i-ésimo como:

$$\begin{aligned} Y_{1j} &= \mu + \gamma X_{1j} + \varepsilon_{1j}, && j = 1, \ldots, n_1, \\ Y_{ij} &= \mu + \alpha_i + \gamma X_{ij} + \delta_i X_{ij} + \varepsilon_{ij}, && i = 2, \ldots, I, \ j = 1, \ldots, n_i, \end{aligned} \tag{1.8}$$

donde μ es el intercepto correspondiente a la regresión simple de Y frente a X en el primer grupo (grupo de referencia), α_i es la diferencia entre el intercepto asociado a la regresión simple de Y frente a X en el grupo i-ésimo, γ es la pendiente de la regresión simple de Y frente a X en el grupo de referencia, δ_i es la diferencia entre la pendiente asociada a la recta de regresión de Y frente a X en el grupo i-ésimo, y los $\varepsilon_{ij} \in N(0, \sigma^2)$ son los errores, que se supondrán independientes tanto dentro del grupo como entre ellos. El modelo (1.8) es equivalente a:

$$Y_{ij} = \mu_i + \gamma_i X_{ij} + \varepsilon_{ij}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, n_i,$$

donde μ_i y γ_i son, respectivamente, el intercepto y la pendiente de la regresión de Y sobre X para el grupo i-ésimo. A continuación, con el objetivo de ver como encaja el ANCOVA lineal con interacción en el modelo (1.2) se tomará de nuevo una muestra de tamaño n bajo diseño fijo y se procederá a exponerlo en forma matricial:

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 & x_{1,1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \cdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 & x_{1,n_1} & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & x_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & 0 & x_{2,n_2} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \cdots & 0 & 1 & \vdots & \vdots & \cdots & 0 & x_{I,1} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 & 0 & \cdots & 0 & x_{I,n_I} \end{pmatrix} \cdot \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \\ \gamma_1 \\ \vdots \\ \gamma_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \varepsilon_{I1} \\ \vdots \\ \varepsilon_{In_I} \end{pmatrix}. \tag{1.9}$$

En este caso, como los estimadores son iguales a los que se obtendrían al realizar una regresión simple en cada uno de los grupos, ya no se explicitarán.

Contraste del efecto de la interacción

Con la introducción del modelo (1.8) surge la duda de cuándo se debe plantear un modelo con interacción y cuándo no. Para responder a estas preguntas se lleva a cabo el siguiente contraste sobre el modelo (1.8):

$$\begin{cases} H_0 : \delta_i = 0 \ \forall\ i, \\ H_a : \exists\ j \text{ tal que } \delta_j \neq 0. \end{cases} \tag{1.10}$$

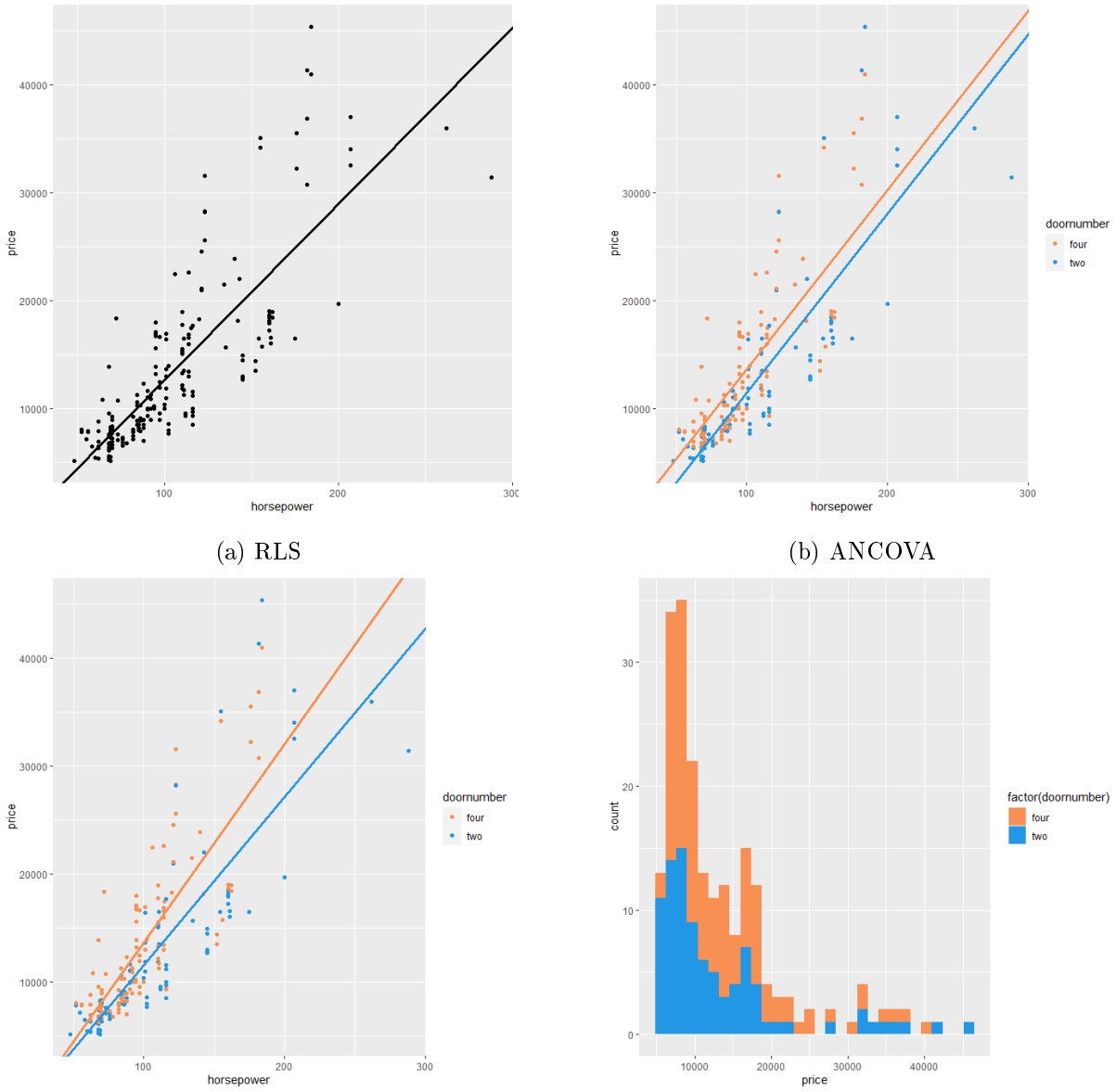
En este contraste, H_0 representa el caso en el que no hay efecto del grupo sobre la pendiente (ANCOVA sin interacción) y H_a el caso en que sí hay efecto del grupo sobre la pendiente (ANCOVA con interacción).

Para llevarlo a cabo se empleará el denominado test F, basado en el siguiente estadístico:

$$\frac{(RSS_0 - RSS)/(I-1)}{RSS/(n-2I)} \in F_{I-1,n-2I},$$

donde RSS_0 es el RSS de la sección anterior y RSS es la suma de I sumas residuales de cuadrados correspondientes a las I regresiones simples de Y sobre X. En consecuencia, RSS_0 y RSS representan la suma de los residuos al cuadrado a los que daría lugar un ANCOVA sin interacción y un ANCOVA con interacción respectivamente. En este caso, RSS es el error al que daría lugar si los datos se ajustasen mediante una recta para cada grupo por regresión simple, es decir, si el grupo no solo influyese en la ordenada en el origen sino también en la pendiente. En consecuencia, el test justifica si la reducción del error que se produce al emplear el ANCOVA con interacción en lugar del ANCOVA sin interacción vale la pena, teniendo en cuenta los grados de libertad.

A continuación, para ver la utilidad de los modelos (1.4) y (1.8) frente al modelo de regresión lineal simple en algunas situaciones, se pondrá un ejemplo del conjunto de datos [Manish Kumar, 2019]. Este consiste en un listado de coches del mercado americano y sus principales características (caballos de potencia, tamaño del motor, número de puertas...) las cuales se adjuntan con la intención de predecir el precio de un coche a partir de las mismas. En la Figura 1.1 se presentan unas gráficas que, al intentar explicar el precio de los coches (*price*) en función de la potencia (*horsepower*) y el número de puertas (*doornumber*), reflejan las ventajas de los modelos anteriormente mencionados respecto al modelo lineal general. En la Figura 1.1a se puede observar que la aproximación lineal, sin tener en cuenta la variable *doornumber*, es bastante buena. No obstante, tras separar las observaciones por colores en función del número de puertas, sería razonable considerar el efecto de la variable discreta. Es por ello que se muestran también la Figura 1.1b y la Figura 1.1c, las cuales representan un modelo ANCOVA, y un modelo ANCOVA con interacción respectivamente. En base a lo expuesto en estas gráficas, se intuye que la variable discreta tiene un claro efecto en el intercepto y al menos un pequeño efecto en la pendiente y, para ver si son significativos, se realizaron los contrastes (1.7) y (1.10), los cuales devolvieron p-valores de 0.0009 y 0.0986 respectivamente. En consecuencia, aunque hay efecto de la variable discreta sobre el intercepto, su efecto sobre la pendiente no es suficientemente significativo, por lo que podemos afirmar que estamos ante un caso de rectas paralelas. Por último, en la Figura 1.1d se puede observar un histograma del precio en donde las observaciones aparecen separadas en función del número de puertas.



(a) RLS

(b) ANCOVA

(c) ANCOVA con interacción

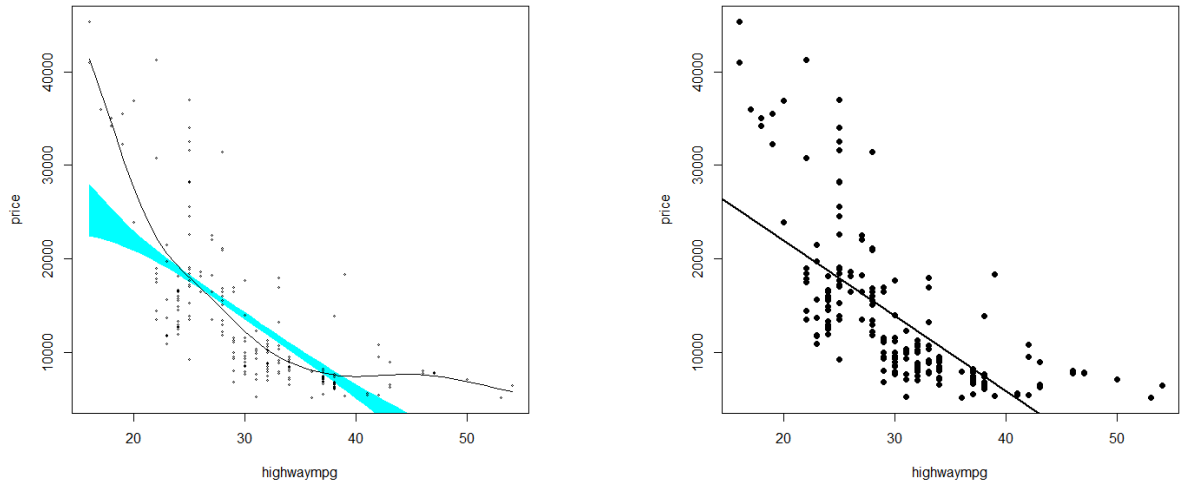
(d) Histograma del precio

Figura 1.1: Diagramas de dispersión del precio con respecto las distintas variables de estudio. Histograma del precio, separando las observaciones en función del número de puertas.

1.4. Objetivos y organización del trabajo

Ahora que se entiende en qué situaciones es útil emplear el ANCOVA lineal, surge la siguiente pregunta. ¿Qué sucede si la relación entre la variable explicativa y la respuesta no es lineal? Esta pregunta es muy natural, de hecho, en el conjunto de datos introducido previamente, la relación

entre la variable precio y las millas por galón en autopista (*highwaympg*) es no lineal. Esto se puede observar, tanto en la Figura 1.2a, pues la función de regresión no está contenida en el área azul, como en la Figura 1.2b, en la cual es evidente que el ajuste lineal no es adecuado. Es por ello que surge el ANCOVA no paramétrico que se expondrá en el siguiente capítulo. Con la introducción de este nuevo modelo se tiene como objetivo obtener una mejor estimación de la función de regresión en conjuntos de datos en los que el ajuste lineal no es adecuado. No obstante, antes de presentar este nuevo modelo, se van a proponer primero los métodos para estimar la función de regresión de forma no paramétrica (sin imponer una forma paramétrica concreta). Después, se presentarán un test de no efecto y el modelo ANCOVA no paramétrico, con sus contrastes asociados. Por último, en el tercer capítulo, se llevará a cabo un estudio de simulación con distintos tipos de datos generados por ordenador, con múltiples tamaños muestrales y con diferentes varianzas y distribuciones para el error (que recordemos, se suele suponer normal). Con estas simulaciones se intentará poner de manifiesto en qué casos produce un mejor ajuste emplear el ANCOVA no paramétrico frente al modelo lineal y viceversa. De esta forma, no solo se introducen el modelo y sus ventajas de forma teórica, sino que se llevan a cabo simulaciones que permiten obtener una intuición práctica de cuando un modelo es más favorable que el otro. Por último, con el objetivo de demostrar la utilidad del modelo no paramétrico en un caso real, se usarán los contrastes sobre el conjunto de datos [Manish Kumar, 2019] introducido previamente.



(a) `sm.regression` (b) Ajuste lineal

Figura 1.2: Gráfico del test de linealidad realizado por la función `sm.regression` del paquete `sm` de `R` que se encuentra en [Bowman y Azzalini, 2021] y diagrama de dispersión con ajuste lineal.

Capítulo 2

ANCOVA no paramétrico

En este capítulo se presentará detalladamente el ANCOVA no paramétrico, pero antes, con la intención de facilitar el correcto estudio del mismo, se expondrán algunos estimadores no paramétricos de la regresión, destacando entre ellos al estimador local lineal, el cual será el utilizado a la hora de ajustar el modelo ANCOVA no paramétrico.

En consecuencia, este capítulo se dividirá en dos secciones: en la primera sección se introducirán los estimadores no paramétricos de la regresión, concretamente, se hará hincapié en el Nadaraya-Watson, el Gasser-Müller y por último, el más empleado en el ANCOVA no paramétrico, el estimador local lineal. En esta sección también se expondrán sus ventajas e inconvenientes, así como algunas de sus propiedades más importantes. En la segunda sección, se estudiará en detalle el ANCOVA no paramétrico y se explicitarán los procedidmientos de estimación y sus contrastes asociados.

2.1. Estimador no paramétrico de la función de regresión

En esta sección se introducirán algunos estimadores de la función de regresión para así poder ver sus ventajas e inconvenientes en el contexto del modelo ANCOVA no paramétrico. No obstante, antes de introducir los estimadores habrá que plantear el modelo de regresión no paramétrico, véase [Bowman y Azzalini, 1997]. Por simplicidad, se planteará para una única variable explicativa X y bajo diseño fijo, de esta forma, dada una muestra de observaciones $(x_1, Y_1), ..., (x_n, Y_n)$, siendo Y la variable respuesta, el modelo se puede expresar como:

$$Y = m(X) + \varepsilon, \tag{2.1}$$

con $m(x) = \mathbb{E}(Y|X = x)$ la función de regresión, y donde ε representa el error, siendo los errores de las distintas observaciones independientes entre sí y tales que $\mathbb{E}(\varepsilon) = 0$ y $\mathrm{Var}(\varepsilon) = \sigma^2$.

Una vez planteado el modelo surge el problema de cómo estimar $m(x)$. Esta misma cuestión ha sido abordada por múltiples autores a lo largo de los años, por lo que, a continuación, se presentan tres de los estimadores más relevantes.

El primer estimador que se introducirá será el estimador de Nadaraya-Watson, el cual se formula de la siguiente manera, véase [Nadaraya, 1964]. Supongamos el modelo (2.1), el estimador de Nadaraya-Watson de la función de regresión es:

$$\hat{m}_{NW,h}(x) = \frac{\sum_{j=1}^{n} Y_j K_h(x - x_j)}{\sum_{j=1}^{n} K_h(x - x_j)}, \tag{2.2}$$

donde h es el parámetro de suavizado o ancho de banda y su elección afecta considerablemente al resultado de la estimación. Además, $K_h(x - x_j) = \frac{1}{h} K\left(\frac{u - x}{h}\right)$, siendo K la denominada función núcleo que, según [Wand y Jones, 1994], es una función de densidad unimodal, simétrica y centrada en cero. Por último, es importante destacar que según [Wand y Jones, 1994] este estimador fue concebido originalmente para diseño aleatorio, lo que provoca que sus propiedades se obtengan condicionadas a la muestra.

El siguiente estimador es el de Gasser-Muller, al que de ahora en adelante se denotará como $\hat{m}_{GM,h}$. Este fue introducido en [Gasser y Müller, 1979] y tiene la siguiente expresión:

$$\hat{m}_{GM,h}(x) = \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} K_h(u - x)\, du Y_j, \tag{2.3}$$

con $s_j = (x_{j+1} + x_j)/2 \;\; \forall\, j \in \{1, \ldots, n-1\}$, $s_0 = -\infty$, $s_n = \infty$ y $K_h(u - x) = \frac{1}{h} K\left(\frac{u - x}{h}\right)$. De nuevo, h es el parámetro de suavizado o ancho de banda que se deberá elegir adecuadamente para obtener un estimador con el menor error posible.

