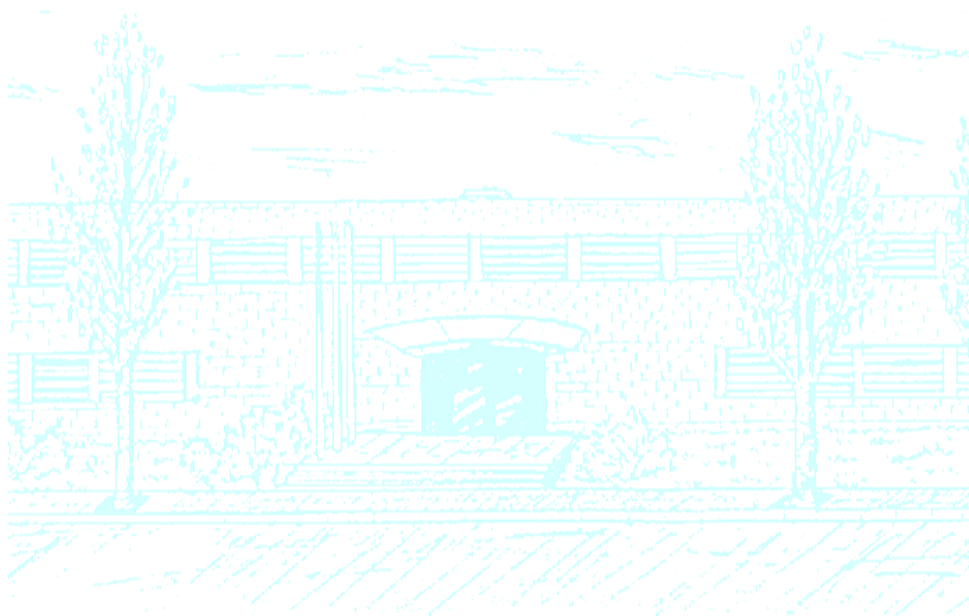


# Diplomatura de Estadística

**Título:** Modelos GEE para el análisis de la evolución de las lesiones acnéicas tratadas con *Nadixa*.

**Autores:** David Agis Cherta  
Lourdes Cirugeda Bolós

**Directores:** Erik Cobo Valeri  
Josep A. Sánchez Espigares



Facultat de Matemàtiques  
i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA



*En memoria de mi madre (1956-2005)*

*y de mi tía (1959-2008).*

*David.*

## **AGRADECIMIENTOS**

A Ilonka Zsolt del Departamento Médico de Laboratorios Ferrer, por habernos dado la oportunidad de realizar este proyecto facilitándonos la base de datos.

Al Dr. Erik Cobo, profesor del Departamento de Estadística e Investigación Operativa de la UPC, por la paciencia y el apoyo moral y académico que siempre recibimos de él.

A Josep Anton Sánchez, profesor del Departamento de Estadística e Investigación Operativa de la UPC, por el tiempo que no tenía y aun así nos dedicó.

## RESUMEN

El profesor Bakar Bouadjar del departamento de dermatología (CHU de Bab-el-Oued, Argelia) y Laboratorios Ferrer quieren estudiar la evolución de los pacientes tratados con *Nadixa crema 1%* para el acné a través de un estudio longitudinal no comparativo.

Para el estudio se seleccionaron 30 pacientes con lesiones inflamatorias de acné vulgar con una gravedad leve-moderada. En la semana basal se realizó un recuento de lesiones y en las semanas 2, 4, 8 y 12 se evaluaron la eficacia y la tolerabilidad cutánea. De los pacientes, 22 son mujeres y 7 son hombres. La edad media ( $\pm$  SD) de los pacientes es de 21,7 años ( $\pm$  4,45) y la duración media ( $\pm$  SD) de la enfermedad es de 48,19 meses ( $\pm$  25,04).

Se ha observado un descenso en el número de lesiones inflamatorias (*pápulas*, *pústulas* y *nódulos*) del -80,6% (IC<sub>95%</sub> [-81,6% ; -79,9%]). En cambio, el número de lesiones no inflamatorias (*quistes*, *lesiones residuales* y *cicatrices*), se mantiene prácticamente constante en el tiempo.

Respecto a la tolerabilidad al medicamento, la probabilidad de tener *eritema* y *prurito* disminuye en los individuos a medida que transcurre el tiempo: se pasa de tener una probabilidad del 65,1% (IC<sub>95%</sub> [50,5% ; 77,3%]) y del 48,7% (IC<sub>95%</sub> [33,6% ; 64,1%]), respectivamente, en la semana 2, a tener 0,9% (IC<sub>95%</sub> [0,1% ; 6,7%]) y 0,3% (IC<sub>95%</sub> [0,02% ; 5,3%]) al final del tratamiento. Sin embargo, la probabilidad de que se produzca *descamación* y *sequedad* no decrece de forma clara a medida que avanzan las semanas: se pasa de tener una probabilidad del 44,7% (IC<sub>95%</sub> [30,8% ; 59,5%]) y del 52% (IC<sub>95%</sub> [37,6% ; 66,2%]), respectivamente, en la semana 2, a tener 21,8% (IC<sub>95%</sub> [10,6% ; 39,8%]) y 26,4% (IC<sub>95%</sub> [13,6% ; 44,9%]) en la semana 12.

En conclusión, después de doce semanas de tratamiento con *Nadixa*, se ha observado una disminución general en las lesiones inflamatorias causadas por el acné. También se ha observado que, en general, la probabilidad de padecer las reacciones adversas *eritema* y/o *prurito* disminuye de forma muy rápida, ya que se produce desde las primeras semanas.

# ÍNDICE

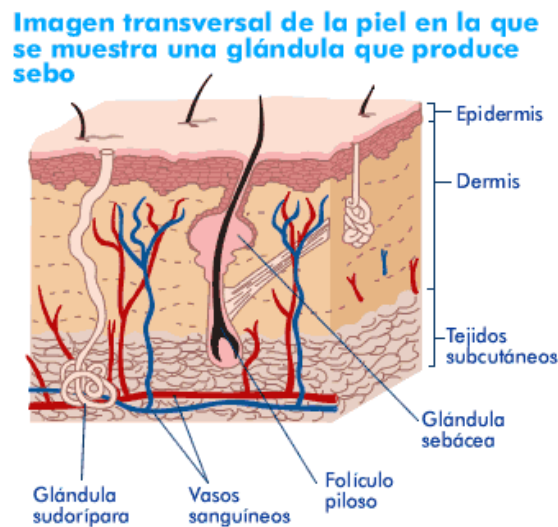
<b>1. INTRODUCCIÓN</b> .....	<b>8</b>
1.1. Acné.....	8
<b>2. OBJETIVOS Y MOTIVACIONES</b> .....	<b>12</b>
2.1. Objetivos.....	12
2.2. Motivaciones.....	12
<b>3. MATERIAL Y MÉTODOS</b> .....	<b>13</b>
3.1. Protocolo del estudio .....	13
3.2. Variables recogidas .....	14
3.3. Variables creadas .....	15
3.4. Metodología estadística descriptiva.....	15
3.5. Modelos de regresión lineal .....	15
3.5.1. Premisas del modelo lineal .....	17
3.5.2. Estimación de los parámetros.....	17
3.5.3. Propiedades de los estimadores .....	18
3.5.4. Residuos.....	20
3.6. Intervalo de confianza para el parámetro de la Distribución Binomial ...	20
3.7. GEE ( <i>Generalized Estimation Equations</i> ).....	21
3.7.1. Formulación .....	22
3.7.2. Estimación de los parámetros.....	24
3.7.3. Estructura de correlación.....	27
3.7.4. Datos faltantes ( <i>missings</i> ) .....	28
3.7.5. Ajuste del modelo GEE en R.....	28
3.7.6. Software.....	29
<b>4. RESULTADOS</b> .....	<b>30</b>
4.1. Descriptiva .....	30
4.2. Regresión lineal simple: variables cuantitativas.....	34
4.3. Modelos GEE: variables cuantitativas .....	41
4.4. Intervalos de confianza de las variables cualitativas .....	49
4.5. Modelos GEE: variables cualitativas.....	51

<b>5. CONCLUSIONES</b> .....	<b>56</b>
5.1. Conclusiones aplicadas .....	56
5.2. Conclusiones de los métodos estadísticos utilizados .....	58
5.3. Posibles extensiones .....	60
5.4. Limitaciones del estudio.....	60
BIBLIOGRAFIA.....	61
ANEXOS.....	62
Anexo I. Escala de gravedad del acné.....	63
Anexo II. Plantilla recogida de datos .....	65
Anexo III. Principales estructuras de correlación.....	69
Anexo IV. Cálculo de la variación porcentual semanal.....	70
Anexo V. Scripts en R.....	71
Anexo VI. Salidas y gráficos de los modelos GEE en R.....	77

# 1. INTRODUCCIÓN

## 1.1. Acné

El acné, también conocido como acné común (*acné vulgaris*), es una enfermedad de la piel caracterizada por la inflamación de las glándulas pilosebáceas que cursa con la obstrucción de poros y la aparición de diferentes lesiones en la piel.



Es una patología dermatológica que tiene su origen en un funcionamiento anormal del aparato pilosebáceo y origina una serie de lesiones elementales que se producen en la cara, cuello, pecho, hombros y espalda.

El acné es la enfermedad de la piel más común: personas de todas las razas y edades la pueden padecer. Así mismo, el acné es más frecuente en adolescentes y en jóvenes. Alrededor de un 80% de las personas de entre 11 y 30 años tienen brotes en algún momento de su vida. Aun así, algunas personas tienen acné después de los 40-50 años.



## Causas

No se conocen las causas exactas del acné, pero los médicos creen que es debido a diversos factores. El factor principal es tener un nivel alto de hormonas. Las hormonas sexuales masculinas, llamadas andrógenos, aumentan en la pubertad. Un mayor número de andrógenos implica un mayor tamaño de las glándulas sebáceas, aumentando la producción de sebo. Otro factor relacionado con el acné es el hereditario o genético. La tendencia a desarrollar acné puede heredarse de los padres.

Hay diversos factores que pueden empeorar el acné. Por ejemplo, los niveles hormonales en las niñas y en las mujeres pueden empeorarlo días antes del comienzo del ciclo menstrual. Los cambios hormonales debidos al embarazo, el estrés, algunos medicamentos, el maquillaje graso, los irritantes ambientales... también pueden empeorarlo.

## Síntomas

Las lesiones del acné reciben el nombre de *comedones*, folículos de piel que han aumentado de tamaño al estar taponados con sebo y bacterias. Si no se pueden ver a simple vista, se suelen llamar *microcomedones*.

Si el *comedón* queda debajo de la piel, se le llama *comedón cerrado* o *punto blanco*. Los *puntos blancos* son pequeños bultos en la superficie de la piel. Un *comedón* que llega a la superficie de la piel y se abre, se llama *punto negro*.

En general, se pueden dividir las lesiones en dos grandes grupos, según se mantenga la fase inflamatoria, en la que existe una reacción “viva” del organismo, o se haya pasado a la fase residual.

Lesiones inflamatorias:

- *Pápulas*: lesiones inflamatorias dolorosas a la palpación que se originan principalmente a partir de un *comedón abierto* y, en alguna ocasión, de un *comedón cerrado*.



- *Pústulas*: lesiones inflamatorias llenas de pus que pueden tener una base rojiza.



- *Nódulos*: lesiones inflamatorias profundas, a veces más palpables que visibles, recubiertas de piel normal o piel con *eritema*, que producen dolor.



### Lesiones no inflamatorias:

- *Quistes*: lesiones profundas y llenas de pus que pueden causar dolor y *cicatrices*.



- *Cicatrices*: Hay dos tipos de *cicatrices*, las llamadas *cicatrices pigmentadas* y las que vulgarmente se llaman “*agujeros*” o “*pozos*”.
  - *Cicatrices pigmentadas*: son manchas de color rosa o café que persisten una vez el grano desaparece.
  - “*Agujeros*” o “*Pozos*”: son literalmente agujeros pequeños en la piel.
- *Lesiones residuales*: son secuelas de las lesiones inflamatorias, manchas rojas que pueden durar varios meses y cicatrices deprimidas.

### **Posibles reacciones adversas al medicamento**

- *Eritema*: enrojecimiento o inflamación de la piel.



- *Prurito*: hormigueo o irritación incómoda de la piel que provoca el deseo de rascarse en el área afectada. Comúnmente se llama picazón.
- *Descamación*: desprendimiento o exfoliación visible de las capas externas de la piel.
- *Sequedad*: bajo nivel de concentración hídrica.

## 2. OBJETIVOS Y MOTIVACIONES

### 2.1. Objetivos

El objetivo de este trabajo es modelar la evolución del acné durante 12 semanas después de tratarlo con *Nadixa* crema al 1%. Para ello, primero se aplicó regresión lineal simple, pero al no cumplirse algunas premisas de dicha regresión, se modelaron las respuestas discretas y ordinales longitudinales mediante los modelos GEE.

Las variables principales analizadas han sido el número total de lesiones inflamatorias (*ppn*) y el número total de lesiones no inflamatorias (*crs*<sup>1</sup>). Además, se han analizado las variables ordinales *eritema*, *descamación*, *sequedad* y *prurito*.

### 2.2. Motivaciones

Ante la necesidad de modelar las variables respuesta en función del tiempo, nos surgió la pregunta: ¿cuál es el modelo estadístico que conocemos? En la diplomatura, y más concretamente en la asignatura Modelos Lineales, se nos enseñó a trabajar con el modelo de regresión lineal simple para poder hacer este tipo de análisis. Pero al aplicarlo en el presente estudio, constatamos que los residuos no cumplían la premisa de independencia, ya que el modelo lineal empleado ignoraba que el mismo paciente era observado a lo largo del tiempo. Por este motivo, ajustar un modelo poblacional mediante regresión lineal simple puede no ser útil, porque no se tiene en cuenta la correlación entre las observaciones de un mismo individuo. Ahora bien, si quisiéramos ajustar un diseño mediante este modelo, se podría ajustar un modelo por cada individuo, pero obtendríamos un modelo sobreparametrizado. Por todo esto, decidimos introducirnos en una metodología alternativa que no hemos visto a lo largo de la diplomatura: el método de las *Ecuaciones de Estimación Generalizadas* (GEE), ya que era una solución razonable al problema de este proyecto. Desde nuestro punto de vista, nos hizo falta una formación teórica en este tipo de modelos y una formación práctica en un software estadístico (R v.2.7.2.) que permitiese estimarlos y en la interpretación de los resultados obtenidos.

---

<sup>1</sup> *ppn* y *crs* son variables creadas a partir de las originales.

### 3. MATERIAL Y MÉTODOS

Los datos fueron proporcionados por los Laboratorios Ferrer, del estudio “*Clinical efficacy and safety Nadixa (Nadifloxacin) cream 1% in mild to moderate inflammatory forms of acné vulgaris: an open study in Algeria*”. Dicho estudio fue dirigido por el profesor Bakar Bouadjar, del departamento de dermatología, CHU de Bab-el-Oued, Argelia.

#### 3.1. Protocolo del estudio

El tamaño del estudio se fijó en 30 pacientes con lesiones inflamatorias de acné vulgar con una gravedad de leve a moderada<sup>2</sup>.

Antes del comienzo del estudio, a cada paciente/tutor se le entregó el formulario del consentimiento informado, indicando las directrices y el objetivo del estudio. Dicho consentimiento informado debía ser firmado por los pacientes/tutores antes del inicio del tratamiento.

Los pacientes no podían recibir medicación tópica durante los 14 días previos al inicio del tratamiento, ni tratamientos con antibióticos durante los 30 días previos.

El tratamiento consistió en aplicaciones tópicas dos veces al día de *Nadifloxacin* durante un período de doce semanas.

En la semana basal, se realizó una primera valoración de las lesiones acnéicas. La eficacia y tolerabilidad cutánea se evaluaron en las semanas 2, 4, 8 y 12.

---

<sup>2</sup> Para ver la escala de gravedad del acné véase Anexo I.

### 3.2. Variables recogidas

La situación basal y la eficacia fueron determinadas por un dermatólogo mediante el conteo de las lesiones inflamatorias (*pápulas, pústulas y nódulos*) y las lesiones no inflamatorias (*quistes, lesiones residuales y cicatrices*).

La tolerabilidad cutánea fue determinada mediante el examen de los signos de *eritema, descamación, sequedad y prurito*. Todas las evaluaciones de la tolerabilidad fueron realizadas por la calificación de diversos síntomas en una escala del 0 al 3.

La recogida de los datos se realizó a través de una plantilla (véase Anexo II), anotándose la evolución del paciente durante el tratamiento. Posteriormente, en las instalaciones de la UPC se introdujeron en una base de datos.

En resumen, las variables recogidas fueron de dos tipos: demográficas y las relativas al acné.

#### Variables demográficas:

- *Date*: fecha en la que se inició el tratamiento.
- *Age*: edad del paciente al iniciar el tratamiento. Expresada en años.
- *Disease Duration*: duración de la enfermedad. Expresada en meses.
- *Sex*: sexo del paciente. F = Female, M = Male y \* = Unknown.

#### Variables relativas al acné:

- *Week*: semanas de seguimiento del tratamiento (0, 2, 4, 8, 12).
- *Patient*: número de paciente (identificador).
- *Papules*: número de pápulas.
- *Pustules*: número de pústulas.
- *Nodules*: número de nódulos.
- *Cysts*: número de quistes.
- *Retentional Lesions*: número de lesiones residuales.
- *Scar Lesions*: número de cicatrices.

- *Erythema*: variable cualitativa que evalúa el grado de eritema<sup>3</sup>.
- *Scaling*: variable cualitativa que evalúa el grado de descamación.
- *Dryness*: variable cualitativa que evalúa el grado de sequedad.
- *Pruritus*: variable cualitativa que evalúa el grado de prurito.

### 3.3. Variables creadas

Las siguientes variables fueron creadas a partir de las variables originales siguiendo las definiciones de la gravedad de las lesiones:

- *ppn*: suma de las variables *papules*, *pustules* y *nodules* (lesiones inflamatorias).
- *crs*: suma de las variables *cysts*, *retentional lesions* y *scar lesions* (no inflamatorias).

### 3.4. Metodología estadística descriptiva

Se realizó un análisis descriptivo de las variables de interés detalladas en el apartado anterior. Para las variables cuantitativas se obtuvieron los estadísticos de tendencia central y dispersión habituales y se realizaron boxplots. Las variables cualitativas se resumieron mediante tablas de frecuencias y gráficos de barras o de sectores.

### 3.5. Modelos de regresión lineal

El análisis de regresión lineal es un procedimiento para analizar las relaciones de asociación entre una variable dependiente y una o más variables independientes  $x_1, x_2, \dots, x_p$ . Su objetivo es encontrar la función lineal de las variables independientes que mejor predice la respuesta  $y$ . Se trata de una de las técnicas estadísticas más utilizadas.

Una regresión en la que se emplea una sola variable explicativa es conocida como regresión lineal simple, y con dos o más se conoce como regresión lineal múltiple. En el presente trabajo, únicamente se ha empleado el tiempo como variable independiente, por lo que sólo se explicará la teoría de la regresión simple.

---

<sup>3</sup> Categorías de las variables de tolerabilidad: 0 (“none”), 1 (“mild”), 2 (“moderate”) y 3 (“severe”).

La regresión se utiliza para lo siguiente:

- determinar si las variables independientes explican una variación estadísticamente significativa en la variable dependiente, es decir, si existe relación o no.
- determinar cuánto de la variación de la variable dependiente se explica con las variables independientes, es decir, se mide la fuerza de la relación.
- determinar la estructura o la forma de la relación: la ecuación matemática que relaciona las variables independientes con la dependiente.
- pronosticar los valores de la variable dependiente.
- controlar otras variables independientes al evaluar las aportaciones de una o varias variables específicas.

