FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE GEOMETRÍA Y TOPOLOGÍA

GRADO EN ESTADÍSTICA

Trabajo Fin de Grado

Título

Enfoque geométrico en Regresión

Realizado por:
Ana Borrego Fernández

Supervisado por:
Desamparados Fernández Ternero y
Carmen Márquez García

30 de Junio de 2021

Resumen

El objetivo de este proyecto es presentar un enfoque geométrico del análisis de regresión estadístico, basados en la distribución normal, de la forma más sencilla posible. La técnica estadística de regresión es un procedimiento útil y beneficioso para la obtención de conocimiento de un fenómeno a partir del estudio, en este caso, de dos características de forma que permite modelar el valor esperado de una variable dependiente a través del valor de una variable independiente. Este trabajo se centra en la regresión polinomial puesto que es un procedimiento diseñado para construir un modelo que permita describir el impacto de un sólo factor, o característica, sobre una variable dependiente, de forma que el modelo ajustado dependerá de la variable independiente y de las potencias de dicha variable.

Abstract

This project aims to introduce a geometric approach to statistical regression analysis in the simplest way possible, based on the normal distribution. The statistical regression technique is a useful and beneficial procedure for obtaining knowledge of a phenomenon as from the study, in this case, of two characteristics. It allows modelling the expected value of a dependent variable through the value of an independent variable. This work focuses on polynomial regression, since it is designed to build a model that allows describing the impact of a single factor, or characteristic, on a dependent variable. The adjusted model will depend on the independent variable and the powers of said variable.

Índice general

Introducción

En este trabajo se aborda una técnica estadística desde el punto de vista de la Geometría, como es la regresión. La Geometría permite representar de forma gráfica el problema que resuelve la regresión y expresar de forma sencilla las herramientas necesarias para su mejor comprensión. Para ello, se presentan como introducción conceptos estadísticos que se utilizaran a lo largo del proyecto.

Las técnicas de regresión tienen como objetivo el estudio de la relación existente entre dos o más características que se estudian sobre una población, de manera que proporciona una organización de los datos útil para los posteriores estudios. En este proyecto, se estudiarán relaciones entre dos características, pero todo lo explicado se puede generalizar para el estudio de más de dos, de manera que se pretende conocer de qué forma influyen ciertas variables dependientes en función de las independientes. Además, a través de la regresión, se puede explicar un fenómeno bajo estudio, como por ejemplo la variación del peso a través de una dieta, y, una vez se conoce el comportamiento del fenómeno, permite hacer predicciones futuras con cierto nivel de confianza, siguiendo el ejemplo, nos permite predecir cuánto tiempo se debe continuar con la dieta para alcanzar el peso óptimo.

Por todo esto, el análisis en regresión permite la toma de decisiones en diferentes ámbitos, y organizar estrategias beneficiosas para el objetivo marcado. En el anterior ejemplo presentado, permitirá el conocimiento de cómo afecta al peso de un individuo la dieta bajo estudio de modo que se podrá tomar decisiones sobre la variación de alimentos en la dieta para su mejora en función del objetivo marcado.

La estructura del proyecto se centra en conocer el procedimiento de regresión polinomial desde el punto de vista geométrico. Se analiza la regresión polinomial, pero se podrían considerar diferentes tipos de funciones que ajustaran los datos.
En el Capítulo 1, se presenta la regresión lineal, que al tratarse de la regresión más sencilla dentro de las polinomiales, nos servirá de base para una mejor

comprensión del Capítulo 2, en el que se expone la regresión polinomial en general.

En este segundo capítulo, se muestra el procedimiento a seguir cuando tenemos una población sobre las que se consideran dos características y pretendemos estudiar la relación polinomial existente entre ellas. Se hará una diferenciación según los datos estudiados, dividiendo así el estudio en ajustes con término de error puro y en ajustes con términos de error no puro.

Para ejemplificar lo explicado en el Capítulo 2, se presenta finalmente en los Capítulos 3 y 4 dos ejemplos, uno para cada procedimiento teórico descrito según el tipo de ajuste. Aunque en la memoria se indican los valores redondeados, los cálculos han sido realizados con mayor exactitud a través de RStudio. Además, todas las figuras han sido realizadas con GeoGebra.

Estadística - Conceptos básicos

La Estadística es una ciencia, con base matemática, que engloba procedimientos de recogida, análisis e interpretación de datos, que tiene como finalidad el conocimiento de fenómenos aleatorios permitiendo así realizar predicciones y tomar decisiones. Usualmente, la estadística se centra en estudios de características, de cualquier tipo, sobre una población.

Para entender esta ciencia se debe conocer diferentes conceptos que se utilizan en un estudio estadístico:

- **Población objetivo**: conjunto de individuos del que se está interesado en sacar conclusiones.
- **Individuo**: cada elemento que forma la población objetivo.
- **Muestra**: subconjunto de los individuos que forman la población, usualmente se escogen los individuos de forma aleatoria para así asegurar la representatividad de la muestra con respecto a la población.
- **Variable aleatoria**: es aquella que cuantifica la característica que se quiere estudiar sobre la población objetivo. Por ejemplo, el peso.
- **Parámetro**: cantidad numérica calculada sobre la población, por ejemplo la media, que es el valor medio esperado de la variable.
- **Estadístico**: cantidad numérica calculada sobre la muestra, usualmente utilizado para aproximar un parámetro. También denominado, estimador.
- **Varianza**: cuantifica la dispersión de los datos respecto a su respectiva media aritmética, obtenida como la media de las desviaciones que se producen en cada dato con respecto de la media.
- **Distribución de probabilidad**: función que asigna a cada suceso, o valor que toma la variable, una probabilidad de que dicho suceso ocurra, o que dicha variable tome dicho valor.

Existen muchas distribuciones conocidas, en este estudio trabajamos con datos poblaciones distribuidos según una ley Normal. A partir de ellos calcularemos estadísticos, a través de las muestras, que seguirán distribuciones sobradamente conocidas, puesto que aparecen con frecuencia en los análisis

muestrales. Se expondrán las distribuciones utilizadas en este estudio a continuación. Como es usual, para estas distribuciones utilizaremos tablas donde se recogen los valores de las funciones de distribución.
Además, se especificarán los correspondientes puntos críticos de las distribuciones tabuladas, para ello recordamos el concepto de cuantil:

- **Cuantil o punto crítico**: corresponde a un valor de la variable que divide la distribución de una variable aleatoria en dos intervalos. Sea X una variable aleatoria que sigue cualquier distribución, el punto crítico de la distribución de X a un nivel de significación de α será aquel que cumpla
$$P(X \leq x) = \alpha$$

Distribución normal

Un gran número de fenómenos aleatorios continuos se pueden modelar con la distribución Normal. Se dirá que una variable aleatoria X sigue una distribución normal con parámetros μ y σ^2, si su función de densidad es,

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \ , \ x \in \mathbb{R}$$

Los parámetros que diferencian las variables que siguen una distribución normal son $\mu = \mathbb{E}(X)$ y $\sigma^2 = Var(X)$. La distribución normal es simétrica con respecto de la media y por ello

$$P(X \leq \mu - x) = P(X \geq \mu + x)$$

Otra propiedad fundamental de la distribución normal es su *reproductividad*, es decir, teniendo $X \sim N(\mu_X, \sigma_X^2)$ independiente de $Y \sim N(\mu_Y, \sigma_Y^2)$, entonces, para cualesquiera a , b , $c \in \mathbb{R}$, se cumple que

$$(aX + bY + c) \ \sim \ N(a\mu_X + b\mu_Y + c \ , \ a^2\sigma_X^2 + b^2\sigma_Y^2) \ ,$$

y, bajo las mismas condiciones,

- $X + Y \ \sim N(\mu_X + \mu_Y \ , \ \sigma_X^2 + \sigma_Y^2)$ y
- $X - Y \ \sim N(\mu_X - \mu_Y \ , \ \sigma_X^2 + \sigma_Y^2)$.

En particular, si se tienen $X_1, \ldots, X_n$ variables independientes idénticamente distribuidas según una ley Normal $N(\mu, \sigma^2)$, para $i = 1, \ldots, n$, se tiene que

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ \sim \ N\left(\mu, \frac{\sigma^2}{\sqrt{n}}\right) .$$

La distribución normal de media 0 y varianza 1 se conoce como la distribución normal estándar, siendo la distribución normal más sencilla. Se puede conseguir una distribución normal estándar a partir de cualquier distribución normal utilizando la tipificación, que consiste en el siguiente procedimiento

$$X \sim N(\mu, \sigma^2) \implies Z = \frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1) .$$

Esta distribución, como hemos dicho, está tabulada. Es simétrica respecto al origen, puesto que $\mu = 0$, por tanto se verifica que

$$P(Z \leq -x) = P(Z \geq x) .$$

Cuantil o punto crítico de la distribución normal estándar: Dado un $\alpha \in [0, 1]$, mediante Z_α representamos aquel valor real tal que

$$P(Z \leq Z_\alpha) = \alpha ,$$

y se denomina *punto crítico de la distribución* $N(0, 1)$ *a nivel de significación* α.
Como Z es simétrica respecto del origen, se cumple que $Z_\alpha = -Z_{1-\alpha}$.

Distribución Chi-Cuadrado

Sean $X_1, \ldots, X_n$ variables aleatorias independientes idénticamente distribuidas según una ley Normal estándar, $X_i \sim N(0, 1)$, para $i = 1, \ldots, n$. Si consideramos la variable aleatoria

$$Y = (X_1^2 + X_2^2 + \cdots + X_n^2) ,$$

a la distribución de Y se le denomina *distribución chi-cuadrado con n grados de libertad*, lo que se representa como $Y \sim \chi_n^2$.

Distribución t-Student

Sean X e Y dos variables independientes de forma que $X \sim N(0,1)$ e $Y \sim \chi^2_n$. Si definimos la variable aleatoria

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \ ,$$

a la distribución de T se denomina *distribución t de Student con n grados de libertad*, y se representa como $T \sim t_n$.

Esta distribución tiene propiedades similares a la distribución normal estándar. La distribución t-Student es simétrica con respecto a su media, pero su media es 0, por tanto se verifica que $P(T \leq -t) = P(T \geq t)$.

Los cuantiles o puntos críticos, a nivel de significación α, de esta distribución se representan como $t_{n,\alpha}$ y, por la simetría respecto del origen, se cumple que: $t_{n,\alpha} = -t_{n,1-\alpha}$.

Distribución de Fisher-Snedecor

Sean X e Y dos variables aleatorias independientes tales que $X \sim \chi^2_n$ e $Y \sim \chi^2_m$. Si definimos la variable aleatoria

$$F = \frac{\frac{X}{n}}{\frac{Y}{m}} \ ,$$

a la distribución de F se le denomina *distribución F-Snedecor con n y m grados de liberdad*, y se representa como $F \sim F_{n,m}$.

Los cuantiles o puntos críticos de la distribución F-Snedecor con n y m grados de libertad se representan como $F_{n,m,\alpha}$.

Estimación por intervalos de confianza

Sea X una variable aleatoria con función de distribución que depende de un parámetro desconocido, θ, y sea $X_1, \ldots, X_n$ una muestra aleatoria procedente de la población descrita por la variable X. Se dirá que $I_{1-\alpha}(\theta)$ es un *intervalo aleatorio a nivel de significación* α, ó lo que es lo mismo, *intervalo aleatorio a nivel de confianza* $1-\alpha$, para cualquier valor de $\alpha \in [0,1]$, si se cumple:

$$P(\theta \in I_{1-\alpha}(\theta)) \geq 1 - \alpha \ ,$$

ó lo que es lo mismo

$$P(\theta \in I_{1-\alpha}(\theta)) < \alpha \ .$$

Se puede observar que el intervalo aleatorio de confianza corresponde a calcular dos cantidades numéricas: el extremo inferior y el extremo superior. Al intervalo de extremos dichas cantidades es lo que se denomina *intervalo de confianza al* $(1-\alpha) * 100\,\%$. Para el cálculo de estos extremos se utiliza el *método de la cantidad pivotal*, que consiste en formar una nueva variable, Y, a partir de la información que se tiene, que contenga el parámetro a estimar y que siga una distribución conocida, pero que sólo dependa de dicho parámetro, y considerar los extremos a y b tales que

$$P(a \leq Y \leq b) = 1 - \alpha$$

y operar hasta tener en el centro de la probabilidad el parámetro a estimar.

Interpretación de los intervalos de confianza

Si se pudieran obtener todos los intervalos de confianza al 95 % a partir de todas las posibles muestras que se pudieran extraer de la población objetivo, se sabría que el 95 % de todos esos intervalos contienen el valor verdadero del parámetro, y sólo el 5 % restante, no.

Contrastes de Hipótesis

Una hipótesis estadística será cualquier afirmación que se realiza sobre la distribución de la población o sobre algún parámetro de la misma. El problema consiste en decidir, a partir de la información muestral, si la información es correcta considerando cierto nivel de confianza o significación.
En los estudios de hipótesis se consideran las hipótesis nulas, H_0, las cuales se definen con la afirmación realizada.
Para resolver los contraste de hipótesis, se necesita una regla de decisión o procedimiento para rechazar o aceptar H_0. A esta regla se le denomina *test estadístico*. Normalmente, se rechaza una hipótesis cuando el "comportamiento" de la muestra dista del esperado bajo la hipótesis nula, en ese caso, podemos afirmar que existen evidencias muestrales para suponer la afirmación como errónea y aceptaríamos la hipótesis alternativa.

Siguiendo el razonamiento de un test de hipótesis, podrían ocurrir dos tipos de errores, *error tipo I*, el cuál será rechazar la hipótesis nula siendo ésta cierta, y el *error tipo II*, el cuál será aceptar la hipótesis nula siendo ésta falsa. Lo ideal sería tener un test que minimizara ambos errores para fallar lo menos posible, pero esto es imposible. Por tanto, lo que se hace es acotar

la probabilidad de ocurrencia de uno de ellos, el de tipo I. A dicha cota se le denomina *nivel de significación,* α, y una vez fijado, se trata de encontrar una regla de decisión, es decir, un test que minimice la probabilidad de error tipo II.

Relación entre intervalos de confianza y contrastes de hipótesis paramétricos

En los contrastes de hipótesis paramétricos, se considera que la variable aleatoria bajo estudio sigue una distribución conocida, si bien que se desconoce alguno de los parámetros que la determinan, es decir, el contraste de hipótesis está referido a algún parámetro, θ, que se desconoce y se quiere contrastar si toma el valor θ_0. Luego el contraste sería,

$$\begin{cases} H_0 : \ \theta \ = \theta_0 \\ H_1 : \ \theta \ \neq \theta_0 \, . \end{cases}$$

Existe una estrecha relación entre estos contrastes y los intervalos de confianza, de hecho es demostrable la siguiente equivalencia:

La hipótesis $H_0 : \theta = \theta_0$ es rechazada con un nivel de significación α si, y sólo si θ_0 no es un valor del intervalo de confianza para θ al nivel $1 - \alpha$.

Análisis de Regresión

El objetivo de la regresión en términos geométricos es sustituir la nube de puntos que representan los datos recogidos, por una línea que se adapte lo mejor posible a la nube de puntos. De forma que queda simplificado el comportamiento de las características estudiadas en el experimento.
En términos estadísticos, se pretende hacer una estimación de los valores de una característica, Y, que llamaremos dependiente, a partir de los valores de la característica X, que llamaremos independiente.

Por tanto, consiste en buscar una función, h, de la variable independiente que nos ayude a ajustar, predecir o aproximar los valores de Y. Para ello, utilizaremos el método de mínimos cuadrados, de forma que nuestro objetivo será buscar la función h que minimice el cuadrado de las diferencias entre los valores observados y_i y los valores ajustados por la función. Es decir, buscar h de forma que minimice

$$\sum_{i=1}^{n} (y_i - h(x_i))^2 \ .$$

La función h puede ser una función lineal, polinomial, logarítmica, ... En este trabajo se presenta la regresión lineal y la regresión polinomial a través del enfoque geométrico.

Enfoque geométrico

El procedimiento geométrico para la búsqueda del modelo se basará en definir un espacio construido a partir de las observaciones con un sistema de generadores ortonormales, que constituirá el espacio modelo M y completar el espacio hasta dimensión n, ampliando con vectores de nuevo ortonormales, teniendo en cuenta los errores de ajuste que se realizarán.
Se ajustarán modelos a través de vectores calculados con las observaciones de la variable independiente, X, que consideremos, con el fin de poder obtener información sobre la variable dependiente a través de la muestra considerada.
Para realizar una regresión polinomial desde el punto de vista de la geometría se seguirán los siguientes pasos:

1. Se estudiará cuál es el mayor orden polinomial que se pueda ajustar.

2. Se constituirá el espacio modelo y el espacio de errores, utilizando el vector de la variable independiente, de forma que se consiga una base ortonormal del espacio modelo que se amplía a una base ortonormal n-dimensional.

3. Se ajustará el modelo ortgonal mediante las proyecciones del vector dependiente sobre el espacio n-dimensional.

4. Se utilizará la descomposición de Pitágoras para calcular la suma de cuadrados de las proyecciones,

$$\|y - \bar{y}\|^2 = \|P_{U_1} y\|^2 + \cdots + \|P_{U_n} y\|^2$$

5. A través de los cuantiles/puntos críticos se estudiarán las componentes significativas.

6. Se ajustará el modelo adecuado y se presentará la información que éste proporciona.

Cabe destacar que los procedimientos explicados en los posteriores capítulos se basan en procedimientos estadísticos habituales, con la diferencia de que se trabaja con vectores y no con observaciones individuales.

Capítulo 1

Fundamentos geométricos en la regresión lineal

Se pretende estudiar si dos características o variables X e Y de una población están relacionadas siguiendo un modelo lineal a partir de n datos obtenidos de cada una de ellas de forma simultánea, es decir, de mediciones de las dos características sobre los n individuos de la población que constituirán la muestra con la que trabajaremos:

$$y_1, y_2, ..., y_n \text{ y } x_1, x_2, ..., x_n.$$

Por tanto, la idea es aplicar los métodos geométricos para interpretar el problema planteado y darle una solución.
Los vectores observación y de valores se construirán de la siguiente forma:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \text{ y } x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}.$$

El modelo de regresión lineal asume que la relación entre las variables X e Y es una línea recta $Y = \alpha + \beta X$ donde α corresponderá a la ordenada en el origen y β a la pendiente de la recta que relaciona las variables y la cuál se pretende ajustar.

Dentro del ajuste de los datos a un modelo de regresión nos podemos encontrar diferentes casos:

1. **Variables no relacionadas**: Si las variables X e Y no están relacionadas, entonces $\beta = 0$ y nuestro vector observación será parte de una nube de puntos centrada en el punto $[\alpha, \alpha, \alpha, \alpha, \alpha, ...]$.

2. **Variables relacionadas**: Si las variables están relacionadas, entonces $\beta \neq 0$ y nuestro vector observación será parte de una nube de puntos centrada en el punto final del vector al considerarlo en el origen de coordenadas:

$$\begin{bmatrix} \alpha+\beta x_1 \\ \alpha+\beta x_2 \\ \alpha+\beta x_3 \\ \vdots \\ \alpha+\beta x_n \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}.$$

Deducimos así que, si $\beta = 0$, la nube de puntos se centra en un punto que está sobre la línea equiangular, mientras que, si $\beta \neq 0$, la nube de puntos se aleja de la línea equiangular en una dirección del espacio bidimensional generado por los vectores $J = [1, 1, 1, 1, 1, ...]^t$ y $x = [x_1, x_2, ..., x_n]^t$.
Por tanto, el objetivo en primer lugar será decidir si existe esa relación de tipo lineal entre nuestras variables, contrastando la hipótesis $H_0 : \beta = 0$.

Una vez se tiene la existencia o no de la relación lineal, el siguiente objetivo será estimar los parámetros del modelo considerado (α y β) en caso de que existiera tal relación.

Al considerar el modelo de regresión lineal, supondremos lo siguiente en el estudio:

- Consideraremos el modelo lineal como $Y = \beta_0 + \beta_1(X - \bar{x})$ en lugar del modelo lineal expuesto anteriormente $Y = \alpha + \beta X$.

  Las relaciones entre estos dos modelos son, claramente; $\alpha = \beta_0 + \beta_1\bar{x}$ y $\beta = \beta_1$.

- La media de las observaciones de Y dependerá del valor de X fijado, x, con el que está asociado, a través de la relación lineal $E(Y) = \beta_0 + \beta_1(x - \bar{x})$, donde β_0 y β_1 son los parámetros desconocidos a estimar del modelo y $\bar{x}$ es la media muestral de los n valores de X.

- Para cada valor de x, Y se distribuye según una *Ley Normal* con la media indicada y varianza común σ^2. Así la distribución de Y para un valor de X dado es:

$$Y \sim N(\beta_0 + \beta_1(x - \bar{x}), \sigma).$$

- Los errores son valores independientes de una distribución Normal $N(0, \sigma)$.

Como el modelo considerado es $Y = \beta_0 + \beta_1(X - \bar{x})$, tenemos que el vector modelo resultante es:

$$\beta_0 + \beta_1(x - \bar{x}) = \beta_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix},$$

donde el espacio modelo M es un subespacio bidimensional del espacio n-dimensional de observaciones y el espacio de errores, que es el espacio ortogonal al modelo, tiene dimensión $n - 2$.

Una vez tenemos esto claro, podemos calcular un vector unitario U_1 a partir de la dirección J de la forma:

$$U_1 = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Necesitaríamos un segundo vector unitario U_2, contenido en el plano definido por U_1 y por el vector de valores de X. Además, lo calcularemos de forma que sea ortogonal a U_1 para así tener una primera base ortonormal del espacio modelo.