Este estimador, en lugar de hacer depender los distintos pesos de los coeficientes, de la altura de la función núcleo reescalada, los hace depender del área bajo dicha función, consiguiendo así unas propiedades más interesantes que las de (2.2), como se presentará en breve. Otra diferencia entre el estimador (2.2) y el (2.3) es que este último sí fue concebido originalmente para diseño fijo, véase [Gasser y Müller, 1979]. Por último, antes de hacer un resumen de sus propiedades más importantes, se hará una introducción del estimador lineal local. Este estimador consiste en el ajuste de una recta de regresión de manera local, asignándole un peso distinto a cada observación en función de la distancia al punto del cual se quiere estimar la respuesta. Este peso se asigna mediante una función núcleo y, el estimador, tal y como se ve en [Wasserman, 2006], se obtiene de la siguiente manera:

$$\hat{m}_{LL}(x) = \hat{\beta}_0(x), \tag{2.4}$$

siendo $\hat{\boldsymbol{\beta}}(x) = \begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = \left(\boldsymbol{X}_x^T \boldsymbol{W}_{x,h} \boldsymbol{X}_x\right)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_{x,h} \boldsymbol{Y}$,

con $\boldsymbol{X}_x = \begin{pmatrix} 1 & x_1 - x \\ 1 & x_2 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}$ y $\boldsymbol{W}_{x,h} = \begin{pmatrix} K_h(x_1 - x) & 0 & \cdots & \cdots & 0 \\ 0 & K_h(x_2 - x) & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & K_h(x_n - x) \end{pmatrix}$.

La matriz $\boldsymbol{W_{x,h}}$ está determinada por h, al que se denota como parámetro de suavizado o de ancho de banda, por lo que, al igual que en el caso de los estimadores previos, la elección de este parámetro es determinante para el resultado de la estimación. En la Tabla 2.1 se pueden ver en detalle las expresiones de los sesgos y varianzas de los distintos estimadores mencionados anteriormente:

MÉTODO	SESGO	VARIANZA
Nadaraya-Watson	$\left(m''(x) + \dfrac{2m'(x)f'(x)}{f(x)}\right) b_n$	V_n
Gasser-Müller	$m''(x)b_n$	$1.5V_n$
Lineal Local	$m''(x)b_n$	V_n

Tabla 2.1: Comparación de sesgos y varianzas de los distintos estimadores, véase [Fan y Gijbels, 1996].
En la Tabla 2.1 la notación empleada fue: $b_n = \dfrac{h^2}{2} \int_{-\infty}^{\infty} u^2 K(u)du$, $V_n = \dfrac{\sigma^2}{f(x)nh} \int_{-\infty}^{\infty} K^2(u)du$.
y $f(x)$ la función de densidad de X.

Se concluye entonces que el estimador que se usará de ahora en adelante será el estimador lineal local, pues al contrario que el Nadaraya-Watson o el Gasser-Müller, en vez de ajustar localmente rectas horizontales, ajusta rectas con pendientes no necesariamente nulas, lo que le confiere mejores propiedades. Además, a la vista de la Tabla 2.1, si se compara con el estimador de Nadaraya-Watson, el lineal local tiene un sesgo menor, pues el sesgo del Nadaraya-Watson aumenta mucho si f'/f es un valor grande, siendo f la función de densidad de la variable X. Es importante resaltar que, si $f(x)$ fuese nula, entonces indica que no hay densidad de puntos de la explicativa (o de las explicativas, en un contexto multidimensional), por lo que no tiene

sentido calcular la esperanza condicionada. Además, técnicamente se puede ver que la baja o nula densidad dificulta o imposibilita algunos cálculos, por lo que no tendría sentido calcular $m(x)$ en dichos puntos. Por otra parte, aunque el de Gasser-Müller tenga el mismo sesgo que el lineal local, tiene mayor variabilidad cuando no se trabaja sobre diseño fijo. Por último, ambos estimadores, tanto el (2.2) como el (2.3), sufren un aumento importante en el sesgo de aquellos puntos situados en la frontera del soporte de la variable explicativa, mientras que el lineal local no, por lo que, en general, es un estimador con mejores propiedades. No obstante, sí es posible emplear los estimadores (2.2) y (2.3) para estimar la función de regresión del ANCOVA no paramétrico pero, el primero, al tener un sesgo que depende de la distribución de la variable explicativa, no permitirá simplificar los estadísticos de contraste, y el segundo, a pesar de no tener este problema, posee peores propiedades que el local lineal, por lo que no se suelen emplear ninguno de ellos para este fin. En consecuencia, tal como se dijo anteriormente, el estimador local lineal será el que se usará de ahora en adelante.

Como para cualquier estimador de la función de regresión, elegir un valor adecuado de h es esencial para obtener un buen ajuste, pero esto no es tarea fácil, pues tanto valores pequeños como grandes de h, proporcionan un ajuste inadecuado. Los valores pequeños de h harán que la función de regresión sea demasiado rugosa, es decir, que se ajuste demasiado a las observaciones. Esto genera un bajo sesgo, pero una alta varianza, ya que cuanto menor sea el h, más se parecerá la regresión a una interpolación. Si por el contrario, h presenta un valor grande, la función de regresión será más suave y estará más alejada de las observaciones, por lo que tendrá un mayor sesgo, pero una menor varianza. De hecho, si el h toma valores muy grandes, la función de regresión se aproxima a una regresión lineal. Es por estos motivos que interesa emplear el h óptimo, es decir, un parámetro de suavizado que haga un balance adecuado entre sesgo y varianza para minimizar el error. Según [Fan, 1992] el h óptimo para reducir el MISE (Mean Integrated Squared Error) del estimador lineal local es:

$$h_{opt} = \left(\frac{R(K) \int_{-\infty}^{\infty} f^{-1}(x)\sigma^2 dx}{\left(\int_{-\infty}^{\infty} u^2 K(u)\right)^2 \int_{-\infty}^{\infty} [m''(x)]^2 \, dx} \right)^{1/5} n^{-1/5}, \text{ con } R(K) = \int_{-\infty}^{\infty} K^2(u)du.$$

Sin embargo, aunque de forma teórica este es el parámetro de suavizado óptimo, es difícilmente calculable en la práctica, porque tanto la función $f(x)$, como $m(x)$, son en general desconocidas. En consecuencia, en la práctica se emplea un h escogido por validación cruzada, una técnica que se detallará a continuación. Dada una muestra de observaciones con sus correspondientes respuestas, se lleva a cabo la predicción de cada observación empleando todos los datos de la muestra excepto el de dicha observación y se calcula el error entre esta predicción y el valor conocido de la respuesta. A continuación, se calcula la suma de estos errores al cuadrado y se escoge el parámetro de suavizado que minimiza dicha suma. Este método de elección del parámetro será el empleado en el estudio de simulación del Capítulo 3 y se puede sintetizar de

la siguiente forma:

$$h_{cv} = \min_h \sum_{j=1}^{n} \left(Y_j - \hat{Y}_{j(j)}\right)^2,$$

donde h_{cv} es el h obtenido por validación cruzada, y_j es la respuesta de la observación j-ésima, e $\hat{y}_{j(j)}$ es la predicción de la j-ésima observación de la muestra, x_j, sobre el modelo que excluye a la j-ésima observación.

2.2. Test de no efecto

Este test se denomina test de no efecto, ya que comprueba si la variable respuesta y la variable explicativa están relacionadas entre sí, es decir, comprueba si hay efecto de la explicativa en la respuesta, véase [Bowman y Azzalini, 1997]. La formulación de este test se plantea de la siguiente manera:

$$\begin{cases} H_0 : Y_j = \mu + \varepsilon_j, \\ H_a : Y_j = m(X_j) + \varepsilon_j, \text{ con } m(X_j) \neq \mu \text{ para algún } j \in \{1, \ldots, n\}. \end{cases} \tag{2.5}$$

En este contraste, H_0 representa el caso en que no hay efecto de la variable explicativa sobre la respuesta y H_a el caso en que sí hay efecto de la variable explicativa sobre la respuesta. Además, $\varepsilon_j \in N\left(0,\ \sigma^2\right)$ independientes para todo $j \in \{1, \ldots, n\}$, con $n = \sum_1^I n_i$.

El estadístico empleado para este test es, exceptuando que no tiene en cuenta los grados de libertad, análogo al empleado en el caso paramétrico, y se describe de la siguiente forma:

$$F = \frac{RSS_0 - RSS}{RSS},$$

siendo $RSS_0 = \sum_{j=1}^{n} \left(Y_j - \bar{Y}\right)^2$, $RSS = \sum_{j=1}^{n} \left[Y_j - \hat{m}(X_j)\right]^2$ y $\hat{m}$ es el estimador lineal local. Como se ha comentado previamente, la idea detrás del estadístico es muy similar a la del test (1.6) del modelo paramétrico. Consiste en valorar si la reducción del error que se produce al emplear la regresión no paramétrica como estimación en lugar de la media, compensa el aumento en la complejidad. Esto se hace calculando el cociente del aumento de la suma de los errores al cuadrado al pasar de H_0 a H_a, con respecto a la suma de los residuos al cuadrado bajo H_a. En consecuencia, el estadístico ya no requiere una división entre sus grados de libertad pues ya tiene la escala de σ. A continuación, se explicitará la distribución del estadístico bajo H_0, pero para ello será útil escribir las sumas de residuos al cuadrado en forma matricial:

$$RSS_0 = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{L})^T(\boldsymbol{I} - \boldsymbol{L})\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{L})\boldsymbol{Y},$$

$$RSS = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})\boldsymbol{Y},$$

donde $\boldsymbol{L} \in \mathcal{M}_{n\times n}(\mathbb{R})$ es una matriz en la que todos los elementos son $1/n$, y $\boldsymbol{S}$ es la matriz de suavizado que se requiere para obtener $\{\hat{m}(X_j)\}$, y que se explicitará en la sección (2.3.1). En consecuencia, el estadístico se puede expresar como:

$$F = \frac{\boldsymbol{Y^T B Y}}{\boldsymbol{Y^T A Y}}, \tag{2.6}$$

con $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{S})^T(\boldsymbol{I} - \boldsymbol{S})$ y $\boldsymbol{B} = \boldsymbol{I} - \boldsymbol{L} - \boldsymbol{A}$, por consiguiente el p-valor se calcula como:

$$p = \mathbb{P}\left(\frac{\boldsymbol{Y^T B Y}}{\boldsymbol{Y^T A Y}} > F_{Obs}\right) = \mathbb{P}\left[\boldsymbol{Y^T}\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{Y} > 0\right],$$

donde F_{Obs} es el valor observado del estadístico (2.6). Se ha obtenido entonces, una forma cuadrática en variables normales del tipo $\boldsymbol{z^T M z}$, donde $\boldsymbol{M} \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica, ya que, tanto $\boldsymbol{A}$ como $\boldsymbol{B}$ son simétricas. Sin embargo, el resultado de [Johnson et al. 1995] que permite dar una distribución aproximada para el estadístico, requiere que la media de $\boldsymbol{z}$ (variable normal) sea $\boldsymbol{0}$. Como bajo H_0 podemos sustituir $\boldsymbol{Y}$ por $\boldsymbol{\mu} + \boldsymbol{\varepsilon}$, se puede ver que:

$$\boldsymbol{Y}^T\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{Y} = \boldsymbol{\mu}^T\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{\mu} + \boldsymbol{\varepsilon}^T\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{\varepsilon}$$

y por construcción de las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ el primer sumando desaparece, lo que permite expresar p como:

$$p = \mathbb{P}\left[\boldsymbol{\varepsilon}^T\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{\varepsilon} > 0\right],$$

por lo que ahora $\boldsymbol{\varepsilon}^T\left(\boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}\right)\boldsymbol{\varepsilon}$ es una forma cuadrática de variables normales con $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ y $\boldsymbol{M} = \boldsymbol{B} - \boldsymbol{A}\cdot F_{Obs}$ simétrica, estando así en las condiciones del resultado previamente mencionado. Esto permite afirmar que, bajo H_0, la distribución puede aproximarse por la de una χ^2 reescalada y recentrada, equivalentemente, $a\chi^2(b) + c$, pues esta suele ser adecuada para estadísticos con formas cuadráticas. Consecuentemente, con el fin de escoger los valores de a, b y c idóneos para aproximar el estadístico, se obligará a la distribución a compartir el valor de los 3 primeros cumulantes de la distribución original. Para ello se emplea un resultado de [Johnson et al. 1995], el cual afirma que el k-ésimo cumulante de la distribución de una forma cuadrática como la mencionada anteriormente viene dado por:

$$\nu_k = 2^{(k-1)}(k-1)!\mathrm{tr}\left[\left(\boldsymbol{VM}\right)^k\right], \text{ donde } \boldsymbol{V} = \mathrm{cov}(\boldsymbol{z})$$

De esta forma, se obtiene una distribución con la misma media, varianza y asimetría que la distribución real de $\boldsymbol{\varepsilon}^T\boldsymbol{M}\boldsymbol{\varepsilon}$ y los parámetros a, b y c se calculan de la siguiente manera:

$$a = \frac{|\nu_3|}{4\nu_2} \qquad b = \frac{8\nu_2^3}{\nu_3^2} \qquad c = \nu_1 - ab,$$

lo que permite calcular el p-valor como: $p = P(a\chi^2(b) + c \geq 0)$.

2.3. Introducción al ANCOVA no paramétrico

Tras la introducción del ANCOVA lineal en el Capítulo 1, es natural preguntarse qué sucede cuando existe una variable categórica que influye en la función de regresión si la relación entre la variable explicativa continua y la respuesta no es lineal y además, no puede asumirse un modelo paramétrico. Es ahí donde surge el modelo no paramétrico, ya que permite tener en cuenta, tanto el efecto de la variable categórica, como el de la continua, sea cual sea su relación con la respuesta, haciéndolo así mucho más versátil. El modelo, tal como se plantea en [Young y Bowman, 1995] puede adoptar cualquiera de las siguientes tres formas, en función de como afecte la variable discreta a la respuesta. La primera de ellas es:

$$Y_{ij} = m_i(x_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \in N\left(0,\ \sigma^2\right), \tag{2.7}$$

y se adoptará si la variable categórica tiene algún efecto sobre la respuesta, es decir, si el modelo es un ANCOVA no paramétrico. Cabe destacar que, dentro de este caso se puede distinguir un caso particular, el cual se formula de la siguiente manera:

$$Y_{ij} = \alpha_i + m(x_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \in N\left(0,\ \sigma^2\right). \tag{2.8}$$

Este caso representa las situaciones en las que el efecto del grupo se traduce en la suma de una constante a la función de regresión, pero no produce ningún efecto en la forma de la misma. En otras palabras, serían funciones de regresión paralelas. Asimismo, es importante recalcar que, tanto $m(x_{ij})$, como $m_i(x_{ij})$ representan a $\mathbb{E}\left(Y|X = x_{ij}\right)$, con la diferencia de que, para llevar a cabo la media, $m(x_{ij})$ emplea todas las observaciones, mientras que $m_i(x_{ij})$ tan solo las del grupo i-ésimo.

Por último, la tercera y última forma posible que puede admitir el modelo es:

$$Y_{ij} = m(x_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \in N\left(0,\ \sigma^2\right),$$

y esta es la que seguirá cuando no hay un efecto de la variable categórica sobre la respuesta, es decir, si en realidad el modelo, más que un ANCOVA no paramétrico, es en realidad un modelo de regresión como el planteado en (2.1). A continuación, con el objetivo de determinar cual de los modelos previamente introducidos es el que mejor ajusta una muestra de observaciones, se desarrollarán los estadísticos de contraste.

Dada una muestra de observaciones de una variable respuesta Y y dos variables explicativas X y X_g como las del capítulo anterior, con el objetivo de saber cuál de los modelos enumerados anteriormente se ajusta mejor a los datos, se plantean los siguientes contrastes. En primer lugar, se lleva a cabo un test para comprobar si hay efecto de la variable categórica en la función

de regresión, el cual es análogo al contraste (1.7). En caso de que el test concluya que existen pruebas significativas para afirmar que la variable categórica produce un efecto relevante, tendrá sentido plantear el segundo test. Este consiste en comprobar si el efecto de la variable categórica se traduce en el paralelismo de las funciones de regresión o si modifica la forma de estas. A continuación, se expondrán los contrastes planteados y se detallarán tanto los estadísticos de los test como sus distribuciones.

2.3.1. Test de igualdad

Este test se denomina test de igualdad, pues comprueba la igualdad de las funciones de regresión planteadas por separado para cada uno de los grupos definidos por la variable categórica. Es por ello que, en el fondo, la comprobación es una versión no paramétrica del test (1.7) del ANCOVA lineal. En caso de que el test no detecte una diferencia significativa entre las funciones de regresión por grupos, se asume que se puede plantear una única función de regresión para toda la muestra. Este test se plantea de la siguiente manera:

$$\begin{cases} H_0 : Y_{ij} = m(X_{ij}) + \varepsilon_{ij}, \\ H_a : Y_{ij} = m_i(X_{ij}) + \varepsilon_{ij}. \end{cases} \tag{2.9}$$

En este contraste, H_0 representa el caso en el que no hay efecto del grupo sobre la función de regresión y H_a el caso en que sí hay efecto del grupo sobre la función de regresión. Además, $\varepsilon_{ij} \in N\left(0,\ \sigma^2\right)$ independientes para todo $i \in \{1, \ldots, I\} \quad j \in \{1, \ldots, n_i\}$.