La ecuación teórica de regresión lineal simple se expresa:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Las  $\beta_i$  se llaman coeficientes de regresión y están determinadas por la estructura de la relación. La  $\varepsilon$  es una contribución aleatoria, también conocida como *residuo* o *error*. Se puede interpretar como la idiosincrasia de esa unidad, es decir, aquello que es característico de ese caso y no se comparte con el resto.

La ecuación del modelo ajustado de regresión simple es de la forma:

$$\hat{y} = b_0 + b_1 x_1 + e$$

El primer objetivo es obtener estimaciones, es decir, valores numéricos de los coeficientes  $b_0$  y  $b_1$  en función de los datos de la muestra.

La consistencia entre los datos y el modelo no implica la consistencia entre el modelo y la realidad. Lo único que se puede afirmar es que los supuestos del investigador no son contradictorios y, por lo tanto, pueden ser válidos. Pero el "ser válidos" no quiere decir que sean la única explicación del fenómeno objeto de estudio, ya que es posible que otros modelos también se adapten a los mismos datos. Por tanto, a diferencia de un modulo experimental que puede estimar efectos, el modelado no permite realizar inferencia causal sin premisas adicionales.



### 3.5.1. Premisas del modelo lineal

Para poder formular un modelo de regresión lineal es necesario que se cumplan varias premisas:

- que la relación entre las variables sea lineal.
- que los errores tengan varianza constante, es decir, que la varianza no dependa de los valores que tome  $x_j$ .
- que los errores sigan una Distribución Normal.
- que los errores sean independientes.
- que las variables independientes no sean aleatorias.

Para utilizar la inferencia en regresión, los datos deben satisfacer los supuestos en que se basa el modelo de regresión. Para comprobar si los datos cumplen las premisas, se han de analizar los residuos.

### 3.5.2. Estimación de los parámetros

Las estimaciones de los parámetros se suelen hacer por el método de los mínimos cuadrados, que consiste en minimizar la suma de los cuadrados de los residuos, es decir, minimizar:

$$S(\beta_0, \beta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Los residuos miden la discrepancia entre el valor real y el valor que se ha aproximado. Contienen toda la información entre la  $y$  y las  $x$ 's que el modelo ajustado no es capaz de explicar.

Residuo observado:

$$e_i = y_i - \hat{y}_i$$

Para minimizar esta suma de cuadrados, se ha de igualar a cero las primeras derivadas parciales respecto a los coeficientes de regresión. Las ecuaciones del sistema se llaman *ecuaciones normales*,  $b_0$  es el término constante y  $b_1$  es la estimación del coeficiente de regresión:

$$\begin{cases} \frac{\partial \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\partial b_0} = 0 \\ \frac{\partial \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\partial b_1} = 0 \end{cases}$$

De aquí obtenemos las ecuaciones para poder calcular  $b_0$  y  $b_1$ :

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov(x,y)}{s_x^2}$$

El parámetro  $\beta_0$  representa el valor medio de  $y$  cuando  $x$  vale cero, y  $\beta_1$  representa el incremento que experimenta la media de  $y$  cuando  $x$  aumenta en una unidad.

### 3.5.3. Propiedades de los estimadores

- De la estimación del estimador  $b_0$ , se deduce que la recta de regresión pasa por el punto  $(\bar{x}, \bar{y})$ , que es el centro geométrico de la nube de datos.
- Utilizando las hipótesis de normalidad e independencia, se obtiene que la distribución del estimador  $b_1$  es una Normal de media  $\beta_1$  y varianza  $\frac{\sigma^2}{ns_x^2}$ . Esto es,

$$b_1 \sim N \left( \beta_1, \frac{\sigma^2}{ns_x^2} \right)$$

Por lo tanto, la  $V(b_1)$ :

- disminuye al aumentar  $n$ .
- disminuye al aumentar  $s_x^2$ .
- disminuye al disminuir  $\sigma^2$ .

- El estimador  $b_0$  indica el valor de la ordenada en la recta de regresión estimada para  $x=0$ . Tiene menor importancia y, en muchos casos, no tiene una interpretación práctica. La distribución de  $b_0$  es una Normal de media  $\beta_0$  y varianza  $\frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{ns_x^2}$ . Esto es,

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \left(1 + \frac{\bar{x}^2}{s_x^2}\right)\right)$$

Por tanto, la  $V(b_0)$ : - disminuye al disminuir  $V(b_1)$ .  
- disminuye al disminuir  $\bar{x}^2$ .

- Como estimador de  $\sigma^2$  se utiliza la varianza residual  $s_R^2$ , definida como la suma de residuos al cuadrado dividida por el número de grados de libertad:

$$\hat{s}_R^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

$\hat{s}_R^2$  es un estimador consistente e insesgado.

- La distribución de la varianza residual viene dada por:

$$\frac{(n-2)\hat{s}_R^2}{\sigma^2} \sim \chi_{n-2}^2$$

A partir de este estadístico, se pueden obtener intervalos de confianza de la varianza poblacional  $\sigma^2$  con nivel de confianza  $1 - \alpha$ .

- Los estimadores  $b_0$  y  $b_1$  no son variables aleatorias independientes, ya que su covarianza viene dada por:

$$Cov(b_0, b_1) = \frac{-\bar{x}\sigma^2}{ns_x^2}$$

Por lo tanto, si  $\bar{x}$  es positiva, la  $Cov(b_0, b_1)$  es negativa, por lo que, al crecer la estimación  $b_1$ , disminuye  $b_0$ .

### 3.5.4. Residuos

Los residuos, en un modelo lineal, contienen toda aquella información sobre la respuesta que contienen los datos, pero que el modelo ajustado no refleja.

Los residuos son el error de predicción que indica en cuánto se equivoca el modelo a la hora de predecir  $y$ .

No se puede dar un modelo por razonable hasta que no se esté seguro de que en los residuos no queda información sobre la relación entre la  $y$  y la  $x$ , que se pueda aprovechar para mejorar el modelo.

Los objetivos del análisis de los residuos son:

- Verificar que se cumplen las premisas del modelo.
- Sugerir mejoras del modelo.
- Detectar puntos especiales.

Para compararlos entre ellos, suele ser más ilustrativo estandarizarlos, es decir, dividir cada residuo por una estimación de su desviación típica.

Las herramientas gráficas para realizar el análisis de los residuos, y así comprobar la adecuación del modelo, serán el histograma y el *Normal Probability Plot* (para comprobar su normalidad). El gráfico de los residuos contra los valores previstos y el de los residuos contra el orden, son útiles para estudiar la independencia y la homocedasticidad de los residuos.

### 3.6. Intervalo de confianza para el parámetro de la Distribución Binomial

Dado que el interés es conocer la tolerabilidad al medicamento, se observa la evolución de las diferentes reacciones adversas en el tiempo. Para ello, se estudia cómo evoluciona la categoría 0 (“none”), es decir, se estudia si a medida que avanzaban las semanas, el porcentaje de pacientes en dicha categoría aumentaba o no. Con este enfoque, la interpretación de los resultados se simplificó a la ausencia/presencia de reacciones adversas (recategorización de las variables cualitativas)<sup>4</sup>.

---

<sup>4</sup> De 4 categorías (0, 1, 2, 3) a 2 (0, 1).

Para poder ver la tendencia de las reacciones adversas durante el tratamiento, se utiliza intervalos de confianza al 95%. Al tratarse de una variable binaria (0=ausencia, 1=presencia), dichos intervalos se calcula con el método exacto basado en la Distribución Binomial, ya que el reducido tamaño muestral ( $n=30$ ) no permite utilizar los resultados asintóticos basados en la Distribución Normal.

Para construir el intervalo de confianza para  $p$ , de la Distribución Binomial, con un nivel  $(1 - \alpha)$  de confianza, se deben determinar los límites inferior ( $x_1 = L_I$ ) y superior ( $x_2 = L_S$ ) resolviendo las siguientes ecuaciones:

$$1. P[Bin(n; p) \leq x_1] = \alpha/2 \Rightarrow \sum_{k=0}^{x_1} P(Bin = k) = \alpha/2$$

$$2. P[Bin(n; p) \geq x_2] = \alpha/2 \Rightarrow 1 - \sum_{k=0}^{x_2-1} P(Bin = k) = \alpha/2$$

donde  $Bin(n; p)$  es una variable aleatoria Binomial, el valor de  $p$  corresponde a la frecuencia relativa observada y el de  $n$  corresponde al número de casos considerados para cada semana.

### 3.7. GEE (*Generalized Estimation Equations*)

Los datos longitudinales son observaciones repetidas de una variable respuesta, recogidas en diferentes instantes de tiempo, en un mismo individuo. Para realizar un análisis correcto de dichos datos, se debe considerar la posible correlación entre las medidas de cada sujeto. En el modelo lineal se ignora dicha correlación, por lo que la varianza de los estimadores sería incorrecta y, por lo tanto, se llegarían a conclusiones erróneas respecto a la significación estadística.

Una herramienta adecuada para el análisis de este tipo de datos es introducida por Liang y Zeger (1986), a partir de las llamadas *Ecuaciones de Estimación Generalizadas* (GEE). Los GEE modelan la esperanza marginal o poblacional, incorporando la correlación entre las observaciones correspondientes a un mismo individuo y asumiendo independencia entre los individuos. Además, es interesante considerar los GEE porque admiten la posibilidad de que la variable respuesta siga una distribución diferente de la Gaussiana, pudiendo incluir distribuciones de la familia exponencial (Binomial, Poisson, Gamma,...).

Estos modelos consideran una ecuación de estimación que es escrita en dos partes: la primera, estima los parámetros de regresión y la segunda, los de correlación.

Los modelos GEE son bastante flexibles, ya que para hacer una buena estimación de los parámetros del modelo sólo necesitan una transformación o función “link”, una función de la varianza y una estructura de correlación.

Los GEE funcionan mejor cuando:

- el número de observaciones por sujeto es pequeño y el número de sujetos es grande.
- se trata de estudios longitudinales, obteniéndose las medidas siempre en el mismo instante de tiempo para todos los sujetos.

Dado que los datos de la muestra estudiada cumplen estas dos características, decidimos aplicar los modelos GEE, aun escapándose de nuestros conocimientos adquiridos en la Diplomatura.

### 3.7.1. Formulación

En la formulación del modelo se distinguen dos partes:

1. *Parte sistemática*: relativa a la esperanza de una observación a través de una función “link”  $g$  que conecta el predictor lineal con la respuesta:

$$g(E[Y_{ij}]) = g(\mu_{ij}) = x'_{ij} \beta, \quad \text{donde } i = 1, \dots, n \\ j = 1, \dots, n_i$$

Para los datos del presente estudio, se necesitan dos tipos de función “link”, una para las variables respuesta tipo recuento y otra para las ordinales, previamente dicotomizadas:

- La Distribución Poisson es la más utilizada cuando se trabaja con recuentos. Es de la forma:

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \dots$$

donde  $E(Y|X) = V(Y|X) = \mu$

La función “link” es:

$$\log E(Y_{ij}) = x'_{ij} \beta$$

- Como las respuestas ordinales tienen muy pocos casos con valores mayores que 1, se recategoriza para tener una respuesta binaria, siendo 0 ausencia de reacciones adversas y 1 presencia de reacciones adversas al tratamiento.

La función “link” es:

$$\text{logit } P(Y_{ij} = 1|X_1, X_2, \dots, X_k) = \log \left( \frac{P(Y_{ij} = 1|X_1, X_2, \dots, X_k)}{1 - P(Y_{ij} = 1|X_1, X_2, \dots, X_k)} \right) = x'_{ij} \beta$$

donde  $E(Y|X) = p$

$$V(Y|X) = p(1 - p)$$

2. *Parte aleatoria*: especifica cómo la varianza se relaciona con la esperanza matemática mediante una función de la varianza.

$V(Y_{ij}) = v(\mu_{ij}) \phi$ , donde  $v$  es la función de la varianza y  $\phi$  el parámetro de escala.

Además se necesita especificar la estructura de correlación (véase 3.7.3.). Esto se hace de la siguiente manera:

- i. se identifican los grupos donde las observaciones están correlacionadas.
- ii. se elige para estos grupos una estructura de la matriz de correlaciones  $R(\alpha)$ , llamada *working correlation matrix* (es un matriz cuadrada completamente especificada por un vector de parámetros  $\alpha$ ).

Con estos tres componentes especificados (función “link”, función de la varianza y *working correlation matrix*), se pueden derivar las ecuaciones para calcular las  $\beta$ 's.

Nótese que con los modelos GEE no es necesaria la especificación de un modelo estadístico, es decir, una distribución de probabilidad  $f(y; \theta)$ . De este modo, los GEE son bastante flexibles, necesitando sólo unas pocas suposiciones. Pero el precio por dicha flexibilidad es:

- la estimación de las  $\beta$ 's no tiene por qué ser la mejor posible.
- la inferencia (tests) está basada asintóticamente.
- métodos de validación más complicados.

### 3.7.2. Estimación de los parámetros

Los principales parámetros a estimar en los modelos GEE son:

- los parámetros de regresión ( $\beta$ ), que expresan la relación entre los predictores y la respuesta.
- el factor de escala ( $\phi$ ), que analiza la relación media-varianza de la variable respuesta. Indica la variabilidad de los datos.
- el parámetro de correlación ( $\alpha$ ), que expresa la correlación intra clúster, es decir, entre las observaciones de un mismo individuo. El o los parámetros de correlación se calculan a partir del tipo de matriz de correlaciones especificada por el usuario.

La forma de estimar las  $\beta$ 's es muy similar al método de estimación de las observaciones independientes. En otras palabras, la estimación de los GEE es muy similar a la estimación de mínimos cuadrados.



Las estimaciones de las  $\beta$ 's son unas buenas estimaciones en el sentido que:

- el estimador es asintóticamente consistente, esto es, cuando el tamaño de la muestra tiende a infinito, las  $\hat{\beta}$ 's son prácticamente las verdaderas  $\beta$ 's.
- el estimador es asintóticamente Normal,  $\hat{\beta} \sim N(\beta, \hat{V}(\hat{\beta}))$ . El problema está en  $\hat{V}(\hat{\beta})$ .

La normalidad asintótica se mantiene aunque:

- la función de la varianza esté mal especificada.
- la *working correlation matrix* no sea la matriz apropiada.

Si tenemos una buena estimación de  $\hat{V}(\hat{\beta})$  podremos:

- hacer test de hipótesis.
- construir contrastes.
- hacer intervalos de confianza.
- etc.

Para estimar  $\beta$  se considera la *ecuación de estimación generalizada*:

$$S(\beta) = \sum_{i=1}^n \left( \frac{\partial \mu_i}{\partial \beta} \right)' V_i^{-1} (Y_i - \mu_i) = 0$$

donde  $V_i$  es la matriz de covarianza  $Y_i$

$$V_i = \phi A^{1/2} R_i(\alpha) A^{1/2}$$

y  $A$  es una matriz diagonal  $n_i \times n_i$  con  $v(\mu_{ij})$  como el elemento  $j$ -ésimo de la diagonal.

Si  $R_i(\alpha)$  es la verdadera matriz de correlación de  $Y_i$ , entonces  $V_i$  será la verdadera matriz de covarianza de  $Y_i$ .

La estimación del parámetro de escala es

$$\hat{\phi} = \frac{1}{(\sum_{i=1}^n n_i) - p} \sum_{i=1}^n \sum_{j=1}^{n_i} e_{ij}^2$$

siendo  $p$  el número de covariables. Si  $\hat{\phi} \cong 1$ , la distribución de probabilidad escogida será la adecuada. En cambio, si  $\hat{\phi}$  es mayor o menor que 1, indicará que hay sobredispersión o infradispersión, respectivamente. En nuestro caso, se va a trabajar con Poisson y Binomial, y en estos dos modelos la  $\hat{\phi}$  debería ser 1 por construcción del modelo probabilístico.

En muchas ocasiones, la distribución de probabilidad asumida no se cumple, por lo que se suele verificar la hipótesis a priori. Para ello se comprueba el cociente entre la varianza y la media:

$$\frac{s^2}{\bar{x}} \begin{cases} > 1 \Rightarrow \text{Distribución sobredispersa} \\ \cong 1 \Rightarrow \text{Distribución adecuada} \\ < 1 \Rightarrow \text{Distribución infradispersa} \end{cases}$$

Si se da la distribución infradispersa/sobredispersa, esto podría indicar que otra distribución sería más adecuada. Esta infradispersión/sobredispersión se corroborará con la  $\hat{\phi}$  obtenida en la salida de R.

En los GEE existen dos formas de estimar la verdadera matriz de varianza-covarianza  $\Sigma$ , dependiendo de la forma de la *working correlation matrix*:

1. Estimador *robusto*, *empírico* o *sandwich*. Esta estimación tiende asintóticamente a la verdadera  $\Sigma$ , aunque la *working correlation matrix* no esté bien especificada. Se estimaría de la siguiente manera:

$$\hat{V}(\hat{\beta}) = I_0^{-1} I_1 I_0^{-1}$$

donde

$$I_0 = \sum_{i=1}^n \frac{\partial \mu_i'}{\partial \beta} V(Y_i)^{-1} \frac{\partial \mu_i}{\partial \beta}$$

$$I_1 = \sum_{i=1}^n \frac{\partial \mu_i'}{\partial \beta} V(Y_i)^{-1} (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)' V(Y_i)^{-1} \frac{\partial \mu_i}{\partial \beta}$$

2. Si la *working correlation matrix* es correcta, el mejor estimador es el *naive* o *model-based*. Se estimaría de la siguiente manera:

$$\hat{V}(\hat{\beta}) = I_0^{-1}$$

En su trabajo original, Liang y Zeger (1986) muestran que solamente la eficiencia del estimador de  $\beta$ , y no la consistencia, es afectada cuando la matriz de correlación especificada es incorrecta. En lo que respecta a la precisión del estimador de  $\beta$ , fue mencionado que una manera de protegerse ante una estructura de correlación incorrecta es utilizar el estimador *robusto*.

### 3.7.3. Estructura de correlación

Las estructuras de correlación más conocidas son: *independiente*, *exchangeable*, *autoregresiva (AR(1))* y *desestructurada* (Anexo III). La *working correlation matrix* la escoge el analista según su criterio. Aunque no existen unas normas claras que indiquen en qué caso se debe escoger cada estructura de correlación, sí se dispone de una recomendación: la mejor estructura debería ser la más simple con la que se ajusten los datos.

Algunos autores sugieren utilizar la estructura de correlación *independiente* cuando el interés es la estimación de  $\beta$  porque en muchas situaciones produce resultados satisfactorios, con estimadores tan eficientes como los obtenidos con estructuras más complejas (Lipsitz et al. 1991; Miller et al. 1993). Aunque también ocurre que la pérdida de eficiencia bajo independencia puede ser muy alta, por lo que se sugiere un esfuerzo en la modelización de la correlación, aun cuando se considere una estructura de correlación incorrecta (Fitzmaurice 1995). Por este motivo, en el presente trabajo, se utiliza el tipo de matriz de correlaciones *exchangeable*. Esta  $R(\alpha)$  asume que todas las observaciones de un mismo sujeto tienen la misma correlación entre ellas:  $\text{Corr}(y_{ki}, y_{kj}) = \alpha$ .

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

Cabría esperar que la estructura adecuada fuese la  $AR(1)$ , pero ésta se utiliza con la condición de que los datos sean longitudinales y equidistantes, por lo que no es la estructura adecuada con los datos disponibles.