Para obtener dicho vector se calcula la diferencia entre el vector x y su proyección sobre la dirección calculada U_1, es decir:

$$x - P_{U_1}x = x - (x \cdot U_1)U_1 = x - \bar{x}$$

y dividiendo por su módulo, tenemos $U_2 = \frac{x - \bar{x}}{||x - \bar{x}||}$, donde $\bar{x}$ aquí denota el vector media de X.

Así, una base ortonormal del espacio modelo M está constituida por los vectores U_1 y U_2, $M = \langle U_1, U_2 \rangle$.

La idea ahora es ampliar dicha base hasta obtener una base ortonormal de dimensión n, así los nuevos $n - 2$ vectores constituirán una base del espacio de errores.

En primer lugar, ampliaremos hasta obtener una base n-dimensional y posteriormente, por ejemplo mediante el método de Gram-Schmidt, la transformamos en una base ortogonal que, finalmente, convertimos en una base ortonormal $\{U_1, U_2, U_3, ..., U_n\}$.
A partir de la base ortonormal que se obtiene, se puede conseguir la siguiente descomposición del vector observación y en componentes ortogonales, cada una de las cuales corresponde a una de las direcciones de la base ortonormal obtenida.

La descomposición ortogonal sería de la forma:

$$y = P_{U_1}y + P_{U_2}y + P_{U_3}y + \cdots + P_{U_n}y.$$

$$y = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + \cdots + (y \cdot U_n)U_n.$$

Podemos observar que:

- $P_{U_1}y = (y \cdot U_1)U_1 = \bar{y}$
- $P_{U_2}y = (y \cdot U_2)U_2 = \dfrac{y \cdot (x - \bar{x})}{\|x - \bar{x}\|}\dfrac{x - \bar{x}}{\|x - \bar{x}\|} = \dfrac{y \cdot (x - \bar{x})}{\|x - \bar{x}\|^2}(x - \bar{x}) = b(x - \bar{x})$

  donde $b = \dfrac{y \cdot (x - \bar{x})}{\|x - \bar{x}\|^2} = \dfrac{y_1(x_1 - \bar{x}) + \cdots + y_n(x_n - \bar{x})}{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}$.

Haciendo la diferencia:

$$y - P_{U_1}y - P_{U_2}y = y - \bar{y} - b(x - \bar{x}),$$

obtenemos el vector error:

$$e = P_{U_3}y + \cdots + P_{U_n}y = (y \cdot U_3)U_3 + \cdots + (y \cdot U_n)U_n.$$

Así, podemos ver la descomposición ortogonal anterior como esta otra descomposición ortogonal del vector observación en tres vectores: el *vector media*, $\bar{y}$, el *vector pendiente*, $b(x - \bar{x})$, y el *vector error*, e.

y	$=$	$\bar{y}$	$+$	$b(x - \bar{x})$	$+$	e
vector observación	$=$	vector media	$+$	vector pendiente	$+$	vector error

1.1. Distribuciones de las longitudes de las proyecciones

Para contrastar la hipótesis de interés necesitamos conocer cómo se distribuyen las longitudes de las proyecciones del vector observación, $Y \cdot U_i$, $i = 1, \ldots, n$. Conocer estas distribuciones también nos servirá para obtener estimaciones de los parámetros de nuestro modelo.

Se tiene que la media de $Y \cdot U_1$ es $\sqrt{n}\beta_0$:

$$y \cdot U_1 = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \cdot \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{y_1 + y_2 + \cdots + y_n}{\sqrt{n}}$$

y, sustituyendo la ecuación del modelo, $y_i = \beta_0 + \beta_1(x_i - \bar{x})$, vemos que tiene media

$$\frac{(\beta_0 + \beta_1(x_1 - \bar{x})) + (\beta_0 + \beta_1(x_2 - \bar{x})) + \cdots + (\beta_0 + \beta_1(x_n - \bar{x}))}{\sqrt{n}} =$$

$$= \frac{n\beta_0 + \beta_1(x_1 + x_2 + \cdots + x_n - n\bar{x})}{\sqrt{n}} = \frac{n\beta_0 + \beta_1(n\bar{x} - n\bar{x})}{\sqrt{n}} = \sqrt{n}\beta_0.$$

Análogamente, se obtiene que la media de $Y \cdot U_2$ es $\beta_1\|x - \bar{x}\|$, puesto que:

$$y \cdot U_2 = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \cdot \frac{(x - \bar{x})}{\|x - \bar{x}\|} = \frac{y_1(x_1 - \bar{x}) + y_2(x_2 - \bar{x}) + \cdots + y_n(x_n - \bar{x})}{\|x - \bar{x}\|},$$

así, la media es:

$$\frac{(\beta_0 + \beta_1(x_1 - \bar{x}))(x_1 - \bar{x}) + \cdots + (\beta_0 + \beta_1(x_n - \bar{x}))(x_n - \bar{x})}{\|x - \bar{x}\|} =$$

$$= \frac{\beta_0(x_1 + x_2 + \cdots + x_n - n\bar{x}) + \beta_1(x - \bar{x})(x - \bar{x})}{\|x - \bar{x}\|} =$$

$$= \frac{\beta_0(n\bar{x} - n\bar{x}) + \beta_1\|x - \bar{x}\|^2}{\|x - \bar{x}\|} = \beta_1\|x - \bar{x}\|.$$

También se tiene que la media de $Y \cdot U_e$ es nula, donde $U_e = [e_1, e_2, \ldots, e_n]^t$ es una dirección cualquiera en el espacio de errores.

$$Y \cdot U_e = y_1 e_1 + \cdots + y_n e_n$$

Despejando la ecuación del modelo $y_i = \beta_0 + \beta_1(x_i - \bar{x})$, tenemos que

$$y_1 e_1 + \cdots + y_n e_n = (\beta_0 + \beta_1(x_1 - \bar{x}))e_1 + \cdots + (\beta_0 + \beta_1(x_n - \bar{x}))e_n =$$

$$= \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = 0$$

Esto es, como U_e es el vector del espacio de errores, es ortogonal a la dirección J y a $(x - \bar{x})$, por tanto sus productos escalares son iguales a 0.

Así las longitudes de las proyecciones se distribuyen normalmente con varianza común σ^2. Las distribuciones de $Y \cdot U_1$ e $Y \cdot U_2$ se centran en cantidades potencialmente distintas de cero, mientras que las distribuciones $Y \cdot U_3, \ldots, Y \cdot U_n$ siempre están centradas en cero. O sea,

- $Y \cdot U_1 \sim N(\sqrt{n}\beta_0, \sigma)$,
- $Y \cdot U_2 \sim N(\beta_1\|x - \bar{x}\|, \sigma)$ y
- $(Y \cdot U_3), \ldots, (Y \cdot U_n) \sim N(0, \sigma)$.

1.1.1. Estimación de los parámetros del modelo

Al conocer las distribuciones de las longitudes de las proyecciones, podemos utilizarlas para estimar las constantes β_0 y β_1 del modelo.

Sabemos que $Y \cdot U_1$ sigue una ley normal de parámetros $\sqrt{n}\beta_0$ y σ y que:

$$y \cdot U_1 = \frac{1}{\sqrt{n}}(y_1 + y_2 + \cdots + y_n) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} y_i\right) = \frac{n}{\sqrt{n}}\bar{y} = \sqrt{n}\bar{y}\,.$$

Por otra parte, sabemos que la media de $Y \cdot U_1$ es $\sqrt{n}\beta_0$, podemos estimar β_0 y obtener:

$$\sqrt{n}\bar{y} = \sqrt{n}\beta_0 \implies \hat{\beta}_0 = \frac{\sqrt{n}}{\sqrt{n}}\bar{y} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Un estimador de β_0, $\hat{\beta}_0$, es la media muestral del vector observación y.

Para estimar ahora β_1 vamos a utilizar la longitud de la proyección en U_2, que sabemos que sigue una ley normal de media $\beta_1\|x - \bar{x}\|$ y varianza σ^2. Por tanto, despejando obtenemos una estimación de β_1:

$$y \cdot U_2 = \hat{\beta}_1\|x - \bar{x}\| \implies \hat{\beta}_1 = \frac{y \cdot U_2}{\|x - \bar{x}\|} = \frac{y_1(x_1 - \bar{x}) + \cdots + y_n(x - \bar{x})}{(x - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}.$$

Mientras que las direcciones $U_3, \ldots, U_n$ las utilizaremos para estimar la varianza de la población σ^2:

$$\hat{\sigma}^2 = s^2 = \frac{(y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2}{(n-2)} = \frac{\|e\|^2}{(n-2)}.$$

Como conocemos las relaciones que tienen β_0 y β_1 con los parámetros del modelo de regresión, con sus estimaciones podemos obtener los estimadores de la ordenada en el origen y de la pendiente de la recta de regresión:

- La pendiente coincide con el parámetro β_1, por tanto, la estimación de la pendiente será $\hat{\beta} = \hat{\beta}_1$.
- La ordenada en el origen se relaciona con los parámetros de nuestro modelo de la forma $\alpha = \beta_0 + \beta_1\bar{x}$, luego su estimación será $\hat{\alpha} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$.

1.2. Contraste de hipótesis

Para contrastar la hipótesis $H_0 : \beta_1 = 0$ la dirección de interés es U_2, ya que la longitud media de la proyección $Y \cdot U_2$ depende directamente de si β_1 es cero o no.
Por ello, comparamos el cuadrado de la longitud de la proyección sobre el vector pendiente, $(y \cdot U_2)^2$, y el promedio de los cuadrados de las longitudes sobre el espacio de los errores:

$$\frac{(y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2}{n-2}.$$

Utilizando la descomposición de Pitágoras asociada a la descomposición ortogonal del vector observación y:

$$\|y\|^2 = (y \cdot U_1)^2 + (y \cdot U_2)^2 + (y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2,$$

$$\|y\|^2 = (y \cdot U_1)^2 + (y \cdot U_2)^2 + \|e\|^2.$$

Tenemos, por tanto, que el test estadístico es:

$$F = \frac{(y \cdot U_2)^2}{[(y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2]/(n-2)} = \frac{(y \cdot U_2)^2}{\|e\|^2/(n-2)} \tag{1.1}$$

Si la hipótesis nula es cierta, $\beta_1 = 0$, las longitudes de las proyecciones $Y \cdot U_2, Y \cdot U_3, \ldots, Y \cdot U_{n-1}$ e $Y \cdot U_n$ siguen una distribución $N(0, \sigma)$ y, así:

Como $(Y \cdot U_i) \sim N(0, \sigma)$ para $i = 2, \ldots, n$, si tipificamos:

$$\frac{(Y \cdot U_i) - 0}{\sqrt{\sigma^2}} = \frac{(Y \cdot U_i)}{\sigma} \sim N(0,1), i = 2, \ldots, n$$

Tenemos variables normales estándares, que sumadas al cuadrado dan lugar a una distribución chi-cuadrado. Separando los cuadrados de las longitudes de la proyección sobre el vector pendiente y de las proyecciones sobre el espacio de errores tenemos que:

$$\frac{(Y \cdot U_2)^2}{\sigma^2} \sim \chi^2_1$$

e

$$\frac{(Y \cdot U_3)^2 + \cdots + (Y \cdot U_n)^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Por tanto, para formar el estadístico estamos dividiendo dos distribuciones Chi-cuadrado entre ellas y entre sus grados de libertad, de la forma:

$$F = \frac{\dfrac{(y \cdot U_2)}{\sigma^2}}{\dfrac{(y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2}{\sigma^2(n-2)}} \sim F_{1,n-2}$$

Simplificamos σ^2 y tenemos la expresión del test (1.1):

$$F = \frac{(y \cdot U_2)^2}{\dfrac{(y \cdot U_3)^2 + \cdots + (y \cdot U_n)^2}{(n-2)}} \sim F_{1,n-2}.$$

Una vez conocemos la distribución del estadístico, pasamos a estudiar si el valor de F observado es grande o pequeño comparando con los percentiles 90, 95 y 99 de la distribución de F-Snedecor con 1 y $n-2$ grados de libertad.

1.2.1. Contraste mediante el coeficiente de correlación

También podemos contrastar si la pendiente se anula, es decir, si $\beta_1 = 0$, mediante el *coeficiente de correlación lineal de Pearson*, r, que se define como el coseno del ángulo entre los vectores $x - \bar{x}$ e $y - \bar{y}$:

$$r = \frac{(x - \bar{x}) \cdot (y - \bar{y})}{\|x - \bar{x}\| \|y - \bar{y}\|}.$$

El coeficiente de correlación nos indica si existe relación de tipo lineal entre las variables, y, por tanto, si $\beta_1 \neq 0$, teniendo en cuenta que cuanto más próximo esté $|r|$ de 1, más fuerte será la correlación lineal entre las variables.
Además, el signo de r nos indica si la correlación entre dichas variables es directa o inversa, es decir, si la característica Y crece o decrece a medida que aumentan los valores de la característica X.

Capítulo 2

Fundamentos geométricos en la regresión polinomial

En el Capítulo 1, se ha supuesto que la relación existente entre dos características es de tipo lineal, pero esto no tiene por qué ocurrir, en muchos casos, esta suposición es claramente falsa correspondiendo la relación a una función matemática más compleja. Existen diferentes tipos de dependencia aún considerando las funciones elementales, las funciones polinómicas, que puede ser lineales (grado polinomial 1), cuadráticas (grado polinomial 2), cúbicas (grado polinomal 3),...

En este capítulo, se pretende estudiar las diferentes dependencias existentes entre dos características mediante la elección del grado polinomial que mejor ajuste los datos.

En la realidad, no siempre se da realmente una relación polinomial entre dos características bajo estudio, pero vamos a suponer que existe un ajuste polinomial perfecto para cada conjunto de datos bajo estudio.
La elección del modelo polinomial apropiado para ciertos datos es, hasta cierto punto, arbitrario. Normalmente, se prefieren los modelos polinomiales de menor grado a modelos polinomiales de mayor grado, puesto que estos son más complejos.

Se observarán dos situaciones a la hora de analizar los datos: experimentos donde no hay dos valores iguales de Y para un mismo valor de X y experimentos donde existen dos o más valores de Y para un mismo valor de X. A estas situaciones se les denominaran, respectivamente, *ajuste con término de error no puro* y *ajuste con término de error puro.*

Se expondrá el procedimiento a seguir correspondiente a la situación en la que nos encontremos, y se aplicarán a los experimentos que se detallan en los Capítulos 3 y 4.

2.1. Modelo

Tanto si tenemos término de error puro como si no, una vez tenemos los datos podremos ajustar hasta un modelo polinomial de orden igual al número de valores diferentes que tenemos en X disminuido en 1 unidad. Es decir, si estuviéramos en un caso de término de error no puro, en el que la variable X toma 5 valores distintos, podremos ajustar un modelo polinomial de orden hasta 4, y si tuviéramos una situación de término de error puro y tuviéramos 15 valores para X, pero sólo 7 fueran distintos, podríamos ajustar un modelo polinomial de hasta orden 6.

En general, si X toma k valores distintos, se podría ajustar modelos polinomiales de orden $0, 1, 2, \ldots, (k-1)$ y la secuencia de los $k-1$ modelos polinomiales posibles son:

$$
\begin{array}{ll}
\text{Modelo orden 0:} & y = \alpha_0 \\
\text{Modelo orden 1:} & y = \alpha_0 + \alpha_1(x - \bar{x}) \\
\text{Modelo orden 2:} & y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 \\
\text{Modelo orden 3:} & y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \alpha_3(x - \bar{x})^3 \\
 & \vdots \\
\text{Modelo orden } k-1: & y = \alpha_0 + \alpha_1(x - \bar{x}) + \cdots + \alpha_k(x - \bar{x})^{k-1}
\end{array}
$$

donde los valores de α_i no son los mismos en cada modelo, es decir, el α_0 del modelo lineal es distinto del α_0 del modelo constante, por ejemplo.

Nótese que si no tenemos distintos valores de Y para un mismo valor de X, $k = n$.

Para el orden polinomial que nos proporcionaría el ajuste perfecto, asumimos las siguientes hipótesis:

- Los valores del vector Y están distribuidos según una Ley Normal con varianza constante alrededor de la curva ajustada.
- Los errores de estimación cometidos en el ajuste son independientes.

2.1.1. Ortogonalización del modelo

Si convertimos la secuencia de posibles modelos polinomiales a una secuencia de posibles modelos ortogonales, obtenemos que:

$$
\begin{array}{ll}
\text{Modelo orden 0:} & y = \beta_0 \\
\text{Modelo orden 1:} & y = \beta_0 + \beta_1 p_1(x) \\
\text{Modelo orden 2:} & y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) \\
\text{Modelo orden 3:} & y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_3(x) \\
 & \vdots \\
\text{Modelo orden } k-1: & y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) + \cdots + \beta_{k-1} p_{k-1}(x) \quad ,
\end{array}
$$

donde $p_0(x), p_1(x), p_2(x), p_3(x), \ldots, p_k(x)$ serán las componentes constante, lineal, cuadrática, cúbica, etc, del modelo que están predeterminadas por la sucesión de valores de X.

Es conveniente convertir las secuencias de los posibles modelos en una secuencia de modelos polinomiales ortogonales, ya que nos proporciona la ventaja de que en el modelo ortogonal los coeficientes polinomiales, $\beta_i, i = 0, \ldots, k-1$, no variarán entre los distintos modelos. Por ello, es suficiente ajustar el modelo completo para posteriormente eliminar componentes que no sean explicativas en el modelo.

Cuando escribimos de forma vectorial el modelo no ortogonal completo queda de la forma:

$$Y = \alpha_0 X_1 + \alpha_1 X_2 + \alpha_2 X_3 + \alpha_3 X_4 + \cdots + \alpha_{k-1} X_k \,,$$

esto es porque, en términos vectoriales, $X_1 = 1$, $X_2 = (x - \bar{x})$, $X_3 = (x - \bar{x})^2$, $X_4 = (x - \bar{x})^3$, $\ldots$, $X_k = (x - \bar{x})^{k-1}$.

Para ortogonalizar el modelo, en primer lugar se calculan los correspondientes vectores del modelo $X_1, X_2, \ldots, X_{k-1}$. Así, el conjunto de vectores X_i, serán, respectivamente,

$$
\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} , \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ x_3 - \bar{x} \\ \vdots \\ x_k - \bar{x} \end{bmatrix} , \begin{bmatrix} (x_1 - \bar{x})^2 \\ (x_2 - \bar{x})^2 \\ (x_3 - \bar{x})^2 \\ \vdots \\ (x_k - \bar{x})^2 \end{bmatrix} , \ldots , \begin{bmatrix} (x_1 - \bar{x})^{k-1} \\ (x_2 - \bar{x})^{k-1} \\ (x_3 - \bar{x})^{k-1} \\ \vdots \\ (x_k - \bar{x})^{k-1} \end{bmatrix} .
$$

Una vez calculados estos vectores, utilizaremos el método de Gram-Schmidt para ortogonalizarlos. Este método consiste en hacer que cada vector sea ortogonal a sus predecesores restándole al vector su proyección sobre el espacio generado por los predecesores.

Gram-Schmidt

De cada vector, se calculará el vector ortogonal T_i, y el correspondiente vector unitario, U_i, sobre las que se realizarán las proyecciones. Sabemos que el vector $X_1 = 1$ y éste es ortogonal a $X_2 = (x - \bar{x})$, por tanto, los dos primeros vectores ortogonales T_1 y T_2 ya los tenemos calculados, sólo faltarían los correspondientes vectores unitarios,

$$T_1 = X_1 \ , \ T_2 = X_2 \, ,$$

y los vectores unitarios se calculan como,

$$U_i = \frac{1}{\|T_i\|} T_i \implies U_1 = \frac{1}{\|T_1\|} T_1 \ , \ U_2 = \frac{1}{\|T_2\|} T_2 \, .$$

El vector $X_1 = 1$, por tanto, su módulo será igual a la raíz cuadrada del número de observaciones n del estudio, así

$$U_1 = \frac{1}{\|T_1\|} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \, .$$

Para el resto de vectores, siguiendo el procedimiento, se calculará T_i y U_i de la forma,

$$T_i = X_i - \sum_{j=1}^{i-1} P_{U_j} X_i \ , \text{ e } \ U_i = \frac{1}{\|T_i\|} T_i \, .$$

Por tanto, ya tendríamos calculada una base del espacio modelo, $T_1, \ldots, T_k$. A partir de aquí, la forma de ajustar un modelo polinomial cambia en función de si tenemos términos de error puro o no puro.

- Si tuviéramos término de error no puro, es decir, no existen varios valores de Y para un mismo valor de X ($k = n$). Los vectores unitarios ortogonales calculados por Gram-Schmidt, generarán el espacio modelo, $M = \langle U_1, \ldots, U_n \rangle$, y éste completa el espacio de dimensión n necesario para el ajuste.