El estadístico empleado en [Young y Bowman, 1995] para el test es:

$$\mathrm{TS}_I = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} \left[\hat{m}_i(X_{ij}) - \hat{m}(X_{ij})\right]^2}{\hat{\sigma}^2}, \tag{2.10}$$

donde $\hat{m}(x)$ y $\hat{m}_i(x)$ son los estimadores lineales locales de la regresión global y de la regresión en cada grupo respectivamente, y $\hat{\sigma}^2$ es un estimador de la varianza que se explicitará más adelante. La ventaja que posee tanto este estimador lineal local, como el de Gasser-Müller, frente al Nadaraya-Watson es que, tal como se expuso en la Tabla 2.1, el sesgo del Nadaraya Watson es $\left(m''(x) + \dfrac{2m'(x)f'(x)}{f(x)}\right) b_n$ mientras que el sesgo de los otros dos estimadores es $m''(x)b_n$. En consecuencia, bajo H_0, donde se supone que las funciones de regresión son iguales en todos los grupos, se puede observar que, en el numerador del estadístico (2.10) los términos que corresponden al sesgo se cancelan, al menos asintóticamente, para los estimadores (2.3) y (2.4). Esto se debe a que, el sesgo de estos estimadores únicamente depende de b_n y de $m_i(x)$, siendo b_n igual en todos los grupos, y siendo $m(x) = m_i(x) \quad \forall i \in \{1, \ldots, I\}$ por hipótesis. Sin embargo, esto no sucede con el estimador (2.2), a menos que las distribuciones de la X sean idénticas en todos sus grupos, pues como su sesgo depende de $f(x)$, la función de densidad de X,

en cada grupo no tiene por qué ser la misma, consecuentemente, a menos que todos los grupos tengan la misma distribución, los términos del sesgo no serán cancelables. A mayores, como se ha estudiado en la Sección 2.1 que el estimador de Gasser-Müller posee peores propiedades que el lineal local, tiene sentido escoger este último para estimar $m(x)$ y $m_i(x)$, pues es sin duda el más adecuado.

Por otra parte, se puede observar que el estadístico (2.10) requiere calcular $\hat{\sigma}^2$, por lo que a continuación se explicitará como se obtiene este estimador de la varianza:

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i-1)\,\hat{\sigma}_i^2, \tag{2.11}$$

con $\hat{\sigma}_i^2$ como en [Rice, 1984], donde tiene la siguiente expresión:

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i-1)}\sum_{j=1}^{n_i-1}\left(y_{i,[j+1]}-y_{i[j]}\right)^2,$$

siendo n el número total de observaciones de la muestra, o equivalentemente, $n=\sum_{i=1}^{I} n_i$. Además, $Y_{i[j]}$ denota el valor de la respuesta correspondiente a $X_{i[j]}$, con $X_{i[j]}$ el j-ésimo valor más grande en X del grupo i-ésimo. Es importante destacar que, para poder emplear esta estimación de la varianza, se ha de asumir la homocedasticidad de los datos. Por otra parte, si bien este estimador es totalmente válido, según [Bowman y Azzalini, 1997] este valor de la varianza puede verse inflado por la forma de la función de regresión, y en consecuencia, sería mejor tomar otro estimador de $\hat{\sigma}^2$. La propuesta que hacen para evitar el problema de la inflación de la varianza es emplear un estimador propuesto previamente en [Gasser et al. 1986], el cual corrige este problema empleando pseudoresiduos, que son cantidades basadas en la diferencia entre el $Y_{i[j]}$ y la recta que une los puntos $\left(X_{i,[j-1]},Y_{i,[j-1]}\right)$ y $\left(X_{i,[j+1]},Y_{i,[j+1]}\right)$ evaluada en $X_{i,[j]}$. Veamos a continuación como se definen:

$$\tilde{\varepsilon}_{i,[j]} = \frac{X_{i,[j+1]}-X_{i,[j]}}{X_{i,[j+1]}-X_{i,[j-1]}}Y_{i,[j-1]}+\frac{X_{i,[j]}-X_{i,[j-1]}}{X_{i,[j+1]}-X_{i,[j-1]}}Y_{i,[j+1]}-Y_{i,[j]} = a_{i,[j]}Y_{i,[j-1]}+b_{i,[j]}Y_{i,[j+1]}-Y_{i,[j]}.$$

Partiendo de estos términos, se expresa el estimador de la varianza del grupo i-ésimo como:

$$\hat{\sigma}_i^2 = \frac{1}{n_i-2}\sum_{j=2}^{n_i-1} c_{i[j]}^2\tilde{\varepsilon}_{i[j]}^2, \tag{2.12}$$

con $c_{i[j]}^2 = (a_{i[j]}^2+b_{i[j]}^2+1)^{-1}$. Finalmente, empleando las varianzas de cada grupo, se puede estimar la varianza global de la siguiente forma:

$$\hat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}(n_i-2)\hat{\sigma}_i^2.$$

Una vez establecido como obtener el estadístico (2.10), se buscará qué distribución sigue. Es fácil ver que, tanto el numerador como el denominador son formas cuadráticas, pues se pueden

expresar en forma matricial de la siguiente manera. Sea $\boldsymbol{Y_i}$ el vector de las observaciones del grupo i-ésimo para la variable respuesta, y $\hat{\boldsymbol{m}}_{\boldsymbol{i}}$ el vector de valores ajustados para el grupo i-ésimo. Cada uno de los elementos de este vector de valores ajustados se puede expresar como:

$$\hat{\boldsymbol{m}}_{\boldsymbol{i}} = \boldsymbol{S_i Y_i}, \text{ donde } \boldsymbol{S_i} \in \mathcal{M}_{n_i \times n_i}(\mathbb{R}),$$

en concreto:

$$\boldsymbol{S_i} = \begin{pmatrix} \left[\left(\boldsymbol{X}^T_{\boldsymbol{x_{i,1}}} \boldsymbol{W}_{\boldsymbol{x_{i,1}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{i,1}}}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x_{i,1}}} \boldsymbol{W}_{\boldsymbol{x_{i,1}},\boldsymbol{h}}\right]_{1,\bullet} \\ \left[\left(\boldsymbol{X}^T_{\boldsymbol{x_{i,2}}} \boldsymbol{W}_{\boldsymbol{x_{i,2}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{i,2}}}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x_{i,2}}} \boldsymbol{W}_{\boldsymbol{x_{i,2}},\boldsymbol{h}}\right]_{1,\bullet} \\ \vdots \\ \left[\left(\boldsymbol{X}^T_{\boldsymbol{x_{i,n_i}}} \boldsymbol{W}_{\boldsymbol{x_{i,n_i}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{i,n_i}}}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x_{i,n_i}}} \boldsymbol{W}_{\boldsymbol{x_{i,n_i}},\boldsymbol{h}}\right]_{1,\bullet} \end{pmatrix},$$

siendo

$$\boldsymbol{X_x} = \begin{pmatrix} 1 & x_{i,1} - x \\ 1 & x_{i,2} - x \\ \vdots & \vdots \\ 1 & x_{i,n_i} - x \end{pmatrix} \text{ y } \boldsymbol{W_{x,h}} = \frac{1}{h} \begin{pmatrix} K\left(\frac{x_{i,1}-x}{h}\right) & 0 & \cdots & \cdots & 0 \\ 0 & K\left(\frac{x_{i,2}-x}{h}\right) & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & K\left(\frac{x_{i,n_i}-x}{h}\right) \end{pmatrix}$$

y donde $\left[\left(\boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}} \boldsymbol{X_x}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}}\right]_{1,\bullet}$ denota la primera fila de la matriz correspondiente, equivalentemente, $\left[\left(\boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}} \boldsymbol{X_x}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}}\right]_{1,\bullet} = (1,0) \cdot \left[\left(\boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}} \boldsymbol{X_x}\right)^{-1} \boldsymbol{X}^T_{\boldsymbol{x}} \boldsymbol{W_{x,h}}\right]$.

De esta forma, si se denota por $\hat{\boldsymbol{m}}$ a la colección de todos los valores ajustados, y por $\boldsymbol{Y}$ al vector con las observaciones de todos los grupos para la variable respuesta, entonces:
$\hat{\boldsymbol{m}} = \boldsymbol{S_d Y}$, donde $\boldsymbol{S_d} \in \mathcal{M}_{n \times n}(\mathbb{R})$ sería una matriz con I bloques, cada uno correspondiéndose con un grupo.

$$\boldsymbol{S_d} = \left(\begin{array}{c|c|c|c|c|c} \boldsymbol{S_1} & \boldsymbol{0} & \cdots & \cdots & \cdots & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{S_2} & \boldsymbol{0} & \cdots & \cdots & \boldsymbol{0} \\ \hline \vdots & \boldsymbol{0} & \boldsymbol{S_3} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \hline \vdots & \vdots & \boldsymbol{0} & \ddots & \ddots & \boldsymbol{0} \\ \hline \vdots & \vdots & \vdots & \ddots & \ddots & \boldsymbol{0} \\ \hline \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{S_I} \end{array}\right).$$

Por otra parte, el vector de valores ajustados bajo H_0 se puede escribir en forma matricial

como $\boldsymbol{S_s Y}$, siendo $\boldsymbol{S_s} \in \mathcal{M}_{n\times n}(\mathbb{R})$, en concreto:

$$\boldsymbol{S_s} = \begin{pmatrix} \left[\left(\boldsymbol{X}_{\boldsymbol{x_{1,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{1,1}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{1,1}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{1,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{1,1}},\boldsymbol{h}}\right]_{1,\bullet} \\ \vdots \\ \left[\left(\boldsymbol{X}_{\boldsymbol{x_{1,n_1}}}^T \boldsymbol{W}_{\boldsymbol{x_{1,n_1}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{1,n_1}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{1,n_1}}}^T \boldsymbol{W}_{\boldsymbol{x_{1,n_1}},\boldsymbol{h}}\right]_{1,\bullet} \\ \left[\left(\boldsymbol{X}_{\boldsymbol{x_{2,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{2,1}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{2,1}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{2,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{2,1}},\boldsymbol{h}}\right]_{1,\bullet} \\ \vdots \\ \left[\left(\boldsymbol{X}_{\boldsymbol{x_{2,n_2}}}^T \boldsymbol{W}_{\boldsymbol{x_{2,n_2}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{2,n_2}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{2,n_2}}}^T \boldsymbol{W}_{\boldsymbol{x_{2,n_2}},\boldsymbol{h}}\right]_{1,\bullet} \\ \vdots \\ \left[\left(\boldsymbol{X}_{\boldsymbol{x_{I,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{I,1}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{I,1}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{I,1}}}^T \boldsymbol{W}_{\boldsymbol{x_{I,1}},\boldsymbol{h}}\right]_{1,\bullet} \\ \vdots \\ \left[\left(\boldsymbol{X}_{\boldsymbol{x_{I,n_I}}}^T \boldsymbol{W}_{\boldsymbol{x_{I,n_I}},\boldsymbol{h}} \boldsymbol{X}_{\boldsymbol{x_{I,n_I}}}\right)^{-1} \boldsymbol{X}_{\boldsymbol{x_{I,n_I}}}^T \boldsymbol{W}_{\boldsymbol{x_{I,n_I}},\boldsymbol{h}}\right]_{1,\bullet} \end{pmatrix}, \tag{2.13}$$

donde la notación empleada es idéntica a la empleada en (2.4). Nótese la diferencia con la notación utilizada para este mismo estimador bajo H_a, ya que, tanto el tamaño de las matrices, como los valores de X son distintos. A mayores, cabe destacar también que, como $\mathbf{S_d}$ y $\mathbf{S_s}$ son matrices de suavizado, ambas requieren de la elección de un parámetro h. En el caso de $\mathbf{S_s}$, este se escogerá por validación cruzada, y se usará el mismo h en $\mathbf{S_d}$ para hacer posibles las simplificaciones en el numerador del estadístico que se detallarán más adelante.

Una vez escogidos los h, se concluye que el numerador de (2.10) se puede reescribir como:

$$\boldsymbol{Y}^T \left[\boldsymbol{S_d} - \boldsymbol{S_s}\right]^T \left[\boldsymbol{S_d} - \boldsymbol{S_s}\right] \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{Q} \boldsymbol{Y}, \text{ siendo } \boldsymbol{Q} = \left[\boldsymbol{S_d} - \boldsymbol{S_s}\right]^T \left[\boldsymbol{S_d} - \boldsymbol{S_s}\right].$$

Con el objetivo de ver la distribución del estadístico (2.10) bajo H_0, reemplazamos $\boldsymbol{Y}$ por $\boldsymbol{m}+\boldsymbol{\varepsilon}$, en consecuencia, al haber escogido los parámetros de suavizado de S_d y de S_s iguales, varios términos del numerador desaparecen asintóticamente gracias a las propiedades del sesgo del estimador local lineal y el numerador se simplifica a $\boldsymbol{\varepsilon}^T \boldsymbol{Q} \boldsymbol{\varepsilon}$. Por otra parte, $\hat{\sigma}^2 = \boldsymbol{Y}^T \boldsymbol{B} \boldsymbol{Y}$, siendo la $\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\boldsymbol{\varepsilon}^T \boldsymbol{B} \boldsymbol{\varepsilon}\right) + \boldsymbol{m}^T \boldsymbol{B} \boldsymbol{m}$ con $\boldsymbol{B} \in \mathcal{M}_{n\times n}(\mathbb{R})$ una matriz simétrica, en concreto, se puede expresar la matriz $\boldsymbol{B}$ de las dos siguientes formas, tal como se ve en [Alonso-Pena, 2019]:
Si se emplea el estimador (2.11), la matriz $\boldsymbol{B}$ está compuesta por I bloques de tamaño $n_i \times n_i$ en la que cada uno de los bloques es de la siguiente forma:

$$\frac{1}{2n-2I}\begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & & \vdots \\ 0 & -1 & 2 & -1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

Sin embargo, si el estimador empleado es el (2.12), la matriz $\boldsymbol{B} = \boldsymbol{A}^T\boldsymbol{A}$, siendo $\boldsymbol{A}$ una matriz formada por I bloques de tamaño $n_i \times n_i$ en la que el i-ésimo bloque es:

$$\frac{1}{\sqrt{n-I}}\begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ a_{i[2]}/c_{i[2]} & -1/c_{i[2]} & b_{i[2]}/c_{i[2]} & \ddots & & & \vdots \\ 0 & a_{i[3]}/c_{i[3]} & -1/c_{i[3]} & b_{i[3]}/c_{i[3]} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & a_{i[n_i-1]}/c_{i[n_i-1]} & -1/c_{i[n_i-1]} & b_{i[n_i-1]}/c_{i[n_i-1]} \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \end{pmatrix}.$$

Además, el último término del sumando, el cual representa la suma de las diferencias sucesivas al cuadrado de la función $m(x)$, es pequeño en comparación con el primero, por lo que, tal como se expone en [Young y Bowman, 1995], ignorar este término unicamente hace el test más robusto. En consecuencia, el p-valor es casi equivalente a:

$$p = \mathbb{P}\left(\frac{\boldsymbol{\varepsilon}^T\boldsymbol{Q}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T\boldsymbol{B}\boldsymbol{\varepsilon}} > F_{Obs}\right) = \mathbb{P}\left[\boldsymbol{\varepsilon}^T\left(\boldsymbol{Q} - \boldsymbol{B}\cdot F_{Obs}\right)\boldsymbol{\varepsilon} > 0\right],$$

donde F_{Obs} es el valor observado del estadístico (2.10). Finalmente, se ha obtenido una forma cuadrática en variables normales del tipo $\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}$, donde $\mathbb{E}(\boldsymbol{z}) = \boldsymbol{0}$ y $\boldsymbol{A} \in \mathcal{M}_{n\times n}(\mathbb{R})$ simétrica. Esto permite afirmar que bajo H_0, la distribución que sigue el estadístico es, aproximadamente, una χ^2 reescalada y recentrada, equivalentemente, $a\chi^2(b) + c$, por lo que de forma análoga a la de la sección (2.2), se obtiene una distribución con la misma media, varianza y asimetría que la del estadístico del test y que permite calcular el p-valor de la forma usual.

2.3.2. Test de paralelismo

Este test se denomina test de paralelismo, pues comprueba si las funciones de regresión que se calculan para cada uno de los grupos de la variable categórica son paralelas o no, por lo que,

en el fondo, la comprobación es una versión no paramétrica del test (1.10) del ANCOVA lineal. En caso de que el test detecte una diferencia significativa entre las formas de las funciones de regresión por grupos, se asume que no se pueden plantear funciones paralelas, sino que habrá que plantear funciones independientes para cada grupo de la muestra. Este test se plantea de la siguiente manera:

$$\begin{cases} H_0 : Y_{ij} = \alpha_i + m(X_{ij}) + \varepsilon_{ij}, \qquad \alpha_1 = 0, \\ H_a : Y_{ij} = m_i(X_{ij}) + \varepsilon_{ij}, \text{ con } m_j(\cdot) \neq m_k(\cdot) + \alpha \text{ para algún } j,k \in \{1,\ldots,I\}, \forall\ \alpha \in \mathbb{R}. \end{cases} \tag{2.14}$$

Aquí, H_0 representa el caso en el que todas las funciones son paralelas y H_a el caso en que algunas de estas no son paralelas entre sí. Además, $\varepsilon_{ij} \in N\left(0,\ \sigma^2\right)$ independientes para todo $i \in \{1,\ldots,I\} \quad j \in \{1,\ldots,n_i\}$.