#### 3.7.4. Datos faltantes (*missings*)

Es deseable no tener o tener la mínima cantidad de datos perdidos, ya que su presencia puede complicar el análisis. El análisis mediante GEE sólo proporciona resultados consistentes en los casos de datos MCAR<sup>5</sup> (Liang and Zeger (1986)).

#### 3.7.5. Ajuste del modelo GEE en R

Para el ajuste del modelo GEE, se han empleado las funciones *geeglm* y *gee* de los paquetes de R *Geepack* y *Gee*, respectivamente.

La función *geeglm* proporciona:

- la estimación, el error estándar robusto y el test de Wald (con el p-valor) de las  $\beta$ 's.
- $\hat{\phi}$  y  $\hat{\alpha}$ , con sus respectivos errores estándar.
- el número de clúster y el número máximo de casos de los clústers.

La función *gee* proporciona:

- una descriptiva de los residuos.
- la estimación, el *Naive S.E.* y el *Robust S.E.* (con sus respectivos valores z) de las  $\beta$ 's.
- $\hat{\phi}$  y  $R(\hat{\alpha})$ .

La función *gee* permite ajustar el modelo observando las semejanzas entre los errores estándar de los estimadores *naive* y *robusto*. Ahora bien, si existen diferencias entre ambos, aún se podría validar el modelo si el estadístico de Wald (de la función

---

<sup>5</sup> Missing Completely At Random: el dato faltante es independiente de los datos observados y de los no observados.

*geeglm*) y los valores  $z$  (de la función *gee*) son equivalentes en cuanto a la significación estadística.

Para calcular de los intervalos de confianza se ha utilizado la función *predict* con la opción *se.fit* para obtener el error estándar de las predicciones.

### **3.7.6. Software**

La descriptiva y la regresión lineal simple se han realizado con el paquete estadístico Minitab 15. Los  $IC_{95\%}$  de las variables cualitativas se han calculado con Microsoft Excel 2007. Para los modelos GEE, se ha utilizado el programa R v.2.7.2.

## 4. RESULTADOS

Este apartado se estructura en tres grandes bloques. En el primero, se realiza una breve descriptiva de las variables de interés; en el segundo, se exponen los resultados obtenidos de las variables cuantitativas; y en el tercero, los resultados de las variables cualitativas.

### 4.1. Descriptiva

Antes de empezar a modelar los datos, se realiza una breve descriptiva de las diferentes variables del estudio.

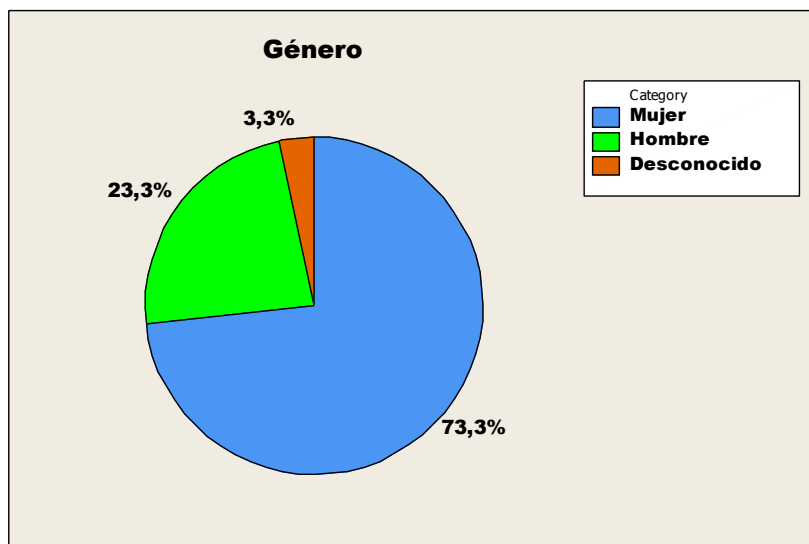


Figura 1. Pie Chart del Género.

La muestra consta de 30 pacientes, de los cuales 22 son mujeres y 7 son hombres. El género de uno de los individuos (paciente 14) no se ha podido identificar.

Variable	N	N*	Mean	StDev	Min	Q1	Median	Q3	Max	Range	IQR
Age	30	0	21,7	4,45	13	18	21	24	32	19	6
Disease duration	26	4	48,19	25,04	5	24	42	72	84	79	48

Tabla 1. Descriptiva de la edad de los pacientes y de la duración de la enfermedad.

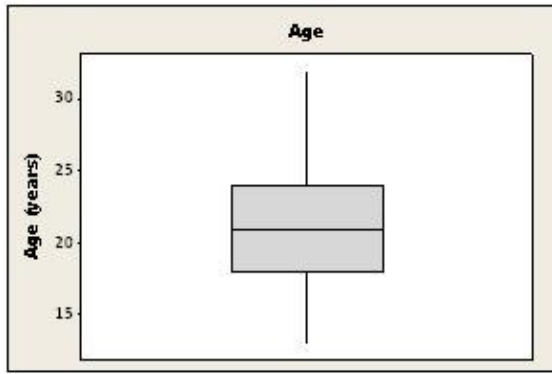


Figura 2. Boxplot edad (años).

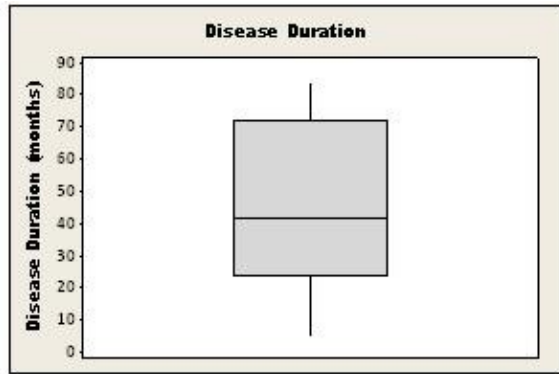


Figura 3. Boxplot duración enfermedad (meses).

La edad media de los pacientes estudiados es aproximadamente de 22 años ( $\pm 4,45$ ) y la duración media de la enfermedad es de unos 48 meses (25,04), es decir, unos 4 años.

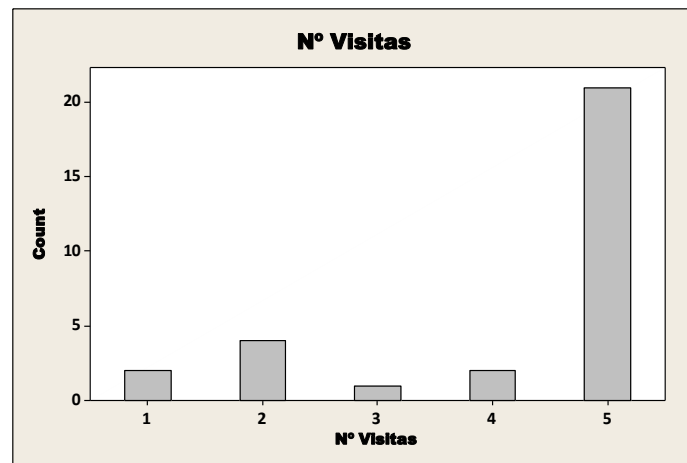


Figura 4. Número de visitas.

De los 30 pacientes que iniciaron el estudio, sólo 21 de ellos completaron el seguimiento hasta la semana 12.

En las Tablas 2, 3 y Figura 5, se puede observar la evolución de las diferentes lesiones acnéicas y de las reacciones adversas al tratamiento durante los diferentes periodos de seguimiento.

		ALL (n=30)						ALL (n=30)			
		Week	n	Mean	St.Dev.			Week	n	Mean	St.Dev.
<b>Papules</b>	<b>0</b>	30	22,33	23,88	<b>Cysts</b>	<b>0</b>	30	1,33	2,60		
	<b>2</b>	28	16,82	16,09		<b>2</b>	28	1,25	2,59		
	<b>4</b>	24	10,50	9,73		<b>4</b>	24	1,04	2,07		
	<b>8</b>	23	7,30	6,44		<b>8</b>	23	1,09	2,11		
	<b>12</b>	21	3,90	3,03		<b>12</b>	21	1,19	2,18		
<b>Pustules</b>	<b>0</b>	30	7,60	8,91	<b>Retentional Lesions</b>	<b>0</b>	30	7,53	8,40		
	<b>2</b>	28	4,57	6,94		<b>2</b>	28	6,61	6,00		
	<b>4</b>	24	2,63	5,08		<b>4</b>	24	6,08	5,86		
	<b>8</b>	23	0,87	2,45		<b>8</b>	23	5,91	5,53		
	<b>12</b>	21	0,00	0,00		<b>12</b>	21	5,33	4,87		
<b>Nodules</b>	<b>0</b>	30	5,60	7,53	<b>Scars</b>	<b>0</b>	30	3,03	4,98		
	<b>2</b>	28	4,61	6,08		<b>2</b>	28	2,89	4,56		
	<b>4</b>	24	2,58	3,26		<b>4</b>	24	1,96	3,21		
	<b>8</b>	23	1,83	2,41		<b>8</b>	23	2,61	3,65		
	<b>12</b>	21	1,67	2,41		<b>12</b>	21	2,86	4,05		
<b>ppn</b>	<b>0</b>	30	35,53	37,68	<b>crs</b>	<b>0</b>	30	11,90	13,00		
	<b>2</b>	28	26,00	26,44		<b>2</b>	28	10,75	10,73		
	<b>4</b>	24	15,71	16,64		<b>4</b>	24	9,08	9,17		
	<b>8</b>	23	10,00	9,58		<b>8</b>	23	9,61	9,60		
	<b>12</b>	21	5,57	4,62		<b>12</b>	21	9,38	9,35		

Tabla 2. Evolución de las lesiones del acné.

En la Tabla 2, se observa una tendencia decreciente en las lesiones inflamatorias (columna izquierda) con el paso del tiempo. Sin embargo, el número de lesiones no inflamatorias (columna derecha) parece mantenerse prácticamente constante durante el periodo del tratamiento.



		ALL (n=30)					
		Week	n	None	Mild	Moderate	Severe
Erythema	2	28	10 (35,7%)	10	7	1	
	4	24	15 (62,5%)	8	1	0	
	8	23	21 (91,3%)	2	0	0	
	12	21	21 (100%)	0	0	0	
Scaling	2	28	19 (67,9%)	4	4	1	
	4	24	12 (50%)	8	4	0	
	8	23	13 (56,5%)	10	0	0	
	12	21	19 (90,5%)	2	0	0	
Dryness	2	28	16 (57,1%)	8	4	0	
	4	24	11 (45,8%)	8	5	0	
	8	23	14 (60,9%)	8	1	0	
	12	21	17 (81%)	4	0	0	
Pruritus	2	28	14 (50%)	6	6	2	
	4	24	19 (79,2%)	5	0	0	
	8	23	22 (95,7%)	1	0	0	
	12	21	21 (100%)	0	0	0	

Tabla 3. Evolución de las reacciones adversas al tratamiento.

Se puede observar en la Tabla 3 que las categorías *moderate* y *severe* tienen muy pocos casos con valores mayores que 1, por lo que se colapsarán junto a la modalidad *mild* en la nueva categoría ‘presencia de reacciones adversas’, ya que cuando hay pocas observaciones en alguna de las clases no se va a poder ajustar un buen modelo. De este modo, se trabajará con una respuesta binaria: 0 = ausencia y 1 = presencia de reacciones adversas al tratamiento (reategorización).

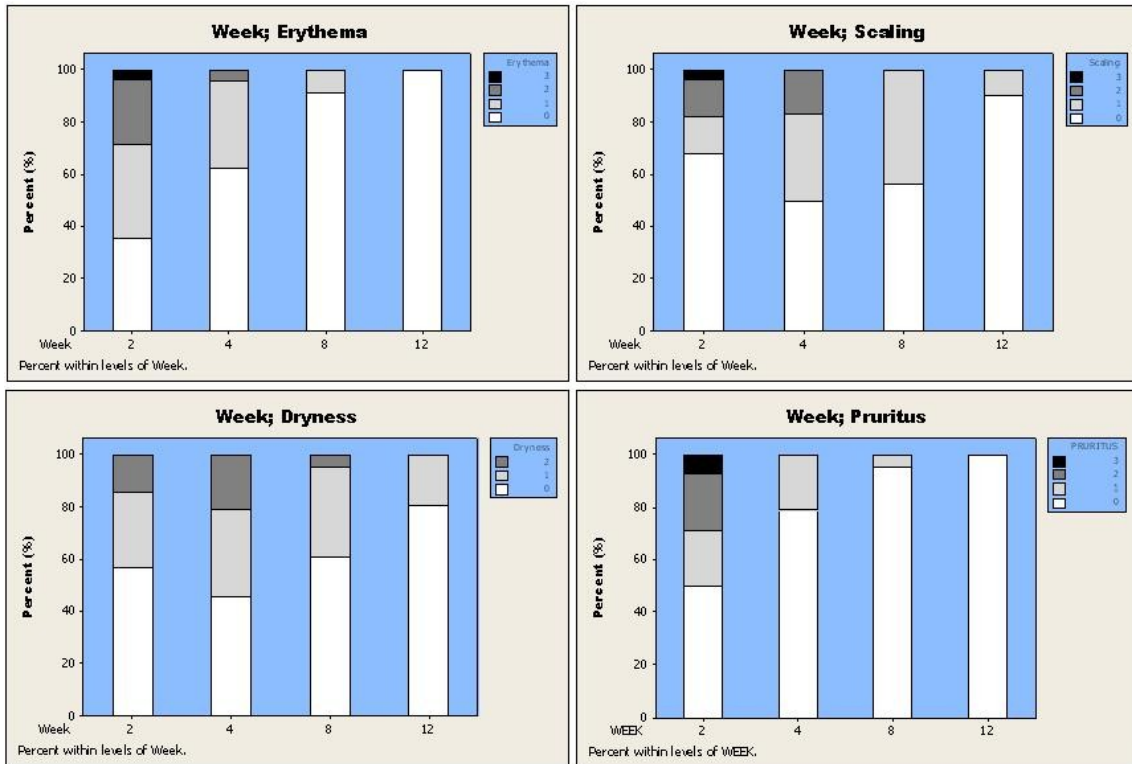


Figura 5. Evolución de las reacciones adversas.

En la Figura 5, se puede apreciar gráficamente la posible existencia de tendencia en las diferentes reacciones adversas.

#### 4.2. Regresión lineal simple: variables cuantitativas

##### Variable *ppn*:

##### Análisis exploratorio

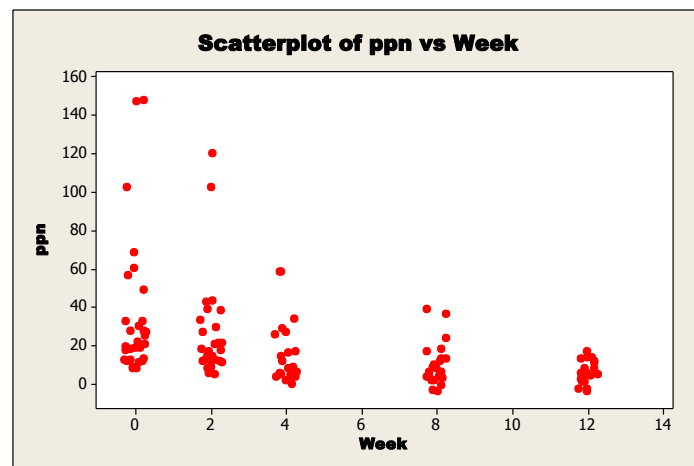


Figura 6. *ppn* vs. *Week*.

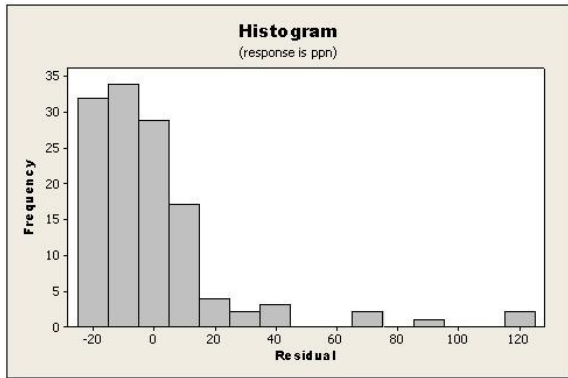


Figura 7. Histograma de los residuos de *ppn*.

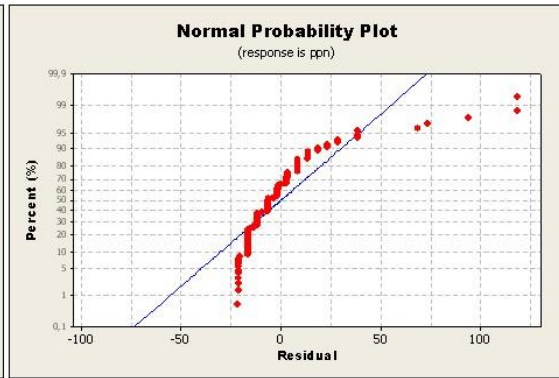


Figura 8. Normal Probability Plot de los residuos de *ppn*.

En la Figura 6, se puede observar que la relación entre las variables *ppn* y *week* no es lineal, y en la Figura 7 y 8 que la distribución es asimétrica, por lo que los residuos no cumplen la premisa de normalidad. A pesar de no ser premisa, es conveniente que la variable respuesta sea Normal para que el modelo presente un mejor ajuste. Por ello, se aconseja una transformación que la haga Normal.

Se parte de un problema donde las premisas de la regresión lineal no son adecuadas.

## Premisas

Para intentar que se cumplan las premisas de linealidad y normalidad de los residuos, se transforma la variable respuesta. La transformación aplicada es el  $\log(ppn+1)$ .

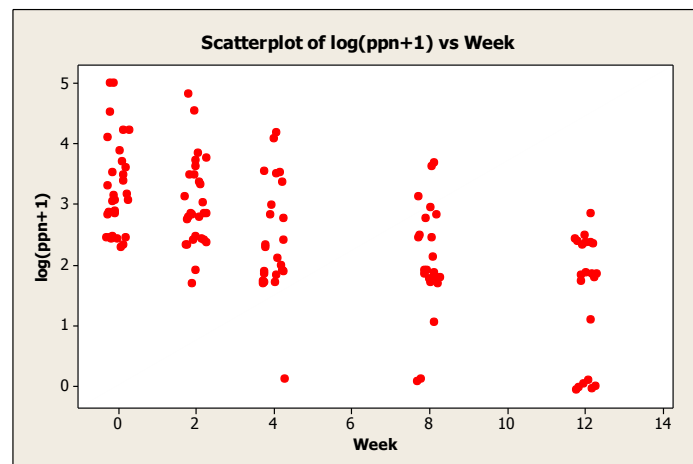


Figura 9.  $\log(ppn+1)$  vs. *Week*.

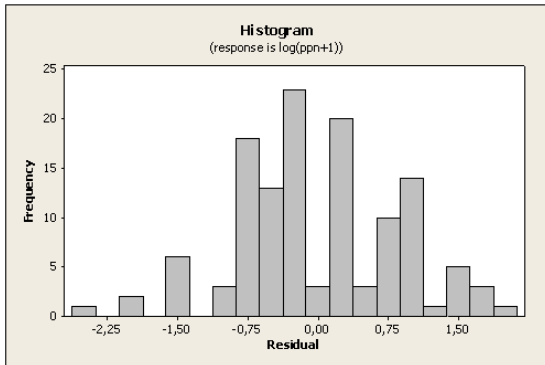


Figura 10. Histograma de los residuos de  $\log(ppn+1)$ .