- Si tuviéramos término de error puro, es decir, existen varios valores de Y para un mismo valor de X ($k \neq n$), los vectores unitarios ortogonales calculados formarán el espacio modelo, pero no completarán el espacio de dimensión n necesario. Por tanto, la base obtenida, debe ser ampliada con lo que será el espacio de error puro.

$$M = \langle U_1, \ldots, U_k \rangle \ , \ E = \langle U_{k+1} \ldots, U_n \rangle$$

La descomposición ortogonal del modelo completo de grado máximo que ajusta los datos será la proyección del vector y sobre el espacio modelo y el espacio de errores:

$$y = P_{U_1} y + P_{U_2} y + \cdots + P_{U_k} y + P_{U_{k+1}} y + \cdots + P_{U_n} y \ .$$

Sabiendo que la proyección del vector y sobre la dirección U_1 es el vector media de las observaciones, la descomposición de Pitágoras será:

$$\|y - (y \cdot U_1)^2\|^2 = \|P_{U_2} y\|^2 + \|P_{U_3} y\|^2 + \cdots + \|P_{U_k} y\|^2 + \|e\|^2 \, ,$$

donde $\|P_{U_i} y\|^2 = (y \cdot U_i)^2$ para $i = 1, \ldots, k$, y $\|e\|^2 = (y \cdot U_{k+1})^2 + \cdots + (y \cdot U_n)^2$, por tanto, tenemos que:

$$\|y - (y \cdot U_1)^2\|^2 = (y \cdot U_2)^2 + \cdots + (y \cdot U_k)^2 + \cdots + (y \cdot U_n)^2 \, .$$

Componentes polinomiales

En el modelo ortogonal, T_1 es el vector de valores para la componente constante, $p_0(x) = 1$, y T_2 es el vector de valores para la componente de orden 1 o componente lineal, $p_1(x) = (x - \bar{x}) = T_2$.
Luego, $T_3 = X_3 - P_{U_1} X_3 - P_{U_2} X_3 = X_3 - (X_3 \cdot U_1) U_1 - (X_3 \cdot U_2) U_2$ será el vector de valores para la componente cuadrática del modelo que, con las expresiones obtenidas para $p_0(x)$ y $p_1(x)$, y teniendo en cuenta cómo está definido X_3, será

$$T_3 = X_3 - \frac{(X_3 \cdot U_1)}{\|T_1\|} T_1 - \frac{(X_3 \cdot U_2)}{\|T_2\|} T_2 \Rightarrow$$

$$\Rightarrow p_2(x) = (x - \bar{x})^2 - \frac{(X_3 \cdot U_1)}{\|T_1\|} p_0(x) - \frac{(X_3 \cdot U_2)}{\|T_2\|} p_1(x) =$$

$$= (x - \bar{x})^2 - \frac{(X_3 \cdot U_1)}{\|T_1\|} 1 - \frac{(X_3 \cdot U_2)}{\|T_2\|} (x - \bar{x}) \ .$$

Este procedimiento se puede generalizar para el cálculo de cualquier componente polinomial de la siguiente forma. Siguiendo el procedimiento, la componente cúbica, de orden 3, y la componente de orden 4 serán, respectivamente,

$$p_3(x) = (x - \bar{x})^3 - \frac{X_4 \cdot U_1}{\|T_1\|} p_0(x) - \frac{X_4 \cdot U_2}{\|T_2\|} p_1(x) - \frac{X_4 \cdot U_3}{\|T_3\|} p_2(x) \ ,$$

$$p_4(x) = (x-\bar{x})^4 - \frac{X_5 \cdot U_1}{\|T_1\|} p_0(x) - \frac{X_5 \cdot U_2}{\|T_2\|} p_1(x) - \frac{X_5 \cdot U_3}{\|T_3\|} p_2(x) - \frac{X_5 \cdot U_4}{\|T_3\|} p_3(x) \ .$$

Así, tenemos que de forma general la componente polinomial de orden i, se obtiene según

$$p_i(x) = (x - \bar{x})^i - \left[\sum_{j=1}^{i-1} \frac{X_{i+1} \cdot U_j}{\|T_j\|} \, p_{j-1}(x) \right] \ , \ i = 0, 1, \ldots$$

2.2. Procedimiento de elección del modelo adecuado

Una vez tenemos calculado los vectores que generan el espacio n-dimensional, y de ellos sabemos cuáles generan el correspondiente espacio modelo, M, y el de errores, para el modelo de mayor grado que podemos aproximar, de grado $k-1$, pasaremos a ver los diferentes procedimientos para la elección del modelo que mejor se adecúe a los datos. Los procedimientos son distintos dependiendo en qué caso nos encontremos, es decir, dependiendo si estamos en el caso de un ajuste con término de error puro o, en su defecto, en el caso de ajuste con término de error no puro.

En este punto, el *objetivo* es eliminar las componentes que no sean significativas del modelo de mayor orden ajustable, con el fin de obtener el modelo polinomial más sencillo que proporcione el mejor ajuste de los datos.

2.2.1. Ajuste con término de error no puro

Sabemos que en este caso, el espacio modelo M completa el espacio n-dimensional debido a que no existen varias observaciones de Y para un mismo valor de X ($k = n$). Por tanto, para el modelo completo no se tendrá en cuenta ningún espacio de errores.

$$M = \langle U_1, U_2, \ldots, U_k \rangle = \langle U_1, U_2, \ldots, U_n \rangle$$

En primer lugar, para el modelo completo se presentará, además de su ecuación, la correspondiente tabla ANOVA reflejada en la Tabla 2.1.

La tabla ANOVA para un modelo polinomial de orden $k-1$ está formada normalmente por 4 columnas, la primera se refiere a las componentes polinomiales de cada orden que participan en el modelo, y para cada una de dichas componentes se calculan los grados de libertad (gl), la suma cuadrática, SC, y la suma media de cuadrados, MC, que no es más que la suma cuadrática dividida por los grados de libertad. Los grados de libertad son iguales a 1 ya que estos coinciden con la dimensión de cada componente.

La tabla termina con el cálculo de la suma de cuadrados total para el modelo.

Orden polinomial	grados de libertad (gl)	SC	$MC = SC/(gl)$
Orden 1	1	$(y \cdot U_2)^2$	$(y \cdot U_2)^2$
Orden 2	1	$(y \cdot U_3)^2$	$(y \cdot U_3)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden i	1	$(y \cdot U_{i+1})^2$	$(y \cdot U_{i+1})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden $k-1$	1	$(y \cdot U_k)^2$	$(y \cdot U_k)^2$
Total	$k-1$	$\sum_{i=2}^{k}(y \cdot U_i)^2$	

Tabla 2.1: Cálculo y forma de la tabla ANOVA en ajustes con término de error no puro

Se observará que la componente suma de cuadrados decrece rápidamente con el incremento del orden polinomial en el caso de que el ajuste mediante el modelo de grado máximo no fuera el correcto.

El procedimiento de selección del orden polinomial adecuado es el siguiente:

1. Empezaremos considerando la constante polinomial e iremos incrementando el orden hasta que la siguiente componente polinomial se considere no representativa.

2. Una componente se considera no representativa sólo cuando es no significativa frente a:

   a) un estimador de la varianza, s^2, obtenido a través de la media de las sumas de cuadrados de las componentes de grado mayor al grado considerado.

b) un estimador de la varianza, obtenido a través de la media de las sumas de cuadrados de las componentes polinomiales de mayor grado exceptuando a la componente justo posterior a la considerada.

La condición (b) se incluye para salvaguardar las ocasiones en las que el estimador obtenido en (a) se ve inflado por un valor alto de la componente siguiente a la que estamos estudiando.

Este procedimiento irá acompañado de una tabla resumen con la estructura que se presenta en la Tabla 2.2.

		Condición *a*)			Condición *b*)		
Componente bajo estudio	SC	s^2	gl	F estad	Var *b)*	gl	F estad
Orden 1	$(y \cdot U_2)^2$	s_1^2	$n-2$	SC_1/s_1^2	b_1^2	$n-3$	SC/b_1^2
Orden 2	$(y \cdot U_3)^2$	s_2^2	$n-3$	SC_2/s_2^2	b_2^2	$n-4$	SC/b_2^2
Orden 3	$(y \cdot U_4)^2$	s_3^2	$n-4$	SC_3/s_3^2	b_3^2	$n-5$	SC/b_3^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden $k-2$	$(y \cdot U_{k-1})^2$	s_{k-2}^2	$n-k+1$	$\frac{SC_{k-2}}{s_{k-2}^2}$	-	-	-
Orden $k-1$	$(y \cdot U_k)^2$	-	-	-	-	-	-

Tabla 2.2: Tabla resumen del procedimiento

En dicha tabla, se obtienen estadísticos F tanto en el estudio de la condición *a*) como en la de *b*), estos estadísticos nos ayudarán a decidir cuáles de las componentes son significativas. Se puede estudiar esto de dos formas:

- Comparando si las sumas de cuadrados son suficientemente grandes con respecto al estadístico F obtenido.
- Calculando la distribución del estadístico F y comparándolo con los cuantiles correspondientes según el nivel de significación que se quiera asignar.

Todo esto se verá de forma detallada en el ejemplo que se presenta en el Capítulo 3.

Una vez se decide el orden polinomial que mejor ajusta los datos, las correspondientes componentes de ordenes superiores pasarán a formar el espacio de errores para este ajuste, manteniendo así la correspondiente dimensión del espacio en el que se está trabajando, n.

2.2.2. Ajuste con término de error puro

En este caso tenemos un espacio n-dimensional mayor que el espacio modelo M, puesto que tenemos un espacio de errores constituido por los vectores unitarios calculados por Gram-Schmidt para completar el espacio dimensional del problema.

$$M = \langle U_1, U_2, \ldots, U_k \rangle \ , \ E = \langle U_{k+1}, \ldots U_n \rangle \ .$$

Al igual que en el anterior ajuste, se obtendrá en primer lugar la correspondiente tabla ANOVA para el modelo completo de la misma forma que la anterior, con la diferencia de que ahora tenemos un espacio de error puro que también debemos considerar.

Orden polinomial	grados de libertad (gl)	SC	$MC = SC/(gl)$
Orden 1	1	$(y \cdot U_2)^2$	$(y \cdot U_2)^2$
Orden 2	1	$(y \cdot U_3)^2$	$(y \cdot U_3)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden i	1	$(y \cdot U_{i+1})^2$	$(y \cdot U_{i+1})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden k	1	$(y \cdot U_{k+1})^2$	$(y \cdot U_{k+1})^2$
Error Puro	$n-k$	SC_e	$MC_e = s^2$
Total	$n-1$	$\sum_{i=2}^{n}(y \cdot U_i)^2$	

Tabla 2.3: Cálculo y forma de la tabla ANOVA en ajustes con término de error puro

SC_e será la suma de los términos $(y \cdot U_i)^2$ correspondientes al espacio de errores e y MC_e será la media de la suma cuadrática de los errores, calculada dividiendo la suma de cuadrados debido al error, SC_e, por los correspondientes grados de libertad, los cuáles equivalen a la dimensión de dicho espacio de errores.

El procedimiento de elección del modelo polinomial más sencillo que mejor ajuste los datos se basa en una ampliación de la tabla ANOVA, añadiéndole un primer estadístico, y un estudio del llamado "déficit de ajuste", el cuál se refiere a la falta de ajuste producida en los distintos valores de X debido a que tienen asignados varios valores de Y.
A partir de la tabla ANOVA, se calcula un primer estadístico F para cada componente, dividiendo las sumas de cuadrados entre el estimador s^2. Por tanto, sabemos que todos esos estadísticos seguirán la misma distribución F-Snedecor con 1 y $n-k$ grados de libertad.
Por otro lado, se calcula la suma cuadrática media del déficit de ajuste para cada componente polinomial de forma que para cada componente, esta suma equivaldrá al promedio de la suma de cuadrados de las componentes de orden mayor. A partir de esa suma cuadrática media del déficit, se calculará un segundo estadístico, F' a partir de la división de esta suma cuadrática media entre s^2. Estos segundos estadísticos no seguirán la misma distribución F-Snedecor, puesto que dependerán del número de componentes que participen en el numerador de F'.

Todo esto, se recogerá en una tabla resumen tal y como se muestra en la Tabla 2.4.

Una vez tenemos todo lo anterior calculado, el procedimiento consiste en empezar estudiando desde la componente de orden polinomial 0 e ir aumentando el orden del polinomio hasta que, tanto el estadístico $F_{1,n-k}$, como el nuevo estadístico F' referido a la parte del déficit de ajuste sean no significativos, cuándo se compara con el término de error puro.

Comp. polinomial	gl	SC	$F_{1,n-k}$	Comp. polinomial	Déficit de ajuste gl	MC_d	F'
Orden 1	1	SC_1	SC_1/s^2	Orden 0	$k-1$	$MC_{d,0}$	F'_1
Orden 2	1	SC_2	SC_2/s^2	Orden 1	$k-2$	$MC_{d,1}$	F'_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Orden $k-1$	1	SC_{k-1}	$\frac{SC_{k-1}}{s^2}$	Orden $k-2$	1	$MC_{d,k-2}$	F'_{k-2}
Error	$n-k$	SC_e			$n-k$	$MC_e = s^2$	
Total	$n-1$	$\sum_{i=2}^{n}(y \cdot U_i)^2$					

Tabla 2.4: Resumen del procedimiento de elección en situaciones de término de error puro.

El procedimiento se explica detalladamente en el ejemplo del Capítulo 4.

2.3. Modelo que mejor ajusta los datos

Una vez hemos decidido cuál es el modelo polinomial que mejor ajusta los datos, se retoma la ecuación del modelo completo

$$y = P_{U_1}y + P_{U_2}y + \cdots + P_{U_k}y + P_{U_{k+1}}y + \cdots + P_{U_n}y\,,$$

se puede reescribir como

$$y = \frac{y \cdot U_1}{\|T_1\|}T_1 + \frac{y \cdot U_2}{\|T_2\|}T_2 + \cdots + \frac{y \cdot U_{k-1}}{\|T_k\|}T_k + P_e y$$

de donde podemos obtener la ecuación del modelo completo, ya que T_1 es el vector de valores de $p_0(x)$, T_2 el de $p_1(x)$, ..., así

$$y = \frac{y \cdot U_1}{\|T_1\|}p_0(x) + \frac{y \cdot U_2}{\|T_2\|}p_1(x) + \cdots + \frac{y \cdot U_{k-1}}{\|T_k\|}p_k(x)\,.$$

Finalmente, de esta última descomposición podemos sacar una fórmula para calcular una estimación de los coeficientes polinomiales β_i, $\hat{\beta}_i$,

$$\hat{\beta}_i = \frac{y \cdot U_{i-1}}{\|T_{i-1}\|}\ ,\ i = 1, \ldots, k\ .$$

Cabe destacar que, como el modelo ortogonal tiene la ventaja que los coeficientes β_i son los mismos para cualquier orden, solo debemos calcular los coeficientes que participen en el modelo decidido en el punto anterior.

2.4. Volviendo a la Geometría

Si escogemos un modelo polinomial de orden j, nos decantaríamos por un nuevo espacio modelo de dimensión $j+1$, y un espacio de errores de dimensión $n-(j+1)$. El modelo ajustado sería:

$$\begin{array}{lccccc}
 & y & = & P_{U_1}y + P_{U_2}y + \cdots + P_{U_j}y & + & \text{vector error}\ , \\
\text{ó} & y & = & (y \cdot U_1)U_1 + \cdots + (y \cdot U_j)U_j & + & \text{vector error}\ , \\
\text{ó} & y & = & \hat{y} & + & (y - \hat{y})\ , \\
 & \text{Vector} & & \text{Vector} & & \text{Vector} \\
 & \text{observación} & & \text{valores ajustados} & & \text{error}\ .
\end{array}$$

Se observa que el vector de valores ajustados es igual a la proyección de y sobre el espacio modelo y el vector error es igual a la diferencia entre las observaciones, y, y los correspondientes valores ajustados, $\hat{y}$. El vector error también puede ser calculado como la suma de las proyecciones de y sobre los vectores unitarios que generan el espacio de errores.

Una vez se tiene ajustado el modelo, se presenta la tabla ANOVA final para el modelo ajustado, con el correspondiente valor de la suma de cuadrados del espacio de errores que se forma, que tiene la misma estructura que la Tabla 2.3.

2.5. Comprobación de las hipótesis

Una vez ajustado el modelo, faltaría comprobar que se cumplen las suposiciones realizadas para hacer el ajuste, es decir,

- las observaciones son independientes y se distribuyen según una ley normal con varianza constante alrededor de la curva de ajuste, y
- la independencia de los errores.

La independencia de las observaciones se comprueban según la recogida de datos realizada para el experimento, si la característica se estudia en cada individuo de forma independiente y aleatoria, se podrá asumir la independencia de las observaciones.
La normalidad en los datos se estudia realizando un histograma de los errores, donde se comprueba que si los errores se ajustarían a la distribución normal que seguiría, podríamos asumir que las observaciones se distribuyen según una normal al igual que los errores procedentes de los ajustes realizados para cada una de ellas. Se busca que el histograma sea más o menos simétrico. Esto es debido a la reproductividad de la normal, puesto que los errores son calculados a partir de las observaciones y los valores ajustados, a partir de las observaciones, luego para que lo errores siguieran la normal obligatoriamente las observaciones deben seguir una normal.
Para la independencia de los errores, basta con representar los errores cometidos en cada valor ajustado frente al correspondiente error ajustado. Si se observara un patrón, los errores serían dependientes. Además se podrá observar si la varianza es constante o no, observando la dispersión de los puntos del gráfico.

2.5.1. La geometría en términos de error puro

Cabe destacar que en este caso, teníamos un déficit de ajuste y un término de error puro al principio del estudio, por tanto, nuestro nuevo vector error calculado al decidir el modelo polinomial que ajustamos se podrá descomponer en la suma del vector de error puro, que consiste en las desviaciones verticales de las observaciones con la curva ajustada, y el vector error del déficit de ajuste, que consiste en las desviaciones que se producen entre las medias de los tratamientos, x, y los valores apropiados a la curva ajustada. De esta forma,

$$\begin{array}{ccccc} \text{Vector} & & \text{Vector error} & & \text{Vector error} \\ \text{error} & = & \text{déficit de ajuste} & + & \text{puro} \\ (y-\hat{y}) & & (\bar{y_{i\cdot}}-\hat{y}) & & (y-\bar{y_{i\cdot}}) \end{array} ,$$

donde $\bar{y_{i\cdot}}$ denota la media muestral de las observaciones correspondientes a cada valor distinto de x.

Por tanto, la suma de cuadrados del término error del modelo ajustado será igual a

$$\begin{array}{cccccc} & SC_e & = & SC \text{ Error de déficit de ajuste} & + & SC \text{ Error puro} \\ \text{ó} & SC_e & = & SC_{ed} & + & SC_{ep} \\ \text{ó} & \|y-\hat{y}\|^2 & = & \|\bar{y_{i\cdot}}-\hat{y}\|^2 & + & \|y-\bar{y_{i\cdot}}\|^2 \end{array} .$$

2.6. Coeficiente de correlación múltiple, R, y coeficiente de determinación R^2

Como vimos en el Capítulo 1, en la regresión simple o lineal el *coeficiente de correlación*, r, es el coseno del ángulo entre el vector $y-\bar{y}$ y el vector $x-\bar{x}$, y toma valores entre -1 y 1. Esto de forma numérica es lo mismos que el coseno del ángulo, θ entre $y-\bar{y}$ y el vector ajustado corregido, que en la regresión simple es

$$\hat{y}-\bar{y} = (y \cdot U_2)U_2 = \hat{\beta}_1 T_2 = \hat{\beta}_1(x-\bar{x}) .$$

Este coeficiente se interpretaba según el signo del valor resultante de forma que para valores positivos la relación era directa y para valores negativos era inversa.

Para regresiones polinomiales con orden mayor a 1, se define el *coeficiente de correlación múltiple*, R, calculado igual que el coeficiente de correlación

pero donde el vector de los valores ajustados corregido está constituido por más de una componente.
Para regresiones polinomiales de orden mayor a 1, no podemos atribuir una interpretación a la dirección de la regresión con más de una variable independiente, por tanto, el coeficiente R se define siempre como no negativo. Luego, por definición, θ estará contenido en el intervalo de $0°$ a $90°$ y R tendrá un valor de 0 a 1.

Por otro lado, el *coeficiente de determinación*, R^2, es el cuadrado del coeficiente de correlación múltiple, y nos proporciona una forma de reescribir la descomposición de Pitágoras del modelo, lo que se verá con detalle en el Capítulo 3.
La interpretación del coeficiente de determinación R^2 es la proporción de la suma total de cuadrados que se explica en el modelo de regresión ajustado, de modo que si R^2 es 1, el ajuste será perfecto, ya que explicaría la suma de cuadrados total de las observaciones reales.

A partir del coeficiente de determinación se puede obtener bastante información sobre el ajuste realizado, al relacionarlo con el estadístico F de la tabla ANOVA del modelo, y utilizándolo como herramienta para tener una medida del porcentaje de varianza explicado por el modelo.

2.7. Intervalos de confianza

Para los dos tipos de ajuste que podemos realizar, ajustes con términos de error puro o no puro, los intervalos de confianza siguen la misma estructura.

$$\text{estimador} \mp t - \text{valor} \times se(\text{estimador}) \,,$$

donde el $se(estimador)$ define el error estándar que se produce en la estimación, y tendrá una expresión diferente dependiendo del parámetro que se quiera ajustar a través de los estimadores.

Estos dos últimos apartados son "información extra" por lo que podemos prescindir de ellos y el estudio estaría completo. En el Capítulo 3 se explican de forma detallada y se podría calcular de manera análoga para cualquier ejemplo.