El estadístico empleado para el test es:

$$\mathrm{TS}_P = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left[\widehat{\alpha}_i + \hat{m}(X_{ij}) - \hat{m}_i(X_{ij})\right]^2}{\hat{\sigma}^2}, \tag{2.15}$$

donde $\hat{\sigma}^2$ puede escogerse como en la sección anterior. Cabe destacar que, antes de calcular el valor del estadístico (2.15) ha de conocerse tanto $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_2,\ldots,\widehat{\alpha}_I)^T$ como $\hat{\boldsymbol{m}}$ y $\hat{\boldsymbol{m}}_{\boldsymbol{i}}$. Con dicho fin, se escribirá el modelo bajo la hipótesis nula en forma vectorial como sigue:

$$\boldsymbol{Y} = \boldsymbol{D\alpha} + \boldsymbol{m} + \boldsymbol{\varepsilon}, \tag{2.16}$$

donde $\boldsymbol{D} \in \mathcal{M}_{n\times(I-1)}(\mathbb{R})$ es una matriz de diseño con $0s$ y $1s$. Si ahora se supone $\boldsymbol{\alpha}$ conocido, se puede escribir lo siguiente:

$$\hat{\boldsymbol{m}} = \boldsymbol{S}\left(\boldsymbol{Y} - \boldsymbol{D\alpha}\right), \tag{2.17}$$

donde $\boldsymbol{S}$ es una matriz de suavizado. A continuación, sin más que sustituir $\hat{\boldsymbol{m}}$ en (2.16) y derivar la expresión resultante para aplicar el método de mínimos cuadrados, se obtiene el siguiente estimador de $\boldsymbol{\alpha}$:

$$\widehat{\boldsymbol{\alpha}} = \left[\boldsymbol{D}'\left(\boldsymbol{I_n} - \boldsymbol{S_1}\right)'\left(\boldsymbol{I_n} - \boldsymbol{S_1}\right)\boldsymbol{D}\right]^{-1}\boldsymbol{D}'\left(\boldsymbol{I_n} - \boldsymbol{S_1}\right)'\left(\boldsymbol{I_n} - \boldsymbol{S_1}\right)\boldsymbol{Y} = \boldsymbol{AY},$$

donde $\boldsymbol{S_1}$ es una matriz de suavizado diferente de la empleada en la estimación de $\hat{\boldsymbol{m}}$. Por ser $\boldsymbol{S_1}$ una matriz de suavizado, esta requiere de un parámetro de suavizado h, en cual debe escogerse para minimizar el sesgo de $\widehat{\boldsymbol{\alpha}}$. Con este propósito, en [Young y Bowman, 1995], se recomienda emplear distintos parámetros de suavizado, no obstante, también se afirma que $2R/n$ suele ser una buena elección para h, siendo R el rango de los puntos de diseño. Cabe destacar que, en [Speckman, 1988] se afirma que este estimador de $\boldsymbol{\alpha}$ es asintóticamente normal con sesgo

despreciable, lo que permitirá posteriormente hacer la simplificación necesaria para obtener la forma cuadrática análoga a la del test (2.9).

Una vez se ha obtenido el estimador $\widehat{\boldsymbol{\alpha}}$, este puede sustituir a $\boldsymbol{\alpha}$ en la expresión (2.17), obteniendo:

$$\hat{\boldsymbol{m}} = \boldsymbol{S}\left(\boldsymbol{Y} - \boldsymbol{D}\widehat{\boldsymbol{\alpha}}\right),$$

siendo $\boldsymbol{S}$ una matriz de suavizado que use un h obtenido por validación cruzada. Por último, bajo H_a se puede expresar el vector de valores ajustados para los puntos del diseño como:

$$\hat{\boldsymbol{m}} = \boldsymbol{S_d}\boldsymbol{Y},$$

donde $\boldsymbol{S_d}$ es una matriz como la descrita en el test de igualdad, pero empleando un parámetro de suavizado que, con el fin de poder cancelar posteriormente los términos del sesgo en el numerador, ha de ser igual al escogido en H_0.

Finalmente, al igual que para el test (2.9), el estadístico (2.15) se puede expresar como una forma cuadrática, donde el numerador será:

$$\begin{bmatrix} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \boldsymbol{\varepsilon} \end{bmatrix}^T \left[\left(\boldsymbol{I_n} - \boldsymbol{S_s}\right)\boldsymbol{X}\left(\boldsymbol{S_s} - \boldsymbol{S_d}\right)\right]^T \left[\left(\boldsymbol{I_n} - \boldsymbol{S_s}\right)\boldsymbol{X}\left(\boldsymbol{S_s} - \boldsymbol{S_d}\right)\right] \begin{bmatrix} (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \\ \boldsymbol{\varepsilon} \end{bmatrix},$$

donde los vectores entre corchetes son unos vectores con dos bloques. Además, como $\mathbb{E}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right)$ es despreciable, el numerador puede volver a escribirse como $\boldsymbol{\varepsilon}^T\boldsymbol{Q}\boldsymbol{\varepsilon}$, por lo que empleando los mismos procedimientos que para el test (2.9) se consigue obtener una distribución aproximada del estadístico (2.15) bajo H_0, la cual será de la forma $a\chi^2(b) + c$.

Capítulo 3

Estudio de simulación y ejemplo de datos reales

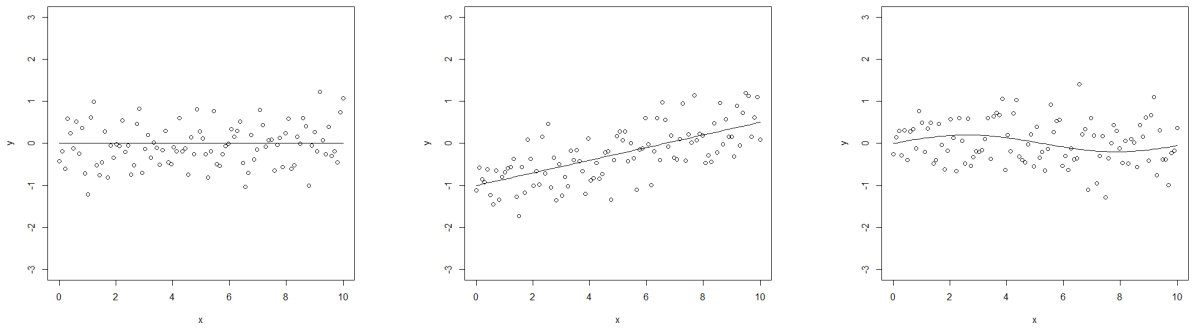
En este capítulo se llevará a cabo un estudio de simulación de los tres test no paramétricos presentados en el capítulo anterior. Para el contraste de no efecto se simularán datos bajo H_0 y bajo H_a para dos modelos distintos, mientras que, para los test de igualdad y paralelismo, se simularán 2 grupos distintos de datos bajo cada una de las hipótesis del test y para cada uno de los distintos modelos. Además, para el modelo A de todos los test se hará también el contraste asociado al modelo lineal, expuesto en el Capítulo 1, con el objetivo de comparar con el test no paramétrico y observar cuanta potencia se pierde si se emplea el test no paramétrico trabajando en un contexto lineal. Para todos los contrastes se emplearán errores bajo distribución normal con distintas desviaciones típicas ($\sigma = 0.5,\ \sigma = 1,\ \sigma = 1.5$). A mayores, se emplearán varios tamaños muestrales ($n = 100,\ n = 200,\ n = 500$), diferentes métodos para la obtención de los puntos del diseño (diseño fijo equidistante, grupos idénticos determinados por una sola muestra aleatoria de una distribución $U(0, 10)$ y grupos distintos cada uno determinado por una muestra aleatoria diferente de una distribución $U(0, 10)$). Por último, el estudio comprueba que el test sea válido para los niveles de significación más empleados en la práctica ($\alpha = 0.1,\ \alpha = 0.05,\ \alpha = 0.01$). Cabe destacar que todos los test no paramétricos se llevaron a cabo empleando un único parámetro de suavizado, el cual fue obtenido por validación cruzada. Por otra parte, a pesar de que bajo las hipótesis necesarias para emplear los contrastes, los errores han de ser normales, se introducirán errores exponenciales para observar como afecta este fallo en el cumplimiento de las hipótesis al desempeño de los mismos. En consecuencia, en la Sección 3.4 se llevarán a cabo los mismos test, pero en el caso en el cual los errores son exponenciales, y se simularán con distintas tasas de ocurrencia ($\lambda = 1,\ \lambda = 2,\ \lambda = 3$). Por último, con el objetivo de demostrar la verdadera utilidad de los contrastes, en la sección 3.5 se usarán todos los test anteriormente mencionados sobre el conjunto de datos reales introducido en el Capítulo 1.

3.1. Test de no efecto

A continuación, se estudiará la potencia y calibrado del test de no efecto tal como se describió anteriormente. El cálculo se realizó con la función `sm.regression` del paquete `sm` de `R` que se encuentra en [Bowman y Azzalini, 2021] donde dicho test está implementado. Se considerarán los dos siguiente modelos:

	H_0	H_a
Modelo A	$Y = \varepsilon$	$Y = -1 + 0.15X + \varepsilon$
Modelo B	$Y = \varepsilon$	$Y = \frac{1}{5}\sin\left(\frac{3X}{5}\right) + \varepsilon$

donde ε es el error de distribución normal con las varianzas indicadas en la presentación de este capítulo y los diagramas de dispersión y las funciones de regresión reales se pueden observar en la Figura 3.1 que aparece a continuación.



(a) Modelos A y B bajo H_0 (b) Modelo A bajo H_a (c) Modelo B bajo H_a

Figura 3.1: Diagramas de dispersión de los datos simulados para el test de no efecto a partir de los modelos A y B con $n = 100$ y con $\sigma = 0.5$, junto con las funciones de regresión reales.

En cada uno de los modelos, la primera columna corresponde a la situación en la que la hipótesis nula es verdadera, es decir, cuando no hay efecto de la variable X sobre la respuesta. La segunda columna corresponde a la situación en la cual es la hipótesis alternativa la que se cumple, en otras palabras, en la que sí existe efecto de la variable X sobre la respuesta, en un caso lineal y en otro no lineal. El test se llevó a cabo sobre 500 muestras de los datos simulados y se registraron los porcentajes de rechazo para los niveles de significación $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$.

Modelo A: En la Tabla 3.1 se pueden ver los resultados del test bajo H_0, tanto para el Mo-

delo A como para el Modelo B. Se puede observar que bajo diseño equiespaciado, para todos los α considerados, los porcentajes de rechazo son bastante más altos que los niveles de significación (aproximadamente el doble) y, aunque a medida que n aumenta estos se hacen más pequeños, no llegan a unos valores aceptables, por lo que se puede afirmar que el test está mal calibrado. Este error en el calibrado del test se debe muy probablemente a la elección del parámetro de suavizado, el cual se ha escogido por validación cruzada. Por otra parte, los resultados del test bajo diseño a partir de $U(0, 10)$ son muy similares, de nuevo duplican los niveles de significación, lo que sugiere que el test no está bien calibrado. Cabe destacar que, en base a los resultados obtenidos no se puede afirmar que el test funcione mejor en un diseño que en el otro. No obstante, en este caso el contraste lineal sí obtiene mejores resultados, pues sus porcentajes de rechazo están en unos valores aceptablemente cercanos a los niveles de significación

A mayores, en la Tabla 3.3 se pueden observar los porcentajes de rechazo del test bajo H_a. En este caso, para ambos diseños podemos ver que la potencia del test es óptima, pues rechaza el 100 % de las veces cuando $n = 500$. Además, cabe destacar que, para todos los tamaños muestrales y varianzas, el test obtiene resultados muy similares para los dos diseños y no pierde potencia con respecto al contraste lineal. No obstante, se observa que, cuando $\sigma = 1.5$ y $n = 100$, los porcentajes se reducen considerablemente y este fenómeno se acentúa cuando el nivel de significación es más bajo. Esto se debe a que, si α toma valores muy pequeños y σ toma valores muy grandes, se le está exigiendo un nivel de confianza muy grande al test, con observaciones que se alejan mucho de la función de regresión real, por lo que es más fácil que la confunda con una recta horizontal.

Modelo B: Como se ha explicado anteriormente, en la Tabla 3.1 se pueden ver los resultados del Grupo 1 tanto para el Modelo A como para el Modelo B, por lo que al tener ambos modelos los mismos resultados, se remite al lector a lo expuesto previamente. A continuación, se analizará el contenido de la Tabla 3.2, donde se observan los porcentajes de rechazo bajo H_a. Tanto bajo diseño equiespaciado, como bajo diseño a partir de una uniforme $U(0, 10)$, el test tiene una potencia más que aceptable cuando $\sigma = 0.5$, con porcentajes de rechazo que se aproximan al 100 % cuando n aumenta. Sin embargo, estos valores empeoran mucho si se aumentan el nivel de confianza exigido y la varianza del error, llegando a unos porcentajes de rechazo de entorno al 20 % cuando $n = 500$ y de incluso un 5 % cuando $n = 100$. Parece además que, de nuevo, el test no empeora al introducir un diseño aleatorio en lugar del diseño fijo.

Test de no efecto. Modelos A y B bajo H_0

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	0.222	0.160	0.204	0.106	0.090	0.116	0.028	0.032	0.030
	L	0.106	0.094	0.104	0.048	0.060	0.058	0.006	0.006	0.016
$n = 200$	NP	0.174	0.168	0.180	0.098	0.084	0.092	0.026	0.020	0.030
	L	0.088	0.116	0.104	0.034	0.044	0.046	0.010	0.010	0.008
$n = 500$	NP	0.154	0.148	0.164	0.078	0.098	0.088	0.016	0.022	0.024
	L	0.082	0.084	0.104	0.042	0.050	0.048	0.012	0.004	0.002
Diseño a partir de $U(0, 10)$										
$n = 100$	NP	0.176	0.174	0.192	0.100	0.092	0.112	0.026	0.018	0.022
	L	0.108	0.098	0.094	0.040	0.040	0.052	0.008	0.004	0.012
$n = 200$	NP	0.176	0.158	0.158	0.094	0.074	0.080	0.014	0.022	0.022
	L	0.118	0.098	0.100	0.062	0.052	0.066	0.004	0.020	0.018
$n = 500$	NP	0.198	0.192	0.136	0.078	0.106	0.076	0.018	0.016	0.020
	L	0.120	0.086	0.078	0.060	0.050	0.042	0.012	0.012	0.010

Tabla 3.1: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de no efecto bajo H_0 de los Modelos A y B basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de no efecto. Modelo B bajo H_a

	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado									
$n = 100$	0.816	0.416	0.318	0.726	0.282	0.186	0.504	0.114	0.084
$n = 200$	0.970	0.564	0.338	0.948	0.448	0.234	0.860	0.212	0.096
$n = 500$	1.000	0.884	0.598	1.000	0.822	0.466	1.000	0.624	0.234
Diseño a partir de $U(0, 10)$									
$n = 100$	0.848	0.394	0.296	0.762	0.270	0.184	0.534	0.114	0.056
$n = 200$	0.980	0.582	0.352	0.960	0.458	0.256	0.878	0.248	0.084
$n = 500$	1.000	0.914	0.610	1.000	0.844	0.470	0.998	0.674	0.244

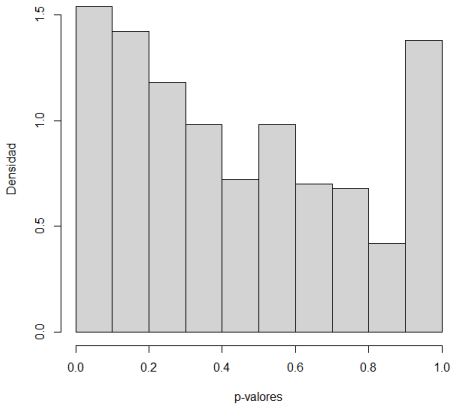
Tabla 3.2: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de no efecto bajo H_a del Modelo B basado en 500 muestras simuladas.

Test de no efecto. Modelo A bajo H_a

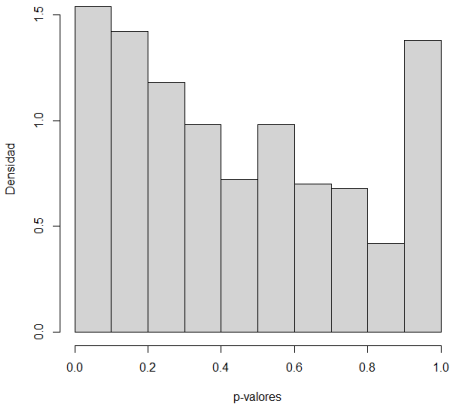
		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	1.000	0.992	0.890	1.000	0.988	0.792	1.000	0.958	0.596
	L	1.000	0.990	0.878	1.000	0.988	0.784	1.000	0.962	0.582
$n = 200$	NP	1.000	1.000	0.990	1.000	1.000	0.974	1.000	1.000	0.924
	L	1.000	1.000	0.996	1.000	1.000	0.980	1.000	1.000	0.936
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Diseño a partir de $U(0, 10)$										
$n = 100$	NP	1.000	0.992	0.864	1.000	0.972	0.768	1.000	0.922	0.528
	L	1.000	0.998	0.862	1.000	0.980	0.772	1.000	0.932	0.530
$n = 200$	NP	1.000	1.000	0.996	1.000	1.000	0.984	1.000	1.000	0.948
	L	1.000	1.000	0.996	1.000	1.000	0.986	1.000	1.000	0.962
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.3: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de no efecto bajo H_a del Modelo A basado en 500 muestras simuladas.NP hace referencia al test no paramétrico y L al test paramétrico lineal.

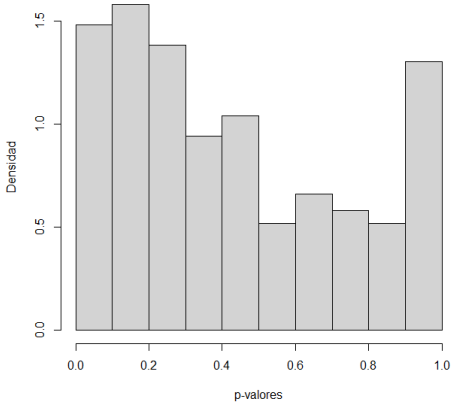
Por último, como es claro que bajo H_0 ha de cumplirse que $\mathbb{P}\{X \leq \alpha\} = \alpha$, la distribución que sigue la probabilidad de rechazo del test bajo H_0 respecto a α debería ser una $U(0, 1)$. Sin embargo, en la Figura 3.2 se puede ver que, tanto para el Modelo A, como para el Modelo B, a los α más pequeños se les ha asignado una probabilidad mayor de la que deberían tener, haciendo que otros valores tengan una probabilidad menor que la esperada. Esto es coherente con los porcentajes de rechazo que se obtuvieron anteriormente y quiere decir que el test no está bien calibrado. Esto se debe principalmente a la elección del parámetro de suavizado, que al no ser del tamaño adecuado altera el resultado del test. No obstante, a pesar de que el test no está bien calibrado, el estadístico (2.6) parece seguir una distribución χ^2 reescalada y recentrada, tal y como se puede apreciar en la Figura 3.3, por lo que es posible que el mal calibrado se deba a la aproximación de los parámetros de la distribución. En dicha figura, además del histograma, aparece superpuesta en rojo una estimación de la función de densidad, la cual fue obtenida empleando el método de validación cruzada para la elección del parámetro de suavizado, véase [Silverman, 2018].