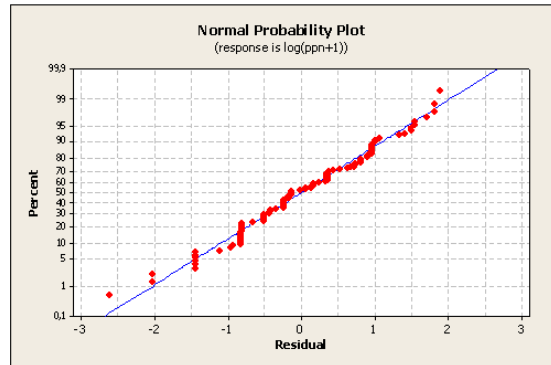


Figura 11. Normal Probability Plot de los residuos de  $\log(ppn+1)$ .

En las Figuras 10 y 11, se observa que el modelo Normal presenta ahora un mejor ajuste. Por ello, se decide seguir trabajando con la variable transformada.

**Modelo obtenido**

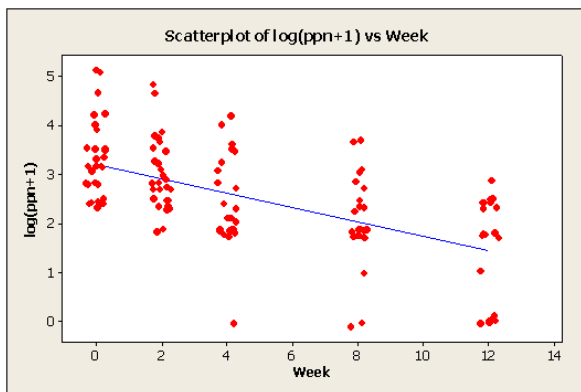


Figura 12. Recta de regresión.

**Regression Analysis: log(ppn+1) versus Week**

The regression equation is  
 $\log(ppn+1) = 3,21 - 0,147 \text{ Week}$

Predictor	Coef	SE Coef	T	P
Constant	3,2063	0,1153	27,81	0,000
Week	-0,14654	0,01832	-8,00	0,000

S = 0,868411    R-Sq = 34,0%    R-Sq(adj) = 33,5%

Tabla 4. Modelo obtenido.

En la Tabla 4, se puede ver que tanto  $b_0$  como  $b_1$  son estadísticamente significativos ( $p$ -valor  $< 0,0001$ ), por lo que el modelo obtenido es:

$$\log(ppn_{week} + 1) = 3,2063 - 0,1465 \cdot week + e \quad , e \sim N(0; 0,8684)$$

$$ppn_{week} = (e^{3,2063} \cdot e^{(-0,14654 \cdot week)}) - 1$$

## Validación

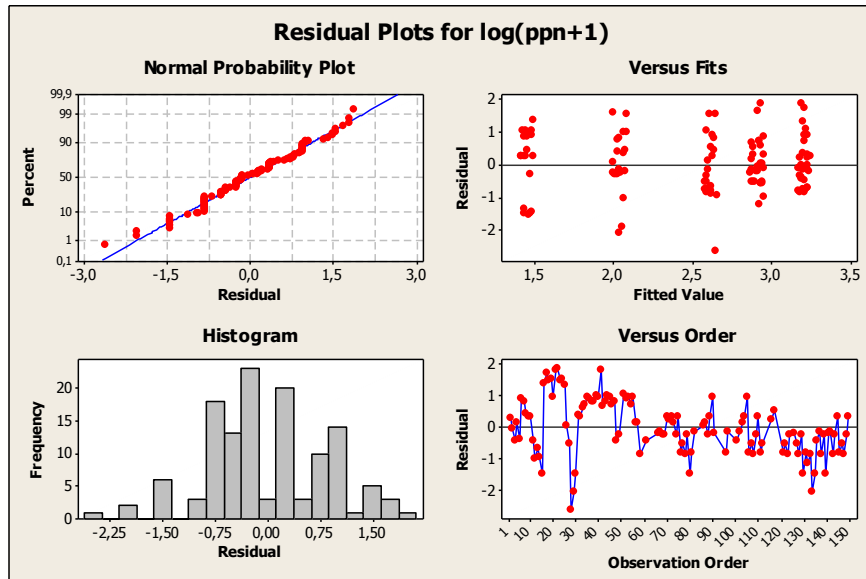


Figura 13. *Four in One* de los residuos.

La bondad de ajuste del modelo se basa en los gráficos de la Figura 13. En el histograma y en el *Normal Probability Plot* se puede ver que la premisa de normalidad de los residuos se cumple.

La premisa de homocedasticidad también se cumple, ya que la variabilidad se mantiene aproximadamente constante a lo largo de las semanas (véase gráfico *Residual Versus Fits*).

En el gráfico *Residual Versus Order*, se aprecian rachas de valores positivos seguidas de rachas de valores negativos, esto indica que no hay aleatoriedad en los residuos. Esto es debido a las medidas repetidas, por lo que la premisa de independencia en los residuos no se cumple.

**Evolución de la variable *ppn* según el modelo obtenido**

Del modelo ajustado, el parámetro de interés es  $b_1 = -0,14654$ . Siguiendo las indicaciones del Anexo IV:

$$\frac{ppn_{week+1} - ppn_{week}}{ppn_{week}} = [(e^{-0,14654 \cdot 1}) - 1] \cdot 100 = -13,63\%, \text{ descenso porcentual semanal de } ppn.$$

Variable	Variación porcentual	IC <sub>95%</sub> L.I.	IC <sub>95%</sub> L.S.
<i>ppn</i>	-13,63	-16,67	-10,47

Tabla 5. Estimación puntual e IC<sub>95%</sub> de la variación porcentual semanal.

**Variable *crs*:**

**Análisis exploratorio**

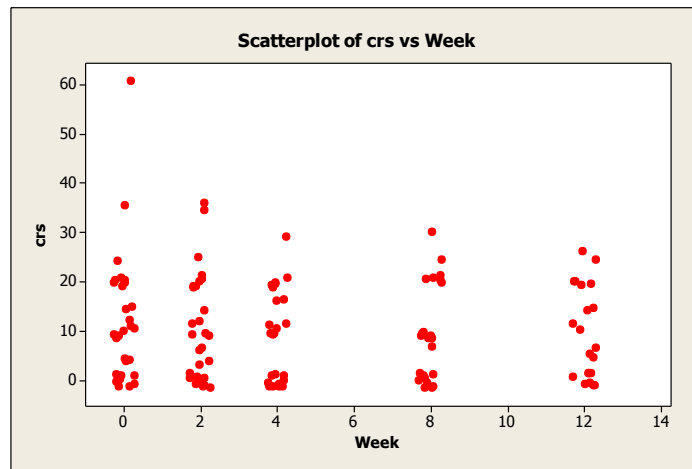


Figura 14. *crs* vs. *Week*.

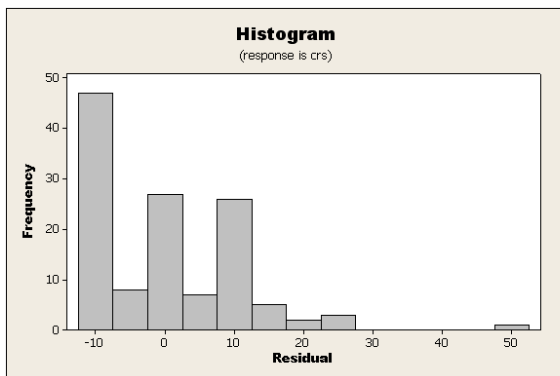


Figura 15. Histograma de los residuos de *crs*.

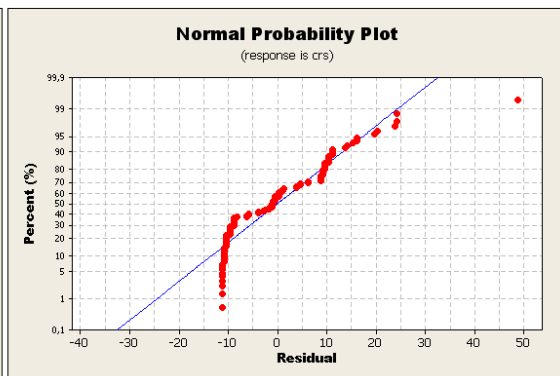


Figura 16. Normal Probability Plot de los residuos de *crs*.

Al igual que en la variable  $ppn$ , las premisas de la regresión lineal no se cumplen de forma razonable (véanse las Figuras 14, 15 y 16).

## Premisas

Como en la variable  $ppn$ , se transforma la variable respuesta:  $\log(crs+1)$ .

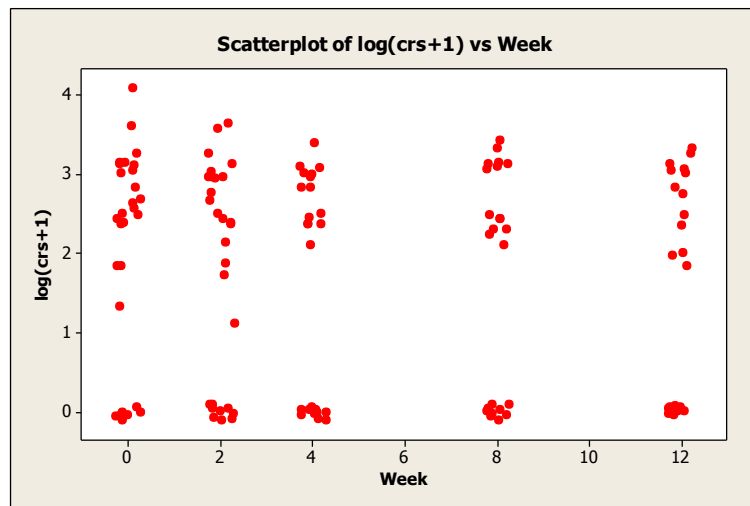


Figura 17.  $\log(crs+1)$  vs. Week.

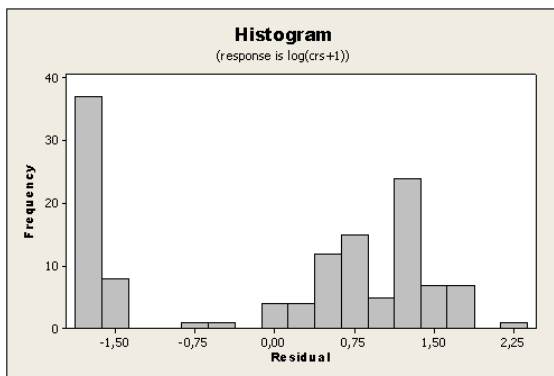


Figura 18. Histograma de los residuos de  $\log(crs+1)$ .

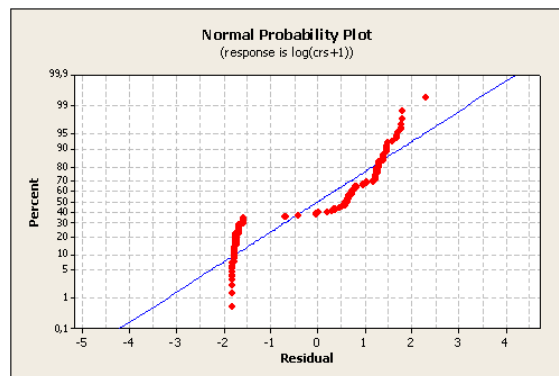


Figura 19. Normal Probability Plot de los residuos de  $\log(crs+1)$ .

En la Figura 18, se aprecia asimetría en la distribución de los residuos, esto es debido a que hay bastantes observaciones en el cero. En este caso, no se observa un mejor ajuste del modelo Normal (Figura 19). Sin embargo, con la variable respuesta transformada, se ha conseguido un comportamiento más lineal (véase Figura 17).

Modelo obtenido

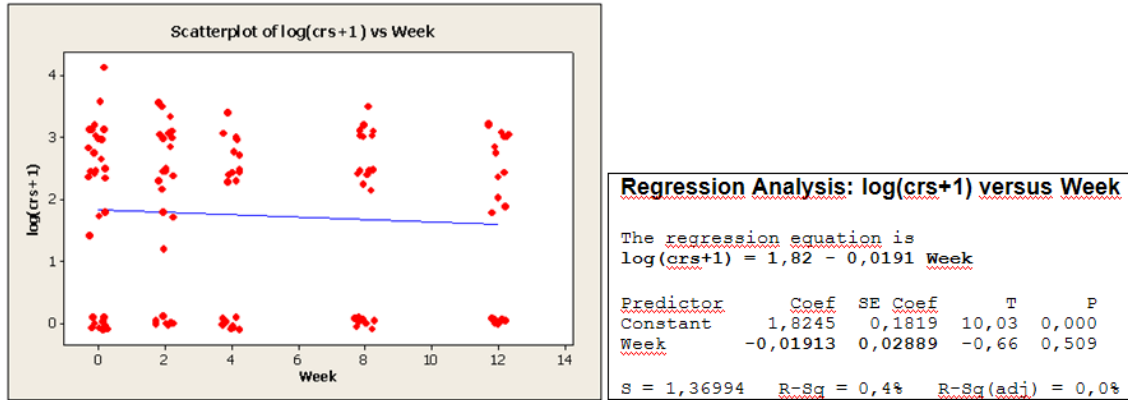


Figura 20. Recta de regresión.

Tabla 6. Modelo obtenido.

En la Tabla 6, se observa que sólo  $b_0$  es estadísticamente significativo (p-valor < 0,0001), por lo que el modelo obtenido es:

$$\log(crs + 1) = 1,8245 + e \quad , e \sim N(0; 1,3699)$$

$$crs = e^{1,8245} - 1$$

Validación

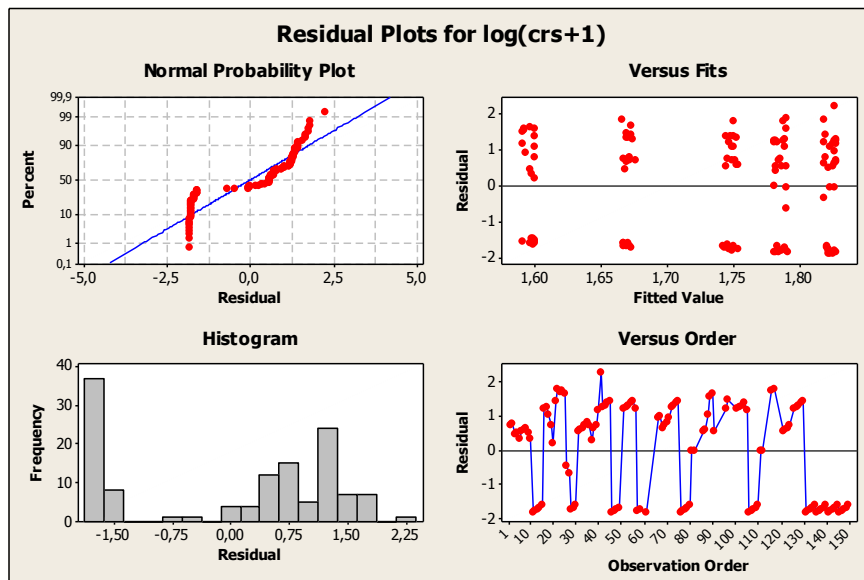


Figura 21. Four in One de los residuos.

En el histograma y en el *Normal Probability Plot* se observa que la premisa de normalidad de los residuos no se cumple.

La premisa de variabilidad constante de los residuos, en este caso se cumple (véase gráfico *Residual Versus Fits*).



En el gráfico *Residual Versus Order*, igual que en la variable *ppn*, no se aprecia aleatoriedad en los residuos, por lo que la premisa de independencia en los residuos no se cumple.

### Evolución de la variable *crs* según el modelo obtenido

Realizando los cálculos del Anexo IV, se observa un descenso semanal estadísticamente no significativo del 1,89% en las lesiones no inflamatorias. La variable semana no es estadísticamente significativa, por lo que *crs* se puede considerar prácticamente constante a lo largo del tiempo.

$$\frac{crs_{\text{week} + 1} - crs_{\text{week}}}{crs_{\text{week}}} = [e^{-0,01913} - 1] \cdot 100 = -1,89\%, \text{ descenso porcentual semanal de } crs \text{ no significativo.}$$

Variable	Variación porcentual	IC <sub>95%</sub> L.I.	IC <sub>95%</sub> L.S.
<i>crs</i>	-1,89	-7,30	3,82

Tabla 7. Estimación puntual e IC<sub>95%</sub> de la variación porcentual semanal.

Se observa que el IC<sub>95%</sub> de la variable *crs* contiene el cero.

### 4.3. Modelos GEE<sup>6</sup>: variables cuantitativas

En nuestro caso, el cociente  $\frac{s^2}{\bar{x}}$  fue mayor a 1, tanto en *ppn* como en *crs*. Esto podría indicar que otra distribución sería más adecuada, pero aun así se decidió trabajar con una Poisson sobredispersa, ya que los modelos GEE lo permiten y como primer contacto con dichos modelos, se consideró que ya era apropiado para los conocimientos adquiridos durante la diplomatura.

<sup>6</sup> En los Anexos V y VI, se encuentran la sintaxis y las salidas de R, respectivamente.

## Variable *ppn*:

### Análisis exploratorio

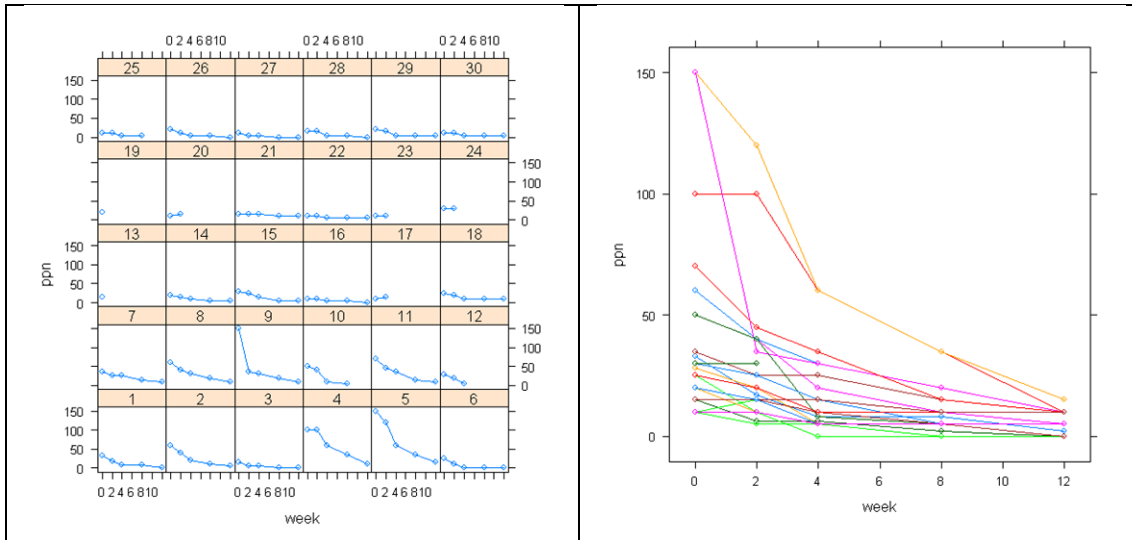


Figura 22. Evolución de las lesiones inflamatorias por paciente: no agregado (izquierda) y agregado (derecha).

En la Figura 22 (izquierda), se observa el comportamiento de las lesiones inflamatorias en cada uno de los pacientes a lo largo de los diferentes periodos de seguimiento. En general, se produce un descenso en el número de lesiones.

La Figura 22 (derecha) facilita la misma información que el gráfico anterior, pero en lugar de mostrar los individuos por separado, los muestra todos juntos en un mismo gráfico. Este plot agregado va bien para hacerse una idea del comportamiento de la media de la variable *ppn* en cada una de las semanas de seguimiento.