Capítulo 3

Ejemplo Error No Puro

Se han recogido datos sobre el aumento de peso, en gramos por oveja y día, de cinco rebaños de ovejas gamosas (viejas y desdentadas) alimentados con distintos niveles de asignación de pastos, medidos en kilogramos de materia seca (DM) por oveja y día. Las observaciones recogidas son las siguientes:

Nivel de asignación de pastos (X)	1.05	1.20	1.35	1.50	2.00
Peso ganado (Y)	-65	-42	-19	-8	20

Para facilitar la visualización de cómo están distribuidos los datos, vamos a representar el diagrama de dispersión, en la Figura 3.1, del aumento de peso frente al nivel de asignación de los pastos utilizados para cada rebaño

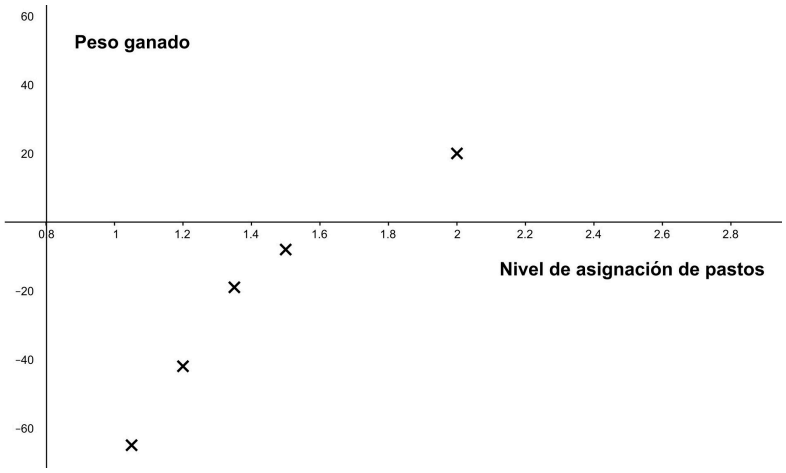

Figura 3.1: Diagrama de dispersión

3.1. Modelo

Como no tenemos dos valores de Y para un mismo valor de X, es decir, tenemos 5 valores distintos de la variable X, estamos en un caso con término de error no puro.
Por tanto, al tener 5 valores distintos de X, no podremos ajustar un modelo polinomial de grado mayor a 4, ya que no podemos ajustar un modelo de grado mayor al número de valores distintos de X reducido en una unidad.

La secuencia de posibles modelos polinomiales es desde el modelo constante al modelo de grado 4:

Modelo constante:	$y = \alpha_0$
Modelo lineal:	$y = \alpha_0 + \alpha_1(x - \bar{x})$
Modelo cuadrático:	$y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2$
Modelo cúbico:	$y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \alpha_3(x - \bar{x})^3$
Modelo de grado 4:	$y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \cdots + \alpha_4(x - \bar{x})^4$

donde los valores α_i, $i = 1, \ldots, 4$, varían de cada línea a la siguiente línea.

Asumimos que existe un grado polinomial que aporta un ajuste perfecto a los datos y para este grado polinomial realizamos las suposiciones:

- La variable Y se distribuye según una ley normal con varianza constante, σ^2, alrededor de la media real de la curva a la que se ajusta.
- Los errores de muestreo obtenidos en el ajuste son independientes.

Es conveniente convertir la secuencia de posibles modelos en una secuencia de modelos polinomiales ortogonales.

3.1.1. Ortogonalización del modelo

La secuencia de modelos polinomiales ortogonales posibles es:

Modelo constante:	$y = \beta_0$
Modelo lineal:	$y = \beta_0 + \beta_1 p_1(x)$
Modelo cuadrático:	$y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x)$
Modelo cúbico:	$y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_3(x)$
Modelo de grado 4:	$y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_3(x) + \beta_4 p_4(x)$

donde $p_0(x) = 1$, $p_1(x)$, $p_2(x)$, $p_3(x)$ y $p_4(x)$ serán la componente constante, lineal, cuadrática, cúbica y de orden 4, y vienen determinadas por el patrón de valores de la variable X.

La ventaja de utilizar este modelo es que los coeficientes estimados, β_0, β_1, β_2, β_3 y β_4, no variarán de un grado polinomial al siguiente. Por ello, es suficiente ajustar primero el modelo completo, de orden 4 en este caso, y después eliminar las componentes que no sean necesarias para el ajuste.

Vamos a escribir en forma vectorial el valor de los vectores, X_1, X_2, X_3, X_4 y X_5, y el modelo completo:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} -65 \\ -42 \\ -19 \\ -8 \\ 20 \end{bmatrix}, X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

$$X_2 = (x - \bar{x}) = \begin{bmatrix} 1.05 - 1.42 \\ 1.2 - 1.42 \\ 1.35 - 1.42 \\ 1.5 - 1.42 \\ 2 - 1.42 \end{bmatrix} = \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix},$$

$$X_3 = (x - \bar{x})^2 = \begin{bmatrix} (1.05 - 1.42)^2 \\ (1.2 - 1.42)^2 \\ (1.35 - 1.42)^2 \\ (1.5 - 1.42)^2 \\ (2 - 1.42)^2 \end{bmatrix} = \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix},$$

$$X_4 = (x - \bar{x})^3 = \begin{bmatrix} (1.05 - 1.42)^3 \\ (1.2 - 1.42)^3 \\ (1.35 - 1.42)^3 \\ (1.5 - 1.42)^3 \\ (2 - 1.42)^3 \end{bmatrix} = \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix},$$

$$X_5 = (x - \bar{x})^4 = \begin{bmatrix} (1.05 - 1.42)^4 \\ (1.2 - 1.42)^4 \\ (1.35 - 1.42)^4 \\ (1.5 - 1.42)^4 \\ (2 - 1.42)^4 \end{bmatrix} = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix}.$$

Por tanto el modelo polinomial no-ortogonal de grado 4,

$$y = \alpha_0 X_1 + \alpha_1 X_2 + \alpha_2 X_3 + \alpha_3 X_4 + \alpha_4 X_5,$$

resulta:

$$\begin{bmatrix} -65 \\ -42 \\ -19 \\ -8 \\ 20 \end{bmatrix} = \alpha_0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \alpha_1 \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix} + \alpha_3 \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} + \alpha_4 \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix}.$$

Para ortogonalizar este conjunto de vectores, debemos utilizar el método de Gram-Schmidt. Este método consiste en hacer que cada vector sea ortogonal a sus predecesores restándole al vector su proyección sobre el espacio generado por sus predecesores.

Gram-Schmidt

Llamaremos $T_1 \ldots, T_5$ a los vectores ortogonalizados y $U_1, \ldots, U_5$ a sus respectivos vectores unitarios.
Como X_1 y X_2 ya son ortogonales, los primeros dos vectores serán:

$$T_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, T_2 = \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix}$$

y sus correspondientes vectores unitarios:

$$U_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, U_2 = \frac{1}{\sqrt{0,533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix}$$

Ahora vamos a ortogonalizar X_3, según el método de Gram-Schmidt:

$$T_3 = X_3 - P_{U_1} X_3 - P_{U_2} X_3 = X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2$$

$$(X_3 \cdot U_1) = \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix} \cdot \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{0.532}{\sqrt{5}} = 0,238;$$

$$(X_3 \cdot U_2) = \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix} \cdot \frac{1}{\sqrt{0,533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} = \frac{0.134}{\sqrt{0.533}} = 0,183.$$

Luego,

$$T_3 = \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix} - 0.238U_1 - 0.183U_2 =$$

$$= \begin{bmatrix} 0.137 \\ 0.048 \\ 0.005 \\ 0.006 \\ 0.336 \end{bmatrix} - 0.1064 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - \frac{0.183}{\sqrt{0.533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix}.$$

Por tanto, el vector unitario asociado, U_3, será:

$$U_3 = \frac{T_3}{\|T_3\|} = \frac{1}{\sqrt{0,0438}} \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix}.$$

Repetiremos el proceso para calcular T_4 y T_5.

$$T_4 = X_4 - P_{U_1}X_4 - P_{U_2}X_4 - P_{U_3}X_4 =$$
$$= X_4 - (X_4 \cdot U_1)U_1 - (X_4 \cdot U_2)U_2 - (X_4 \cdot U_3)U_3 =$$

$$(X_4 \cdot U_1) = \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} \cdot \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{0.134}{\sqrt{5}} = 0.0599;$$

$$(X_4 \cdot U_2) = \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} \cdot \frac{1}{\sqrt{0,533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} = \frac{0.13431}{\sqrt{0.533}} = 0.183969;$$

$$(X_4 \cdot U_3) = \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} \cdot \frac{1}{\sqrt{0,0438}} \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix} = \frac{0.01013679}{\sqrt{0.0438}} = 0.0484.$$

Luego,

$$T_4 = \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} - 0,0599U_1 - 0.1894U_2 - 0.0484U_3 =$$

$$= \begin{bmatrix} -0.0507 \\ -0.0106 \\ -0.0003 \\ 0.0005 \\ 0.1951 \end{bmatrix} - \begin{bmatrix} 0.0268 \\ 0.0268 \\ 0.0268 \\ 0.0268 \\ 0.0268 \end{bmatrix} - \frac{0.183969}{\sqrt{0,533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} - \frac{0.0484}{\sqrt{0,0438}} \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix} =$$

$$= \begin{bmatrix} -0.0128 \\ 0.0187 \\ 0.01 \\ -0.0186 \\ 0.0027 \end{bmatrix}.$$

Por tanto, el vector unitario asociado, U_4, será:

$$U_4 = \frac{T_4}{\|T_4\|} = \frac{1}{\sqrt{0.00097}} \begin{bmatrix} -0.0128 \\ 0.0187 \\ 0.01 \\ -0.0186 \\ 0.0027 \end{bmatrix}$$

Por último,

$$T_5 = X_5 - P_{U_1}X_5 - P_{U_2}X_5 - P_{U_3}X_5 - P_{U_4}X_5 =$$
$$= X_5 - (X_5 \cdot U_1)U_1 - (X_5 \cdot U_2)U_2 - (X_5 \cdot U_3)U_3 - (X_5 \cdot U_4)U_4 =$$

$$(X_5 \cdot U_1) = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix} \cdot \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \frac{0.1343}{\sqrt{5}} = 0.06006079$$

$$(X_5 \cdot U_2) = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix} \cdot \frac{1}{\sqrt{0.533}} \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} = \frac{0.58186}{\sqrt{0.533}} = 0.07969937$$

$$(X_5 \cdot U_3) = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix} \cdot \frac{1}{\sqrt{0,0438}} \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix} = \frac{0.0118101}{\sqrt{0,0438}} = 0.05639404$$

$$(X_5 \cdot U_4) = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix} \cdot \frac{1}{\sqrt{0.00097}} \begin{bmatrix} -0.0128 \\ 0.0187 \\ 0.01 \\ -0.0186 \\ 0.0027 \end{bmatrix} = \frac{0.0001114631}{\sqrt{0.00097}} = 0.003581706$$

Luego,

$$T_5 = \begin{bmatrix} 0.01874 \\ 0.00234 \\ 0.00002 \\ 0.00004 \\ 0.11316 \end{bmatrix} - 0.06006079U_1 - 0.07969937U_2 - 0.05639404U_3 - 0.003581706U_4 =$$

$$= \begin{bmatrix} 0.00050 \\ -0.00181 \\ 0.00224 \\ -0.00097 \\ 0.00004 \end{bmatrix}.$$

Por tanto, el vector unitario asociado, U_5, será:

$$U_5 = \frac{T_5}{\|T_5\|} = \frac{1}{\sqrt{0.000009}} \begin{bmatrix} 0.00050 \\ -0.00181 \\ 0.00224 \\ -0.00097 \\ 0.00004 \end{bmatrix}$$

Componentes polinomiales

En el modelo ortogonal, T_1 es el vector de los valores de la componente constante del modelo, $p_0(x) = 1$, mientras que T_2 es el vector de valores para la componente lineal, $p_1(x) = (x - \bar{x}) = (x - 1.42) = T_2$.

Por tanto, $T_3 = X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2$ será el vector de valores para la componente cuadrática del modelo, que con las expresiones obtenidas para $p_0(x)$ y $p_1(x)$, resultará ser:

$$p_2(x) = X_3 - \frac{(X_3 \cdot U_1)}{\|T_1\|}\, p_0(x) - \frac{(X_3 \cdot U_2)}{\|T_2\|}\, p_1(x) =$$

$$= (x - \bar{x})^2 - \frac{0.238}{\sqrt{5}}\, 1 - \frac{0.183}{\sqrt{0.533}}\, (x - \bar{x}) =$$

$$= (x - 1.42)^2 - 0.1064 - 0.251\, (x - 1.42) \ .$$

De la misma forma, calcularemos la componente cúbica y de orden 4 del modelo:

Como $T_4 = X_4 - (X_4 \cdot U_1)U_1 - (X_4 \cdot U_2)U_2 - (X_4 \cdot U_3)U_3$, la componente de orden 3 será:

$$p_3(x) = X_4 - \frac{X_4 \cdot U_1}{\|T_1\|}\, p_0(x) - \frac{X_4 \cdot U_2}{\|T_2\|}\, p_1(x) - \frac{X_4 \cdot U_3}{\|T_3\|}\, p_2(x) =$$

$$= (x - \bar{x})^3 - 0.0268\, p_0(x) - 0.2510\, p_1(x) - 0.2311\, p_2(x) =$$

$$= (x-\bar{x})^3 - 0.0268 - 0.2510(x-1.42) - 0.2311[(x-1.42)^2 - 0.107 - 0.1889(x-1.42)]$$

Como $T_5 = X_5 - (X_5 \cdot U_1)X_5 - (X_5 \cdot U_2)U_2 - (X_5 \cdot U_3)U_3 - (X_5 \cdot U_4)U_4$, la componente de orden 4 será:

$$p_4(x) = X_5 - \frac{X_5 \cdot U_1}{\|T_1\|}\, p_0(x) - \frac{X_5 \cdot U_2}{\|T_2\|}\, p_1(x) - \frac{X_5 \cdot U_3}{\|T_3\|}\, p_2(x) - \frac{X_5 \cdot U_4}{\|T_4\|}\, p_3(x) =$$

$$= (x - \bar{x})^4 - 0.0269\, p_0(x) - 0.1092\, p_1(x) - 0.2693\, p_2(x) - 0.1150\, p_3(x) =$$

$$= (x - \bar{x})^4 - 0.0269 - 0.1092(x - 1.42) - 0.2693\, p_2(x) - 0.1150\, p_3(x).$$

En estos términos, el modelo polinomial ortogonal de orden 4, será:

$$y = \beta_0 p_0(x) + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_1(x) + \beta_4 p_4(x) =$$
$$= \beta_0 + \beta_1(x - 1.42) + \beta_2[(x - 1.42)^2 - 0.107 - 0.251(x - \bar{x})] + \beta_3\, p_3(x) + \beta_4\, p_4(x).$$

Finalmente, tenemos un vector observación $y = [-65, -42, -19, -8, 20]^t$, el espacio modelo generado por los vectores unitarios $M = \langle U_1, \ldots U_5 \rangle$, y las direcciones $U_1, \ldots, U_5$ correspondientes a las componentes ortogonales polinomiales de orden de 0 a 4.

3.2. Ajustar el modelo adecuado

Para el modelo completo, el espacio modelo M completa el espacio dimensional 5, y el sistema de coordenadas ortogonal es $U_1, \ldots, U_5$ como sigue:

$$\frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -0.507 \\ -0.301 \\ -0.096 \\ 0.110 \\ 0.794 \end{bmatrix}, \begin{bmatrix} 0.590 \\ -0.015 \\ -0.400 \\ -0.575 \\ 0.401 \end{bmatrix}, \begin{bmatrix} -0.411 \\ 0.603 \\ 0.319 \\ -0.598 \\ 0.087 \end{bmatrix}, \begin{bmatrix} 0.163 \\ -0.587 \\ 0.727 \\ -0.315 \\ 0.012 \end{bmatrix}.$$

Para ajustar el modelo, se realiza la proyección del vector observación y sobre el espacio modelo $M = \langle U_1, \ldots, U_5 \rangle$:

$$y = P_{U_1} y + P_{U_2} y + P_{U_3} y + P_{U_4} y + P_{U_5} y =$$
$$= (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 + (y \cdot U_4)U_4 + (y \cdot U_5)U_5$$
$$= -50.98U_1 + 62.4U_2 - -17.4438U_3 + 1.872735U_4 - 3.033868U_5\,.$$

Así, la descomposición ortogonal resultante, descrita en la Figura 3.2, es:

$$y = -50.98U_1 + 62.4U_2 - 17.4U_3 + 1.9U_4 - 3.0U_5\,.$$

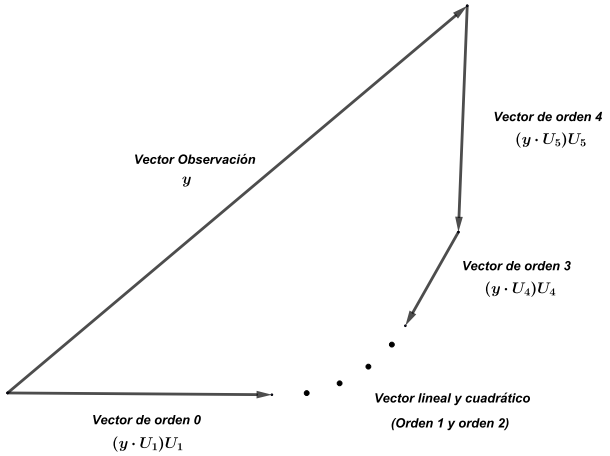

Figura 3.2: Descomposición ortogonal del vector observación

Como sabemos que la proyección de y sobre U_1 es la media de las observaciones, la forma correcta de escribir la descomposición ortogonal será:

$$y - (-50.98U_1) = 62.4U_2 - 17.4U_3 + 1.9U_4 - 3.03U_5$$

Podemos calcular así, la descomposición ortogonal de Pitágoras para el modelo completo:

$$\|y - (y \cdot U_1)^2\|^2 = (y \cdot U_2)^2 + (y \cdot U_3)^2 + (y \cdot U_4)^2 + (y \cdot U_5)^2$$

La correspondiente tabla ANOVA será:

Tipo de ajuste	grados de libertad (gl)	SS $= (y \cdot U_i)^2 =$ $= MS$
Lineal	1	3897.8
Cuadrática	1	304.3
Cúbica	1	3.5
Orden 4	1	9.2
Total	4	4214.8

La componente suma de cuadrados decrece rápidamente con el incremento del orden polinomial.

El procedimiento de selección del orden polinomial es el siguiente:

1. Empezaremos considerando la constante polinomial e iremos incrementando el orden hasta que la siguiente componente polinomial se considere no representativa.

2. Una componente se considera no representativa sólo cuando es no significativa frente a:

   a) un estimador de la varianza, s^2, obtenido a través de la media de las sumas de cuadrados de las componentes de grado mayor al grado considerado.

   b) un estimador de la varianza, obtenido a través de la media de las sumas de cuadrados de las componentes polinomiales de mayor grado exceptuando a la componente justo posterior a la considerada.

   La condición (b) se incluye para salvaguardar las ocasiones en las que el estimador obtenido en (a) se ve inflado por un valor alto de la componente siguiente a la que estamos estudiando.

En nuestro ejemplo:

Componente bajo estudio	*SC*	s^2	gl	*F* estad	Varianza *b)*	gl	*F* estad
Lineal	3897.8	105.7	3	36.9 (**)	6.36	2	613.6
Cuadrática	304.3	6.36	2	47.8 (*)	9.2	1	33.1
Cúbica	3.5	9.2	1	0.4 (ns)	-	-	-
Orden 4	9.2	-	-	-	-	-	-

Cuando comprobamos la componente lineal, la estimación de la varianza, s^2, se ve inflada a 105.7, deduciendo así que la componente cuadrática pertenece al conjunto de errores. Si observamos la tabla, vemos que la suma de cuadrados de la componente lineal, 3897.8, es suficientemente grande en comparación con el estadístico $F_{1,3} = 36.9$, así, la componente lineal es significativa en el modelo (**).

Cuando estudiamos la componente cuadrática, observamos que la suma de cuadrados es suficientemente grande en comparación al estadístico F obtenido, 47.8, deducimos así que la componente cuadrática es significativa para el modelo (*). En este caso, la estimación *(b)* no es tan precisa como la estimación s^2, esto es debido a que los grados de libertad se ven afectados por el bajo orden considerado.

Al estudiar la componente cúbica, nos sale que es no significativa (ns), debido a que la estimación de la varianza s^2 es mayor que la suma de cuadrados para esta componente.

El procedimiento nos está sugiriendo que el modelo que mejor ajusta es, de forma intuitiva, el modelo polinomial de orden 2 o cuadrático, debido a que se observa una gran diferencia entre este y los posteriores grados polinomiales. Se observa que la estimación (b) en este caso no nos proporciona ninguna información y podríamos haber llegado a la misma conclusión sin tener en cuenta dicha estimación. Por otro lado, si la suma de cuadrados de la componente lineal y cuadrática hubieran tenido valores similares, la estimación (b) hubiera sido esencial.

Podemos también estudiar cuál es el modelo adecuado estudiando los estadísticos para cada componente, a través de los percentiles de la distribución F-Snedecor. Como vimos para la regresión lineal en el Capítulo 1, el estimador F sigue una distribución F-Snedecor con grados de libertad iguales a las dimensiones del espacio modelo y espacio de errores considerados. Considerábamos como hipótesis nula que el correspondiente coeficiente era igual a 0, con vistas a saber si existía relación de tipo lineal entre las variables. Podemos generalizar este procedimiento para otros de relaciones, cuadrática, cúbica, etc.

Si consideramos un modelo lineal, tendríamos un valor del estadístico $F_1 = 36.9$ siguiendo una distribución $F_{1,3}$. Si queremos un nivel de significación igual al 0.01, es decir, una confianza del 99 %, compararemos el estadístico F obtenido con el percentil 0.99 de la distribución F-Snedecor con 1 y 3 grados de libertad.

$$F_1 = 36.9 > 34.116 = F_{1,3,0.99}$$

Como el estadístico obtenido es mayor que el percentil, no hay evidencias para negar la existencia de una componente lineal en el modelo, pero sin estudiar las otras componentes aún no podemos decidir que la relación que existe es de tipo lineal. Podemos asumir así, que la componente lineal es significativa en el modelo.