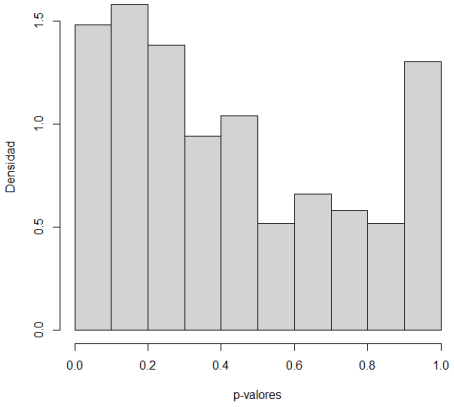
(a) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo A.



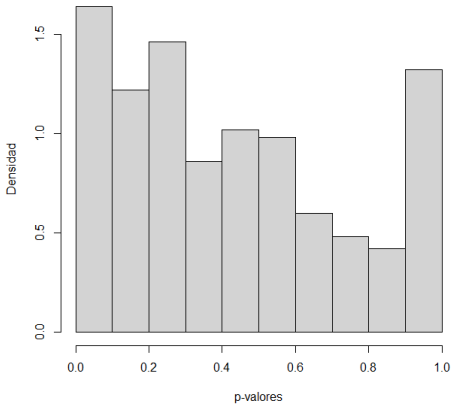
(b) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo B.



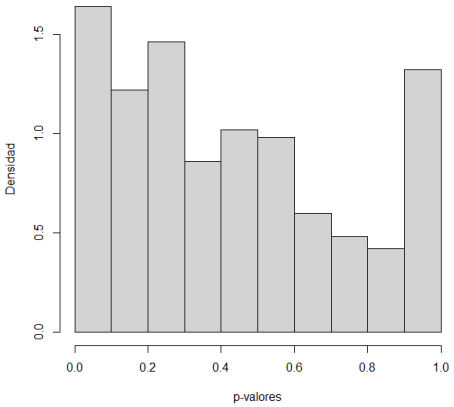
(c) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo A.



(d) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo B.

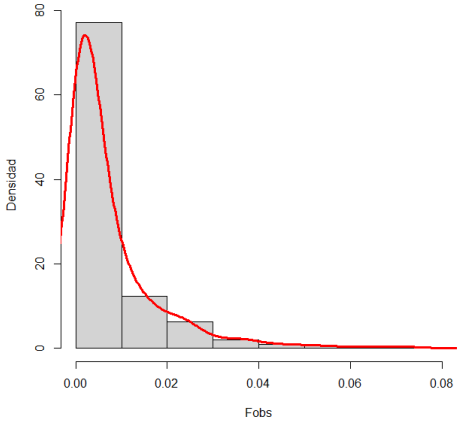


(e) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo A.

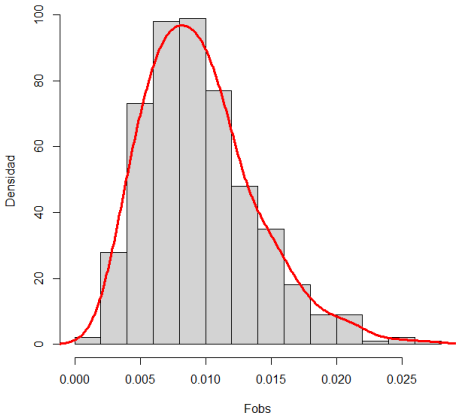


(f) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo B.

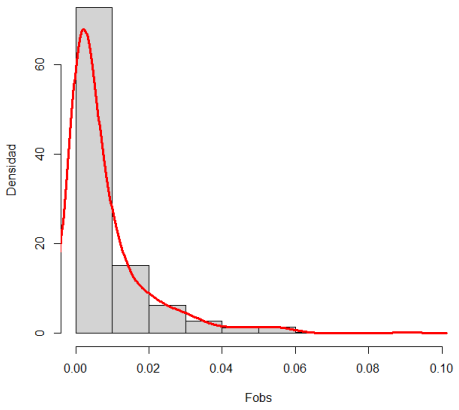
Figura 3.2: Histogramas de la probabilidad de rechazo del test de no efecto bajo H_0 respecto al nivel de significación α, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.



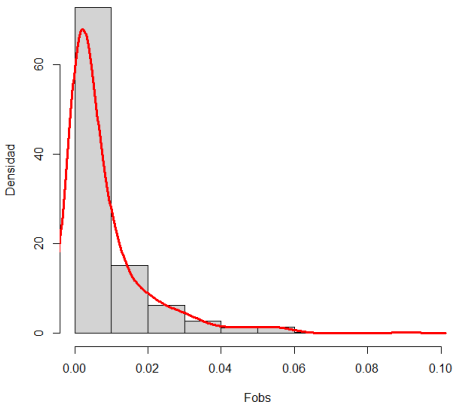
(a) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo A.



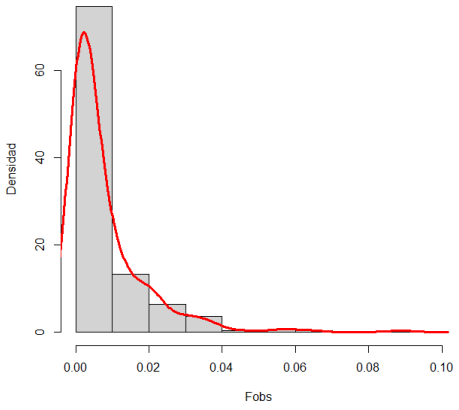
(b) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo B.



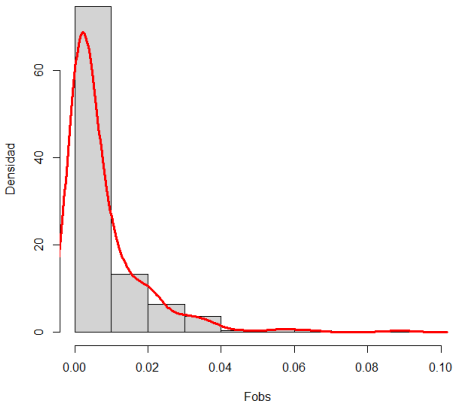
(c) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo A.



(d) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo B.



(e) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo A.



(f) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo B.

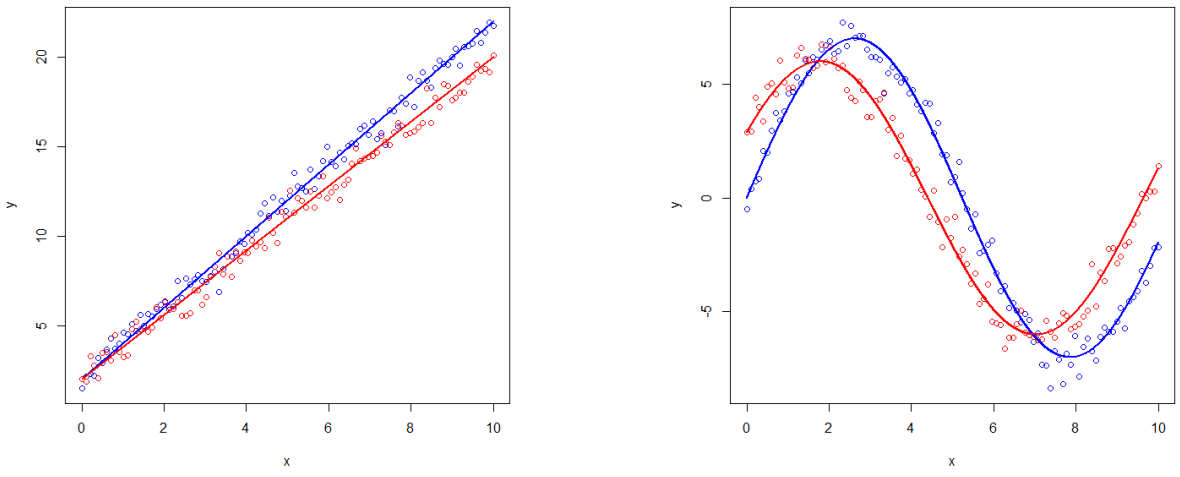
Figura 3.3: Histogramas del estadístico (2.6) para los Modelos A y B bajo H_0 junto con la función de densidad estimada, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

3.2. Test de igualdad

A continuación, se estudiará la potencia y el calibrado del test de igualdad tal como se describió anteriormente. El cálculo se realizó con la función `sm.regression` del paquete `sm` de `R` que se encuentra en [Bowman y Azzalini, 2021] donde dicho test está implementado. Se considerarán los dos siguiente modelos:

	H_0	H_a
Modelo A	Grupo 1: $Y = 2 + 2X + \varepsilon$ Grupo 2: $Y = 2 + 2X + \varepsilon$	Grupo 1: $Y = 2 + 2X + \varepsilon$ Grupo 2: $Y = 2 + \frac{9}{5}X + \varepsilon$
Modelo B	Grupo 1: $Y = 7\sin\left(\frac{3X}{5}\right) + \varepsilon$ Grupo 2: $Y = 7\sin\left(\frac{3X}{5}\right) + \varepsilon$	Grupo 1: $Y = 7\sin\left(\frac{3X}{5}\right) + \varepsilon$ Grupo 2: $Y = 6\sin\left(\frac{3X}{5} + \frac{1}{2}\right) + \varepsilon$

donde ε es el error de distribución normal con las varianzas indicadas en la presentación de este capítulo y los diagramas de dispersión y las funciones de regresión reales se pueden observar en la Figura 3.4 que aparece a continuación.



(a) Test de igualdad. Modelo A.

(b) Test de igualdad. Modelo B.

Figura 3.4: Diagramas de dispersión de los datos simulados para el test de igualdad a partir de los modelos A y B con $n = 100$ y con $\sigma = 0.5$, junto con las funciones de regresión reales.

En el modelo A, se verifica que la hipótesis nula es verdadera cuando ambas rectas son iguales, sin embargo, al cambiar la pendiente de 2 a $\frac{9}{5}$ se verifica la hipótesis alternativa, en otras palabras, las rectas son distintas. En el caso del Modelo B la situación es parecida. Se verifica H_0 cuando las curvas son iguales, y se verifica H_a si se cambia la amplitud de 7 a 6 y se suma $\frac{1}{2}$ dentro del seno, es decir, si las curvas son distintas. El test se llevó a cabo sobre 500 muestras de los datos simulados con n individuos por grupo (siendo n cualquiera de los valores indicados en la presentación de este capítulo) y se registraron los porcentajes de rechazo para los niveles de significación $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$.

Modelo A: En la Tabla 3.4 se exponen los resultados del Modelo A bajo H_0, o dicho de otra forma, los porcentajes de rechazo del test cuando las rectas son iguales. Se puede ver que, tanto bajo diseño equiespaciado, como bajo el diseño obtenido a partir de una o dos uniformes $U(0, 10)$, los porcentajes de rechazo son bastante similares a los niveles de significación para cualquiera de los α considerados, sobre todo a medida que n aumenta, por lo que podemos afirmar que el test está bien calibrado.

Por otra parte, en la Tabla 3.5 se pueden observar los resultados del Modelo A cuando las rectas son distintas, en otras palabras, los porcentajes de rechazo del test bajo H_a. En este caso, bajo diseño equiespaciado podemos ver que la potencia del test es óptima, pues rechaza el 100 % de las veces cuando $n = 200$ o $n = 500$. Este porcentaje solo se ve reducido para $n = 100$ y $\sigma = 1.5$, en donde, al tener tan pocas observaciones y una varianza tan grande, el test no es capaz de detectar el 100 % de las veces que se encuentra bajo H_a. Por otra parte, bajo el diseño obtenido a partir de una o de dos distribuciones $U(0, 10)$, la potencia del test es, de nuevo, casi óptima, rechazando el 100 % de las veces bajo cualquier varianza y para cualquier tamaño muestral a excepción de $n = 100$, $\sigma = 1.5$ y $\alpha = 0.01$. Esto se debe a que, si α toma valores muy pequeños y σ toma valores muy grandes, se le está exigiendo un nivel de confianza muy grande al test, con observaciones que se alejan mucho de la función de regresión real, por lo que es más fácil que confunda ambas rectas entre si. Finalmente, es seguro decir que el test no paramétrico de igualdad no empeora en este caso los resultados del lineal y que, al igual que en el test de no efecto, el resultado no se ve perjudicado por el diseño aleatorio.

Modelo B: En la Tabla 3.6 se pueden observar los resultados del Modelo B cuando las curvas son iguales, es decir, los porcentajes de rechazo del test bajo H_0 para el Modelo B. Se puede ver que, análogamente a lo que sucedía en el Modelo A, tanto bajo diseño equiespaciado, como bajo el diseño obtenido a partir de una o dos uniformes $U(0, 10)$, los porcentajes de rechazo son bastante similares a los niveles de significación para cualquiera de los α considerados, sobre todo a medida que n aumenta.

Por otra parte, en la Tabla 3.7 se pueden observar los resultados del Modelo B cuando las curvas son diferentes, en otras palabras, los porcentajes de rechazo del test bajo H_a. En este caso, la potencia del test es óptima, pues rechaza el 100 % de las veces para todo α y para todo σ en cualquiera de los diseñños planteados. Esto probablemente se deba a que, incluso para la varianza más grande ($\sigma = 1.5$), sigue habiendo regiones de la curva de regresión para las que los puntos de una y otra función están muy separados.

Test de igualdad. Modelo A bajo H_0

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	0.114	0.102	0.120	0.062	0.054	0.072	0.016	0.012	0.020
	L	0.098	0.108	0.084	0.046	0.048	0.050	0.008	0.002	0.018
$n = 200$	NP	0.104	0.120	0.116	0.046	0.062	0.066	0.014	0.012	0.014
	L	0.104	0.098	0.116	0.050	0.048	0.058	0.006	0.014	0.016
$n = 500$	NP	0.100	0.112	0.120	0.052	0.064	0.050	0.016	0.010	0.008
	L	0.110	0.122	0.082	0.054	0.052	0.038	0.004	0.012	0.006
$U(0, 10)$ igual para ambos grupos										
$n = 100$	NP	0.108	0.098	0.096	0.054	0.038	0.052	0.012	0.006	0.022
	L	0.098	0.110	0.084	0.046	0.050	0.050	0.008	0.002	0.018
$n = 200$	NP	0.110	0.132	0.116	0.064	0.070	0.066	0.004	0.016	0.022
	L	0.104	0.100	0.116	0.050	0.048	0.058	0.006	0.014	0.016
$n = 500$	NP	0.116	0.092	0.092	0.046	0.048	0.044	0.006	0.010	0.008
	L	0.110	0.122	0.082	0.054	0.052	0.038	0.004	0.010	0.006
$U(0, 10)$ distinta para cada grupo										
$n = 100$	NP	0.124	0.110	0.104	0.062	0.048	0.058	0.020	0.010	0.020
	L	0.102	0.112	0.082	0.052	0.050	0.044	0.006	0.004	0.018
$n = 200$	NP	0.112	0.102	0.108	0.054	0.048	0.060	0.012	0.008	0.026
	L	0.100	0.102	0.112	0.048	0.046	0.058	0.008	0.014	0.016
$n = 500$	NP	0.106	0.110	0.104	0.046	0.052	0.060	0.002	0.010	0.008
	L	0.112	0.120	0.078	0.054	0.050	0.040	0.004	0.012	0.006

Tabla 3.4: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de igualdad sobre el Modelo A bajo H_0, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de igualdad. Modelo A bajo H_a

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000	0.984
	L	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.980
$n = 200$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ igual para ambos grupos										
$n = 100$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.986
$n = 200$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ distinta para cada grupo										
$n = 100$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.992
$n = 200$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.5: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de igualdad sobre el Modelo A bajo H_a, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de igualdad. Modelo B bajo H_0

	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado									
$n = 100$	0.100	0.096	0.126	0.052	0.044	0.070	0.006	0.012	0.024
$n = 200$	0.084	0.106	0.088	0.044	0.070	0.044	0.006	0.018	0.006
$n = 500$	0.084	0.118	0.102	0.040	0.056	0.052	0.004	0.010	0.008
$U(0, 10)$ igual para ambos grupos									
$n = 100$	0.092	0.084	0.096	0.046	0.046	0.054	0.014	0.014	0.012
$n = 200$	0.096	0.124	0.126	0.040	0.082	0.064	0.012	0.016	0.012
$n = 500$	0.080	0.108	0.088	0.036	0.044	0.050	0.008	0.010	0.010
$U(0, 10)$ distinta para cada grupo									
$n = 100$	0.140	0.100	0.088	0.074	0.050	0.048	0.024	0.006	0.014
$n = 200$	0.098	0.092	0.106	0.050	0.046	0.056	0.006	0.006	0.020
$n = 500$	0.130	0.120	0.108	0.068	0.054	0.058	0.008	0.020	0.010

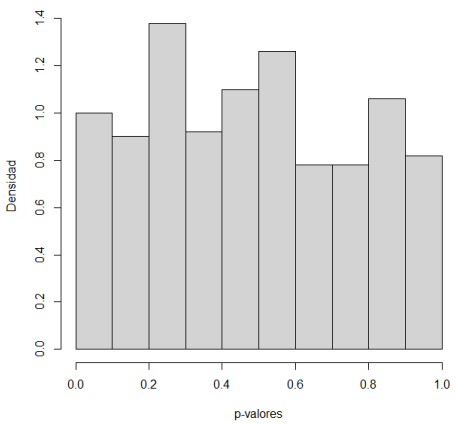
Tabla 3.6: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de igualdad sobre el Modelo B bajo H_0, basado en 500 muestras simuladas.