### Premisas

La estimación del parámetro de escala es 3,15 (Std.err = 0,6646), corroborándose la sobredispersión.

Para trabajar se ha optado por:

- usar la escala logarítmica para la variable semana.
- añadir como variable explicativa la semana basal, ya que se ha observado que el comportamiento es muy diferente en función del número de la variable

respuesta en la primera semana, es decir, las curvaturas cambian en función de la primera semana.

### Modelo obtenido

```

Call:
geeglm(formula = ppn ~ log(basal) + log(week + 1), family = poisson,
       data = skin, id = subject, corstr = "exc", scale.fix = F)

Coefficients:
              Estimate Std.err      Wald Pr(>|W|)
(Intercept)   0.34287   0.17822    3.701   0.0544 .
log(basal)    0.93718   0.04983  353.755 <2e-16 ***
log(week + 1) -0.64015   0.04815  176.721 <2e-16 ***
---
Estimated Scale Parameters:
              Estimate Std.err
(Intercept)   3.148   0.6646

Estimated Correlation Parameters:
              Estimate Std.err
alpha         0.2612   0.1047

```

Tabla 8. Resultados del ajuste del modelo GEE (*geeglm*).

El test de Wald indica que los coeficientes estimados son estadísticamente significativos, es decir, aportan información al modelo, aunque el *Intercept* se encuentra en el límite de la significación. Por lo tanto, el modelo obtenido es:

$$\log(\mu_{ij}) = 0,34287 + 0,93718 \cdot \log(\text{basal}_i) - 0,64015 \cdot \log(\text{week}_j + 1)$$

$$\mu_{ij} = e^{0,34287} \cdot \text{basal}_i^{0,93718} \cdot (\text{week}_j + 1)^{-0,64015}$$

es decir, el modelo para calcular el valor esperado de las *ppn* de un individuo *i*, conociendo su valor basal, en la semana *j* es:

$$\mu_{ij} = \frac{e^{0,34287} \cdot \text{basal}_i^{0,93718}}{(\text{week}_j + 1)^{0,64015}}$$

**Validación**

A continuación, se valida el modelo obtenido con la función *geeglm* mediante la función *gee*. En este caso, el modelo probabilístico Poisson no es del todo adecuado ( $\hat{\phi} = 3,148$ ) y, por lo tanto, hay que trabajar con Poisson sobredispersa. Esta sobredispersión se puede ver en la Figura 23. En la Tabla 9, se observa que ambos *S.E.* (*Naive* y *Robusto*) son muy parecidos, permiten la misma conclusión práctica, y por lo tanto, el modelo es razonable.

Coefficients:					
	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	0.3448	<b>0.21014</b>	1.641	<b>0.17405</b>	1.981
log(basal)	0.9367	<b>0.05203</b>	18.002	<b>0.04898</b>	19.125
log(week + 1)	-0.6402	<b>0.03658</b>	-17.502	<b>0.04785</b>	-13.380

Tabla 9. Resultados del ajuste del modelo GEE (*gee*).

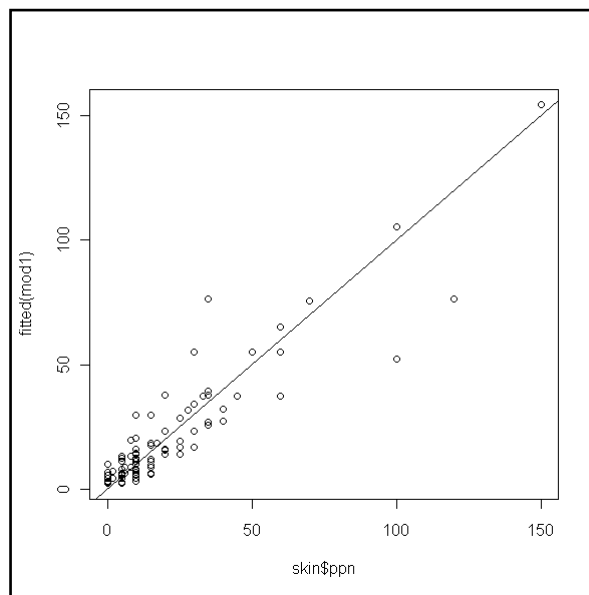


Figura 23. *ppn* observadas vs. *ppn* previstas.

**Evolución de la variable *ppn* según el modelo obtenido**

En la Figura 24, se muestra la curva poblacional de la evolución de las lesiones inflamatorias en el tiempo. Se toma como valor basal la mediana de la semana 0.

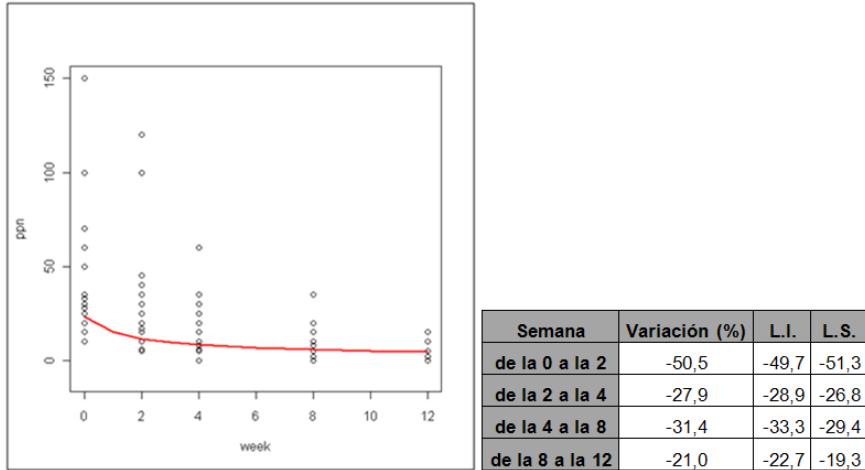


Figura 24. Curva poblacional de la evolución de *ppn*.  
 Tabla 10. Variación porcentual en periodo a periodo.

Tanto en la Tabla 10 como en la Figura 24, se observa la variación porcentual entre periodos de seguimiento. Se observa que decrece muy rápidamente al principio y que se va estabilizando a medida que transcurren las semanas. Dicho descenso se calcula de la siguiente manera:

$$\frac{ppn_{week_j}}{ppn_{week_i}} - 1 = \left( \frac{week_i + 1}{week_j + 1} \right)^{0,64015} - 1$$

donde *i* sería la semana de partida y *j* la semana objetivo. Aplicando la fórmula, obtenemos que de la semana basal a la semana 12 se produce un descenso de las lesiones de un -80,6% (IC<sub>95%</sub> [-81,6% ; -79,9%]).

La estimación del parámetro de correlación es 0,2612 (*Std.err* = 0,1047), siendo estadísticamente significativa, es decir, asumiendo que la correlación es la misma entre todas las visitas, ésta no es despreciable.

Se concluye que, a medida que avanzan las semanas, disminuye el número de lesiones inflamatorias.

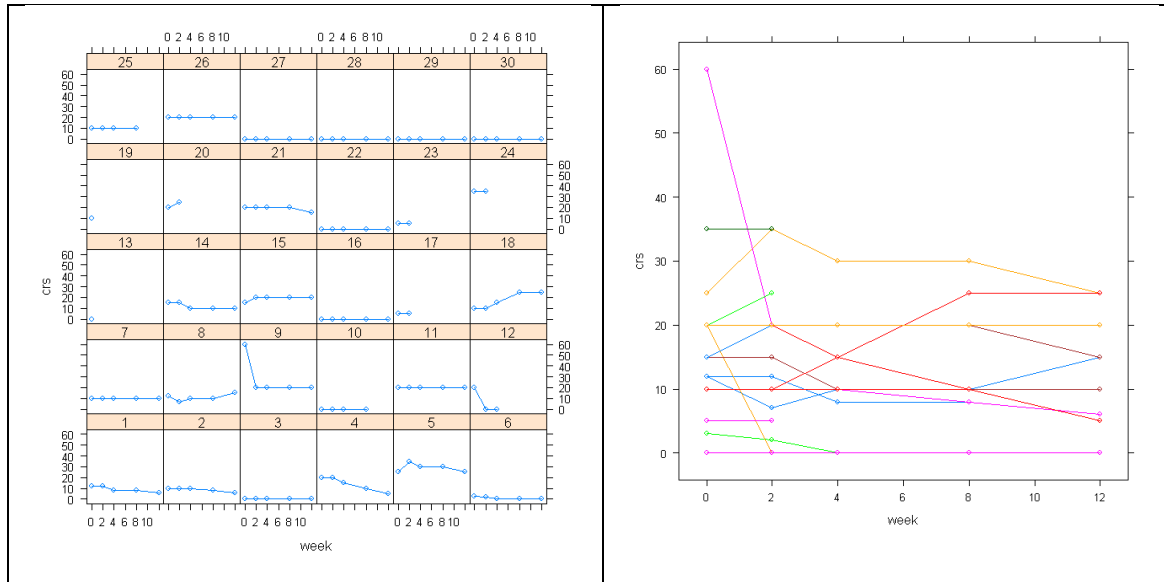
**Variable crs:****Análisis exploratorio**

Figura 25. Evolución de las lesiones inflamatorias por paciente: no agregado (izquierda) y agregado (derecha).

En general, el número de lesiones no inflamatorias se mantiene prácticamente constante durante el tratamiento (véanse Figuras 25).

**Premisas**

La estimación del parámetro de escala es 2,37 (Std.err = 0,866), corroborándose la sobredispersión.

Para trabajar se ha optado por:

- usar la escala logarítmica para la variable semana.
- añadir como variable explicativa la semana basal, ya que se ha observado que el comportamiento es muy diferente en función del número de la variable respuesta en la primera semana, es decir, las curvaturas cambian en función de la primera semana.

## Modelo obtenido

```

Call:
geeglm(formula = crs ~ log(basal1 + 1) + log(week + 1), family = poisson,
       data = skin, id = subject, corstr = "exc", scale.fix = F)

Coefficients:
                Estimate Std.err   Wald Pr(>|W|)
(Intercept)    -0.1693   0.3707   0.21    0.65
log(basal1 + 1)  1.0236   0.1264 65.62  5.6e-16 ***
log(week + 1)  -0.0987   0.0708   1.94    0.16
---
Estimated Scale Parameters:
                Estimate Std.err
(Intercept)     2.37    0.866

Estimated Correlation Parameters:
                Estimate Std.err
alpha           0.612    0.208

```

Tabla 11. Resultados del ajuste del modelo GEE (*geeglm*).

El test de Wald indica que solamente el coeficiente estimado de  $\log(\text{basal}+1)$  es significativo, aporta información al modelo. Por lo tanto, el modelo obtenido es:

$$\log(\mu_{ij}) = 1,0236 \cdot \log(\text{basal}_i + 1)$$

$$\mu_{ij} = (\text{basal}_i + 1)^{1,0236}$$

En este caso, el valor esperado de la variable *crs* sólo depende del valor basal, es decir, no depende de la semana.

## Validación

En este caso, el modelo probabilístico Poisson tampoco es del todo adecuado ( $\hat{\phi} = 2,37$ ) y, por lo tanto, se sigue trabajando con Poisson sobredispersa. Esta sobredispersión se puede ver en la Figura 26. En la Tabla 12, se observa que ambos *S.E.* (*Naive* y *Robusto*) son muy parecidos para el *Intercept* y para el  $\log(\text{basal}+1)$ . En cambio, para  $\log(\text{week}+1)$  el estimador *Naive* indica significación estadística y el estimador *Robusto* no. Se da por válida la no significación del estimador *Robusto*, ya que su *S.E.* coincide con el estimado en la función *geeglm* y es coherente con los resultados obtenidos en la regresión lineal. Por lo tanto, el modelo es razonable.

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.1663	<b>0.3073</b>	-0.541	<b>0.3606</b>	-0.461
log(basall + 1)	1.0226	<b>0.0983</b>	10.401	<b>0.1232</b>	8.301
log(week + 1)	-0.0987	<b>0.0311</b>	-3.167	<b>0.0706</b>	-1.397

Tabla 12. Resultados del ajuste del modelo GEE (*gee*).

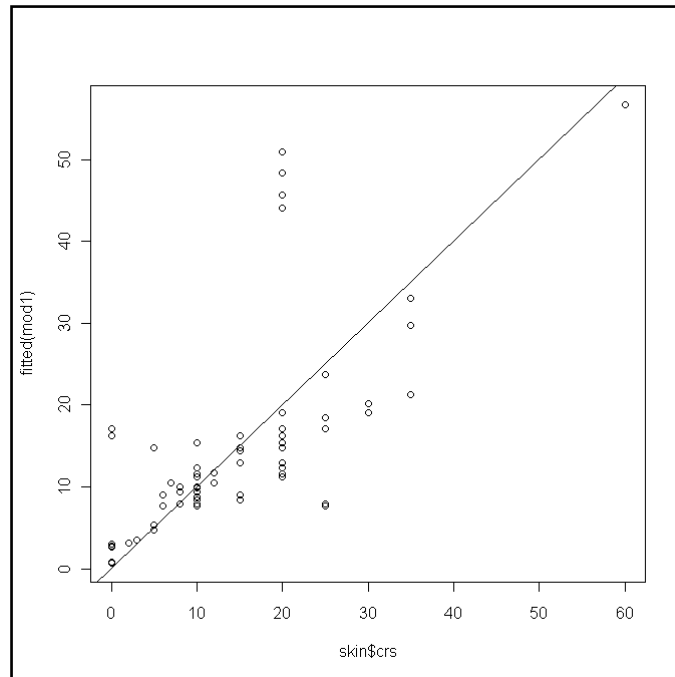
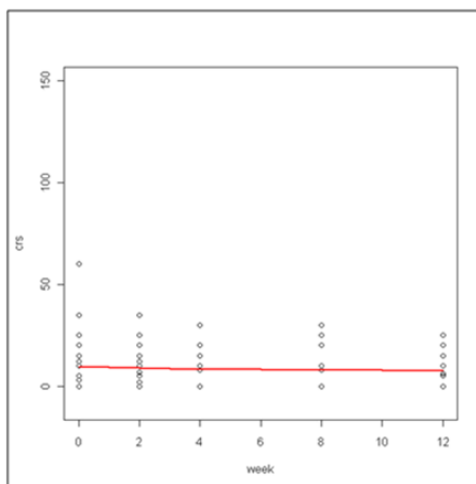


Figura 26. *crs* observadas vs. *crs* previstas.

**Evolución de la variable *crs* según el modelo obtenido**



Semana	Variación (%)	L.I.	L.S.
de la 0 a la 2	-10,3	-5,6	-14,7
de la 2 a la 4	-4,9	-5,2	-4,7
de la 4 a la 8	-5,6	-7,9	-3,3
de la 8 a la 12	-3,6	-5,7	-1,3

Figura 27. Curva poblacional de la evolución de *crs*.

Tabla 13. Variación porcentual periodo a periodo.



Tanto en la Tabla 13, como en la Figura 27, no se aprecia un decrecimiento pronunciado en la variación porcentual a medida que avanzan las semanas. El cálculo de la variación porcentual se ha realizado de la siguiente manera:

$$\frac{ppn_{week_j}}{ppn_{week_i}} - 1 = \left( \frac{week_i + 1}{week_j + 1} \right)^{0,0987} - 1$$

Además, la estimación del parámetro de correlación es 0,612, lo cual indica una alta correlación positiva entre las respuestas de un individuo en dos momentos diferentes. A nivel poblacional, el número de lesiones no inflamatorias se mantiene constante en el tiempo.

#### 4.4. Intervalos de confianza de las variables cualitativas

Las siguientes tablas muestran el porcentaje del cambio por semana de la categoría 0 (“none”), leyéndose todas ellas igual. A modo de ejemplo, se pone la lectura de la primera:

En la semana 2, el porcentaje de pacientes en la categoría 0 (“none”) de *eritema* es 35,71% (10/28). La proporción poblacional de pacientes sin *eritema* está comprendida entre el 18,64 y el 55,93% con un nivel de confianza al 95%.

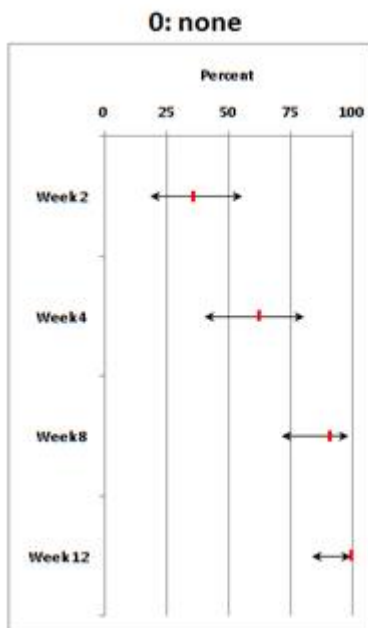


Figura 28. IC<sub>95%</sub> Eritema.

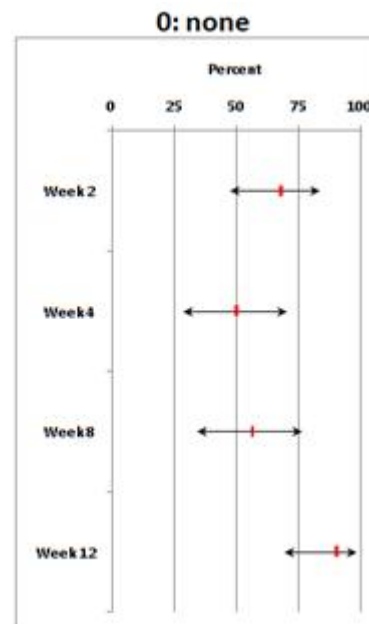


Figura 29. IC<sub>95%</sub> Descamación.

Week	n	none	%	L.L.	U.L.
2	28	10	35,71	18,64	55,93
4	24	15	62,50	40,59	81,20
8	23	21	91,30	71,96	98,93
12	21	21	100	83,89	100

Tabla 14. IC<sub>95%</sub> *Eritema*.

Week	n	none	%	L.L.	U.L.
2	28	19	67,86	47,65	84,12
4	24	12	50,00	29,12	70,88
8	23	13	56,52	34,49	76,81
12	21	19	90,48	69,62	98,83

Tabla 15. IC<sub>95%</sub> *Descamación*.

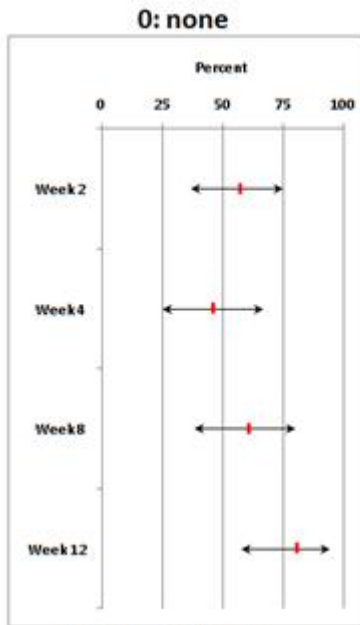


Figura 30. IC<sub>95%</sub> *Sequedad*.

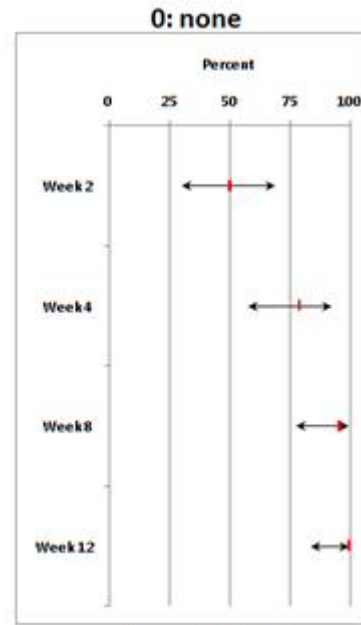


Figura 31. IC<sub>95%</sub> *Prurito*.