Análogamente, estudiamos la componente cuadrática teniendo un estadístico F de valor 47.8, compararemos este estadístico con el percentil 0.99 de la distribución $F_{2,2}$:

$$F_2 = 47.6 > 99.000 = F_{2,2,0.99}$$

Por tanto, la componente de orden 2 o cuadrática también es significativa en el modelo. Comprobaremos ahora la última componente para la cuál es posible el cálculo del estadístico, la componente de orden 3, teniendo una distribución $F_{3,1}$.

$$F_3 = 0.4 < 5403.534 = F_{3,1,0.99}$$

Obteniendo así que la componente de orden 3 es no significativa en el modelo, y, por tanto, no es explicativa, llegando a la misma conclusión que con el procedimiento descrito anteriormente.

Completaremos el análisis asumiendo un modelo polinomial de orden 2.

El modelo polinomial cuadrático

Retomando la descomposición ortogonal del modelo completo,

$$y = -50.98U_1 + 62.4U_2 - 17.4U_3 + 1.9U_4 - 3.0U_5,$$

que lo podemos reescribir de la siguiente forma:

$$y = \frac{-50.98}{\|T_1\|}T_1 + \frac{62.4}{\|T_2\|}T_2 - \frac{17.4}{\|T_3\|}T_3 + \frac{1.9}{\|T_4\|}T_4 - \frac{3.0}{\|T_5\|}T_5.$$

donde T_1 es el vector de valores para $p_0(x)$, sabiendo que $p_0(x) = 1$, T_2 es el vector de valores para $p_1(x) = (x - \bar{x})$, T_3 el de valores para $p_2(x)$, y así sucesivamente. Por tanto, la ecuación del modelo ortogonal completo será:

$$y = -22.8 + \frac{62.4}{\|T_2\|}p_1(x) - \frac{17.4}{\|T_3\|}p_2(x) + \frac{1.9}{\|T_4\|}p_3(x) - \frac{3.0}{\|T_5\|}p_4(x)$$

Con el modelo ortogonal tenemos la ventaja de que los coeficientes β_i, eran los mismos sin importar el orden polinomial que consideremos, luego de esta expresión podemos obtener la estimación de los coeficientes del modelo cuadrático.

Entonces, los coeficientes estimados $\hat{\beta}_i$ del modelo polinomial cuadrático serán :

- $\hat{\beta}_0 = \dfrac{y \cdot U_1}{\|T_1\|} = \dfrac{-50.98}{\sqrt{5}} = -22.8 = \bar{y}$
- $\hat{\beta}_1 = \dfrac{y \cdot U_2}{\|T_2\|} = \dfrac{62.4}{\sqrt{0.533}} = 85.5$
- $\hat{\beta}_2 = \dfrac{y \cdot U_3}{\|T_3\|} = \dfrac{-17.4}{\sqrt{0.0438}} = -83.1$

En conclusión, el modelo polinomial que mejor ajusta los datos es de orden 2 con la siguiente expresión (el ajuste realizado se ve descrito en la Figura 3.3).

$$y = -22.8\, p_0(x) + 85.5\, p_1(x) - 83.1\, p_2(x) =$$
$$y = -22.8 + 85.5(x - \bar{x}) - 83.1[(x - \bar{x})^2 - 0.251(x - \bar{x}) - 0.1064] =$$
$$= -22.8 + 85.5(x - 1.42) - 83.1[(x - 1.42)^2 - 0.251(x - 1.42) - 0.1064]$$

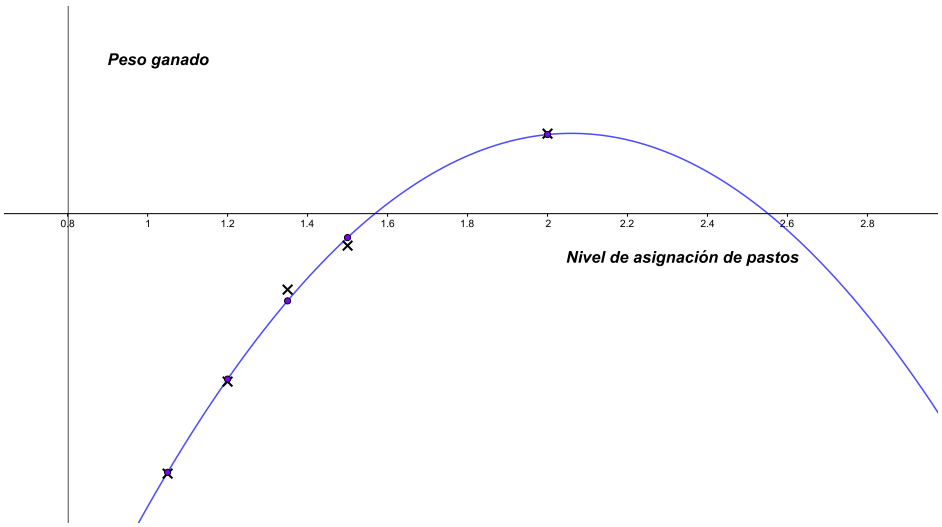

Figura 3.3: Ajuste polinomial de orden 2 a la nube de puntos

3.3. Geometría

Al decidirnos por un modelo cuadrático, hemos establecido que el espacio modelo tenga dimensión tres y, en consecuencia, que el espacio de errores tenga dimensión dos. Nuestro modelo ajustado es,

$$\begin{array}{rcccl} y & = & P_{U_1}y + P_{U_2}y + P_{U_3}y & + & \text{vector error} \\ y & = & (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 & + & \text{vector error} \\ y & = & \hat{y} & + & (y - \hat{y}) \end{array}$$

Es decir,

$$\hat{y} = (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 =$$
$$= \begin{bmatrix} -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \end{bmatrix} + \begin{bmatrix} -31.6 \\ -18.8 \\ -5.99 \\ 6.8 \\ 49.6 \end{bmatrix} + \begin{bmatrix} -10.3 \\ 0.3 \\ 6.98 \\ 10.03 \\ -7.00 \end{bmatrix} = \begin{bmatrix} -64.7 \\ -41.3 \\ -21.8 \\ -5.9 \\ 19.8 \end{bmatrix}$$

y el vector error lo calcularemos de la forma: $e = y - \hat{y}$,

$$e = y - \hat{y} = \begin{bmatrix} -65 \\ -42 \\ -19 \\ -8 \\ 20 \end{bmatrix} - \begin{bmatrix} -64.7 \\ -41.3 \\ -21.8 \\ -5.9 \\ 19.8 \end{bmatrix} = \begin{bmatrix} -0.3 \\ -0.7 \\ 2.8 \\ -2.1 \\ 0.2 \end{bmatrix}.$$

Por tanto,

$$\begin{bmatrix} -65 \\ -42 \\ -19 \\ -8 \\ 20 \end{bmatrix} = \begin{bmatrix} -64.7 \\ -41.3 \\ -21.8 \\ -5.9 \\ 19.8 \end{bmatrix} + \begin{bmatrix} -0.3 \\ -0.7 \\ 2.8 \\ -2.1 \\ 0.2 \end{bmatrix},$$

donde los valores ajustados de y, $\hat{y}$, son los valores obtenidos con el ajuste cuadrático de los valores de X.
Nótese que los correspondientes valores de los errores, $y_i - \hat{y}_i$, pueden verse en el gráfico como las desviaciones verticales de las observaciones a la curva de ajuste.

Corregido el modelo para la media y ampliado en términos de los vectores lineal y cuadrático, el modelo ajustado es:

$$\begin{array}{ccccccc} y - (y \cdot U_1)U_1 & = & (y \cdot U_2)U_2 & + & (y \cdot U_3)U_3 & + & \text{vector error} \\ y - \bar{y} & = & \hat{\beta}_1 T_2 & + & \hat{\beta}_2 T_3 & + & \text{vector error} \\ \text{vector} & & \text{vector} & & \text{vector} & & \\ y \text{ corregido} & & \text{lineal} & & \text{cuadrático} & & \end{array}$$

$$\text{vector } y \text{ corregido} = y - \bar{y} = \begin{bmatrix} -65 \\ -42 \\ -19 \\ -8 \\ 20 \end{bmatrix} - \begin{bmatrix} -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \end{bmatrix} = \begin{bmatrix} -42.2 \\ -19.2 \\ 3.8 \\ 14.8 \\ 42.8 \end{bmatrix}$$

$$\text{vector lineal} = \hat{\beta}_1 T_2 = 85.5 \begin{bmatrix} -0.37 \\ -0.22 \\ -0.07 \\ 0.08 \\ 0.58 \end{bmatrix} = \begin{bmatrix} -31.6 \\ -18.8 \\ -6.0 \\ 6.8 \\ 49.6 \end{bmatrix}$$

$$\text{vector cuadrático} = \hat{\beta}_2 T_3 = -83.1 \begin{bmatrix} 0.123 \\ -0.003 \\ -0.084 \\ -0.120 \\ 0.084 \end{bmatrix} = \begin{bmatrix} -10.3 \\ 0.3 \\ 7.0 \\ 10.0 \\ -7.0 \end{bmatrix}$$

El modelo cuadrático corregido para la media, teniendo en cuenta que el vector error ya lo tenemos calculado, es finalmente:

$$\begin{bmatrix} -42.2 \\ -19.2 \\ 3.8 \\ 14.8 \\ 42.8 \end{bmatrix} = \begin{bmatrix} -31.6 \\ -18.8 \\ -6.0 \\ 6.8 \\ 49.6 \end{bmatrix} + \begin{bmatrix} -31.6 \\ -18.8 \\ -6.0 \\ 6.8 \\ 49.6 \end{bmatrix} + \begin{bmatrix} -10.3 \\ 0.3 \\ 7.0 \\ 10.0 \\ -7.0 \end{bmatrix} + \begin{bmatrix} -0.3 \\ -0.7 \\ 2.8 \\ -2.1 \\ 0.2 \end{bmatrix},$$

donde el vector lineal es el vector de valores de la componente lineal ajustada, $\hat{\beta}_1 p_1(x)$, y el vector cuadrático es el vector de valores de la componente cuadrática ajustada, $\hat{\beta}_2 p_2(x)$.

Esta descomposición se ve ilustrada en la Figura 3.4 y la correspondiente descomposición de Pitágoras está resumida en la Tabla 3.1, donde SC es la suma de cuadrados correspondiente, MC es la media de la suma de cuadrados, y F es el estadístico calculado a partir del cociente entre la suma de cuadrados y el error cuadrático medio.

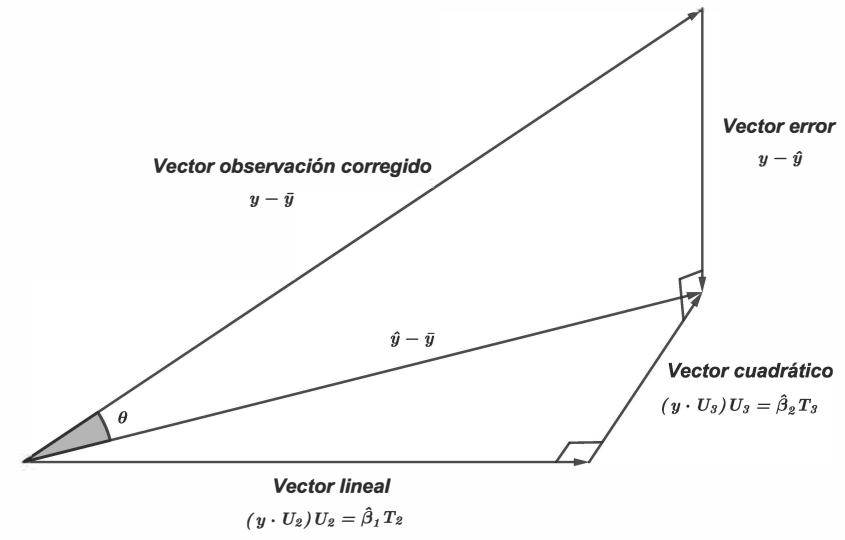

Figura 3.4: Descomposición ortogonal en el modelo cuadrático

Componente	gl	SC	MC	F
Lineal	1	3897.8	3897.8	613.8
Cuadrática	1	304.3	304.3	48.0
Error	2	$3.5 + 9.2 = 12.7$	6.35	
Total	4	4214.8		

Tabla 3.1: ANOVA para el modelo cuadrático

3.4. Comprobación de las hipótesis

Asumiendo que el modelo cuadrático es el más apropiado, hemos asumido que nuestras observaciones son independientes y se distribuyen según una ley normal con varianza constante alrededor de la curva de ajuste.

Para la hipótesis de independencia de las observaciones, analizando el experimento, los cinco rebaños han sido alimentados con cinco tipos diferentes de pastos de forma aleatoria y separada, por tanto, al ser puramente aleatorio la asignación de los pastos, podemos asumir la independencia entre las observaciones.

A su vez, la hipótesis de normalidad se considera razonable sin necesidad de estudiar un histograma puesto que sólo tenemos cinco errores en el estudio. Es decir, debido a que el estudio tiene pocas observaciones, hay pocos errores, por tanto, es considerable que se puedan distribuir según una ley normal.

Para el resto de suposiciones, *la independencia de los errores*, dibujaremos los errores frente a los datos ajustados y observaremos el comportamiento, tal y como se muestra en la Figura 3.5.

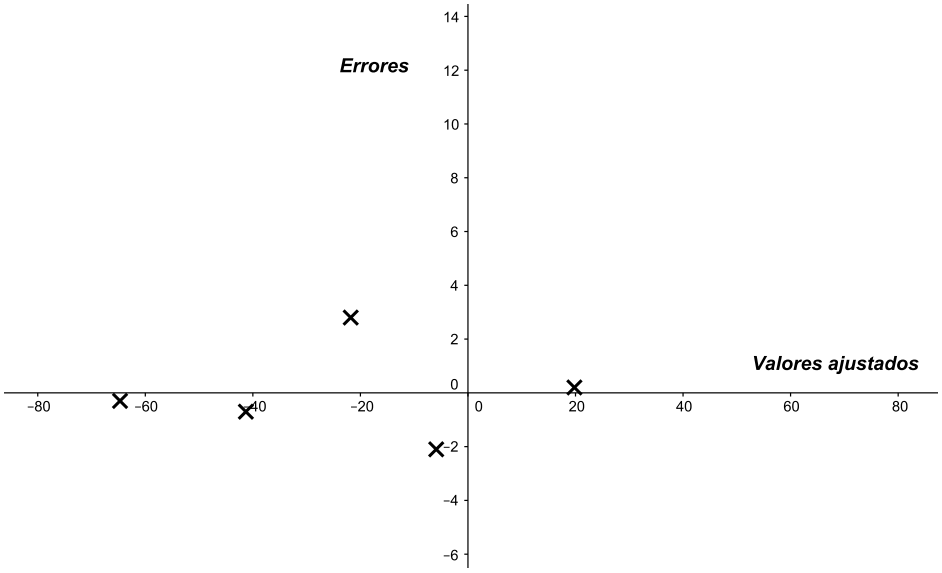

Figura 3.5: Errores frente a los valores ajustados de y

Observamos que no existe tendencia en los puntos, es decir, no se observa ningún tipo de patrón entre los puntos, también se observa que no parecen estar muy dispersos. Con todo esto, podemos aceptar las hipótesis de un modelo cuadrático y, por tanto, aceptar el modelo cuadrático como razonable.

3.5. Intervalos de confianza

Una vez tenemos comprobadas las suposiciones realizadas al principio del estudio y, asumiendo el modelo cuadrático, sabemos que:

1. Existe normalidad con varianza constante e independencia entre las observaciones.

2. Los errores siguen una distribución normal. Al haber sido calculado el espacio de errores como el ortogonal al espacio modelo, seguirán una distribución normal de media 0 y varianza igual a la del espacio modelo.

Recogemos en la Tabla 3.2 las distribuciones necesarias para la obtención de los intervalos de confianza para cada parámetro, asumiendo el modelo cuadrático.

		Media	Varianza
Espacio	$Y \cdot U_1$	$\beta_0 \|T_1\| = \sqrt{5}\beta_0$	σ^2
modelo	$Y \cdot U_2$	$\beta_1 \|T_2\| = \sqrt{0.533}\beta_1$	σ^2
	$Y \cdot U_3$	$\beta_2 \|T_3\| = \sqrt{0.0438}\beta_2$	σ^2
Espacio	$Y \cdot U_4$	0	σ^2
de errores	$Y \cdot U_5$	0	σ^2

Tabla 3.2: Distribuciones de los coeficientes de las proyecciones, $Y \cdot U_i$, asumiendo modelo cuadrático

3.5.1. Intervalos de confianza para los coeficientes polinomiales

Recordamos que la suma cuadrática media MC del espacio de errores es el estimador s^2 de la varianza, la tenemos calculada en la Tabla 3.1, y es igual a 6.35.

En primer lugar, se calculara un intervalo de confianza al 95 % para β_0. Sabemos que $U_1 = \frac{1}{\|T_1\|}T_1$ y que $\hat{\beta}_0 = \frac{(y \cdot U_1)}{\|T_1\|}$. Por tanto, tenemos que: $y \cdot U_1 = \hat{\beta}_0\|T_1\| = \sqrt{5}\hat{\beta}_0$. Luego tenemos que:

$$Y \cdot U_1 = \sqrt{5}\,\hat{\beta}_0 \sim N(\sqrt{5}\beta_0, \sigma^2)\,.$$

Tipificando,

$$\frac{\sqrt{5}\hat{\beta}_0 - \sqrt{5}\beta_0}{\sqrt{\sigma^2}} \sim N(0,1)\,.$$

En nuestro caso, desconocemos la varianza, pero tenemos un estimador de la varianza calculado a partir de la suma cuadrática media del espacio de errores, por tanto, podemos construir así una nueva variable que sigue una distribución t-student con los grados de libertad iguales a la dimensión del espacio de errores, $n - 3$, en este caso 2. Esto es debido a que la suma de cuadrados del espacio de errores seguiría una distribución Chi-cuadrado, ya que los errores siguen una distribución normal y la suma de distribuciones normales al cuadrado da lugar a dicha distribución y, dividiendo una normal estándard $N(0,1)$ entre la raíz de una Chi-cuadrado dividida por sus grados de libertad da lugar a una t-student con esos mismos grados de libertad.

Es decir,

$$SC_e = (Y \cdot U_4)^2 + (Y \cdot U_5)^2 \sim \chi_2$$

$$s^2 = MC_e = \frac{(Y \cdot U_4)^2 + (Y \cdot U_5)^2}{2}$$

$$\frac{\sqrt{5}(\hat{\beta}_0 - \beta_0)}{\sqrt{\frac{(Y \cdot U_4)^2 + (Y \cdot U_5)^2}{2}}} \sim t_2$$

$$\frac{\sqrt{5}(\hat{\beta}_0 - \beta_0)}{\sqrt{s^2}} \sim t_2$$

Ahora tenemos una nueva variable donde sólo se desconoce el parámetro β_0, el cuál es el parámetro a estimar. Por tanto, ya podemos calcular el intervalo de confianza.

Tenemos un nivel de confianza del 95 %, esto quiere decir que $\alpha = 1 - 0.95 = 0.05$. Lo que se pretende calcular es el intervalo $[a, b]$ tal que

$$P\left(a \leq \frac{\sqrt{5}(\hat{\beta}_0 - \beta_0)}{\sqrt{s^2}} \leq b\right) = 1 - \alpha = 0.95\,.$$

Al considerar el intervalo centrado en la media, que es cero, a es equivalente al cuantil de la t-student en $\alpha/2$, es decir, $t_2(0.025)$, y b será en $1-\alpha/2 = 0.975$, $t_2(0.975)$. Entonces, sustituyendo y despejando β_0, tenemos que

$$P\left(t_2(0.025) \leq \frac{\sqrt{5}(\hat{\beta}_0 - \beta_0)}{s} \leq t_2(0.975)\right) = 0.95$$

$$P\left(\frac{t_2(0.025)s}{\sqrt{5}} \leq \hat{\beta}_0 - \beta_0 \leq \frac{t_2(0.975)s}{\sqrt{5}}\right) = 0.95$$

$$P\left(\frac{t_{2,0.025}s}{\sqrt{5}} - \hat{\beta}_0 \leq -\beta_0 \leq \frac{t_{2,0.975}s}{\sqrt{5}} - \hat{\beta}_0\right) = 0.95$$

$$P\left(\hat{\beta}_0 - \frac{t_{2,0.975}s}{\sqrt{5}} \leq \beta_0 \leq \hat{\beta}_0 - \frac{t_{2,0.025}s}{\sqrt{5}}\right) = 0.95$$

Luego,

$$I_{0.95}(\beta_0) = \left[\hat{\beta}_0 - \frac{t_{2,0.975}s}{\sqrt{5}}, \hat{\beta}_0 - \frac{t_{2,0.025}s}{\sqrt{5}}\right]$$

En nuestro ejemplo, $\hat{\beta}_0 = -22.8\,,\ s = \sqrt{6.35}$. Los cuantiles de la distribución t-student están tabulados y, como dicha distribución es simétrica, sabemos que: $t_{2,0.025} = -t_{2,1-0.025} = -t_{2,0.975} = -4.3027$.
Finalmente, los extremos del intervalo de confianza para β_0, será:

$$\hat{\beta}_0 \mp \frac{t_2(0.975)s}{\sqrt{5}} = -22.8 \mp \frac{4.3027\sqrt{6.35}}{\sqrt{5}}$$

Luego,

$$I_{0.95}(\beta_0) = [-27.6489, -17.9511]\,.$$

Por otro lado, para calcular el intervalo de confianza al 95 % para β_1, vamos a realizar el mismo procedimiento que hemos realizado para β_0.