Test de igualdad. Modelo B bajo H_a

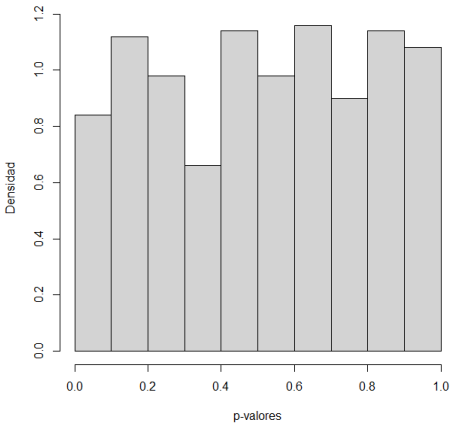
	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ igual para ambos grupos									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ distinta para cada grupo									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.7: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de igualdad sobre el Modelo B bajo H_a, basado en 500 muestras simuladas.

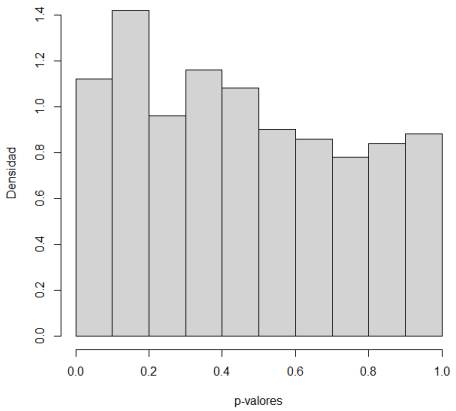
Por último, en la Figura 3.5 se puede ver como, bajo H_0, la distribución que parece seguir la probabilidad de rechazo del test respecto a α es una $U(0,1)$, lo cual es lógico, pues como se ha expuesto anteriormente, $\mathbb{P}\{X \leq \alpha\} = \alpha$. Además, al igual que en la sección anterior, en la Figura 3.6 se ha representado gráficamente la función de densidad estimada del estadístico (2.10), el cual ha de seguir, aproximadamente, una distribución χ^2 reescalada y recentrada. Como se puede apreciar, además de la función de densidad, la cual fue obtenida empleando el método de validación cruzada para la elección del parámetro de suavizado, también se representa el histograma.



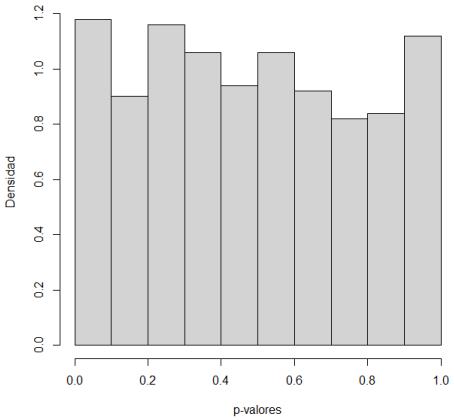
(a) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo A.



(b) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo B.

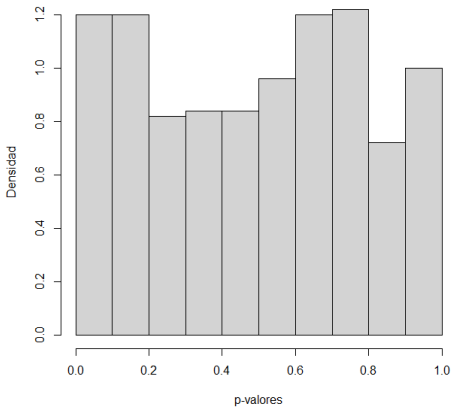


(c) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo A.

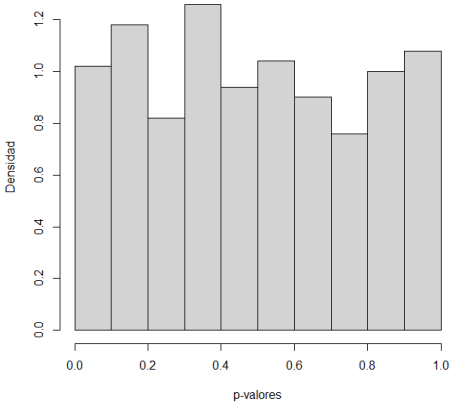


(d) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo B.

Figura 3.5: Histogramas de la probabilidad de rechazo del test de igualdad bajo H_0 respecto al nivel de significación α, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

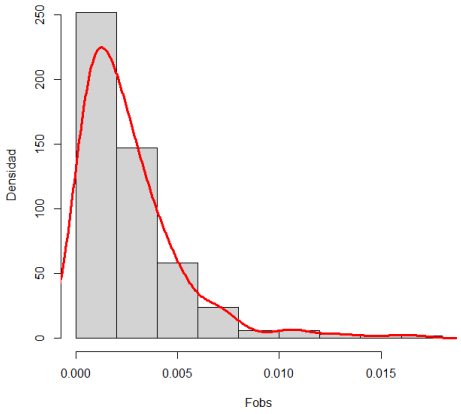


(e) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo A.

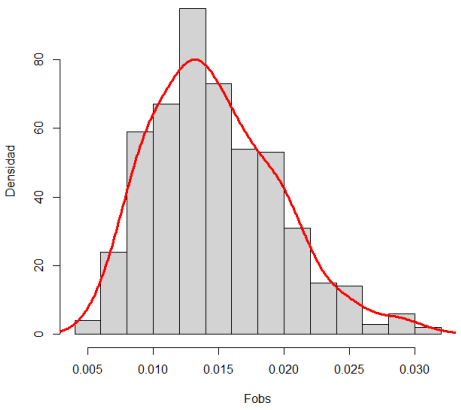


(f) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo B.

Figura 3.5: Histogramas de la probabilidad de rechazo del test de igualdad bajo H_0 respecto al nivel de significación α, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

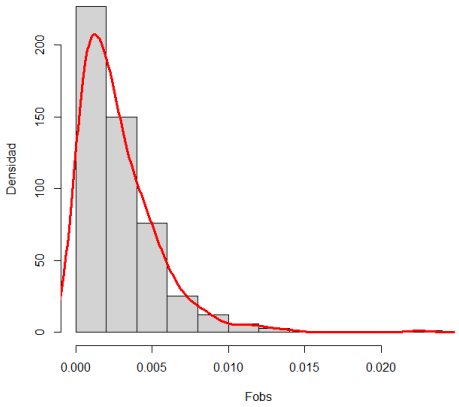


(g) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo A.

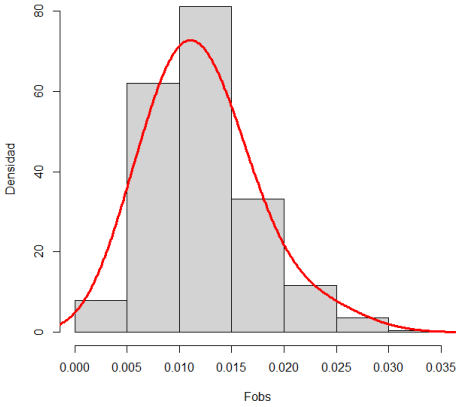


(h) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo B.

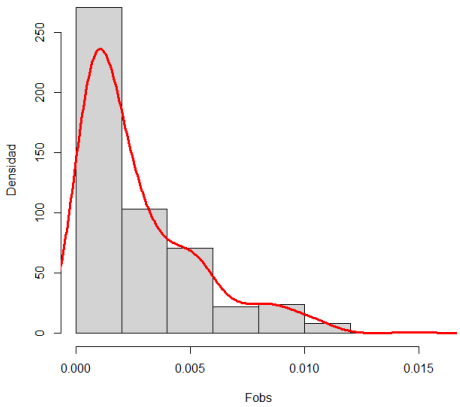
Figura 3.5: Histogramas del estadístico (2.10) para los Modelos A y B bajo H_0, junto con la función de densidad estimada, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.



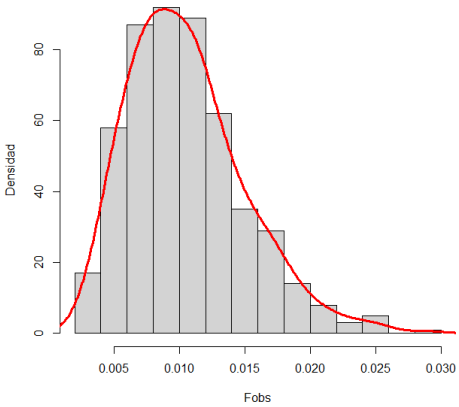
(a) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo A.



(b) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo B.



(c) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo A.



(d) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo B.

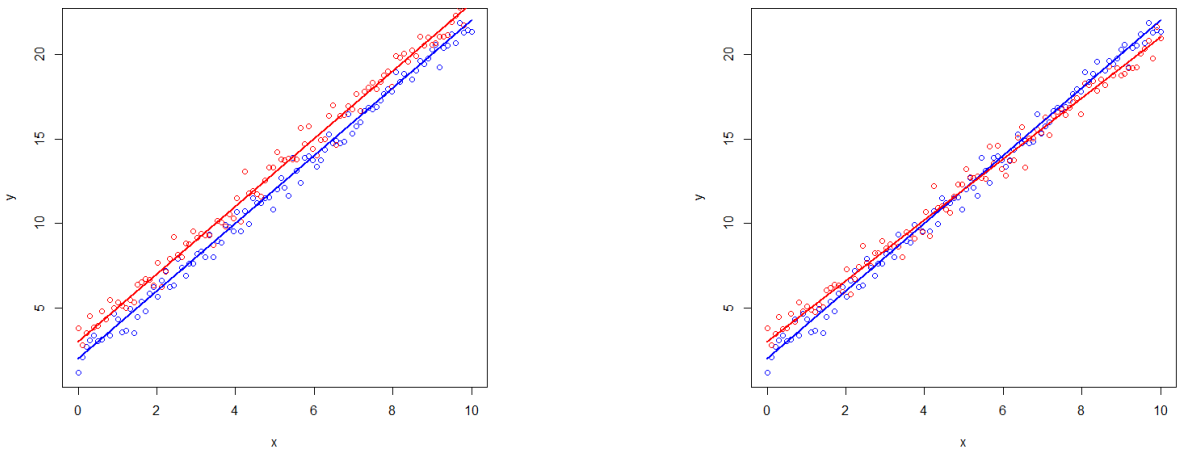
Figura 3.6: Histogramas del estadístico (2.10) para los Modelos A y B bajo H_0, junto con la función de densidad estimada, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

3.3. Test de paralelismo

A continuación, se estudiará la potencia y el calibrado del test de paralelismo tal como se describió anteriormente. El cálculo se realizó con la función `sm.regression` del paquete `sm` de `R`, que se encuentra en [Bowman y Azzalini, 2021] donde dicho test está implementado. Se considerarán los dos siguiente modelos:

	H_0	H_a
Modelo A	Grupo 1: $Y = 2 + 2X + \varepsilon$ Grupo 2: $Y = 3 + 2X + \varepsilon$	Grupo 1: $Y = 2 + 2X + \varepsilon$ Grupo 2: $Y = 3 + \frac{9}{5}X + \varepsilon$
Modelo B	Grupo 1: $Y = 7\sin\left(\frac{3X}{5}\right) + \varepsilon$ Grupo 2: $Y = 7\sin\left(\frac{3X}{5}\right) + 1 + \varepsilon$	Grupo 1: $Y = 7\sin\left(\frac{3X}{5}\right) + \varepsilon$ Grupo 2: $Y = \frac{13}{2}\sin\left(\frac{3X}{5} + \frac{1}{3}\right) + \varepsilon$

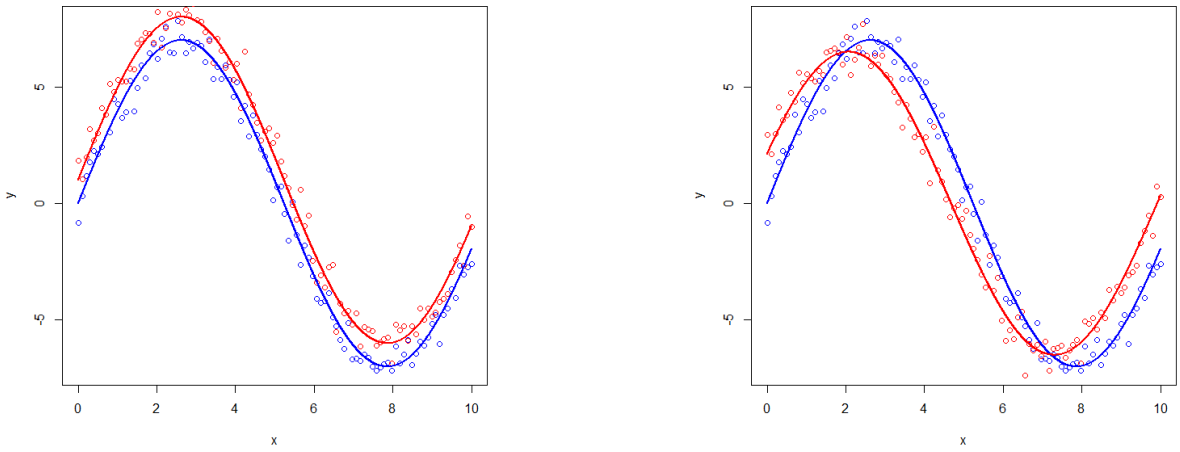
donde ε es el error de distribución normal con las varianzas indicadas en la presentación de este capítulo y los diagramas de dispersión y las funciones de regresión reales bajo H_0 y H_a se pueden observar a continuación en la Figura 3.7 y en la Figura 3.8 respectivamente.



(a) Modelo A bajo H_0

(b) Modelo A bajo H_a

Figura 3.7: Diagramas de dispersión de los datos simulados para el test de paralelismo a partir del modelo A con $n = 100$ y con $\sigma = 0.5$, junto con las funciones de regresión reales.



(a) Modelo B bajo H_0

(b) Modelo B bajo H_a

Figura 3.8: Diagramas de dispersión de los datos simulados para el test de paralelismo a partir del modelo B con $n = 100$ y con $\sigma = 0.5$, junto con las funciones de regresión reales.

En el modelo A, la hipótesis nula es verdadera cuando ambas rectas son paralelas, sin embargo, al modificar la pendiente de 2 a $\frac{9}{5}$ se verifica la hipótesis alternativa, en otras palabras, las rectas dejan de ser paralelas y pasan a ser secantes. En el caso del Modelo B la situación es parecida. Se verifica H_0 cuando las curvas son paralelas y se verifica H_a si estas dejan de serlo, de hecho, con los cambios introducidos en el modelo bajo la hipótesis alternativa las curvas no solo dejan de ser paralelas sino que pasan a ser secantes. El test se llevó a cabo sobre 500 muestras de los datos simulados con n individuos por grupo (siendo n cualquiera de los valores indicados en la presentación de este capítulo) y se registraron los porcentajes de rechazo para los niveles de significación $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$.

Modelo A: En la Tabla 3.8 se pueden observar los resultados del Modelo A cuando las rectas son paralelas, es decir, los porcentajes de rechazo del test bajo H_0 para el Modelo A. Se puede ver que, tanto bajo diseño equiespaciado, como bajo diseño a partir de una o dos uniformes $U(0, 10)$, los porcentajes de rechazo son bastante similares a los niveles de significación para cualquiera de los α considerados, sobre todo a medida que n aumenta.

Por otra parte, en la Tabla 3.9 se pueden observar los resultados del Modelo A cuando las rectas no son paralelas, en otras palabras, los porcentajes de rechazo del test bajo H_a. En este caso, bajo cualquier diseño se puede observar que la potencia del test es casi óptima, pues rechaza entorno al 100 % de las veces cuando $n = 500$. Sin embargo, este porcentaje se ve reducido al re-

ducir el tamaño muestral, especialmente cuando el modelo tiene una desviación típica alta, pues al igual que en el test de igualdad, al tener tan pocas observaciones y una varianza tan grande, el test no es capaz de valorar adecuadamente si las rectas son o no paralelas. En consecuencia, se puede decir que, una vez más, el diseño aleatorio no afecta demasiado en este caso. Finalmente, cabe destacar que, a pesar de que el test lineal obtiene unos resultados ligeramente superiores, no es un aumento considerable.

Modelo B: En la Tabla 3.10 se pueden observar los resultados del Modelo B cuando las curvas son paralelas, es decir, los porcentajes de rechazo del test bajo H_0 para el Modelo B. Se puede ver que, análogamente a lo que sucedía en el Modelo A, tanto bajo diseño equiespaciado, como bajo el diseño obtenido a partir de una o dos uniformes $U(0, 10)$, los porcentajes de rechazo son bastante similares a los niveles de significación para cualquiera de los α considerados, sobre todo a medida que n aumenta.

Por otra parte, en la Tabla 3.11 se pueden observar los resultados del Modelo B cuando las curvas no son paralelas, en otras palabras, los porcentajes de rechazo del test bajo H_a. En este caso, la potencia del test es óptima para todos los diseños considerados, pues rechaza el 100 % de las veces para todo α y para todo σ considerados en el estudio. Esto probablemente se deba a que, incluso para la varianza más grande ($\sigma = 1.5$), sigue habiendo regiones de la curva de regresión para las que los puntos de una y otra función están muy bien diferenciados, permitiendo así juzgar el no paralelismo con mayor precisión. Como consecuencia de que la potencia sea óptima para cualquier diseño, se concluye que el test no sufre cuando se pasa de diseño fijo equidistante a aleatorio uniforme.