Week	n	none	%	L.L.	U.L.
2	28	16	57,14	37,18	75,54
4	24	11	45,83	25,55	67,18
8	23	14	60,87	38,54	80,29
12	21	17	80,95	58,09	94,55

Tabla 16. IC<sub>95%</sub> *Sequedad*.

Week	n	none	%	L.L.	U.L.
2	28	14	50,00	30,65	69,35
4	24	19	79,17	57,85	92,87
8	23	22	95,65	78,05	99,89
12	21	21	100	83,89	100

Tabla 17. IC<sub>95%</sub> *Prurito*.

Se observa que el porcentaje de individuos sin *eritema* (Figura 28 y Tabla 14) aumenta a medida que transcurren las semanas. Esta reacción adversa tiende a desaparecer. En la semana 8 el IC<sub>95%</sub> es [71,96%; 98,93%] y en la semana 12 es [83,89%; 100%]. El porcentaje de no padecer *prurito* se comporta de forma muy similar, pues el IC<sub>95%</sub> es [78,05%; 99,89%] en la semana 8 y [83,89%; 100%] en la semana 12 (Figura 31 y Tabla 17).

El porcentaje de pacientes sin *descamación* no crece de forma clara a medida que avanzan las semanas (Figura 29 y Tabla 15). Este mismo comportamiento se observa en la variable *sequedad* (Figura 30 y Tabla 16).

En general, en la visita final (semana 12), se observa una mejoría en todas las variables (hay un alto porcentaje de individuos en la categoría “none”).

#### 4.5. Modelos GEE: variables cualitativas

En las variables *eritema* y *prurito*, el parámetro de correlación no es estadísticamente significativo con la estructura de correlación *exchangeable*, por lo que llevó a replantear estos dos modelos con la estructura de correlación *independence*. Los resultados obtenidos fueron similares y, por ello, no se han incluido en el presente trabajo. Además, bajo el supuesto de independencia, la pérdida de eficiencia puede ser muy alta, aconsejándose el uso de una estructura que modele la correlación. Los detalles de ajuste están en el Anexo VI.

##### Variable *eritema*

El modelo obtenido es:

$$\text{logit}(p_{ij}) = 1,6702 - 0,5235 \cdot \text{week}_j$$

$$p_{ij} = \frac{e^{1,6702 - 0,5235 \cdot \text{week}_j}}{1 + e^{1,6702 - 0,5235 \cdot \text{week}_j}}$$

En la Figura 32 y en la Tabla 18, se muestran las probabilidades marginales estimadas, es decir, la probabilidad de que un individuo presente *eritema* en un instante determinado.

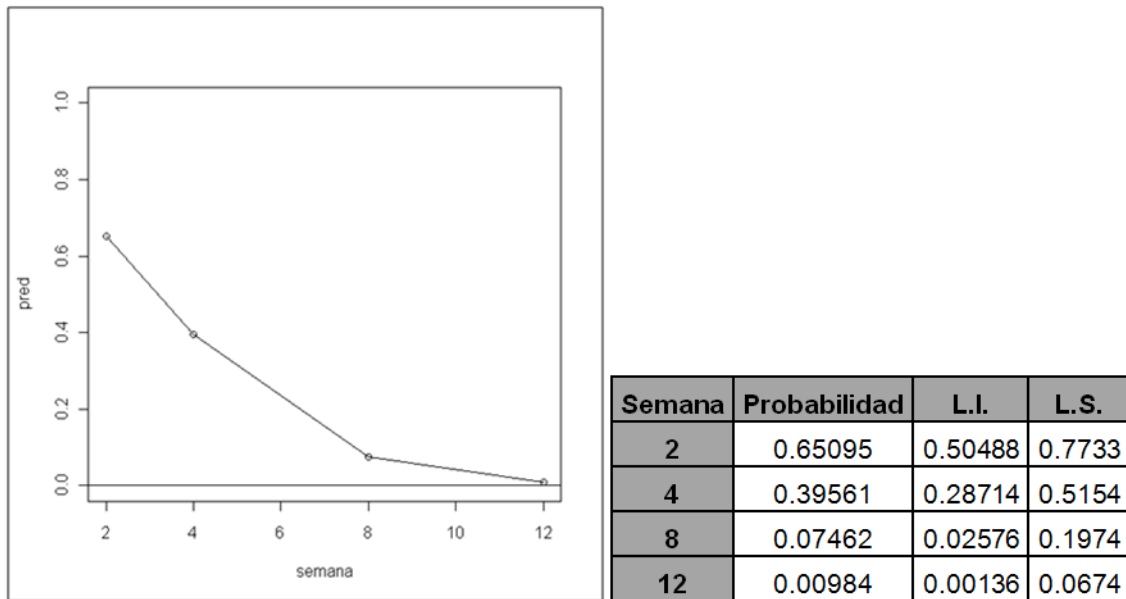


Figura 32 y Tabla 18. Probabilidades predichas por el modelo.

Se observa que, en general, la probabilidad de tener *eritema* disminuye de forma muy rápida, ya que se produce desde las primeras semanas. Este descenso tan marcado se pone de manifiesto en el modelo a través de la significación estadística de la covariable semana ( $p - valor < 0,0001$ ). Esta reacción adversa casi acaba por desapareciendo con el paso del tiempo. De hecho, en la semana 8 la probabilidad de padecer *eritema*, desde el punto de vista poblacional, es del 7,5% (IC<sub>95%</sub> [2,6% ; 19,7%]) y en la semana 12 es prácticamente del 0% (IC<sub>95%</sub> [0,1% ; 6,7%]).

### Variable descamación

El modelo obtenido es:

$$p_{ij} = \frac{1}{1 + e^{0.1061 \cdot w_j}}$$

y las probabilidades poblacionales estimadas se pueden ver en la Figura 33 y en la Tabla 19.

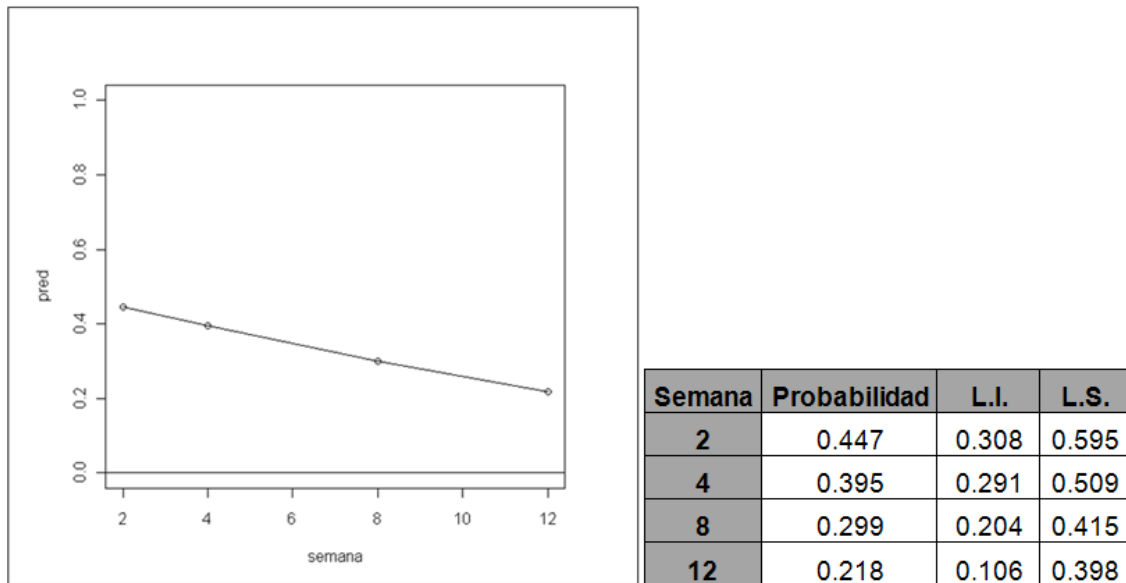


Figura 33 y Tabla 19. Probabilidades predichas por el modelo.

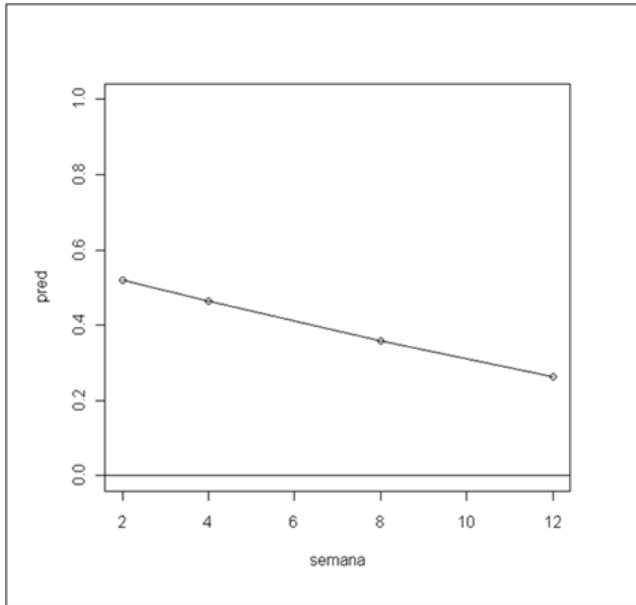
En general, la probabilidad de que se produzca *descamación* no decrece de forma clara a medida que transcurren las semanas.

### **Variable sequedad**

El modelo obtenido es:

$$p_{ij} = \frac{1}{1 + e^{0.1109 \cdot w_j}}$$

y las probabilidades poblacionales estimadas se pueden ver en la Figura 34 y en la Tabla 20.



Semana	Probabilidad	L.I.	L.S.
2	0.520	0.376	0.662
4	0.465	0.355	0.579
8	0.358	0.255	0.476
12	0.264	0.136	0.449

Figura 34 y Tabla 20. Probabilidades predichas por el modelo.

La variable *sequedad* se comporta de forma muy similar a la variable *descamación*.

Observación: parece haber diferencias entre los S.E. proporcionados por la función *gee* (Tabla 21), pero como los z valores, de las funciones *geeglm* y *gee*, son equivalentes en cuanto a la significación estadística, se da por razonable el presente modelo.

Naive S.E.	Robust S.E.
0.3820	0.4392
0.0445	0.0536

Tabla 21. Estimaciones de los S.E.

### Variable *prurito*

El modelo obtenido es:

$$p_{ij} = \frac{1}{1 + e^{0.571 \cdot w_j}}$$

y las probabilidades poblacionales estimadas se pueden ver en la Figura 35 y en la Tabla 21.

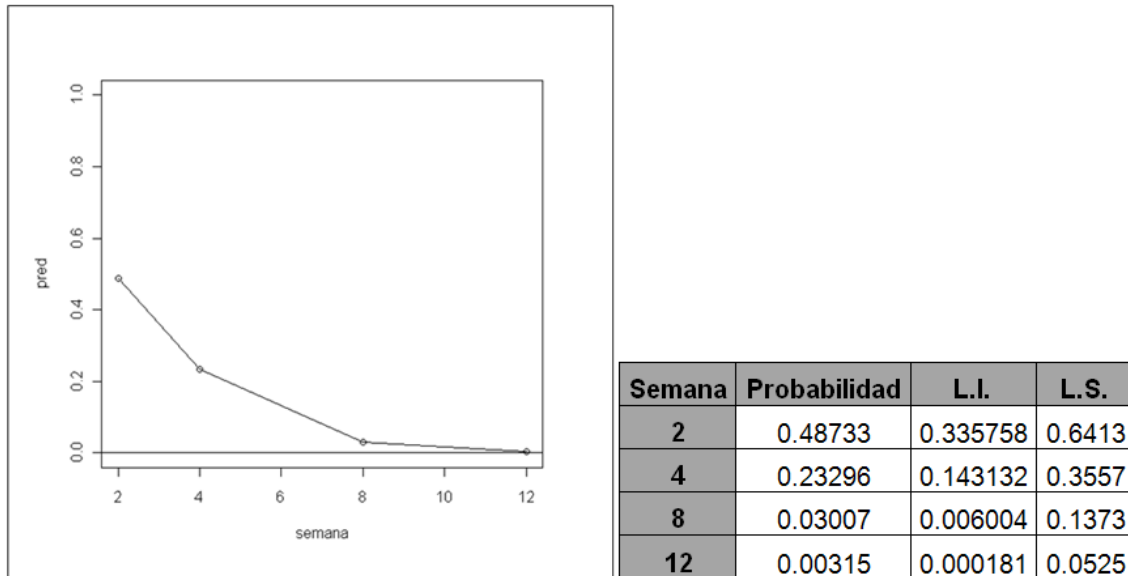


Figura 35 y Tabla 21. Probabilidades predichas por el modelo.

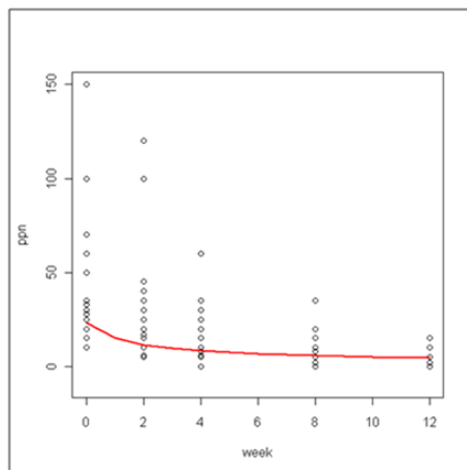
Se aprecia que, en general, la probabilidad de tener *prurito* disminuye en los individuos a medida que transcurre el tiempo (de una probabilidad del 48,7% (IC<sub>95%</sub> [33,6% ; 64,1%]) en la semana 2, se pasa a una del 0,3% (IC<sub>95%</sub> [0,02% ; 5,3%]) en la semana 12).

Esta variable se comporta de forma similar a la variable *eritema*.

## 5. CONCLUSIONES

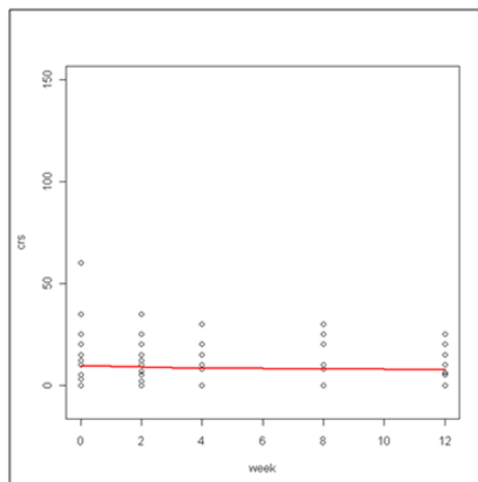
### 5.1. Conclusiones aplicadas

Con el modelo lineal, al imponer un descenso constante en todas las semanas, este se estima en un -13,63% (IC<sub>95%</sub> [-16,67% ; -10,47%]) en las lesiones inflamatorias (*pápulas, pústulas y nódulos*). Al liberar esta premisa con el modelo GEE, el descenso se cuantifica en:



Semana	Variación (%)	L.I.	L.S.
de la 0 a la 2	-50,5	-49,7	-51,3
de la 2 a la 4	-27,9	-28,9	-26,8
de la 4 a la 8	-31,4	-33,3	-29,4
de la 8 a la 12	-21,0	-22,7	-19,3

Sin embargo, en el número de lesiones no inflamatorias (*quistes, lesiones residuales y cicatrices*) no se ha observado un descenso semanal estadísticamente significativo (-1,89%, IC<sub>95%</sub> [-7,30% ; 3,82%]), según el modelo lineal. De acuerdo con el modelo GEE, tampoco se aprecia un decrecimiento:



Semana	Variación (%)	L.I.	L.S.
de la 0 a la 2	-10,3	-5,6	-14,7
de la 2 a la 4	-4,9	-5,2	-4,7
de la 4 a la 8	-5,6	-7,9	-3,3
de la 8 a la 12	-3,6	-5,7	-1,3



Referente a la tolerabilidad al medicamento, y según la metodología para el cálculo del intervalo de confianza basada en el método exacto de la Distribución Binomial, se ha observado que el porcentaje de individuos sin *eritema* aumenta a medida que transcurren las semanas. Es decir, esta reacción adversa tiende a desaparecer: en la semana 8, el porcentaje es del 91,30% (IC<sub>95%</sub> [71,96% ; 98,93%]) y en la semana 12 es del 100% (IC<sub>95%</sub> [83,89% ; 100%]). El porcentaje de no padecer *prurito* se comporta de forma muy similar, pues el porcentaje es del 95,65% (IC<sub>95%</sub> [78,05% ; 99,89%]) en la semana 8 y del 100% (IC<sub>95%</sub> [83,89% ; 100%]) en la semana 12. En cambio, el porcentaje de pacientes sin *descamación* no crece de forma clara a medida que avanzan las semanas. Este mismo comportamiento se observa en la reacción adversa *sequedad*, pudiéndose observar con más detalle en las siguientes tablas:

Week	n	none	%	L.L.	U.L.
2	28	19	67,86	47,65	84,12
4	24	12	50,00	29,12	70,88
8	23	13	56,52	34,49	76,81
12	21	19	90,48	69,62	98,83

*Descamación.*

Week	n	none	%	L.L.	U.L.
2	28	16	57,14	37,18	75,54
4	24	11	45,83	25,55	67,18
8	23	14	60,87	38,54	80,29
12	21	17	80,95	58,09	94,55

*Sequedad.*

En general, en la visita final (semana 12), se observa que el porcentaje de individuos sin estas cuatro reacciones adversas aumenta de forma importante.

Según el modelo GEE, la probabilidad de tener *eritema* y *prurito* disminuye en los individuos a medida que transcurre el tiempo: se pasa de tener una probabilidad del 65,1% (IC<sub>95%</sub> [50,5% ; 77,3%]) y del 48,7% (IC<sub>95%</sub> [33,6% ; 64,1%]), respectivamente, en la semana 2, a tener 0,9% (IC<sub>95%</sub> [0,1% ; 6,7%]) y 0,3% (IC<sub>95%</sub> [0,02% ; 5,3%]) al final del tratamiento. Lo cual significa que estas reacciones adversas claramente decrecen de forma estadísticamente significativa conforme avanza el tratamiento. Sin embargo, la probabilidad de que se produzca *descamación* y/o *sequedad* no decrece de forma tan marcada: se pasa de tener probabilidades del 44,7% (IC<sub>95%</sub> [30,8% ; 59,5%]) y del 52% (IC<sub>95%</sub> [37,6% ; 66,2%]) en la semana 2, a 21,8% (IC<sub>95%</sub> [10,6% ; 39,8%]) y 26,4% (IC<sub>95%</sub> [13,6% ; 44,9%]) en la semana 12, respectivamente.

## 5.2. Conclusiones de los métodos estadísticos utilizados

Los modelos GEE permiten modelar el comportamiento global de la población. Este método no ignora la correlación que existe entre respuestas de un mismo individuo, permitiendo incorporarla en el modelo. De esta manera, se obtiene una mejor estimación de los parámetros y, en particular, de los errores estándar. Para dicha estimación, únicamente necesitamos una función “link”, una función de la varianza y una estructura de correlación. Además, es interesante considerar los modelos GEE porque admiten la posibilidad de que la variable respuesta siga una distribución diferente de la Gaussiana. En nuestro caso, los datos analizados son:

- en las variables cuantitativas, un recuento de lesiones, por lo que, como se nos ha enseñado, puede tener mucho sentido que sigan una Distribución Poisson.
- en las variables cualitativas, valores binarios que siguen una Distribución Binomial.