Sabemos que $U_2 = \frac{1}{\|T_2\|}T_2$ y también sabemos que $\hat{\beta}_1 = \frac{(Y \cdot U_2)}{\|T_2\|}$. Por tanto, tenemos que:

$$(Y \cdot U_2) = \hat{\beta}_1\|T_2\| = \hat{\beta}_1\sqrt{0.533} \sim N(\beta_1\sqrt{0.533}, \sigma^2)\,.$$

Tipificando obtenemos:

$$\frac{\sqrt{0.533}(\hat{\beta}_1 - \beta_1)}{\sqrt{\sigma^2}} \sim N(0,1)\,.$$

Dividimos por el estimador de la varianza para obtener así la t-student tal como vimos anteriormente,

$$\frac{\sqrt{0.533}(\hat{\beta}_1 - \beta_1)}{\sqrt{s^2}} \sim t_2\,.$$

El nivel de confianza, $1 - \alpha = 0.95$, es el mismo que el anterior y los grados de libertad de la distribución también coinciden, el mismo espacio de errores, por tanto, los cuantiles correspondientes de la t-Student serán los mismos,

$$P\left(t_{2,0.025} \leq \frac{\sqrt{0.533}(\hat{\beta}_1 - \beta_1)}{\sqrt{s^2}} \leq t_{2,0.975}\right) = 0.95$$

$$P\left(\frac{t_{2,0.025}\sqrt{s^2}}{\sqrt{0.533}} \leq \hat{\beta}_1 - \beta_1 \leq \frac{t_{2,0.975}\sqrt{s^2}}{\sqrt{0.533}}\right) = 0.95$$

$$P\left(\frac{t_{2,0.025}s}{\sqrt{0.533}} - \hat{\beta}_1 \leq -\beta_1 \leq \frac{t_{2,0.975}s}{\sqrt{0.533}} - \hat{\beta}_1\right) = 0.95$$

$$P\left(\hat{\beta}_1 - \frac{t_{2,0.975}s}{\sqrt{0.533}} \leq \beta_1 \leq \hat{\beta}_1 - \frac{t_{2,0.025}s}{\sqrt{0.533}}\right) = 0.95$$

Como la distribución t-Student es simétrica, los extremos del intervalo de confianza al 95 % será:

$$\hat{\beta}_1 \mp \frac{t_{2,0.975}s}{\sqrt{0.533}}$$

En nuestro ejemplo, tenemos que $\hat{\beta}_1 = 85.5$, $s = \sqrt{6.35}$ y sabemos que $t_{2,0.975} = 4.3027$, por tanto, el intervalo de confianza será:

$$I_{0.95}(\beta_1) = [70.6647, 100.3672]$$

Por último, calcularemos el intervalo de confianza para β_2, de la misma forma:

$$(Y \cdot U_3) = \hat{\beta}_2 \|T_3\| \sim N(\sqrt{\beta_2}, \sigma^2)$$

$$\frac{\sqrt{0.0438}(\hat{\beta}_2 - \beta_2)}{\sqrt{\sigma^2}} \sim N(0,1)$$

$$\frac{\sqrt{0.0438}(\hat{\beta}_2 - \beta_2)}{\sqrt{s^2}} \sim t_2$$

Para un nivel de confianza del 95 %, tenemos que el intervalo de confianza vendrá dado por:

$$P\left(t_{2,0.025} \leq \frac{\sqrt{0.0438}(\hat{\beta}_2 - \beta_2)}{\sqrt{s^2}} \leq t_{2,0.975}\right) = 0.95$$

$$P\left(\hat{\beta}_2 - \frac{t_{2,0.975}\sqrt{s^2}}{\sqrt{0.0438}} \leq \beta_2 \leq \hat{\beta}_2 - \frac{t_{2,0.025}\sqrt{s^2}}{\sqrt{0.0438}}\right) = 0.95$$

$$I_{0.95} = \left[\hat{\beta}_2 \mp \frac{t_2(0.975)\sqrt{s^2}}{\sqrt{0.0438}}\right]$$

Y como sabemos que $\sqrt{s^2} = \sqrt{6.35}$, $\hat{\beta}_2 = -83.1$ y el valor del cuantil $T_{2,0.975} = 4.3027$, tenemos finalmente:

$$I_{0.95}(\beta_2) = [-135.0668, -31.5198]$$

Nótese que para cada coeficiente polinomial, el intervalo de confianza tiene la misma forma:

$$\text{estimador} \mp t - \text{valor} \times se(\text{estimador})\,,$$

donde $se(\text{estimador}) = se(\hat{\beta}_i) = s/\|T_{i+1}\|$, es la estimación de la desviación estándar o típica del estimador. En nuestro ejemplo,

$$se(\hat{\beta}_0) = s/\|T_1\|\,,\ se(\hat{\beta}_1) = s/\|T_2\|\ y\ se(\hat{\beta}_2) = s/\|T_3\|\,.$$

También se observa que si se cambia el nivel de confianza $1-\alpha$, lo único que cambiaría sería $t_{2,1-\alpha/2}$ de las expresiones anteriores.

3.5.2. Intervalo de confianza para un valor ajustado $\hat{y}$

Ahora se quiere obtener un intervalo de confianza para la posición cuadrática de digamos $x_0 = 1.6$ Kg, es decir, queremos saber cuánto valdrá el ajuste cuadrático realizado para un nuevo valor de X.

Es decir, se pretende calcular un intervalo de confianza para $\beta_0 + \beta_1 p_1(x_0) + \beta_2 p_2(x_0) = \beta_0 + \beta_1 p_1(1.6) + \beta_2 p_2(1.6)$, donde $p_1(x_0) = p_1(1.6) = 1.6 - 1.42 = 0.18$ y $p_2(x_0) = p_2(1.6) = (1.6-1.42)^2 - 0.251(1.6-1.42) - 0.1064 = -0.1192$. La correspondiente estimación es:

$$\begin{array}{rcccccccc}
\hat{y} & = & \hat{\beta}_0 & + & \hat{\beta}_1 p_1(x_0) & + & \hat{\beta}_2 p_2(x_0) & = & \\
& = & \hat{\beta}_0 & + & 0.18\hat{\beta}_1 & - & 0.1192\hat{\beta}_2 & = & \\
& = & -22.8 & + & 0.18 \cdot 85.5 & - & 0.1192(-83.1) & = & 2.52
\end{array}$$

Si llamamos $se(B_i) = s/\|T_{i+1}\|$ a la correspondiente estimación de la desviación típica de el estimador β_i y, teniendo en cuenta que está normalmente distribuida de forma independiente, tenemos que

$$\begin{aligned}
Var(\hat{Y}) = Var(B_0) + p_1(x)^2 Var(B_1) + p_2(x)^2 Var(B_2) = \\
= se(B_0)^2 + p_1(x)^2 se(B_1)^2 + p_2(x)^2 se(B_2)^2 = \\
= s^2/\|T_1\|^2 + p_1(x)^2 s^2/\|T_2\|^2 + p_2(x)^2 s^2/\|T_3\|^2 = \\
= s^2(1/5 + 0.18^2/0.533 + (-0.1192)^2/0.0438) = 3.7
\end{aligned}$$

Y por tanto la estimación de la desviación típica para $\hat{y}$ es $se(\hat{Y}) = \sqrt{Var(\hat{Y})} = 1.93$.

En el apartado anterior vimos que los intervalos de confianza se construyen de la forma:

$$\text{estimador} \mp t\text{-valor} \times se(\text{estimador})$$

Luego, para el valor ajustado de y en $x_0 = 1.6$, los extremos del intervalo de confianza al 95 % será

$$\hat{y} \mp 4.3027 \times se(\hat{y})$$
$$2.52 \mp 4.3027 \times 1.93 \text{ y}$$
$$I_{0.95}(y_{x_0}) = [-5.77, 10.81]\,.$$

Nótese que en general la expresión del intervalo de confianza para un valor ajustado de Y es

$$\text{valor ajustado } y \text{ en } x_0 \mp t\text{-value} \times \sqrt{s^2\left(\frac{1}{\|T_1\|^2} + \frac{p_1(x_0)^2}{\|T_2\|^2} + \frac{p_2(x_0)^2}{\|T_3\|^2}\right)}\,.$$

3.5.3. Intervalo de confianza para una predicción y_{pred}

Para obtener un valor "futuro", digamos que queremos obtener el peso para un sexto rebaño de ovejas alimentado con una ración $x_0 = 1.6$ Kg/oveja/día, simplemente añadimos el término "1", de modo que tenemos "$1 + \frac{1}{\|T_1\|^2} + \ldots$" en la expresión calculada en el apartado anterior. Es decir,

$$y_{pred} = \hat{\beta}_0 + \hat{\beta}_1 p_1(x_0) + \hat{\beta}_2 p_2(x_0) = 2.52$$

$$Var(Y_pred) = Var(B_0) + p_1(x)^2 Var(B_1) + p_2(x)^2 Var(B_2) + \sigma^2$$

La expresión de la varianza de la predicción contiene un término de variación de la predicción con respecto al valor real σ^2. Esto nos lleva a la siguiente expresión

$$se(Y_{pred})^2 = se(B_0)^2 + p_1(x)^2 se(B_1)^2 + p_2(x)^2 se(B_2)^2 + s^2$$
$$se(Y_{pred}) = \sqrt{\frac{s^2}{\|T_1\|^2} + p_1(x_0)^2\frac{s^2}{\|T_2\|^2} + p_2(x_0)^2\frac{s^2}{\|T_3\|^2} + s^2} =$$
$$= \sqrt{s^2\left(1 + \frac{1}{\|T_1\|^2} + \frac{p_1(x_0)^2}{\|T_2\|^2} + \frac{p_2(x_0)^2}{\|T_3\|^2}\right)} =$$

$$= \sqrt{s^2(1 + 1/5 + 0.18^2/0.533 + (-0.1192)^2/0.0438)} = 3.17$$

Finalmente, el intervalo de confianza al 95 % para una predicción de la variable Y sobre un valor de X a partir del modelo de orden 2 ajustado será:

$$I_{0.95}(y_{pred}) = [2.52 \mp t_{2,0.975} \times 3.17)] = [-11.1, 16.2]$$

3.6. Coeficiente de correlación múltiple R

En nuestro ejemplo, tenemos una regresión cuadrática entonces el coeficiente de correlación múltiple, R, será el coseno del ángulo, θ, formado por el vector de observaciones corregido $y - \bar{y}$ y el vector de valores ajustados corregido

$$\hat{y} - \bar{y} = (y \cdot U_2)U_2 + (y \cdot U_3)U_3 = \hat{\beta}_1 T_2 + \hat{\beta}_2 T_3$$

El ángulo correspondiente está ilustrado en la Figura 3.4. Por definición, θ estará contenido en el intervalo de 0° a 90°, y R tendrá un valor contenido de 0 a 1.

En nuestro ejemplo, los correspondientes vectores que forman el ángulo bajo estudio para el cálculo del coeficiente de correlación R son:

$$y - \bar{y} = \begin{bmatrix} -42.2 \\ -19.2 \\ 3.8 \\ 14.8 \\ 42.8 \end{bmatrix}$$

$$\hat{y} - \bar{y} = \begin{bmatrix} -64.7 \\ -41.3 \\ -21.8 \\ -5.9 \\ 19.8 \end{bmatrix} - \begin{bmatrix} -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \\ -22.8 \end{bmatrix} = \begin{bmatrix} -41.9 \\ -18.5 \\ 1.0 \\ 16.9 \\ 2.6 \end{bmatrix}.$$

Luego, sabiendo que el coseno del ángulo que forman dos vectores es igual a su producto escalar dividido por la multiplicación de sus módulos,

$$R = \cos\theta = \frac{(y-\bar{y})\cdot(\hat{y}-\bar{y})}{\|y-\bar{y}\|\|\hat{y}-\bar{y}\|} = \frac{4202.10}{64.92\times 64.82} = 0.9985\ ,$$

se observa que el valor de R para nuestro modelo cuadrático es muy cercano a 1, tenemos un coeficiente de correlación múltiple muy alto lo que implica que existe relación cuadrática entre las variables y el ajuste que hemos realizado es casi perfecto.

Se puede observar que el valor del ángulo es $\theta = \cos^{-1}(0.9985) = 3°$, lo que implica que entre esos dos vectores hay un ángulo muy pequeño, y por tanto, los dos vectores son muy próximos.

3.7. Coeficiente de determinación R^2

El *coeficiente de determinación*, R^2, es el cuadrado del coeficiente de correlación múltiple. En nuestro ejemplo: $R^2 = 0.9985^2 = 0.9970$
Como observamos en la Figura 3.6 podemos observar que se puede reescribir la descomposición de pitágoras asociada como:

$$\|y-\bar{y}\|^2 = \|\hat{y}-\bar{y}\|^2 + \|y-\hat{y}\|^2\,. \tag{3.1}$$

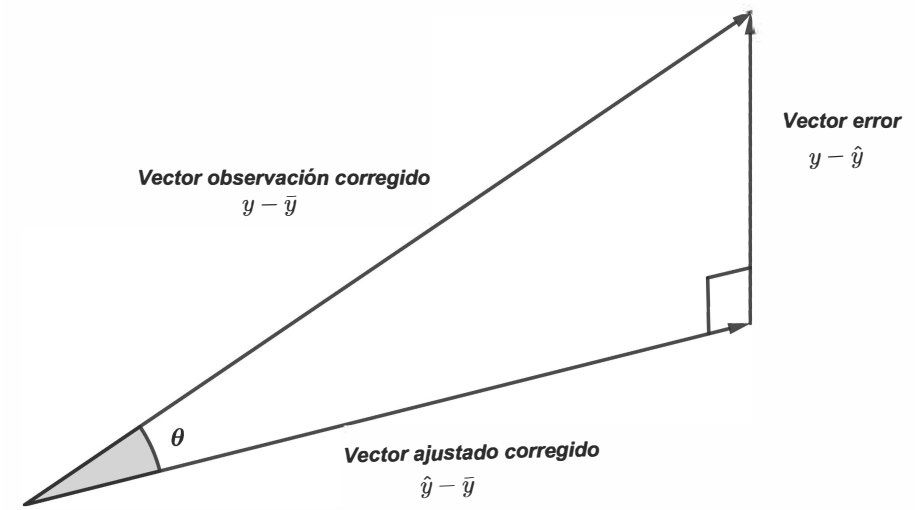

Figura 3.6: Descomposición de orden 2, junto al ángulo bajo estudio θ

A partir de la Figura 3.6, podemos reescribir la descomposición mostrada en (3.1) de la forma

$$\|y - \bar{y}\|^2 = R^2\|y - \bar{y}\|^2 + (1 - R^2)\|y - \bar{y}\|^2 \,,$$

se mostrará a continuación como se ha construido la relación anterior.

Como sabemos que el $R = \cos\theta$ y por definición el $\cos\theta = \dfrac{\|\hat{y} - \bar{y}\|}{\|y - \bar{y}\|} = R$, por tanto:

$$\|\hat{y} - \bar{y}\| = R\|y - \bar{y}\| \,,$$

$$\|\hat{y} - \bar{y}\|^2 = R^2\|y - \bar{y}\|^2 \tag{3.2}$$

Por otro lado, si calculamos el seno de θ tenemos que

$$\operatorname{sen}\theta = \frac{\|y - \hat{y}\|}{\|y - \bar{y}\|} \,,$$

$$\|y - \hat{y}\| = \operatorname{sen}\theta\|y - \bar{y}\| \,,$$

haciendo uso de la fórmula fundamental de la trigonometría, $\operatorname{sen}^2\theta + \cos^2\theta = 1$, siendo en nuestro caso $\operatorname{sen}\theta = \sqrt{1 - cos^2\theta}$, por tanto

$$\|y - \hat{y}\| = \sqrt{(1 - R^2)}\|y - \bar{y}\| \,,$$

$$\|y - \hat{y}\|^2 = (1 - R^2)^2\|y - \bar{y}\|^2 \,. \tag{3.3}$$

Quedando así reflejada la descomposición de pitágoras descrita en (3.1).

3.7.1. Relación entre el estadístico F y el coeficiente de determinación

Los puntos críticos, o cuantiles, de la distribución R^2 bajo la hipótesis nula compuesta $H_0 : \beta_0 = \beta_1 = 0$ están tabulados en ciertos libros de texto. El equivalente estadístico F para esta hipótesis es

$$F = \frac{\text{Suma media de cuadrados del modelo}}{\text{Error cuadrático medio}} = \frac{[(y \cdot U_2) + (y \cdot U_3)]/2}{s^2} =$$

$$= \frac{\|\hat{y} - \bar{y}\|^2/2}{\|y - \hat{y}\|^2/(n-3)} = \frac{R^2\|y - \bar{y}\|^2/2}{(1-R^2)^2\|y - \bar{y}\|^2/(n-3)} ,$$

gracias a (3.2) y en (3.3). Siendo finalmente,

$$F = \frac{(n-3)R^2}{2(1-R^2)} .$$

Donde, el estadístico F seguiría una distribución F-snedecor con 2 y $(n-3)$ grados de libertad.

En la práctica, la prueba F para el contraste compuesto $H_0 = \beta_0 = \beta_1 = 0$ y el equivalente R^2 test, son menos útiles que las pruebas F individuales para las hipótesis simples $H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$.

3.7.2. El coeficiente de determinación para la bondad del ajuste

El coeficiente de determiación R^2, es a veces usado como un indicador de la bondad de ajuste del modelo a través del cálculo de la variabilidad total explicada por el modelo supuesto. Sin embargo, tiene el inconveniente de que siempre aumenta cuando se añade otro término al modelo. En nuestro ejemplo, el valor de R^2 aumenta de 0.9428 a 1.0000 a medida que se incrementa el orden polinomial desde 1 a 4, tal y como se muestra en la tabla 3.3.
Una alternativa preferible al coeficiente R^2, la proporción de la suma de cuadrados obtenida por el modelo de regresión, es el "porcentaje de variabilidad explicado por el modelo de regresión", representado en la Tabla 3.3 como "% variabilidad". Se calcula como la diferencia entre el total de las varianzas estimadas (MC total), y las varianzas estimadas del modelo (el error cuadrático medio), y se expresa como porcentaje, considerando el total de varianza que puede explicar el modelo como el $MCtotal$.

Orden polinomial	SC del modelo de regresión	R^2	s^2	% variabilidad
1	3897.8	0.9248	105.7	90.0 %
2	4202.1	0.9970	6.36	99.4 %
3	4205.6	0.9978	9.2	99.1 %
4	4214.8	1.0000	-	
Total	4214.8		1053.7	$= MC$ total

Tabla 3.3: Estudio de la bondad de los ajustes

Para aclarar el cálculo de la columna de variabilidad explicada, vamos a realizar el cálculo del primer número que aparece en esa columna. En primer lugar consideramos la diferencia entre el MC total y el correspondiente estimador de la varianza $s^2 = 105.7$, obteniendo 948.039. Ahora pasamos este valor a porcentaje mediante una regla de tres, donde MC total $= 1053.7$ es el 100 % y tenemos que calcular qué porcentaje equivale a la diferencia obtenida:

$$\frac{948.039 \times 100}{1053.7} = 89.97238 \simeq 90.0\,\%\,.$$

En nuestro ejemplo, el porcentaje de variabilidad aumenta del 90.0 % al 99.4 % al añadir el término cuadrático, y baja al 99.2 % al añadir el siguiente término. Ambas medidas llegan al 100 % cuando quedan cero grados de libertad para el error.

En general, si se añaden efectos/términos puramente aleatorios al modelo, el porcentaje de varianza explicado por la regresión rondará un valor fijo.

Capítulo 4

Ejemplo Error Puro

El ejemplo que se presenta corresponde a un estudio basado en un experimento realizado para determinar las curvas de crecimiento de la maleza milenrama, bajo cuatro niveles de sombra, concretamente el 100 %, 46.8 %, 23.7 % y 6.4 % de luz diurna. La idea consiste en que si la milenrama resulta ser sensible a la sombra, habría alguna esperanza de suprimir la especie sembrando un cultivo agresivo, como la cebada, en los campos infectados por milenrama.

El experimento original consideraba 24 parcelas con 6 réplicas de los 4 tratamientos principales. Cada parcela constaba de 12 plántulas, dos de las cuales se cosecharon en cada una de las seis fechas de cosecha.

Para realizar el desarrollo del ejemplo se va a considerar únicamente una parcela con una caseta que proporciona un 46.8 % de luz natural que contiene 12 plántulas de milenrama y, a su vez, consideramos que se cosecharon diferentes números de plántulas en las 6 fechas de cosecha.

El *objetivo* de nuestro análisis es ajustar un modelo polinomial que se adecúe a la curva de crecimiento de la planta bajo un 46.8 % de luz natural. Para ello, cada día de cosecha se llevaron al laboratorio plántulas de milenrama seleccionadas al azar, se lavaron, se secaron y se pesaron, obteniendo así los datos que forman la Tabla 4.1

En la misma tabla también aparece el valor del logaritmo natural de los pesos, esto es porque se espera que se incremente el error estándar en proporción a la media y para evitarlo, se debe aplicar una transformación logarítmica. De este modo, tomamos como vector observación a los valores de la correspondiente transformación logarítmica, tomando una precisión de 3 decimales, $y = [0.231, 0.344, \ldots, 2.137]^t$, y el correspondiente vector de valores de X será $x = [7, 7, 7, 13, \ldots, 29, 29, 29]^t$.

	Días desde que empezó el experimento					
	7	13	17	21	25	29
Peso	1.26	2.42	3.34	5.75	6.75	7.93
seco	1.41		3.67	4.97		9.82
	1.07					8.47
ln(Peso)	0.231	0.884	1.206	1.749	1.910	2.071
	0.344		1.300	1.603		2.284
	0.068					2.137

Tabla 4.1: Total de peso seco, y su logaritmo neperiano correspondiente, en gramos para las plántulas en los correspondientes días de cosecha.

El correspondiente diagrama de dispersión de los datos se ve reflejado en la Figura 4.1.