Test de paralelismo. Modelo A bajo H_0

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	0.130	0.094	0.140	0.062	0.046	0.080	0.012	0.014	0.018
	L	0.110	0.100	0.122	0.054	0.042	0.068	0.006	0.008	0.018
$n = 200$	NP	0.106	0.128	0.116	0.052	0.072	0.058	0.014	0.016	0.020
	L	0.096	0.104	0.114	0.042	0.052	0.064	0.010	0.010	0.014
$n = 500$	NP	0.090	0.120	0.142	0.052	0.066	0.074	0.012	0.012	0.008
	L	0.090	0.106	0.116	0.044	0.060	0.052	0.010	0.006	0.006
$U(0, 10)$ igual para ambos grupos										
$n = 100$	NP	0.116	0.084	0.102	0.052	0.036	0.038	0.018	0.008	0.010
	L	0.096	0.082	0.090	0.046	0.034	0.036	0.012	0.004	0.006
$n = 200$	NP	0.140	0.128	0.120	0.060	0.070	0.048	0.008	0.026	0.014
	L	0.124	0.110	0.106	0.068	0.060	0.048	0.004	0.012	0.012
$n = 500$	NP	0.138	0.106	0.112	0.068	0.042	0.048	0.014	0.010	0.014
	L	0.132	0.106	0.084	0.046	0.046	0.044	0.012	0.008	0.006
$U(0, 10)$ distinta para cada grupo										
$n = 100$	NP	0.136	0.112	0.086	0.068	0.060	0.050	0.018	0.018	0.006
	L	0.118	0.102	0.098	0.050	0.052	0.046	0.010	0.010	0.008
$n = 200$	NP	0.108	0.120	0.114	0.052	0.058	0.052	0.012	0.008	0.020
	L	0.100	0.110	0.108	0.042	0.060	0.054	0.008	0.006	0.012
$n = 500$	NP	0.136	0.096	0.134	0.058	0.050	0.068	0.008	0.014	0.010
	L	0.116	0.074	0.108	0.062	0.028	0.066	0.006	0.008	0.008

Tabla 3.8: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de paralelismo sobre el Modelo A bajo H_0, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de paralelismo. Modelo A bajo H_a

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado										
$n = 100$	NP	1.000	0.996	0.840	1.000	0.988	0.754	1.000	0.944	0.552
	L	1.000	0.996	0.862	1.000	0.994	0.782	1.000	0.942	0.568
$n = 200$	NP	1.000	1.000	0.970	1.000	1.000	0.952	1.000	0.998	0.886
	L	1.000	1.000	0.976	1.000	1.000	0.960	1.000	1.000	0.898
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ igual para ambos grupos										
$n = 100$	NP	1.000	0.986	0.796	1.000	0.958	0.690	1.000	0.892	0.456
	L	1.000	0.994	0.828	1.000	0.972	0.726	1.000	0.902	0.454
$n = 200$	NP	1.000	1.000	0.988	1.000	1.000	0.982	1.000	0.998	0.912
	L	1.000	1.000	0.990	1.000	1.000	0.980	1.000	1.000	0.924
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.998
	L	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.998
$U(0, 10)$ distinta para cada grupo										
$n = 100$	NP	1.000	0.992	0.806	1.000	0.966	0.700	1.000	0.900	0.512
	L	1.000	0.994	0.834	1.000	0.980	0.746	1.000	0.916	0.532
$n = 200$	NP	1.000	1.000	0.990	1.000	1.000	0.980	1.000	1.000	0.912
	L	1.000	1.000	0.990	1.000	1.000	0.986	1.000	1.000	0.928
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.9: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de paralelismo sobre el Modelo A bajo H_a, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de paralelismo. Modelo B bajo H_0

	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado									
$n = 100$	0.098	0.088	0.122	0.060	0.044	0.064	0.008	0.010	0.028
$n = 200$	0.082	0.106	0.104	0.044	0.056	0.048	0.010	0.022	0.008
$n = 500$	0.086	0.118	0.104	0.034	0.066	0.046	0.002	0.016	0.014
$U(0, 10)$ igual para ambos grupos									
$n = 100$	0.088	0.098	0.106	0.054	0.044	0.046	0.018	0.014	0.012
$n = 200$	0.096	0.110	0.112	0.046	0.084	0.060	0.010	0.020	0.006
$n = 500$	0.092	0.096	0.082	0.034	0.046	0.052	0.012	0.008	0.014
$U(0, 10)$ distinta para cada grupo									
$n = 100$	0.116	0.088	0.084	0.066	0.042	0.036	0.028	0.010	0.012
$n = 200$	0.116	0.084	0.102	0.058	0.042	0.054	0.006	0.012	0.022
$n = 500$	0.128	0.102	0.092	0.068	0.050	0.060	0.008	0.018	0.014

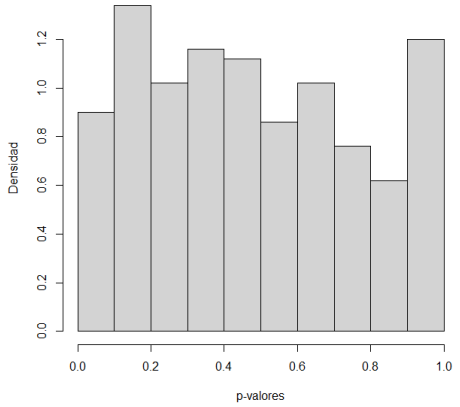
Tabla 3.10: Porcentajes de rechazo (para $\alpha = 0.1, \alpha = 0.05$ y $\alpha = 0.01$) para el test de paralelismo sobre el Modelo B bajo H_0, basado en 500 muestras simuladas.

Test de paralelismo. Modelo B bajo H_a

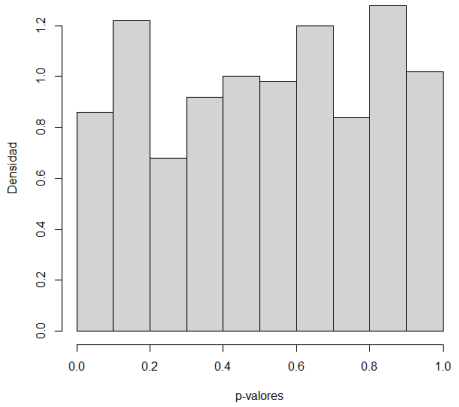
	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
Diseño equiespaciado									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ igual para ambos grupos									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$U(0, 10)$ distinta para cada grupo									
$n = 100$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.11: Porcentajes de rechazo (para $\alpha = 0.1, \alpha = 0.05$ y $\alpha = 0.01$) para el test de paralelismo sobre el Modelo B bajo H_a, basado en 500 muestras simuladas.

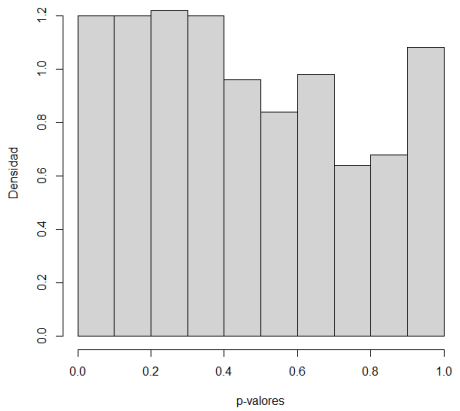
Por último, en la Figura 3.9 se puede ver como, bajo H_0, la distribución que aparenta seguir la probabilidad de rechazo del test respecto a α es una $U(0,1)$. Además, al igual que en la sección anterior, en la Figura 3.10 se ha representado gráficamente la función de densidad estimada del estadístico (2.15), el cual ha de seguir, aproximadamente, una distribución χ^2 reescalada y recentrada. Como se puede apreciar, además de la función de densidad obtenida empleando el método de validación cruzada, también se representa el histograma.



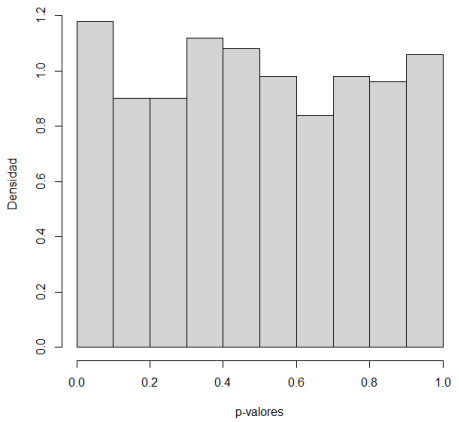
(a) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo A.



(b) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 0.5$. Modelo B.

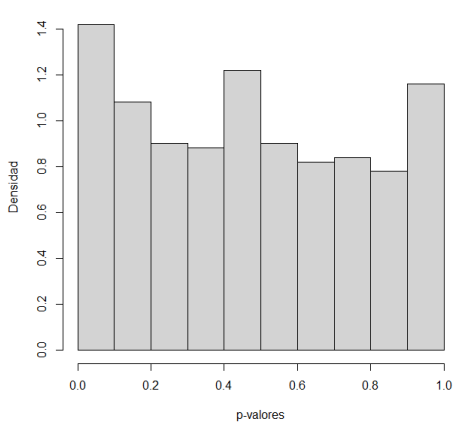


(c) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo A.

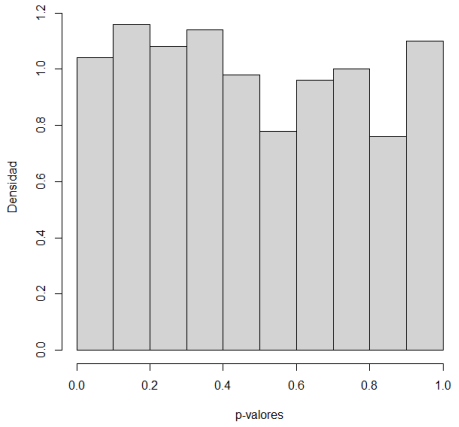


(d) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1$. Modelo B.

Figura 3.9: Histogramas de la probabilidad de rechazo del test de paralelismo bajo H_0 respecto al nivel de significación α, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

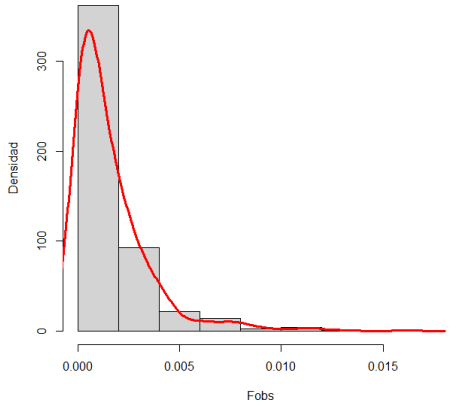


(e) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo A.

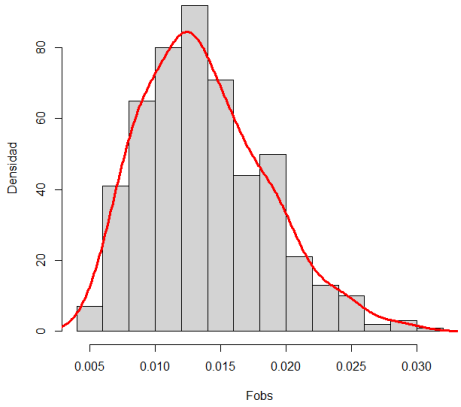


(f) Histograma probabilidad de rechazo, $n = 500$ y $\sigma = 1.5$. Modelo B.

Figura 3.9: Histogramas de la probabilidad de rechazo del test de paralelismo bajo H_0 respecto al nivel de significación α, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

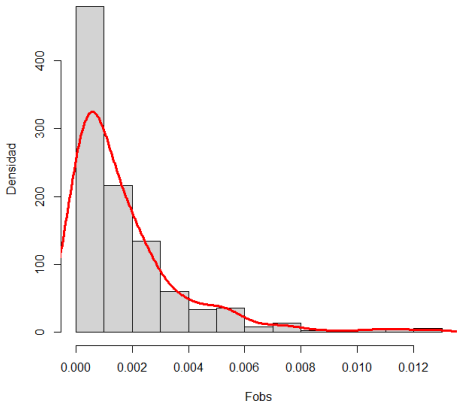


(a) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo A.

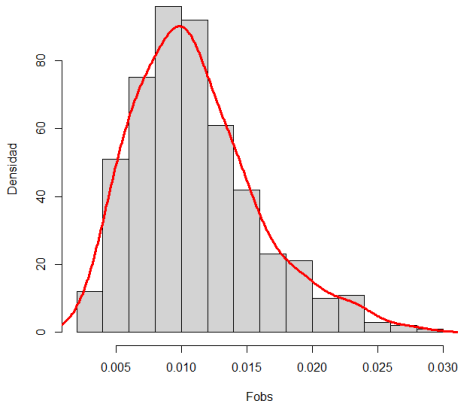


(b) Histograma de estadísticos observados, $n = 500$ y $\sigma = 0.5$. Modelo B.

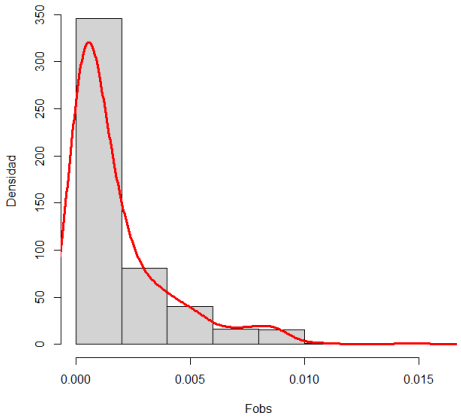
Figura 3.10: Histogramas del estadístico (2.15) para los Modelos A y B bajo H_0 junto con la función de densidad estimada, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.



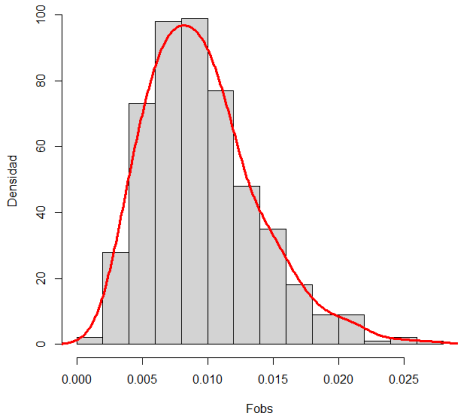
(c) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo A.



(d) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1$. Modelo B.



(e) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo A.



(f) Histograma de estadísticos observados, $n = 500$ y $\sigma = 1.5$. Modelo B.

Figura 3.10: Histogramas del estadístico (2.15) para los Modelos A y B bajo H_0 junto con la función de densidad estimada, $\forall \sigma \in \left\{\frac{1}{2}, 1, \frac{3}{2}\right\}$ y para $n = 500$.

3.4. Otros contextos: errores exponenciales

En la sección anterior se expusieron los resultados de los test para los cuales los errores siguen una distribución normal con varianza constante. Ahora bien, el desempeño de las pruebas de no efecto, igualdad y paralelismo cuando el supuesto de normalidad no se verifica se estudiará en esta sección. Para ello, se repetirán las simulaciones de los modelos considerados en el estudio

anterior manteniendo, tanto los niveles de significación, como el tamaño de los datos, y un diseño equidistante. Sin embargo, en lugar de la distribución normal, los errores seguirán una distribución exponencial: $\varepsilon \sim Exp(\lambda)$, con $\lambda \in \{1, 2, 3\}$, tal y como la que se puede observar en la Figura 3.11:

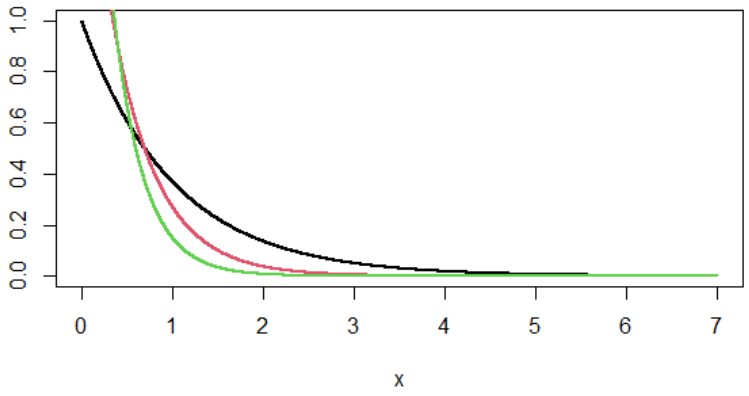


Figura 3.11: La curva en negro es la densidad exponencial con $\lambda = 1$, mientras que la roja y la verde son densidades exponenciales de parámetros $\lambda = 2$ y $\lambda = 3$ respectivamente.

Como se puede observar en la Tabla 3.13 y en la Tabla 3.14, es claro que en ambos test los porcentajes de rechazo bajo H_0 son muy similares al nivel de significación correspondiente en cada caso, por lo que se puede afirmar que el test sigue calibrado a pesar de la introducción de errores exponenciales. Además, para ambos test la potencia también es muy alta, pues el test de igualdad rechaza el 100 % de las veces bajo H_a, para todo n y para todo λ considerados, y el test de paralelismo rechaza el 100 % de las veces bajo H_a, para todo λ considerado si $n = 500$. Esto permite afirmar que, para ambos test, la potencia tiende a 1 cuando el tamaño muestral aumenta. En consecuencia, se puede asegurar que los errores exponenciales no representan un problema a la hora de efectuar dichos test, de hecho parece afectar más a la potencia del test la distribución que sigue el diseño que la de los errores. Por otra parte, en la Tabla 3.12 están los porcentajes de rechazo del test de no efecto. Bajo H_0 el test no paramétrico parece seguir mal calibrado, pues al igual que con los errores normales, sus porcentajes de rechazo bajo la hipótesis nula duplican al nivel de significación. En cuanto a su potencia, para el modelo A es bastante buena, rechaza el 100 % de las veces bajo H_a, para todo λ considerado si $n = 200$ o $n = 500$. Sin embargo, para el modelo B la potencia sufre una disminución considerable, pues cuando $n = 100$ o $n = 200$ los porcentajes de rechazo están, en algunos casos, entorno al $15 - 20\,\%$. Este fenómeno se acentúa cuando el nivel de significación y λ son más bajos, y se debe a que, si α y λ toman valores muy pequeños, se le está exigiendo un nivel de confianza muy grande al test, con

observaciones que se alejan mucho de la función de regresión real, por lo que es más fácil que la confunda con una recta horizontal. Esto sucede también cuando $n = 500$, aunque con menor intensidad. Consecuentemente, se puede afirmar que, con errores exponenciales el test no está bien calibrado y bajo el modelo B no tiene una buena potencia, pero estos problemas también los tenía con errores normales, por lo que no se puede confirmar que los errores exponenciales sean la causa de este mal desempeño del test.

