Anova dos factores

ÍNDICE

0.1 Descripcion del conjunto de datos

El siguiente conjunto de datos fue diseñado para estudiar los efectos de la vitamina C, y el método de suministro de ésta, en el crecimiento de los dientes de cerdos guineanos. La variable respuesta es la longitud de los odontoblastos (que son células responsables del crecimiento de los dientes). Cada animal recibio una dosis de vitamina C en diferentes niveles (0.5,1, y 2 mg/dia) usando uno de dos métodos de suministro disponibles: zumo de naranga (codificada como OJ) o ácido ascórbico(codificada como VC).

La idea principal del estudio de este experimento era estudiar los diferentes niveles de dosis de vitamina C asi como los métodos de suministro con el objetivo de encontrar la mejor manera para incrementar el crecimiento de los dientes de estos animales.

El conjunto de datos se encuentra en el paquete datasets de R. Se puede descargar en la siguiente página web.

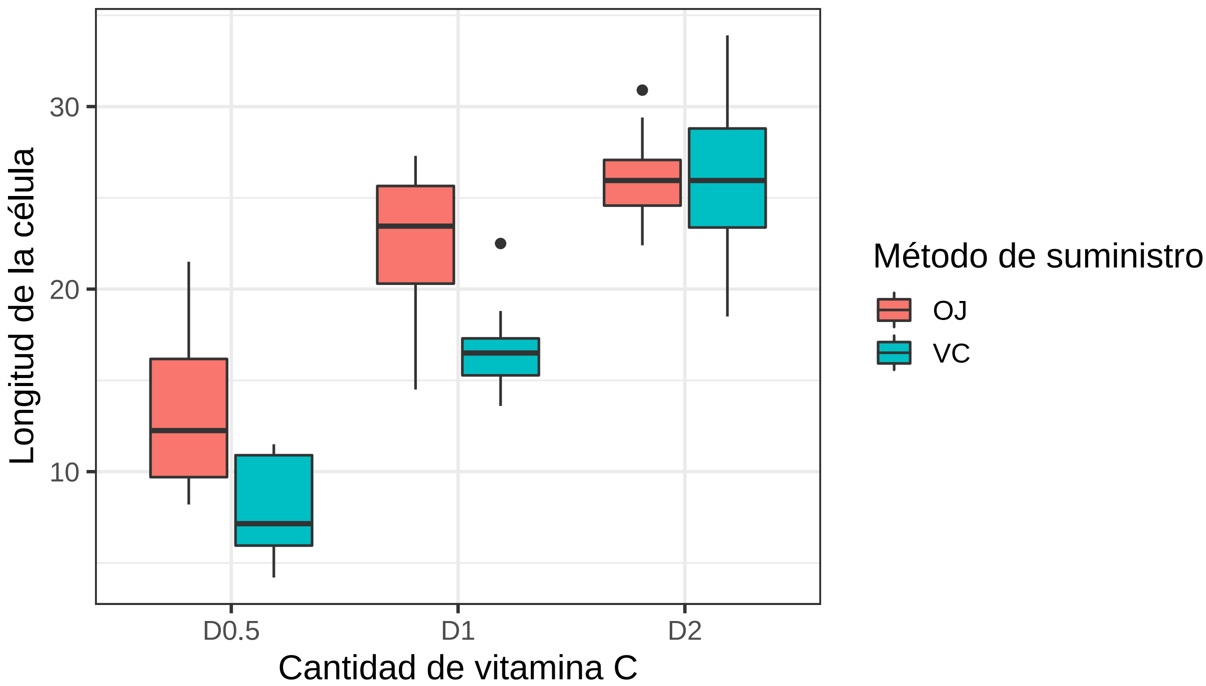
Podemos ver la tabla de frecuencias.

	D0.5	D1	D2
OJ	10	10	10
VC	10	10	10

Tenemos un diseño de celdas 2x3 con dos factores y 10 réplicas para cada tratamiento. Tenemos un diseño balanceado.

0.2 Gráficos preliminares

Es una buena idea examinar los datos gráficamente. La siguiente figura muestra un diagrama de cajas de la longitud de la célula para cada nivel de los dos factores.



Este gráfico indica que, generalmente, cuanto mayor es la cantidad de vitamina C que se suministra al animal, mayor será la longitud de la célula encargada del crecimiento de los dientes. También podemos ver que el método de suministro mediante zumo de naranja afecta en una mayor longitud de la célula sobretodo para dosis pequeñas de vitamina C (cuando la dosis es la mas grande, de 2 mg, parece que el método de suministro es indiferente). Basándonos en este análisis gráfico, intuimos que (1) la cantidad de vitamina C suministrada al animal afecta a la longitud de la célula, y (2) una mayor cantidad de vitamina C suministrada conlleva un mayor crecimiento de los dientes.

0.3 Análisis del problema

En este problema, queremos responder a las siguientes cuestiones:

- ¿ Que efectos tienen la cantidad de vitamina C y el método de suministro en la longitud de la célula encargada del crecimiento de los dientes de los animales?
- ¿ Que combinación de los factores produce una mayor longitud de la célula encargada del crecimiento de los dientes de los animales?