Cabe remarcar que hay un desequilibrio intencionado en este trabajo: la regresión lineal la planteamos lo más simple posible, no aplicamos la transformación logarítmica a la variable *week* ni tampoco incluimos el basal como variable explicativa. En cambio, en los modelos GEE sí. El motivo fue que los modelos lineales no cumplían las premisas de forma razonable, por lo que decidimos no malgastar tiempo en hacer unos modelos muy complicados, ya que sabíamos que teníamos la alternativa del método GEE y sería mejor dedicar un mayor esfuerzo a estos últimos modelos, desconocidos hasta entonces para nosotros. Por lo tanto, si bien las dos metodologías utilizadas no son directamente comparables, lo que queríamos era ilustrar el hecho de que los GEE de alguna manera arreglan el problema de la correlación.

REGRESIÓN LINEAL	PROS	<ul style="list-style-type: none"> <li>- Es simple de calcular.</li> <li>- Los IC se calculan de forma sencilla.</li> <li>- Las predicciones se pueden calcular bajo supuestos muy simples.</li> <li>- La transformación, en este caso la logarítmica, ya linealiza los residuos.</li> <li>- La <math>R^2</math> nos da información de la relación que existe.</li> <li>- Hay mucho software para poder ajustarlo.</li> </ul>
	CONTRAS	<ul style="list-style-type: none"> <li>- No tienen en cuenta la correlación entre las observaciones de un mismo individuo.</li> <li>- Pueden infraestimar los errores estándar.</li> <li>- No tienen en cuenta que lo que estamos trabajando son recuentos.</li> </ul>

MODELOS GEE	PROS	<ul style="list-style-type: none"> <li>- Incluye en el modelo la correlación entre las observaciones intra-individuo.</li> <li>- Permite trabajar con variables respuesta no Gaussianas, es decir, puede incluir respuestas que sigan alguna distribución de la familia exponencial.</li> <li>- Son bastante flexibles, ya que para hacer una buena estimación de los parámetros del modelo sólo necesitan una transformación o función "link", una función de la varianza y una estructura de correlación.</li> <li>- Aun considerando una estructura de correlación incorrecta, las estimaciones de <math>\beta</math> son consistentes.</li> </ul>
	CONTRAS	<ul style="list-style-type: none"> <li>- Precisa especificar una <i>working correlation matrix</i>.</li> <li>- Si se trabaja con la Distribución Poisson, hay que vigilar la infra/sobre dispersión para evitar la infra/sobre estimación de los errores estándar.</li> <li>- La robustez de los resultados se basa en ver si dos errores estándar son similares.</li> <li>- Se necesita un software más especializado.</li> </ul>

### 5.3. Posibles extensiones

Para el conjunto de datos analizado, una extensión de los modelos GEE la podríamos encontrar en los Modelos Lineales Mixtos Generalizados (*GLMM: Generalized Linear Mixed Models*). Los modelos GLMM proporcionan una mejor descripción de los datos, pues caracterizan a cada individuo según los factores o efectos que pueden comportarse de manera distinta para cada individuo, así ya es más preciso particularizarlos. Es decir, estos modelos no sólo dan resultados a nivel poblacional, sino que además dan resultados por individuo. Esta extensión no la hemos podido aplicar porque no disponemos de los conocimientos básicos necesarios para poder abordarla, siendo estos conocimientos más complejos que los necesarios para los modelos GEE.

Respecto al cálculo del intervalo de confianza para el parámetro de la Distribución Binomial, podríamos completarlo con un test de tendencia, es decir, con una regresión en la cual la variable respuesta fuera politómica (datos ordinales de varias categorías) y así mantendríamos la información que hubiera en los diferentes niveles de la variable en cuestión.

### 5.4. Limitaciones del estudio

La principal limitación es no disponer de un grupo control aleatorizado. En el presente estudio, la comparación antes-después estima el efecto de la intervención junto con cualquier otro alias de la misma. Por ejemplo, si el evaluador conoce si el paciente está en el estado pre o post, cualquier tendencia del mismo estará confundida con el efecto de la intervención.

Otra limitación importante es la pérdida de seguimiento de algunos pacientes, lo que provoca la aparición de valores ausentes. El análisis aquí presentado requiere la premisa no contrastable de MCAR.

## BIBLIOGRAFIA

Ballinger G.A. *Using Generalized Estimating Equations for Longitudinal Data Analysis*. Organizational Research Methods 2004, Vol.7, p. 127-150.

Díaz M.P., Nores M.L. *Construcción de Modelos Gee para Variables con Distribución Simétrica*. Revista de la Sociedad Argentina de Estadística. 2005, Vol. 9, p. 43-63.

Diggle P.J., Liang K.Y., and Zeger S.L. *Analysis of longitudinal data*. Oxford : Clarendon Press, 1994.

Halekoh U., Hojsgaard S., and Yan J. *The R package geepack for Generalized Estimating Equations*. Journal of Statistical Software 2006, Vol. 15, Issue 2.

Hardin, J.W. *Generalized estimating equations*. Chapman & Hall/CRC, cop. 2003.

Hojsgaard S. *Generalized Estimating Equations*. Biometry Research Unit, Danish Institute of Agricultural Sciences, January 2006.

Peña D. *Estadística: modelos y métodos*. Alianza 1986-1991.

[www.biostat.harvard.edu/~horton/geereview.pdf](http://www.biostat.harvard.edu/~horton/geereview.pdf)

[http://en.wikipedia.org/wiki/Binomial\\_proportion\\_confidence\\_interval](http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval)

<http://www.sinacne-roche.com.co/general/preguntas.aspx>

<http://es.wikipedia.org/wiki/Acné>

<http://www.todoacne.com/>

<http://www.tuotromedico.com/temas/acne.htm>

<http://www.nlm.nih.gov/medlineplus/spanish/ency/article/000873.htm>

<http://www.acne0.com/conceptos-basicos-clasificacion-tratamiento-y-cuidados/>

[http://www.niams.nih.gov/Portal\\_En\\_Espanol/Informacion\\_de\\_Salud/Acne/acne\\_ff\\_esp\\_esp.pdf](http://www.niams.nih.gov/Portal_En_Espanol/Informacion_de_Salud/Acne/acne_ff_esp_esp.pdf)

<http://www.vademecum.es/medicamento/farmacologia/D/25748/1/nadixa.html>

<http://www.healthbasis.com/spanish%20health%20illustrated%20encyclopedia/5/00327.htm>

<http://www.fisterra.com/mbe/investiga/index.asp>

## **ANEXOS**

## Anexo I. Escala de gravedad del acné

Existen varios criterios:

Si se toma como punto de referencia el tipo de lesiones:

- Grado I: *comedones* y *pápulas* (acné comedónico).
- Grado II: *pápulas* y *pústulas* superficiales.
- Grado III: *pústulas* profundas y *nódulos*.
- Grado IV: *nódulos* y *quistes* (acné nódulo-quístico).

Si el criterio escogido es la gravedad de las lesiones, pueden distinguirse tres tipos de acné:

- Leve: existen *comedones* y sólo algunas *pápulas* y *pústulas*.
- Moderado: el número de lesiones inflamatorias (*pápulas* y *pústulas*) es mayor y éstas acabarán dejando como secuelas algunas *cicatrices*.
- Grave: la patología acnéica se manifiesta con todo tipo de lesiones, incluyendo además gran cantidad de abscesos nodulares, que acabarán causando importantes y extensas *cicatrices* permanentes.

Un tercer criterio es aquel que clasifica los tipos de acné en función de su forma clínica. Según este criterio pueden definirse cinco grandes tipos de acné:

- Acné vulgar o polimorfo: es aquel en el que aparecen todo tipo de lesiones. Según la gravedad con que se manifiesta, pueden distinguirse dentro de él cuatro tipos o grados.
- Acné superficial: como su propio nombre indica, predominan en él las lesiones superficiales, de curso rápido y que desaparecen sin secuelas cicatriciales. Pueden ser de grado I (*comedones* y *pápulas*) o de grado II (*pápulas* y *pústulas*).
- Acné profundo: las lesiones pueden ser de grado III (*pústulas* y *nódulos*) o de grado IV (*nódulos* y *quistes*). Son lesiones más profundas, de evolución mucho más lenta y que al desaparecer dejan *cicatrices* de una cierta consideración.
- Acné conglobata: es una forma severa de acné que cursa con la formación de grandes abscesos supurantes. Es una patología crónica y que deja como consecuencia la formación de *cicatrices* irregulares importantes.
- Acné fulminante: además de las lesiones descritas hasta el momento, esta modalidad cursa con ulceraciones cutáneas muy dolorosas que pueden acompañarse de fiebre y malestar general.

Además de los tipos descritos, existen lo que se puede definir como “reacciones acneiformes”, es decir, la aparición de lesiones acnéicas o agravamiento de lesiones preexistentes como consecuencia de la exposición a agentes etiológicos concretos. Dentro de éstas, encontramos las siguientes: acné inducido por medicamentos (anticonceptivos orales, fenobarbital, isoniazida, sales de litio, cianocobalamina...), acné cosmético (aplicación de formulaciones con excipientes grasos con potencial comedogénico: aceite de almendras, de coco, de germen de trigo, cera de abejas, lanolina...) y acné de origen profesional (exposición a aceites industriales, petróleo crudo, insecticidas, aceites de cocina...).



## Anexo II. Plantilla recogida de datos

University Hospital Center of Bab-El-Oued  
Department of Dermatology  
Professor Bakar Bouadjar  
Algiers, Algeria

Clinical efficacy and safety of Nadixa (Nadifloxacin) cream 1 %  
in mild to moderate inflammatory forms of acne vulgaris :

An open study in Algeria

Clinical label

**Investigator :** .....

**Date :** .....

**Patient :**

**Name :**                      **First name :**

**Sex :**                      **Age :**

**Disease duration :**

**Clinical examination at baseline :**

Lesions topography :

Face :            yes                       no

Trunk :            yes                       no

Skin lesions types

		Number	Topography
Inflammatory lesions	Papules		
	Pustules		
	Nodules		
Others acne lesions	Kystes		
	Retentional les.		
	Scar lesions		

**Week<sub>2</sub>**  
**Efficacy**

		Number	Topography
Inflammatory lesions	Papules		
	Pustules		
	Nodules		
Others acne lesions	Kystes		
	Retentional les.		
	Scar lesions		

**Tolerance**

	0	1	2	3
Erythema				
Scaling				
Dryness				
Pruritus				

(0 : none, 1 : mild, 2 : moderate, 3 : severe)

**Week<sub>4</sub>**  
**Efficacy**

		Number	Topography
Inflammatory lesions	Papules		
	Pustules		
	Nodules		
Others acne lesions	Kystes		
	Retentional les.		
	Scar lesions		

**Tolerance**

	0	1	2	3
Erythema				
Scaling				
Dryness				
Pruritus				

(0 : none, 1 : mild, 2 : moderate, 3 : severe)

**Week<sub>8</sub>**

**Efficacy**

		Number	Topography
Inflammatory lesions	Papules		
	Pustules		
	Nodules		
Others acne lesions	Kystes		
	Retentional les.		
	Scar lesions		

**Tolerance**

	0	1	2	3
Erythema				
Scaling				
Dryness				
Pruritus				

(0 : none, 1 : mild, 2 : moderate, 3 : severe)

**Week<sub>12</sub>**

**Efficacy**

		Number	Topography
Inflammatory lesions	Papules		
	Pustules		
	Nodules		
Others acne lesions	Kystes		
	Retentional les.		
	Scar lesions		

**Tolerance**

	0	1	2	3
Erythema				
Scaling				
Dryness				
Pruritus				

(0 : none, 1 : mild, 2 : moderate, 3 : severe)

## Anexo III. Principales estructuras de correlación

*Independiente:*

Esta estructura asume que todas las observaciones de un mismo sujeto son independientes

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

*Autoregresiva (AR(1)):*

Las observaciones son tomadas una tras otra en una secuencia de tiempo. Esta estructura asume que la correlación entre las observaciones de un mismo individuo decae de forma exponencial a lo largo del tiempo.  $\text{Corr}(y_{ki}, y_{kj}) = \alpha^{|i-j|}$ ,  $|\alpha| < 1$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \alpha^3 \\ \alpha & 1 & \alpha & \alpha^2 \\ \alpha^2 & \alpha & 1 & \alpha \\ \alpha^3 & \alpha^2 & \alpha & 1 \end{bmatrix}$$

*Desestructurada:*

Esta estructura asume que todas las observaciones de un mismo sujeto están correlacionadas pero no se impone ninguna estructura. El número de parámetros  $\alpha_{ij}$  desconocidos son  $\frac{n(n-1)}{2}$ .  $\text{Corr}(y_{ij}, y_{ik}) = \alpha_{ij}$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & 1 & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & 1 \end{bmatrix}$$

## Anexo IV. Cálculo de la variación porcentual semanal

Ecuación de regresión:

$$\log(\text{ppn} + 1) = \alpha + \beta \cdot \text{Week}$$

Se deshace la transformación:

$$\text{ppn} = [\exp(\alpha) \cdot \exp(\beta \cdot \text{Week})] - 1$$

Cálculos de porcentajes:

$x \cdot (1 + y)$  = aumento porcentual del  $y\%$ ,  $(1 + y) > 1 \rightarrow y > 0$

$x \cdot (1 + y)$  = descenso porcentual del  $y\%$ ,  $(1 + y) < 1 \rightarrow y < 0$

**$\exp(\alpha)$  = ppn + 1, en la semana basal**

**$\exp(\beta \cdot \text{Week}) - 1$  = variación porcentual hasta la semana 'Week'**

**Si Week = 1, se obtiene la variación porcentual semanal**

## Anexo V. Scripts en R

```
skin=read.csv("skin2.csv",header=T,sep=";") # Lectura del fichero
skin$subject=as.factor(skin$patient)      # Se asigna como identificador el paciente
summary(skin)                             # Descriptiva de la BD

library(lattice)      # Llamada a la libreria 'lattice'
library(geepack)     # Llamada a la libreria 'geepack'
library(gee)         # Llamada a la libreria 'gee'

#####
###  NUEVA VARIABLE: SUMA PPN  ###
#####

skin$ppn=apply(skin[,3:5],1,sum)          # Suma las variables pápulas, pústulas y nódulos
offsets=skin[skin$week==0,]              # Selecciona la semana basal de cada paciente
row.names(offsets)=offsets$patient       # Selecciona el número de paciente
skin$basal=offsets[as.character(skin$patient),"ppn"] # Crea la "variable" basal de ppn

tabskin=read.csv2("tabskin.dat",header=T)[,-1] # Se ponen como columnas las semanas
cor(tabskin,use="p")                      # Matrix de varianzas-covarianzas de ppn por semanas
cor(log(tabskin+1),use="p")                # Matrix de varianzas-covarianzas de log(ppn+1) por semanas

m=apply(tabskin,2,mean,na.rm=T)           # Media por semana
v=apply(tabskin,2,var,na.rm=T)            # Varianza por semana
rbind(v,m,v/m,v/(m*m))                   # Comprobación para ver si sigue una Poisson

win.graph()
xyplot(ppn~week|subject,skin,type="o")    # Gráfico evolución de ppn de cada paciente

win.graph()
xyplot(ppn~week,skin,type="o",groups=subject) # Gráfico evolución de ppn por paciente

mod1=geeglm(ppn~log(basal)+log(week+1),id=subject,skin,family=poisson,scale.fix=F,corstr="exc")
# Proporciona el modelo
summary(mod1)
mod2=gee(ppn~log(basal)+log(week+1),id=patient,skin,family=poisson,scale.fix=F,corstr="exchangeable")
# Ayuda a validar el modelo
summary(mod2)

win.graph()
plot(fitted(mod1)~skin$ppn) # Gráfico ppn observadas vs. ppn previstas
abline(c(0,1))

win.graph()
plot(ppn~week,skin,ylim=c(-10,150)) # Gráfico curva poblacional de la evolución de ppn
bas=median(skin[skin$week==0,"basal"])
f=function(x) exp(0.34287)*bas^0.93718/((x+1)^0.64015)
lines(0:12,f(0:12),col=2,lwd=2)
p=100*((1-1/(0:11+2))^0.64015-1)# Evolución de la variable según el modelo, semana a
```

```

semana
p
data.frame(semana=(0:11),variación=p)

ps=100*((c(0+1,2+1,4+1,8+1))/c(2+1,4+1,8+1,12+1))^0.64015-1) # Evolución de la variable
según el modelo, "periodo a periodo"
ps
data.frame(semana=(0:3),variación=ps)

## IC ##

mod3=geeglm(ppn~log(basal)+log(week+1),id=patient,skin,family=poisson,corstr="exch")
pr=predict(mod3,newdata=data.frame(basal=rep(bas,5),week=c(0,2,4,8,12)),se.fit=T,dispersion=3.148)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(0,2,4,8,12),exp(pr$fit),exp(li),exp(ls))

vep=100*((exp(pr$fit[c(2,3,4,5)])/exp(pr$fit[c(1,2,3,4)]))-1)
vli=100*((exp(li[c(2,3,4,5)])/exp(li[c(1,2,3,4)]))-1)
vls=100*((exp(ls[c(2,3,4,5)])/exp(ls[c(1,2,3,4)]))-1)

cbind(c(0,1,2,3),vep, vli, vls)

#####
###  NUEVA VARIABLE: SUMA CRS  ###
#####

skin$crs=apply(skin[,6:8],1,sum) # Suma las variables quistes, residuales, cicatrices
offsets=skin[skin$week==0,] # Selecciona la semana basal de cada paciente
row.names(offsets)=offsets$patient # Selecciona el número de paciente
skin$basal1=offsets[as.character(skin$patient),"crs"] # Crea la "variable" basal de crs

tabskin=read.csv2("tabskin.dat",header=T)[-1] # Se ponen como columnas las semanas
cor(tabskin,use="p") # Matrix de varianzas-covarianzas de crs por semanas
cor(log(tabskin+1),use="p") # Matrix de varianzas-covarianzas de log(crs+1) por semanas

m=apply(tabskin,2,mean,na.rm=T) # Media por semana
v=apply(tabskin,2,var,na.rm=T) # Varianza por semana
rbind(v,m,v/m,v/(m*m)) # Comprobación para ver si sigue una Poisson

win.graph()
xyplot(crs~week|subject,skin,type="o") # Gráfico evolución de crs de cada paciente

win.graph()
xyplot(crs~week,skin,type="o",groups=subject) # Gráfico evolución de crs por paciente

mod1=geeglm(crs~log(basal1+1)+log(week+1),id=subject,skin,family=poisson,scale.fix=F,corstr="exc") # Proporciona el modelo

```



```

summary(mod1)
mod2=gee(crs~log(basall+1)+log(week+1),id=patient,skin,family=poisson,scale.fix=F,corstr
="exchangeable")      # Ayuda a validar el modelo
summary(mod2)

win.graph()
plot(fitted(mod1)~skin$crs)  # Gráfico crs observadas vs. crs previstas
abline(c(0,1))

win.graph()
plot(crs~week,skin,ylim=c(-10,150)) # Gráfico curva poblacional de la evolución de crs
bas=median(skin[skin$week==0,"basall"])
f=function(x) exp(-0.1663)*(bas+1)^1.0226/((x+1)^0.0987)
lines(0:12,f(0:12),col=2,lwd=2)

p=100*((1-1/(0:11+2))^0.0987-1)# Evolución de la variable según el modelo, semana a
semana
p
data.frame(semana=(0:11),variación=p)

ps=100*((c(0+1,2+1,4+1,8+1))/c(2+1,4+1,8+1,12+1))^0.098657122-1)# Evolución de la
variable según el modelo, "periodo a periodo"
ps
data.frame(semana=(0:3),variación=ps)

## IC ##

mod4=geeglm(crs~log(basall+1)+log(week+1),id=patient,skin,family=poisson,corstr="exch")
pr=predict(mod4,newdata=data.frame(basall=rep(bas,5),week=c(0,2,4,8,12)),se.fit=T,disper
sion= 2.37)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(0,2,4,8,12),exp(pr$fit),exp(li),exp(ls))

vep=100*((exp(pr$fit[c(2,3,4,5)])/exp(pr$fit[c(1,2,3,4)]))-1)
vli=100*((exp(li[c(2,3,4,5)])/exp(li[c(1,2,3,4)]))-1)
vls=100*((exp(ls[c(2,3,4,5)])/exp(ls[c(1,2,3,4)]))-1)

cbind(c(0,1,2,3),vep, vli, vls)