Test de no efecto con errores exponenciales

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Test de no efecto. Modelo A bajo H_0										
$n = 100$	NP	0.184	0.194	0.222	0.110	0.104	0.130	0.028	0.034	0.030
	L	0.102	0.094	0.118	0.054	0.054	0.060	0.006	0.012	0.008
$n = 200$	NP	0.164	0.180	0.160	0.080	0.090	0.090	0.024	0.022	0.020
	L	0.114	0.132	0.106	0.054	0.056	0.062	0.010	0.008	0.010
$n = 500$	NP	0.158	0.164	0.174	0.088	0.082	0.092	0.028	0.030	0.026
	L	0.116	0.088	0.106	0.054	0.046	0.052	0.010	0.010	0.010
Test de no efecto. Modelo A bajo H_a										
$n = 100$	NP	0.990	1.000	1.000	0.978	1.000	1.000	0.922	1.000	1.000
	L	0.992	1.000	1.000	0.982	1.000	1.000	0.930	1.000	1.000
$n = 200$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Test de no efecto. Modelo B bajo H_0										
$n = 100$		0.184	0.194	0.222	0.110	0.104	0.130	0.028	0.034	0.030
$n = 200$		0.164	0.180	0.160	0.080	0.090	0.090	0.024	0.022	0.020
$n = 500$		0.158	0.164	0.174	0.088	0.082	0.092	0.028	0.030	0.026
Test de no efecto. Modelo B bajo H_a										
$n = 100$		0.422	0.828	0.986	0.294	0.746	0.952	0.126	0.476	0.886
$n = 200$		0.578	0.962	1.000	0.466	0.940	1.000	0.236	0.844	0.992
$n = 500$		0.882	1.000	1.000	0.812	1.000	1.000	0.600	1.000	1.000

Tabla 3.12: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de no efecto con errores exponenciales, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de igualdad con errores exponenciales

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Test de igualdad. Modelo A bajo H_0										
$n = 100$	NP	0.100	0.116	0.122	0.046	0.064	0.048	0.016	0.034	0.016
	L	0.106	0.084	0.104	0.048	0.048	0.054	0.004	0.008	0.012
$n = 200$	NP	0.114	0.108	0.096	0.068	0.046	0.052	0.022	0.014	0.006
	L	0.120	0.082	0.092	0.052	0.050	0.040	0.016	0.010	0.004
$n = 500$	NP	0.122	0.098	0.126	0.064	0.046	0.060	0.008	0.010	0.016
	L	0.072	0.124	0.090	0.028	0.058	0.058	0.002	0.014	0.010
Test de igualdad. Modelo A bajo H_a										
$n = 100$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Test de igualdad. Modelo B bajo H_0										
$n = 100$		0.096	0.136	0.100	0.052	0.088	0.050	0.020	0.026	0.022
$n = 200$		0.106	0.098	0.100	0.052	0.050	0.050	0.012	0.004	0.004
$n = 500$		0.102	0.090	0.116	0.048	0.038	0.062	0.014	0.006	0.010
Test de igualdad. Modelo B bajo H_a										
$n = 100$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.13: Porcentajes de rechazo (para $\alpha = 0.1$, $\alpha = 0.05$ y $\alpha = 0.01$) para el test de igualdad con errores exponenciales, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

Test de paralelismo con errores exponenciales

		$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
		$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Test de paralelismo. Modelo A bajo H_0										
$n = 100$	NP	0.106	0.142	0.112	0.046	0.074	0.054	0.010	0.030	0.012
	L	0.080	0.106	0.116	0.046	0.048	0.052	0.010	0.014	0.010
$n = 200$	NP	0.122	0.120	0.120	0.064	0.056	0.046	0.014	0.016	0.006
	L	0.102	0.104	0.108	0.038	0.050	0.042	0.012	0.008	0.006
$n = 500$	NP	0.128	0.104	0.132	0.078	0.040	0.068	0.022	0.004	0.020
	L	0.112	0.092	0.130	0.062	0.040	0.076	0.024	0.004	0.016
Test de paralelismo. Modelo A bajo H_a										
$n = 100$	NP	0.988	1.000	1.000	0.972	1.000	1.000	0.904	1.000	1.000
	L	0.988	1.000	1.000	0.980	1.000	1.000	0.908	1.000	1.000
$n = 200$	NP	0.998	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000
	L	1.000	1.000	1.000	0.998	1.000	1.000	0.998	1.000	1.000
$n = 500$	NP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	L	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Test de paralelismo. Modelo B bajo H_0										
$n = 100$		0.104	0.144	0.086	0.054	0.098	0.052	0.020	0.030	0.024
$n = 200$		0.072	0.104	0.096	0.036	0.050	0.044	0.012	0.008	0.010
$n = 500$		0.100	0.092	0.132	0.052	0.036	0.064	0.012	0.006	0.006
Test de paralelismo. Modelo B bajo H_a										
$n = 100$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 200$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$n = 500$		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Tabla 3.14: Porcentajes de rechazo (para $\alpha = 0.1, \alpha = 0.05$ y $\alpha = 0.01$) para el test de paralelismo con errores exponenciales, basado en 500 muestras simuladas. NP hace referencia al test no paramétrico y L al test paramétrico lineal.

3.5. Desempeño de los test sobre datos reales

En esta sección, se utilizará el conjunto de datos [Manish Kumar, 2019] que se presentó en el Capítulo 1, en cual recoge diferentes características de 205 coches del mercado americano y se emplearán las variables *higwaympg* y *doornumber* para predecir el precio de los coches. No obstante, antes de comenzar con la estimación de la función de regresión y de realizar los test necesarios, se llevará a cabo un análisis descriptivo de las variables empleadas.

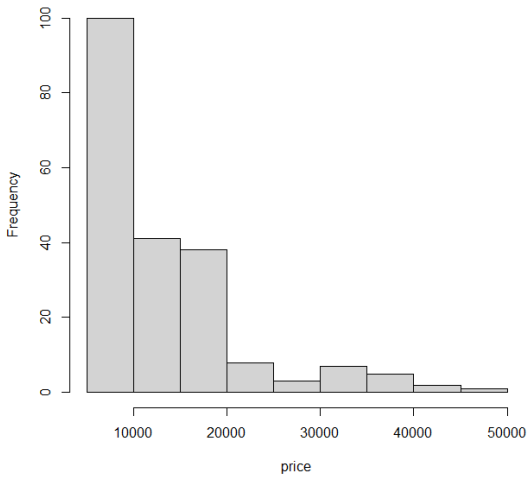
- *highwaympg* es una variable explicativa continua que hace referencia al gasto de combustible del vehículo al circular por la autopista. Concretamente, se refiere a las millas de media que un vehículo es capaz de circular por la autopista con un galón de combustible. Tal como se puede ver en la Figura 3.12b, la gran mayoría de los coches tienen un consumo en autopista que varía entre los 20 mpg y los 40 mpg, por lo que parece que *highwaympg* presenta bastante variabilidad. Además, en la Figura 3.12c se puede observar que la mediana del consumo de los coches de cuatro puertas es mayor que la del consumo de los de dos puertas, sin embargo en el tercer cuartil los papeles se invierten.

- *price* es la variable respuesta continua, y denota el precio de los vehículos en dólares americanos. En este caso, en la Figura 3.12a se puede observar que la mayoría de los coches no superan los 20000 $ de precio y que en el estudio no se han considerado coches por encima de los 50000 $. Además, en la Figura 3.12d se puede ver que la mediana y el tercer cuartil del precio de los coches de cuatro puertas son mayores que los de dos puertas, lo que indica que en general, se pagará más por un coche de cuatro puertas que por uno de dos.

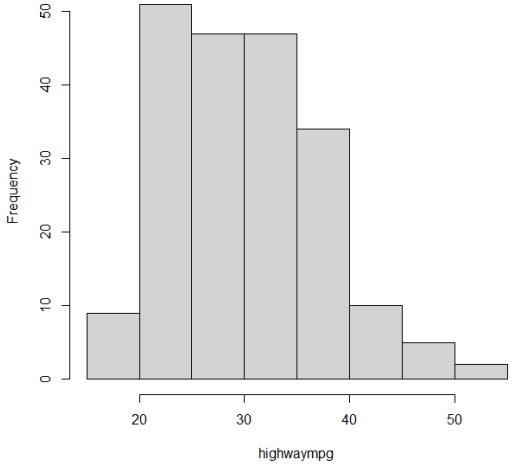
- *doornumber* es una variable explicativa discreta que indica el número de puertas del vehículo. Solo toma dos posibles valores: dos puertas o cuatro puertas. En la base de datos existen 115 vehículos de cuatro puertas y 90 vehículos de dos puertas.

A continuación, aunque en la Figura 3.12e se puede ver a simple vista que la relación entre *highwaympg* y *price* no es lineal, se llevarán a cabo unos test de linealidad de los paquetes `lmtest` y `sm` de `R` que se encuentran en [Zeileis y Hothorn, 2002] y [Bowman y Azzalini, 2021] respectivamente. El test de reset de Ramsey da un p-valor menor que 2.2×10^{-16}, mientras que el test de linealidad del paquete `sm` devuelve directamente un p-valor de 0, por lo que podemos afirmar con seguridad que la relación entre ambas variables no es lineal. Estos resultados sugieren la utilización de una regresión no paramétrica pero, al estar involucrada una variable discreta, además surge la pregunta de si se deberían estimar dos curvas totalmente independientes, de si estas dos curvas son paralelas o de si son directamente la misma curva. Para responder a esta cuestión se emplearán los contrastes no paramétricos estudiados. El test de igualdad devuelve un p-valor de 0.0288 por lo que se puede afirmar que las funciones de regresión son distintas para los niveles de significación de 0.05 y 0.1, en otras palabras, hay evidencias suficientemente significativas como para negar la igualdad de las funciones. A continuación, se realiza el test de paralelismo, el cual devuelve un p-valor de 0.0172 por lo que, de nuevo, se puede afirmar que las funciones de regresión no son paralelas para los niveles de significación de 0.05 y 0.1, dicho de otra forma, hay evidencias suficientemente significativas como para negar el paralelismo de las

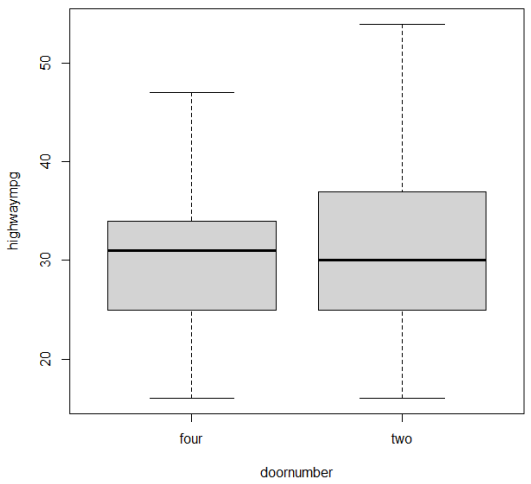
funciones. En consecuencia, se estimarán las funciones de regresión por separado en función del número de puertas del automóvil, tal como se observa en la Figura 3.12f.



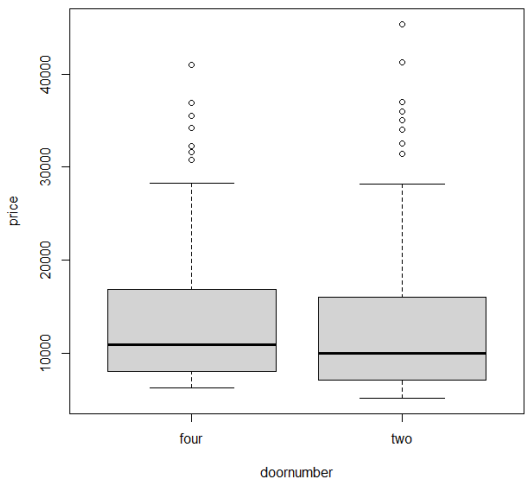
(a) Histograma del precio



(b) Histograma de las millas por galón en autopista

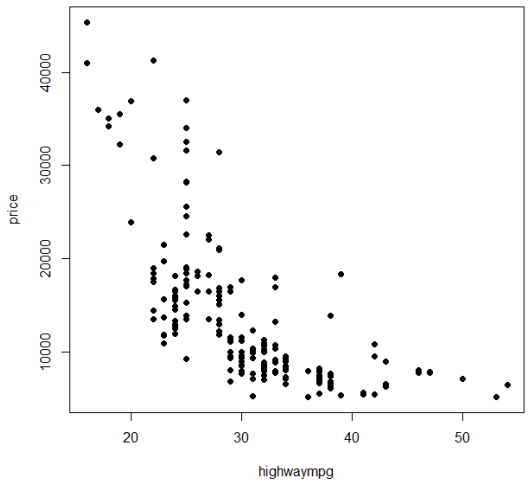


(c) Boxplot del consumo en autopista en funcón del número de puertas

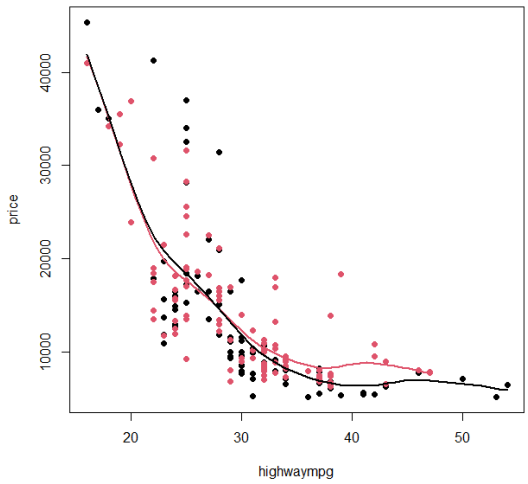


(d) Boxplot del precio en función del número de puertas

Figura 3.12: Gráficas para el análisis descriptivo de los datos.



(e) Diagrama de dispersión del precio en función del consumo en autopista



(f) Diagrama de dispersión del precio en función del consumo en autopista y el número de puertas

Figura 3.12: Gráficas para el análisis descriptivo de los datos.

Capítulo 4

Conclusiones

Tras exponer el modelo ANCOVA no paramétrico y haberlo probado, tanto con datos reales, como simulados, se puede afirmar que con un tamaño muestral suficiente, los resultados obtenidos son bastante satisfactorios. El modelo es capaz de detectar la mayor parte de las veces si las funciones de regresión son iguales, si son paralelas o si son funciones totalmente diferentes. Además, en caso de que el modelo sea lineal y lo estudiemos con el ANCOVA no paramétrico, los contrastes asociados a este modelo apenas pierden potencia con respecto a los del modelo lineal, por lo que no es una mala opción emplearlo si se tienen dudas sobre la linealidad del modelo. A mayores, se obtienen buenos resultados tanto bajo diseño fijo equidistante, como bajo diseño aleatorio uniforme, por lo que es, sin duda, bastante versátil. Asimismo, ninguno de los contrastes estudiados experimenta un empeoramiento de sus resultados al introducir errores exponenciales en lugar de errores normales en las simulaciones, por lo que, en caso de sospechar que los errores puedan ser exponenciales, estos test resultan bastante fiables. A mayores, los contrastes empleados se pueden ejecutar sin dificultad, pues, tal y como se mencionó anteriormente, se llevan a cabo con la función `sm.regression` del paquete `sm` de `R`. Quizás uno de los pocos inconvenientes de estos contrastes es que su potencia, en ciertos casos, disminuye bastante rápido a medida que baja el número de individuos de la muestra, es por ello que tal vez, en según que caso, no son test muy recomendables cuando los tamaños muestrales son de mucho menos de 100 observaciones. En conclusión, el ANCOVA no paramétrico destaca por ser útil a la vez que preciso y fácil de ejecutar, dejando claro que es una buena elección si el número de observaciones es suficiente.

Bibliografía

[Alonso-Pena, 2019] Alonso-Pena, M. (2019). *Nonparametric ANCOVA for Cylindrical and Toroidal Data*. Trabajo de Fin de Máster. Universidade de Santiago de Compostela.

[Bowman y Azzalini, 1997] Bowman, A. W. y Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. OUP Oxford.

[Bowman y Azzalini, 2021] Bowman, A. W. y Azzalini, A. (2021). R package sm:nonparametric smoothing methods (version 2.2-5.7).

[Fan, 1992] Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998-1004.

[Fan y Gijbels, 1996] Fan, J. y Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC.

[Faraway, 2004] Faraway, J. J. (2004). *Linear Models with R*. Chapman and Hall/CRC.

[Gasser y Müller, 1979] Gasser, T. y Müller, H.-G. (1979). Kernel estimation of regression functions. In Gasser, T. and Rosenblatt, M., *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics, pages 23-68. Springer.

[Gasser et al. 1986] Gasser, T., Sroka, L., y Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73(3):625-633.

[Johnson et al. 1995] Johnson, N. L., Kotz, S., y Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Volume 2. John Wiley & Sons.

[Manish Kumar, 2019] Manish Kumar. (2019). Car Price Prediction Multiple Linear Regression. Obtenido el 30/06/2022 de https://www.kaggle.com/datasets/hellbuoy/car-price-prediction.

[Maxwell y Kelley, 1990] Maxwell, Scott E., Delaney, H. D. y Kelley, K. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Chapman and Hall/CRC.

[Nadaraya, 1964] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141-142.

[Rice, 1984] Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12(4):1215-1230.59

[Silverman, 2018] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Chapman and Hall/CRC.

[Speckman, 1988] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(3):413-436.

[Wand y Jones, 1994] Wand, M. P. y Jones, M. C. (1994). *Kernel Smoothing*. Chapman and Hall.

[Wasserman, 2006] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

[Young y Bowman, 1995] Young, S. G. y Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics*, 51(3):920-931.

[Zeileis y Hothorn, 2002] Zeileis, A. y Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7-10.