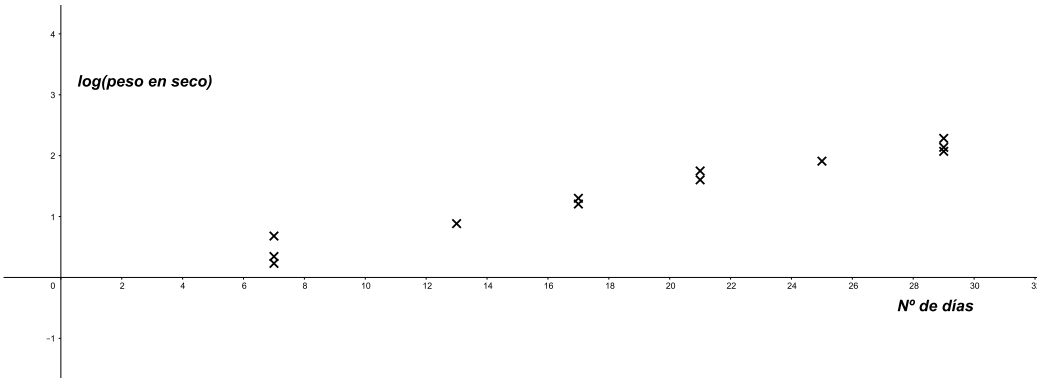

Figura 4.1: Diagrama de dispersión asociado al estudio.

4.1. Modelo

Estamos ante un experimento en el que tenemos varios valores de Y para un mismo valor de X, por tanto, podremos ajustar modelos polinomiales de orden menor que el número de valores distintos que tenemos para X menos 1. En este caso tenemos 6 valores de X, lo que implica que podremos ajustar modelos polinomiales de hasta grado 5.
Por tanto, la forma no ortogonal del modelo tendría la expresión:

$$y = \alpha_0 + \alpha_1(x - \bar{x}) + \alpha_2(x - \bar{x})^2 + \alpha_3(x - \bar{x})^3 + \alpha_4(x - \bar{x})^4 + \alpha_5(x - \bar{x})^5$$

La forma ortogonal del modelo será,

$$y = \beta_0 + \beta_1 p_1(x) + \beta_2 p_2(x) + \beta_3 p_3(x) + \beta_4 p_4(x) + \beta_5 p_5(x)$$

donde aún no tenemos calculadas las componentes polinomiales $p_i(x)$.

Asumiremos, al igual que en el ejemplo anterior, que existe un grado polinomial que aporta un ajuste perfecto a los datos y para ese grado polinomial realizaremos las siguientes suposiciones:

- La variable Y se distribuye según una ley normal con varianza constante, σ^2, alrededor de la media real de la curva a la que se ajusta.
- Los errores de muestreo obtenidos en el ajuste son independientes.

En primer lugar, escribiremos de forma vectorial los vectores y, x y los correspondientes valores de los vectores $X_1, \ldots, X_6$

$$y = \begin{bmatrix} 0.231 \\ 0.344 \\ 0.068 \\ 0.884 \\ 1.206 \\ 1.300 \\ 1.749 \\ 1.603 \\ 1.910 \\ 2.071 \\ 2.284 \\ 2.137 \end{bmatrix}, \; x = \begin{bmatrix} 7 \\ 7 \\ 7 \\ 13 \\ 17 \\ 17 \\ 21 \\ 21 \\ 25 \\ 29 \\ 29 \\ 29 \end{bmatrix},$$

del vector x, calculamos $\bar{x} = 18.5$ días, y calculamos los siguientes vectores X_i,

$$X_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, X_2 = (x - \bar{x}) = (x - 18.5) = \begin{bmatrix} -11.5 \\ -11.5 \\ -11.5 \\ -5.5 \\ -1.5 \\ -1.5 \\ 2.5 \\ 2.5 \\ 6.5 \\ 10.5 \\ 10.5 \\ 10.5 \end{bmatrix},$$

$$X_3 = (x - 18.5)^2 = \begin{bmatrix} 132.25 \\ 132.25 \\ 132.25 \\ 30.25 \\ \vdots \\ 110.25 \\ 110.25 \end{bmatrix}, X_4 = (x - 18.5)^3 = \begin{bmatrix} -1520.9 \\ -1520.9 \\ -1520.9 \\ -166.4 \\ \vdots \\ 1157.6 \\ 1157.6 \end{bmatrix},$$

$$X_5 = (x-18.5)^4 = \begin{bmatrix} 17490.1 \\ 17490.1 \\ 17490.1 \\ 915.2 \\ \vdots \\ 12155.1 \\ 12155.1 \end{bmatrix}, X_6 = (x-18.5)^5 = \begin{bmatrix} -201135.7 \\ -201135.7 \\ -201135.7 \\ -5032.8 \\ \vdots \\ 127628.2 \\ 127628.2 \end{bmatrix}.$$

Luego, el modelo polinomial no ortogonal de grado máximo en este caso sería

$$y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 \,.$$

Para ortogonalizar el modelo aplicaremos Gram-Schmidt, sin embargo, a diferencia del capítulo anterior, ahora estamos trabajando en dimensión 12, porque tenemos 12 observaciones, y aplicando Gram-Schmidt al conjunto de vectores X_i tendríamos dimensión 6. Por tanto, hay que ampliar la base con 6 nuevos vectores que generarán el espacio de errores, de modo que el espacio modelo quedará generado por los vectores ortogonales obtenidos a partir de X_i, y el espacio de errores, por estos 6 nuevos vectores.

4.1.1. Ortogonalización del modelo

Sabemos que X_1 y X_2 son ortogonales, por tanto los primeros dos vectores serán:

$$T_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, T_2 = \begin{bmatrix} -11.5 \\ -11.5 \\ -11.5 \\ -5.5 \\ \vdots \\ 10.5 \\ 10.5 \end{bmatrix},$$

y sus correspondientes vectores unitarios:

$$U_1 = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}, U_2 = \frac{1}{\sqrt{817}} \begin{bmatrix} -11.5 \\ -11.5 \\ -11.5 \\ -5.5 \\ \vdots \\ 10.5 \\ 10.5 \end{bmatrix}.$$

Ahora vamos a ortogonalizar el resto de los vectores X_i según el método de *Gram-Schmidt*, que se basa en hacer cada vector ortogonal a sus predecesores restándole a cada vector su proyección sobre el espacio generado por los anteriores, así:

$$T_i = X_i - \sum_{j=1}^{i-1} P_{U_j} X_i \,.$$

Luego,

$$T_3 = X_3 - P_{U_1} X_3 - P_{U_2} X_3 = X_3 - (X_3 \cdot U_1)U_1 - (X_3 \cdot U_2)U_2 \,.$$

$$P_{U_1} X_3 = \frac{235.8}{\sqrt{12}} T_1 = \begin{bmatrix} 68.08 \\ 68.08 \\ 68.08 \\ 68.08 \\ \vdots \\ 68.08 \\ 68.08 \end{bmatrix}, P_{U_2} X_3 = \frac{-33.48}{\sqrt{817}} T_2 = \begin{bmatrix} 13.47 \\ 13.47 \\ 13.47 \\ 6.44 \\ \vdots \\ -12.30 \\ -12.30 \end{bmatrix}.$$

Por tanto T_3 será, finalmente:

$$T_3 = \begin{bmatrix} 132.25 \\ 132.25 \\ 132.25 \\ 30.25 \\ \vdots \\ 110.25 \\ 110.25 \end{bmatrix} - \begin{bmatrix} 68.08 \\ 68.08 \\ 68.08 \\ 68.08 \\ \vdots \\ 68.08 \\ 68.08 \end{bmatrix} - \begin{bmatrix} 13.47 \\ 13.47 \\ 13.47 \\ 6.44 \\ \vdots \\ -12.30 \\ -12.30 \end{bmatrix} = \begin{bmatrix} 50.70 \\ 50.70 \\ 50.70 \\ -44.28 \\ \vdots \\ 54.47 \\ 54.47 \end{bmatrix},$$

y su vector unitario asociado,

$$U_3 = \frac{T_3}{\|T_3\|} = \frac{1}{\sqrt{34978.7}} \begin{bmatrix} 50.70 \\ 50.70 \\ 50.70 \\ -44.28 \\ \vdots \\ 54.47 \\ 54.47 \end{bmatrix}.$$

De forma análoga, construimos

$$T_4 = X_4 - P_{U_1}X_4 - P_{U_2}X_4 - P_{U_3}X_4\,,$$
$$T_5 = X_5 - P_{U_1}X_5 - P_{U_2}X_5 - P_{U_3}X_5 - P_{U_4}X_5\,,$$
$$T_6 = X_6 - P_{U_1}X_6 - P_{U_2}X_6 - P_{U_3}X_6 - P_{U_4}X_6 - P_{U_5}X_6\,.$$

De forma que se obtiene

$$T_4 = X_4 - \frac{(-276.3)}{\sqrt{12}}\,T_1 - \frac{3209.01}{\sqrt{817}}\,T_2 - \frac{(-220.2)}{\sqrt{34978.7}}\,T_3 = \begin{bmatrix} -90.35 \\ -90.35 \\ -90.35 \\ 478.73 \\ \vdots \\ 122.67 \\ 122.67 \end{bmatrix},$$

y su respectivo vector unitario será: $U_4 = \frac{1}{\|T_4\|}T_4 = \frac{1}{\sqrt{640558.2}}T_4$.
Para $i = 5$

$$T_5 = X_5 - \frac{26478.4}{\sqrt{12}}\,T_1 - \frac{(-7479.0)}{\sqrt{817}}\,T_2 - \frac{24423.5}{\sqrt{34978.7}}\,T_3 -$$

$$-\frac{(-695.2)}{\sqrt{640558.2}}\,T_4 = \begin{bmatrix} 138.5 \\ 138.5 \\ 138.5 \\ -1969.9 \\ \vdots \\ 252.7 \\ 252.7 \end{bmatrix},$$

el vector unitario asociado será $U_5 = \frac{1}{\|T_5\|}\,T_5 = \frac{1}{\sqrt{10943605}}\,T_5$.

Con T_6 tendremos ya formado el espacio modelo,

$$T_6 = X_6 - \frac{-61710.8}{\sqrt{12}}\,T_1 - \frac{387048.2}{\sqrt{817}}\,T_2 - \frac{-52052.9}{\sqrt{34978.7}}\,T_3 - \frac{117975.5}{\sqrt{640558.2}}\,T_4 -$$

$$-\frac{-449.2}{\sqrt{10943605}}\,T_5 = \begin{bmatrix} -151.4 \\ -151.4 \\ -151.4 \\ 4099.3 \\ \vdots \\ 372.7 \\ 372.7 \end{bmatrix},$$

por tanto, el vector unitario asociado será $U_6 = \frac{1}{\|T_6\|}T_6 = \frac{1}{\sqrt{151116047}}\,T_6$.

Tenemos calculado ya el espacio modelo, $M = \langle U_1, \ldots, U_6\rangle$, vamos ahora a ampliar la base hasta dimensión 12 con los vectores $U_7, U_8, \ldots, U_{12}$ unitarios y ortogonales entre ellos y al espacio modelo.

Para ello, nos fijamos en los vectores unitarios que tenemos en el espacio modelo, las tres primeras componentes coinciden, puesto que equivalen al primer valor de X, que tiene asignado 3 valores distintos de Y, por tanto construiremos un vector de forma que en su producto escalar solo participen estas componentes y se anulen, completando el resto de las componentes del vector con 0. Seguiremos ese mismo procedimiento para los valores siguientes hasta obtener la dimensión deseada.

Por tanto, la manera más sencilla de construir los vectores unitarios del espacio de errores será:

$$U_7 = \frac{1}{\sqrt{2}}\,[-1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^t$$

$$U_8 = \frac{1}{\sqrt{6}}\,[-1, -1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0]^t$$

$$U_9 = \frac{1}{\sqrt{2}}\,[0, 0, 0, 0, -1, 1, 0, 0, 0, 0, 0, 0]^t$$

$$U_{10} = \frac{1}{\sqrt{2}}\,[0, 0, 0, 0, 0, 0, -1, 1, 0, 0, 0, 0]^t$$

$$U_{11} = \frac{1}{\sqrt{2}} [0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 1, 0]^t$$

$$U_{12} = \frac{1}{\sqrt{6}} [0, 0, 0, 0, 0, 0, 0, 0, 0, -1, -1, 2]^t$$

Finalmente, ya tenemos la base ortonormal con la que vamos a trabajar $\{U_1, U_2, \ldots, U_6, U_7, \ldots, U_{12}\}$.

Nótese que en el modelo ortogonal, T_1 es el vector de valores de la componente constante o de orden 0 del modelo, $p_0(x) = 1$, mientras que T_2 es el vector de valores de la componente lineal o de orden 1, $p_1(x) = (x - \bar{x}) = T_2$. Luego, T_3, T_4, T_5 y T_6 serán los vectores de valores de las componentes de orden $2, 3, 4$ y 5, respectivamente.

4.2. Ajustar el modelo adecuado

Para el modelo de grado máximo, grado 6 en este caso, tenemos el espacio completo con la unión del espacio modelo M generado por los vectores $U_1, \ldots, U_6$ y el espacio de errores E generado por los vectores $U_7, \ldots, U_{12}$. Para ajustar el modelo, se realiza la proyección del vector observación y sobre el espacio modelo y el espacio de errores:

$$y = P_{U_1}y + P_{U_2}y + \cdots + P_{U_6}y + P_{U_7}y + \cdots + P_{U_{12}}y =$$

$$= (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + \cdots + (y \cdot U_6)U_6 + (y \cdot U_7)U_7 + \cdots + (y \cdot U_{12})U_{12} =$$

$$= 4.56U_1 + 2.55U_2 - 0.29U_3 - 0.04U_4 + 0.009U_5 - 0.038U_6 +$$

$$+0.08U_7 + 0.009U_8 + 0.066U_9 - 0.10U_{10} + 0.15U_{11} + 1.66U_{12} \ .$$

Sabiendo que la proyección del vector y sobre la dirección U_1 es la media de las observaciones, la descomposición ortogonal de Pitágoras para el modelo completo será:

$$\|y - (y \cdot U_1)^2\|^2 = \|P_{U_2}y\|^2 + \|P_{U_3}y\|^2 + \cdots + \|P_{U_6}y\|^2 + \|e\|^2 \, .$$

Sabemos que $\|P_{U_i}y\|^2 = (y \cdot U_i)^2$, para $i = 1, \ldots, 6$, y que $\|e\|^2 = (y \cdot U_7)^2 + \cdots + (y \cdot U_{12})^2$, por tanto, tenemos que:

$$\|y - (y \cdot U_1)^2\|^2 = (y \cdot U_2)^2 + \cdots + (y \cdot U_6)^2 + \cdots + (y \cdot U_{12})^2 \, .$$

La Tabla 4.2 muestra la tabla ANOVA para el modelo completo.

Componente bajo estudio	grados de libertad (gl)	SC $= (y \cdot U_i)^2$	MC
Orden 1	1	6.5129	6.5129
Orden 2	1	0.0852	0.0852
Orden 3	1	0.0020	0.0020
Orden 4	1	0.0001	0.0001
Orden 5	1	0.0047	0.0047
Error Puro	6	0.0774	$0.0129 = s^2$
Total	11	6.6823	

Tabla 4.2: Tabla ANOVA para el modelo completo con el correspondiente término de error puro.

Es más, la estimación de la varianza también habría podido ser calculada ajustando seis medias del tratamiento, correspondientes a cada una de las fechas de cosechas, x, y usando la descomposición de pitágoras:

$$\|y - \bar{y_{..}}\|^2 \quad = \quad \|\bar{y_{i.}} - \bar{y_{..}}\|^2 \quad + \quad \|y - \bar{y_{i.}}\|^2$$

$$6.6823 \quad = \quad 6.6049 \quad + \quad 0.0774$$

donde $\bar{y_{..}}$ se refiere a la media muestral del vector observación y, e $\bar{y_{i.}}$ a la media muestral de las observaciones correspondientes a cada valor distinto de x, recogidas en la Tabla 4.3. Los cálculos realizados en la descomposición se muestran a continuación.

$$\|y - \bar{y_{..}}\|^2 = \|y - 1.32\|^2 = 6.6823\,,$$

$$\|\bar{y_{i.}} - \bar{y_{..}}\|^2 = \left\| \begin{bmatrix} 0.214 - 1.32 \\ 0.214 - 1.32 \\ 0.214 - 1.32 \\ 0.884 - 1.32 \\ \vdots \\ 2.164 - 1.32 \end{bmatrix} \right\|^2 = 6.6049 \text{ , y}$$

$$\|y - \bar{y_{i.}}\|^2 = 6.6683 - 6.6049 = 0.0774\,.$$

	Días desde que empezó el experimento					
	7	13	17	21	25	29
ln(Peso)	0.231	0.884	1.206	1.749	1.910	2.071
	0.344		1.300	1.603		2.284
	0.068					2.137
$\bar{y_{i.}}$	0.214	0.884	1.253	1.676	1.910	2.164

Tabla 4.3: Observaciones por días de tratamiento, y sus correspondientes medias.

4.2.1. Procedimiento de elección del modelo adecuado

El procedimiento es sencillo, haremos uso de la suma de cuadrados para cada componente, calculada en la Tabla 4.2, y construiremos el estadístico F para cada modelo utilizando la estimación de la varianza, $s^2 = MC_E = 0.0129$. Además, sabemos que el espacio de errores tiene dimensión 6, por tanto, conocemos la distribución del estadístico a calcular $F \sim F_{1,6}$.
Además de hacer uso de dicho estadístico, calcularemos la suma cuadrática media del "déficit de ajuste" para cada orden polinomial. De forma que para un orden polinomial concreto, la suma media de cuadrados MC correspondiente al déficit de ajuste será el promedio de la suma de cuadrados de las componentes de orden mayor. Todo esto se añadirá a la Tabla 4.2, creando una nueva tabla resumen de los cálculos en la Tabla 4.4.

Una vez tenemos todo lo anterior calculado, el procedimiento consiste en empezar estudiando desde la componente de orden polinomial 0 e ir aumentando el orden del polinomio hasta que tanto el estadístico $F_{1,6}$ como el nuevo

estadístico F referido a la parte del déficit de ajuste sean no significativos, cuándo se comparan con el término de error puro.

Comp. polinomial	gl	SC	$F_{1,6}$	Comp. polinomial	Déficit de ajuste gl	MC	F'
Orden 1	1	6.5129	505.1 (**)	Orden 0	5	1.320	102.5 (**)
Orden 2	1	0.0852	6.61 (*)	Orden 1	4	0.023	1.78 (ns)
Orden 3	1	0.0020	0.16 (ns)	Orden 2	3	0.0023	0.18 (ns)
Orden 4	1	0.0001	0.01 (ns)	Orden 3	2	0.0024	0.19 (ns)
Orden 5	1	0.0047	0.37 (ns)	Orden 4	1	0.0047	0.37 (ns)
Error	6	0.0774			6	0.0129	
Total	11	6.6823					

Tabla 4.4: Descomposición de Pitágoras para la tabla ANOVA y la suma de cuadrados media del déficit de ajuste para cada componente polinomial (para la que se pueda calcular la suma media de cuadrados tal y como la definimos).

En primer lugar, estudiaremos qué componentes son significativas a partir del primer estadístico F, calculado para cada una de ellas. Retomamos el procedimiento utilizado en el capítulo anterior, en este caso podemos mirar directamente la Tabla 4.4 y saber cuáles son significativos y cuáles no, ya que todos siguen una distribución F-Snedecor con 1 y 6 grados de libertad. Si queremos un nivel de significación de 0.5, es decir, un nivel de confianza del 95 %, debemos comparar cada valor del estadístico $F_{1,6}$ de cada componente con el cuantil 0.95 de dicha distribución que tiene un valor igual a $F_{1,6,0.95} = 5.987$.

Recordemos que lo que estudia este test es la hipótesis de que el correspondiente coeficiente sea igual a 0, lo que indicaría que no existe relación del tipo considerado entre las variables y, por tanto, la componente sería no significativa, frente a que el coeficiente fuera distinto a 0, lo que indicaría que sería significativa para el modelo. La intención es simplificar el modelo de componentes no significativas hasta obtener el modelo más sencillo y adecuado para las observaciones.

Como sabemos que el valor del cuantil de la distribución F-Snedecor es igual a 5.987, podemos observar que el primer estadístico F_1 es lo suficientemente

grande en comparación con el cuantil (**), lo que nos indica que no hay evidencias para negar la existencia de la componente lineal en el modelo. Asumimos así, que la componente lineal es significativa para el modelo.

$$F_1 = 505.1 > 5.987 = F_{1,6,0.95} \ .$$

Comprobando ahora la componente cuadrática, o de orden 2, obtenemos un estadístico F_2 mayor que el cuantil considerado (*), por tanto, el correspondiente coeficiente es no nulo, y por ello, la componente cuadrática es significativa para el modelo.

$$F_2 = 6.61 > 5.987 = F_{1,6,0.95} \ .$$

Se observa que las siguientes componentes tienen un valor del estadístico F menor que el valor del cuantil, por tanto, podemos asumir que el resto de componentes son no significativas (ns) con un nivel de confianza del 95 %.

$$F_3 = 0.16 \, , F_4 = 0.01 \, , F_5 = 0.37 \ < 5.987 = F_{1,6,0.95} \ .$$

Para la segunda parte del análisis, tenemos distribuciones de F distintas, por ello debemos estudiar cada estadístico por separado.