```

```

skin$subject=as.factor(skin$patient)
summary(skin)
library(lattice)
library(geepack)
library(gee)

#####
###  ERITEMA  ###
#####

skin$erit=ifelse(as.numeric(skin$erythema)>0,1,0) # Asigna 1 si la observación > 0
skin2=skin
skin2=skin[skin$week!=0,] # Selecciona las semanas diferentes de cero

win.graph()
xyplot(erit~week|subject,skin2,type="o") # Gráfico evolución de eritema de cada paciente

mod1=geeglm(erit~week,id=patient,skin2,family=binomial,corstr="exc") # Proporciona el
modelo
summary(mod1)
mod2=gee(erit~week,id=patient,skin,family=binomial,corstr="exchangeable") # Ayuda a
validar el modelo
summary(mod2)

win.graph()
logitinv=function(x)exp(x)/(1+exp(x)) # Inverso logit
semana=c(2,4,8,12)
pred=logitinv(predict(mod1,newdata=data.frame(week=semana)))
plot(pred~semana,ylim=c(0,1),type="o") # Gráfico probabilidades predichas por el modelo
abline(h=0)
data.frame(week=semana,prob=pred) # Tabla probabilidades predichas por el modelo

## IC ##

mod3=geeglm(erit~week,id=patient,skin2,family=binomial,corstr="exc")
pr=predict(mod3,newdata=data.frame(week=c(2,4,8,12)),se.fit=T,dispersion=0.818)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(2,4,8,12),logitinv(pr$fit),logitinv(li),logitinv(ls))

#####
###  SCALING  ###
#####

skin$scal=ifelse(as.numeric(skin$scaling)>0,1,0) # Asigna 1 si la observación > 0
skin2=skin
skin2=skin[skin$week!=0,] # Selecciona las semanas diferentes de cero

win.graph()
xyplot(scal~week|subject,skin2,type="o")# Gráfico evolución de scaling de cada paciente

```

```

mod1=geeglm(scal~week,id=patient,skin2,family=binomial,corstr="exc") # Proporciona el
modelo
summary(mod1)
mod2=ggee(scal~week,id=patient,skin,family=binomial,corstr="exchangeable") # Ayuda a
validar el modelo
summary(mod2)
win.graph()
logitinv=function(x)exp(x)/(1+exp(x)) # Inverso logit
semana=c(2,4,8,12)
pred=logitinv(predict(mod1,newdata=data.frame(week=semana)))
plot(pred~semana,ylim=c(0,1),type="o") # Gráfico probabilidades predichas por el modelo
abline(h=0)
data.frame(week=semana,prob=pred) # Tabla probabilidades predichas por el modelo

## IC ##

mod3=geeglm(scal~week,id=patient,skin2,family=binomial,corstr="exc")
pr=predict(mod3,newdata=data.frame(week=c(2,4,8,12)),se.fit=T,dispersion=0.98)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(2,4,8,12),logitinv(pr$fit),logitinv(li),logitinv(ls))

#####
### DRYNESS ###
#####

skin$dry=ifelse(as.numeric(skin$dryness)>0,1,0) # Asigna 1 si la observación > 0
skin2=skin
skin2=skin[skin$week!=0,] # Selecciona las semanas diferentes de cero

win.graph()
xyplot(dry~week|subject,skin2,type="o") # Gráfico evolución de dryness de cada paciente

mod1=geeglm(dry~week,id=patient,skin2,family=binomial,corstr="exc")# Proporciona el
modelo
summary(mod1)
mod2=ggee(dry~week,id=patient,skin,family=binomial,corstr="exchangeable") # Ayuda a
validar el modelo
summary(mod2)

win.graph()
logitinv=function(x)exp(x)/(1+exp(x)) # Inverso logit
semana=c(2,4,8,12)
pred=logitinv(predict(mod1,newdata=data.frame(week=semana)))
plot(pred~semana,ylim=c(0,1),type="o") # Gráfico probabilidades predichas por el modelo
abline(h=0)
data.frame(week=semana,prob=pred) # Tabla probabilidades predichas por el modelo

```

```

## IC ##

mod3=geeglm(dry~week,id=patient,skin2,family=binomial,corstr="exc")
pr=predict(mod3,newdata=data.frame(week=c(2,4,8,12)),se.fit=T,dispersion=0.98)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(2,4,8,12),logitinv(pr$fit),logitinv(li),logitinv(ls))

#####
### PRURITUS ###
#####

skin$pru=ifelse(as.numeric(skin$pruritus)>0,1,0) # Asigna 1 si la observación > 0
skin2=skin
skin2=skin[skin$week!=0,] # Selecciona las semanas diferentes de cero

win.graph()
xyplot(pru~week|subject,skin2,type="o") # Gráfico evolución de pruritus de cada paciente

mod1=geeglm(pru~week,id=patient,skin2,family=binomial,corstr="exc")# Proporciona el
modelo
summary(mod1)
mod2=gee(pru~week,id=patient,skin,family=binomial,corstr="exchangeable") # Ayuda a
validar el modelo
summary(mod2)

win.graph()
logitinv=function(x) exp(x)/(1+exp(x)) # Inverso logit
semana=c(2,4,8,12)
pred=logitinv(predict(mod1,newdata=data.frame(week=semana)))
plot(pred~semana,ylim=c(0,1),type="o") # Gráfico probabilidades predichas por el modelo
abline(h=0)
data.frame(week=semana,prob=pred) # Tabla probabilidades predichas por el modelo

## IC ##

mod3=geeglm(pru~week,id=patient,skin2,family=binomial,corstr="exc")
pr=predict(mod3,newdata=data.frame(week=c(2,4,8,12)),se.fit=T,dispersion=0.867)

li=pr$fit-1.96*pr$se.fit
ls=pr$fit+1.96*pr$se.fit
cbind(c(2,4,8,12),logitinv(pr$fit),logitinv(li),logitinv(ls))

```

## Anexo VI. Salidas y gráficos de los modelos GEE en R

### Variable ppn:

```
Call:
geeglm(formula = ppn ~ log(basal) + log(week + 1), family = poisson,
        data = skin, id = subject, corstr = "exc", scale.fix = F)

Coefficients:
                Estimate Std.err   Wald Pr(>|W|)
(Intercept)    0.3429   0.1782    3.7   0.054 .
log(basal)     0.9372   0.0498  353.8 <2e-16 ***
log(week + 1) -0.6402   0.0482  176.7 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:
                Estimate Std.err
(Intercept)    3.15   0.665

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:
                Estimate Std.err
alpha          0.261   0.105
Number of clusters: 30 Maximum cluster size: 5

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
      (Intercept)   log(basal) log(week + 1)
            0.473         0.903         -0.646

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                               Logarithm
Variance to Mean Relation: Poisson
Correlation Structure:               Exchangeable

Call:
gee(formula = ppn ~ log(basal) + log(week + 1), id = patient,
    data = skin, family = poisson, corstr = "exchangeable", scale.fix = F)

Summary of Residuals:
      Min       1Q   Median       3Q      Max
-41.30  -3.78  -1.91    3.83   47.81
```

```

Coefficients:
              Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)    0.345    0.2101    1.64    0.1740    1.98
log(basal)     0.937    0.0520   18.00    0.0490   19.13
log(week + 1) -0.640    0.0366  -17.50    0.0478  -13.38

```

```

Estimated Scale Parameter: 3.22
Number of Iterations: 4

```

```

Working Correlation
      [,1] [,2] [,3] [,4] [,5]
[1,] 1.000 0.258 0.258 0.258 0.258
[2,] 0.258 1.000 0.258 0.258 0.258
[3,] 0.258 0.258 1.000 0.258 0.258
[4,] 0.258 0.258 0.258 1.000 0.258
[5,] 0.258 0.258 0.258 0.258 1.000

```

### Variable crs:

```

Call:
geeglm(formula = crs ~ log(basall + 1) + log(week + 1), family = poisson,
       data = skin, id = subject, corstr = "exc", scale.fix = F)

```

```

Coefficients:
              Estimate Std.err  Wald Pr(>|W|)
(Intercept)   -0.1693  0.3707  0.21    0.65
log(basall + 1)  1.0236  0.1264 65.62 5.6e-16 ***
log(week + 1)  -0.0987  0.0708  1.94    0.16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Estimated Scale Parameters:
              Estimate Std.err
(Intercept)    2.37    0.866

```

```

Correlation: Structure = exchangeable Link = identity

```

```

Estimated Correlation Parameters:
              Estimate Std.err
alpha    0.612    0.208
Number of clusters: 30 Maximum cluster size: 5

```

```

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
      (Intercept) log(basall + 1) log(week + 1)
              -0.0269          0.9747          -0.0941

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

```

```

Model:
  Link:                      Logarithm
  Variance to Mean Relation: Poisson
  Correlation Structure:     Exchangeable

Call:
gee(formula = crs ~ log(basall + 1) + log(week + 1), id = patient,
    data = skin, family = poisson, corstr = "exchangeable", scale.fix = F)

Summary of Residuals:
      Min       1Q   Median       3Q      Max
-30.864  -0.760  -0.291   1.859  17.365

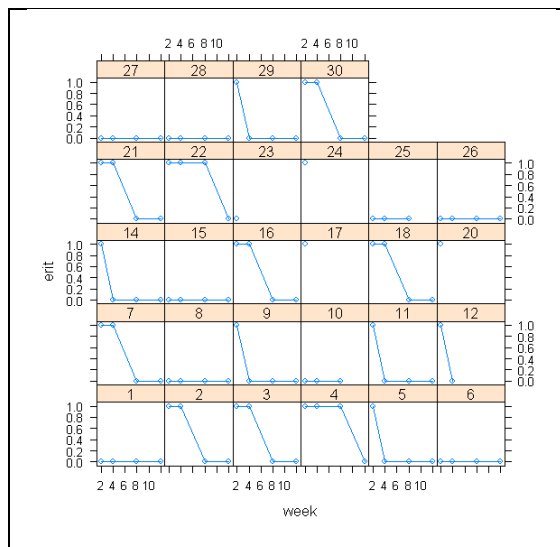
Coefficients:
              Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   -0.1663    0.3073  -0.541    0.3606   -0.461
log(basall + 1)  1.0226    0.0983  10.401    0.1232    8.301
log(week + 1)  -0.0987    0.0311  -3.167    0.0706   -1.397

Estimated Scale Parameter:  2.42
Number of Iterations:  4

Working Correlation
      [,1] [,2] [,3] [,4] [,5]
[1,] 1.000 0.605 0.605 0.605 0.605
[2,] 0.605 1.000 0.605 0.605 0.605
[3,] 0.605 0.605 1.000 0.605 0.605
[4,] 0.605 0.605 0.605 1.000 0.605
[5,] 0.605 0.605 0.605 0.605 1.000

```

**Variable eritema:**



```
Call:
geeglm(formula = erit ~ week, family = binomial, data = skin2,
        id = patient, corstr = "exc")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )	
(Intercept)	1.6702	0.5086	10.79	0.00102	**
week	-0.5235	0.1118	21.92	2.85e-06	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.8185	0.4325

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.2735	0.184

Number of clusters: 28 Maximum cluster size: 4

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)	week
1.6416	-0.5284

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA  
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit  
Variance to Mean Relation: Binomial  
Correlation Structure: Exchangeable

Call:

```
gee(formula = erit ~ week, id = patient, data = skin, family = binomial,
    corstr = "exchangeable")
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.650884	-0.395510	-0.009827	0.349116	0.925427

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	1.6701	0.4288	3.895	0.5019	3.328
week	-0.5236	0.1082	-4.841	0.1079	-4.854

Estimated Scale Parameter: 0.836

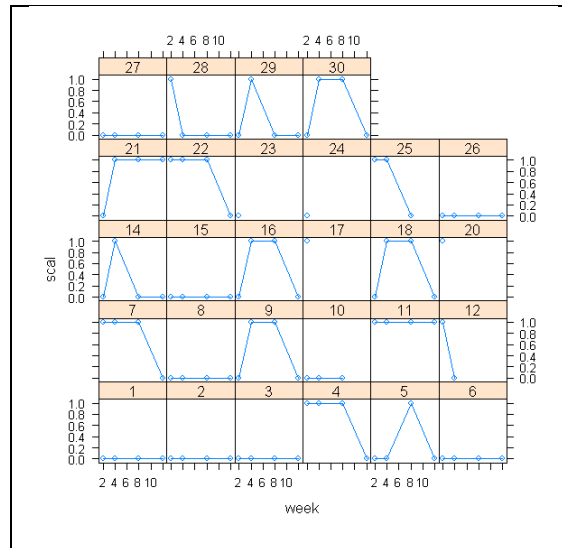


Number of Iterations: 2

Working Correlation

```
      [,1] [,2] [,3] [,4]
[1,] 1.000 0.272 0.272 0.272
[2,] 0.272 1.000 0.272 0.272
[3,] 0.272 0.272 1.000 0.272
[4,] 0.272 0.272 0.272 1.000
```

### Variable descamación:



Call:

```
geeglm(formula = scal ~ week, family = binomial, data = skin2,
        id = patient, corstr = "exc")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	-0.0018	0.3915	2.1e-05	0.996
week	-0.1061	0.0438	5.86	0.016 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.98	0.0766

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.293	0.101

Number of clusters: 28 Maximum cluster size: 4

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27  
running glm to get initial regression estimate

```

(Intercept)      week
      -0.0141    -0.1078

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                Logit
Variance to Mean Relation: Binomial
Correlation Structure: Exchangeable

Call:
gee(formula = scal ~ week, id = patient, data = skin, family = binomial,
    corstr = "exchangeable")

Summary of Residuals:
      Min      1Q  Median      3Q      Max
-0.447 -0.395 -0.218  0.566  0.782

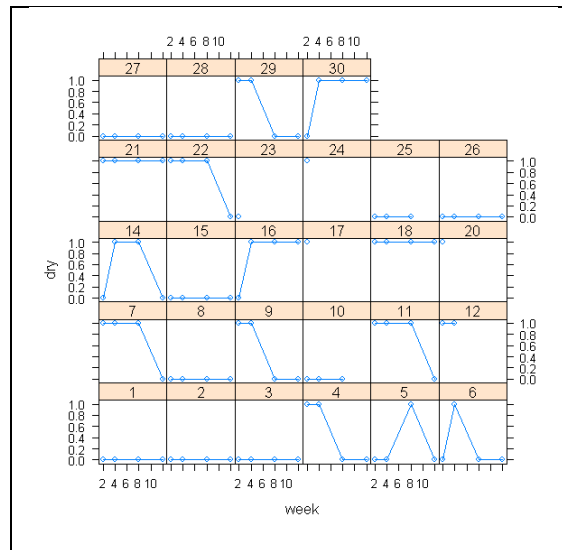
Coefficients:
              Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) -0.00189      0.3905 -0.00485      0.377 -0.00502
week         -0.10609      0.0516 -2.05467      0.042 -2.52558

Estimated Scale Parameter: 1
Number of Iterations: 4

Working Correlation
      [,1] [,2] [,3] [,4]
[1,] 1.000 0.291 0.291 0.291
[2,] 0.291 1.000 0.291 0.291
[3,] 0.291 0.291 1.000 0.291
[4,] 0.291 0.291 0.291 1.000

```

**Variable sequedad:**



```
Call:
geeglm(formula = dry ~ week, family = binomial, data = skin2,
        id = patient, corstr = "exc")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	0.3035	0.4509	0.45	0.501
week	-0.1109	0.0505	4.82	0.028 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.98	0.0222

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	0.439	0.109

Number of clusters: 28 Maximum cluster size: 4

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)	week
0.264	-0.116

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

```

Model:
  Link:                               Logit
  Variance to Mean Relation: Binomial
  Correlation Structure:               Exchangeable

Call:
gee(formula = dry ~ week, id = patient, data = skin, family = binomial,
    corstr = "exchangeable")

Summary of Residuals:
  Min      1Q  Median      3Q      Max
-0.520 -0.465 -0.264  0.535  0.736

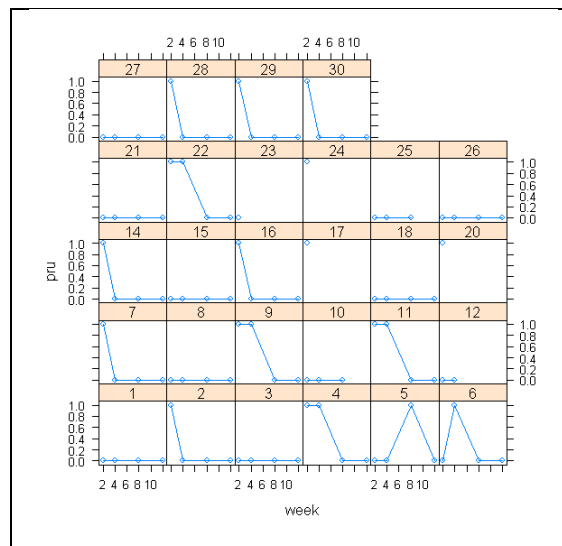
Coefficients:
              Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)   0.303      0.3820  0.794      0.4392    0.69
week          -0.111      0.0445 -2.494      0.0536   -2.07

Estimated Scale Parameter: 1.00
Number of Iterations: 3

Working Correlation
      [,1] [,2] [,3] [,4]
[1,] 1.000 0.436 0.436 0.436
[2,] 0.436 1.000 0.436 0.436
[3,] 0.436 0.436 1.000 0.436
[4,] 0.436 0.436 0.436 1.000

```

**Variable *prurito*:**



```
Call:
geeglm(formula = pru ~ week, family = binomial, data = skin2,
        id = patient, corstr = "exc")
```

Coefficients:

	Estimate	Std.err	Wald	Pr(> W )
(Intercept)	1.090	0.683	2.55	0.1104
week	-0.571	0.182	9.85	0.0017 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:

	Estimate	Std.err
(Intercept)	0.867	1.56

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

	Estimate	Std.err
alpha	-0.00666	0.0539

Number of clusters: 28 Maximum cluster size: 4

---

Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)	week
1.09	-0.57

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

gee S-function, version 4.13 modified 98/01/27 (1998)

Model:

Link: Logit  
Variance to Mean Relation: Binomial  
Correlation Structure: Exchangeable

Call:

```
gee(formula = pru ~ week, id = patient, data = skin, family = binomial,
    corstr = "exchangeable")
```

Summary of Residuals:

Min	1Q	Median	3Q	Max
-0.48733	-0.23296	-0.03007	-0.00315	0.96993

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	1.09	0.576	1.89	0.666	1.64
week	-0.57	0.163	-3.50	0.176	-3.23

Estimated Scale Parameter: 0.886

Number of Iterations: 1

Working Correlation

	[,1]	[,2]	[,3]	[,4]
[1,]	1.00000	-0.00661	-0.00661	-0.00661
[2,]	-0.00661	1.00000	-0.00661	-0.00661
[3,]	-0.00661	-0.00661	1.00000	-0.00661
[4,]	-0.00661	-0.00661	-0.00661	1.00000