Para el modelo de orden 0, o modelo constante, obtenemos un estimador F_0' que seguirá una distribución F-Snedecor con 5 y 6 grados de libertad, puesto que dividimos la suma media cuadrática obtenida a partir de las 5 siguientes componentes de orden superior por el la suma media de cuadrados de los errores que tiene dimensión 6. Compararemos este estadístico con el cuantil 0.95, para mantener el mismo nivel de significación que en la primera parte de la distribución $F_{5,6}$.

$$F_0' = 102.5 > 4.387 = F_{5,6,0.95}$$

Se observa que el estadístico F_0' es lo suficientemente grande, (**), en comparación con el cuantil de la F-Snedecor, por tanto, asumimos que en el déficit de ajuste la componente de orden 0 es significativa frente a los errores. Por lo cuál, seguiremos con el estudio de la siguiente componente.

Para el modelo de orden 1 o modelo lineal, obtenemos un estimador F_1' que seguirá una distribución F-Snedecor con 4 y 6 grados de libertad, debido

a que en la suma media de cuadrados obtenidas para el déficit de ajuste, tenemos una dimensión 4. Lo compararemos con el cuantil $F_{4,6,0.95}$:

$$F_1' = 1.78 < 4.534 = F_{4,6,0.95} \ .$$

Obtenemos que el estimador calculado es menor que el cuantil, luego, en el déficit de ajuste, la componente de orden 1 es no significativa, (ns). En la anterior parte, referida al modelo, se obtuvo que la componente lineal era significativa, porlo que el estudio no ha acabado aún, seguiremos comprobando las demás componentes de orden mayor a 1.

Estudiaremos ahora la componente de orden 2, para ella obtenemos un estimador F_2' que seguirá una distribución F-Snedecor con 3 y 6 grados de libertad, ya que para la suma media de cuadrados del déficit de ajuste utilizamos dimensión 3 frente a la dimensión 6 de la suma de cuadrados media del error.

$$F_2' = 0.18 < 4.757 = F_{3,6,0.95} \ .$$

Se obtiene que la componente de orden 2 no es significativa para el déficit de ajuste, debido a que el estadístico es menor que el cuantil. En el anterior estudio, se obtuvo que esta componente era significativa en el modelo, luego debemos continuar con la siguiente componente en el déficit de ajuste.

La siguiente componente, de orden 3, es la primera considerada no significativa en el estudio de las componentes del modelo. Para el déficit de ajuste, obtenemos un estimador F_3' que seguirá una distribución F-Snedecor con 2 y 6 grados de libertad.

$$F_3' = 0.0024 < 5.143 = F_{2,6,0.95} \ .$$

Efectivamente, se obtiene que la componente de orden 3 es no significativa, (ns), para la parte de déficit de ajuste. Se obtuvo que para el modelo tampoco era significativa. Finalmente, terminamos el procedimiento concluyendo que se puede asumir que a partir de la componente de orden 2 las demás componentes no son significativas en el modelo.

Debido a todo esto, asumiremos un modelo cuadrático, de orden 2, como el modelo polinomial que mejor se ajusta a los datos.

4.2.2. El modelo cuadrático

Como hemos dicho, asumimos que el modelo cuadrático es el que mejor se ajusta a los datos. Retomando la descomposición ortogonal del modelo completo, de orden 5, tenemos que

$$y = 4.56U_1 + 2.55U_2 - 0.29U_3 - 0.04U_4 + 0.009U_5 - 0.038U_6 +$$

$$+0.08U_7 + 0.009U_8 + 0.066U_9 - 0.10U_{10} + 0.15U_{11} + 1.66U_{12}$$

lo que podemos reescribir como,

$$y = \frac{4.56}{\|T_1\|}\, T_1 + \frac{2.55}{\|T_2\|}\, T_2 - \frac{0.29}{\|T_3\|}\, T_3 - \frac{0.04}{\|T_5\|}\, T_4 + \frac{0.009}{\|T_5\|}\, T_5 - \frac{0.038}{\|T_6\|}\, T_6 +$$

$$+0.08U_7 + 0.009U_8 + 0.066U_9 - 0.10U_{10} + 0.15U_{11} + 1.66U_{12}\,,$$

donde T_1 es el vector de valores para $p_0(x)$, sabiendo que $p_0(x) = 1$, T_2 es el vector de valores para $p_1(x) = (x - \bar{x})$, T_3 el vector de valores para $p_2(x)$, y así sucesivamente. Por tanto, podemos obtener la ecuación del modelo ortogonal completo,

$$y = \frac{4.56}{\|T_1\|}p_0(x) + \frac{2.55}{\|T_2\|}p_1(x) - \frac{0.29}{\|T_3\|}p_2(x) - \frac{0.04}{\|T_5\|}p_3(x) - \frac{0.009}{\|T_5\|}p_4(x) - \frac{0.038}{\|T_6\|}p_5(x)\,.$$

Con el modelo ortogonal teníamos la ventaja de que los coeficientes $\hat{\beta}_i$ son los mismos sin importar el orden polinomial que consideramos. Luego, de esta expresión podemos obtener la estimación de los coeficientes que participan en el modelo cuadrático.
Luego, los coeficientes estimados $\hat{\beta}_i$ del modelo polinomial de orden 2 o cuadrático serán:

- $\hat{\beta}_0 = \dfrac{y \cdot U_1}{\|T_1\|} = \dfrac{4.56}{\sqrt{12}} = 1.315 = \bar{y}$
- $\hat{\beta}_1 = \dfrac{y \cdot U_2}{\|T_2\|} = \dfrac{2.55}{\sqrt{817}} = 0.0892$
- $\hat{\beta}_2 = \dfrac{y \cdot U_3}{\|T_3\|} = \dfrac{-0.29}{\sqrt{34978.7}} = -0.00156$

En conclusión, el modelo polinomial que mejor ajusta los datos es de orden 2 con la siguiente expresión:

$$y = 1.315 + 0.0892\; p_1(x) - 0.00156\; p_2(x)\,,$$

sabemos que $p_1(x) = (x - \bar{x}) = (x - 18.5)$, pero la expresión de $p_2(x)$ no la tenemos calculada todavía. El cálculo es sencillo, tenemos que

$$\begin{aligned} p_2(x) &= (x-\bar{x})^2 - \left[\frac{X_3 \cdot U_1}{\|T_1\|}\right] p_0(x) - \left[\frac{X_3 \cdot U_2}{\|T_2\|}\right] p_1(x) = \\ &= (x-\bar{x})^2 - \frac{235.8}{\sqrt{5}} 1 - \frac{(-33{,}48)}{\sqrt{817}} (x-\bar{x}) = \\ &= (x-18.5)^2 - 68.08333 + 1.1713 (x-18.5) . \end{aligned}$$

Por tanto, el modelo polinomial ortogonal de orden 2 que mejor se ajusta a los datos (descrito en la Figura 4.2) tiene la siguiente expresión, teniendo en cuenta la transformación que hicimos para las observaciones,

$$\ln(y) = 1.315 + 0.089\,(x-18.5) - 0.00156\,[\,(x-18.5)^2 - 68.08 + 1.17\,(x-18.5)\,]\,.$$

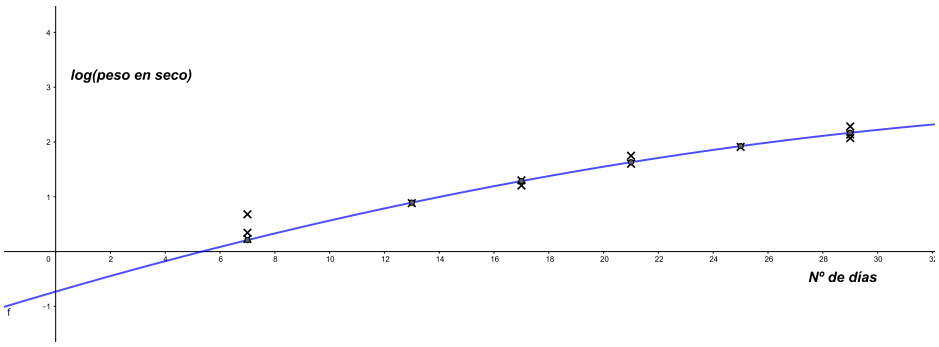

Figura 4.2: Modelo cuadrático ajustado a la nube de puntos

4.3. Geometría

Al escoger un modelo polinomial de orden 2, nos hemos decantado por un espacio modelo de dimensión 3 y un espacio de errores de dimensión 9. Nuestro modelo ajustado es

$$\begin{array}{lcccc} & y & = & (y \cdot U_1)U_1 + (y \cdot U_2)U_2 + (y \cdot U_3)U_3 & + & \text{vector error} \\ \text{ó} & y & = & \hat{y} & + & (y - \hat{y}) \end{array}$$

$$\text{ó} \quad \begin{bmatrix} 0.231 \\ \vdots \\ 2.137 \end{bmatrix} = \begin{bmatrix} 0.2097 \\ \vdots \\ 2.1681 \end{bmatrix} + \begin{bmatrix} 0.0213 \\ \vdots \\ -0.0311 \end{bmatrix}$$

donde los valores ajustados, son los valores para X en el modelo cuadrático, y donde el vector error ha sido calculado restándole a las observaciones, y, los valores ajustados, $\hat{y}$. Estos valores ajustados, se muestran en la Figura 4.2, donde también pueden observarse el valor de los errores, $y_i - \hat{y}_i$, como las desviaciones verticales de las observaciones a la curva ajustada.

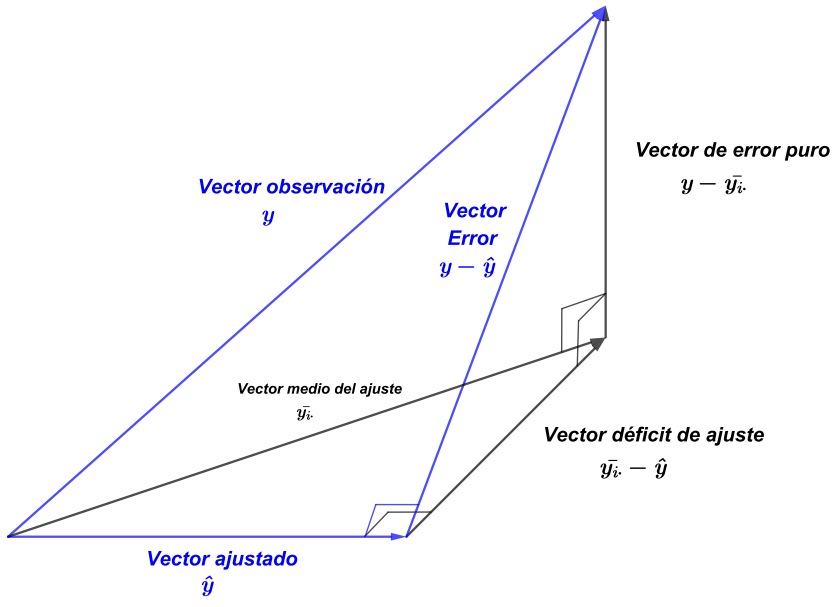

Figura 4.3: Descomposición ortogonal para el modelo cuadrático $y = \hat{y} + (y - \hat{y})$, y para el modelo completo $y = \bar{y_{i\cdot}} + (y - \bar{y_{i\cdot}})$, mostrando una descomposición del vector error.

En el dibujo geométrico de la Figura 4.3 podemos observar que el vector error es la suma del vector de error puro, que consiste en las desviaciones verticales de las observaciones con las medias correspondientes de las distintas fechas de cosecha, x, y el vector error del déficit de ajuste, que consiste en las desviaciones que se producen entre las medias de los tratamientos y los valores apropiados de la curva ajustada. Así,

$$\begin{bmatrix} 0.0213 \\ 0.1343 \\ -0.1417 \\ -0.0096 \\ -0.0811 \\ 0.0129 \\ 0.1183 \\ -0.0277 \\ -0.0144 \\ -0.0971 \\ 0.1159 \\ -0.0311 \end{bmatrix} = \begin{bmatrix} 0.0046 \\ 0.0046 \\ 0.0046 \\ -0.0096 \\ -0.0341 \\ -0.0341 \\ 0.0453 \\ 0.0453 \\ -0.0144 \\ -0.0041 \\ -0.0041 \\ -0.0041 \end{bmatrix} + \begin{bmatrix} 0.0167 \\ 0.1297 \\ -0.1463 \\ 0 \\ -0.0470 \\ 0.0470 \\ 0.0730 \\ -0.0730 \\ 0 \\ -0.0930 \\ 0.1200 \\ -0.0270 \end{bmatrix}$$

$$\begin{array}{ccccc} & & \text{Vector error} & & \text{Vector error} \\ \text{Vector error} & = & \text{déficit de ajuste} & + & \text{puro} \\ (y - \hat{y}) & & (\bar{y_{i\cdot}} - \hat{y}) & & (y - \bar{y_{i\cdot}}) \end{array}$$

Nuestro modelo puede reescribirse como,

$$\begin{array}{rccccccc} & y - \bar{y} & = & (y \cdot U_2) & + & (y \cdot U_3) & + & \text{vector error} \\ \text{ó} & y - \bar{y} & = & \hat{\beta}_1 T_2 & + & \hat{\beta}_2 T_3 & + & \text{vector error} \end{array}$$

Esta es la forma de escribir el modelo que corresponde a la tabla ANOVA asociada al modelo cuadrático, descrita en la Tabla 4.5 , donde el vector error es igual a

$$\begin{aligned} & SC_e = SC \text{ Error de déficit de ajuste} + SC \text{ Error puro} , \\ \text{ó} \quad & SC_e = SC_{ed} + SC_{ep} , \\ \text{ó} \quad & \|y - \hat{y}\|^2 = \|\bar{y_{i.}} - \hat{y}\|^2 + \|y - \bar{y_{i.}}\|^2 , \\ \text{ó} \quad & 0.0842 = 0.0068 + 0.0773 . \end{aligned}$$

Componentes	gl	SC	MC	F
Orden 1	1	6.5129	6.5129	696.13 (**)
Orden 2	1	0.0852	0.0852	9.103 (*)
Error	9	0.0842	0.0094	
Total	11	6.6822		

Tabla 4.5: Tabla ANOVA tradicional para el modelo cuadrático.

Se observa que los estadísticos F han sido comprarados, en la misma tabla ANOVA, con el correspondiente punto crítico, o cuantil, de la distribución F-Snedecor asociada para comprobar si son significativos para el modelo.
La comparación la podemos realizar de la siguiente forma, teniendo en cuenta que los estadísticos F siguen una distribución F-Snedecor con grados de libertad iguales a las dimensiones de los participantes. En este caso, ambas componentes tienen dimensión 1, y están estudiadas frente al espacio de errores con dimensión 9. Luego, el estadístico F, tanto para la componente de orden 1 como para la componente de orden 2, sigue una distribución F-Snedecor con 1 y 9 grados de libertad. Manteniendo el nivel de significación que hemos utilizado a lo largo del estudio, se obtiene que:

$$F_1 = 696.13 \, , \, F_2 = 9.103 > 5.318 = F_{1,9,0.95} \, .$$

4.4. Comportamiento del coeficiente de determinación R^2

Puede ser interesante ver cómo se comportan el coeficiente de determinación, R^2, y el porcentaje de varianza explicado por la regresión, a medida que se aumenta el orden del polinomio de uno a cinco. Su método de cálculo es igual al descrito en el Capítulo 3, obteniendo la Tabla 4.6.

Orden polinomial	SC regresión	R^2	Error MC_e (s^2)	% varianza explicada
1	6.5129	0.975	0.0169	97.2 %
2	6.5981	0.987	0.0094	98.5 %
3	6.6001	0.988	0.0103	98.3 %
4	6.6002	0.988	0.0117	98.1 %
5	6.6049	0.988	0.0129	97.9 %
Total SC = 6.6823			0.6075 = MC total	

Tabla 4.6: Tabla de comportamiento del coeficiente de determiandción, y el porcentaje de varianza explicado por cada modelo de regresión

Observamos que en nuestro ejemplo, el valor del coeficiente de determinación R^2 aumenta hasta 0.988, mientras que el porcentaje de varianza explicado por los modelos de regresión empieza a decrecer una vez que se aumenta el orden polinomial desde el orden 2. Por tanto, el modelo polinomial de orden 2 realiza un buen ajuste, con un coeficiente de determinación de 0.987 y una varianza explicada del 98.5 %.

4.5. Comprobación de las hipótesis

Asumiendo el modelo cuadrático como el modelo que mejor ajusta los datos, hemos asumido también que nuestras observaciones son independientes y se distribuyen según una ley normal con varianza constante alrededor de la curva de ajuste.
La suposición de independencia parece razonable con nuestro experimento, puesto que se estudiaron las plantas separadas unas de otras, todas en las mismas condiciones, y se midieron sus pesos en seco de forma independiente.

Para la hipótesis de normalidad, representamos un histograma de los errores, Figura 4.4, en el cuál se observa que los errores se ajustan en su mayoría a una normal, puesto que la mayoría de las barras están debajo de las curvas, exceptuando un pico final. La hipótesis de normalidad parece razonable.

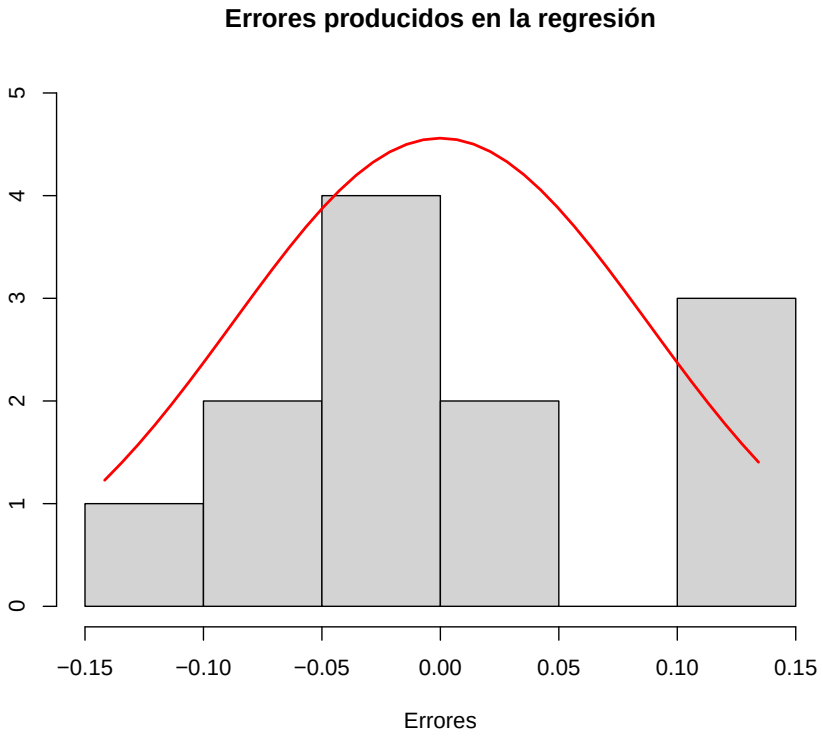

Figura 4.4: Histograma de los errores del ajuste que sigue la línea de distribución normal.

Para el resto de suposiciones, la independencia de los errores, dibujaremos los errores frente a los valores ajustados, y observaremos el comportamiento, tal y como se muestra en la Figura 4.5.

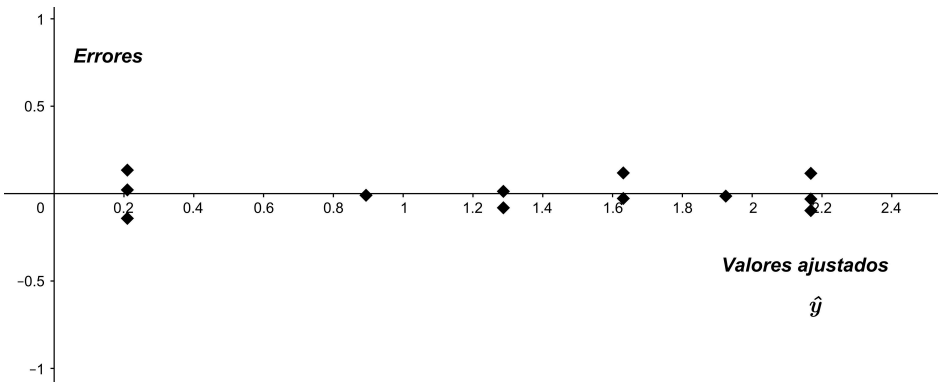

Figura 4.5: Errores frente a los valores ajustados de Y

No se observa ningún patrón, ni parecen estar dispersos. Por tanto, podemos asumir que los errores son independientes y que mantienen varianza constante. Cabe destacar que podemos comprobar la bondad de este modelo en el Gráfico 4.2.
Por todo esto, finalmente lo asumimos como razonable para este conjunto de datos.

4.6. Conclusión final

Se puede afirmar que se ha cumplido el objetivo del experimento, puesto que se ha conseguido calcular una curva que ajusta casi perfectamente los datos, observamos que el valor de los errores cometidos en cada valor que se ajusta, Figura 4.5, son casi nulos. También podemos llegar a la conclusión de que el ajuste es casi perfecto a través del valor del coeficiente de determinación $R^2 = 0.988$, ya que sabemos que equivale al coeficiente de correlación al cuadrado:

$$R^2 = 0.988 \implies R = \sqrt{0.988} = 0.994\,.$$

En el anterior ejemplo, vimos que la interpretación del coeficiente de correlación en un modelo polinomial de orden mayor que 1, era que mientras más cercano fuera a 1, mejor sería el ajuste. En este caso, tenemos un coeficiente de correlación muy cercano a 1 con un valor de 0.994 lo que nos dice que el ajuste es casi perfecto.

Bibliografía

[1] Saville, D. J.; Wood, G. R.: *Statistical Methods: The Geometric Approach*. Springer, 1991.

[2] Saville, D. J.; Wood, G. R.: *Statistical Methods: A Geometric Primer*. Springer, 1996.