Ya que tenemos dos factores (uno con dos niveles y otro con tres niveles), el modelo es un diseño factorial de dos factores, o también llamado ANOVA de dos factores. Los niveles de los factores han sido prefijados por el experimentador, por tanto se trata de un modelo de efectos fijos.

Este diseño es un caso particular del diseño factorial de dos factores. El modelo de efectos fijos puede ser descrito de la siguiente manera

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \qquad i = 1, ..., a \quad j = 1, ..., b \quad k = 1, 2, ..., n$$

donde μ es el efecto medio global, τ_i es el efecto sobre la media del $i$ésimo nivel del factor A , β_j es el efecto medio del $j$ésimo nivel del factor B, $(\tau\beta)_{ij}$ es el efecto sobre la media de la interacción entre τ_i and β_j , y ϵ_{ijk} es el término aleatorio de error. Se asume que ambos factores son **fijos**.

En nuestro problema, el factor Cantidad de vitamina C tiene tres niveles, por tanto $a = 3$. El factor método de suministro tiene dos niveles, por tanto $b = 2$. Tenemos 10 réplicas para cada tratamiento (10 observaciones para cada combinación de los niveles de los factores), por tanto, $n = 10$. Por último, hay un total de $abn = 3 * 2 * 10 = 60$ observaciones.

Estamos interesados en contrastar la igualdad de los efectos de los tratamientos, es decir,

$$H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0$$
$$H_1 : \tau_i \neq 0 \text{ para al menos un } i$$

donde rechazaremos H_0 si el valor del estadístico de contraste es mayor que el valor del estadístico teórico, es decir, si $F_A = \frac{MS_A}{MS_E} > F_{\alpha,(a-1),ab(n-1)}$.

o

$$H_0 : \beta_1 = \beta_2 = ... = \beta_b = 0$$
$$H_1 : \beta_j \neq 0 \text{ para al menos un } j$$

donde rechazaremos H_0 si el valor del estadístico de contraste es mayor que el valor del estadístico teórico, es decir, si $F_A = \frac{MS_B}{MS_E} > F_{\alpha,(b-1),ab(n-1)}$.

También estamos interesados en determinar cuando los tratamientos interactúan, es decir,

$$H_0 : (\tau\beta)_{ij} = 0 \text{ para todo } ij$$
$$H_1 : (\tau\beta)_{ij} \neq 0 \text{ para al menos un par } ij$$

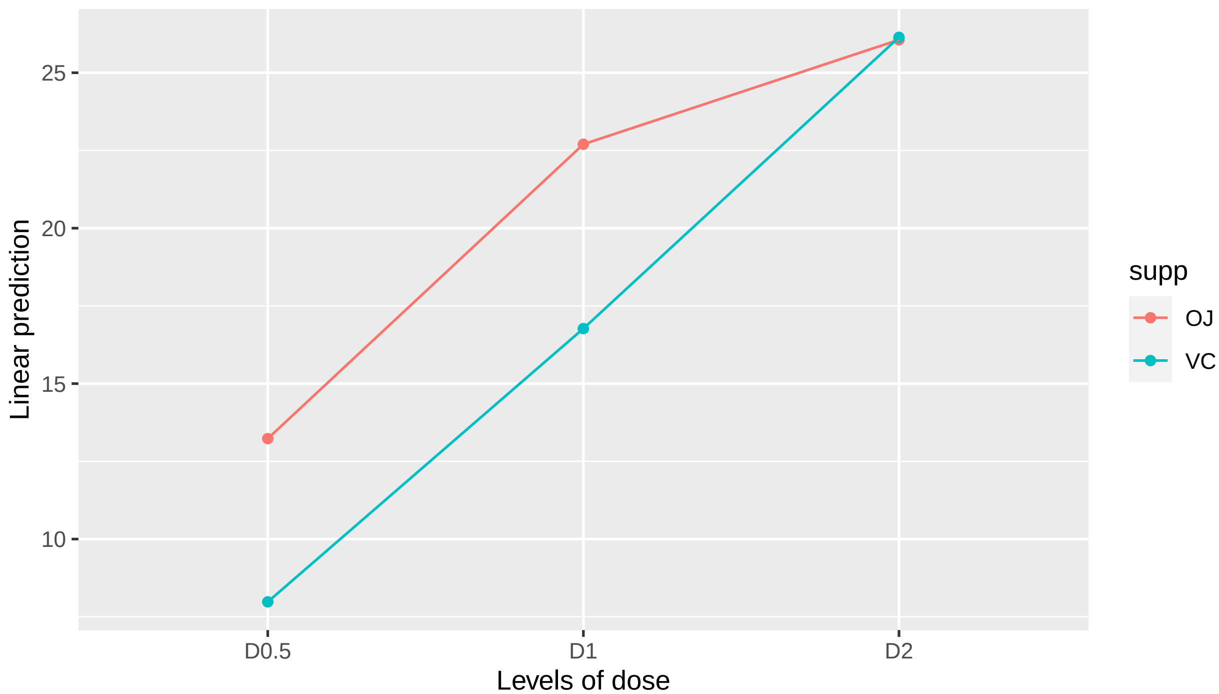
donde rechazaremos H_0 si el valor del estadístico de contraste es mayor que el valor del estadístico teórico, es decir, si $F_{AB} = \frac{MS_{AB}}{MS_E} > F_{\alpha,(a-1)(b-1),ab(n-1)}$.

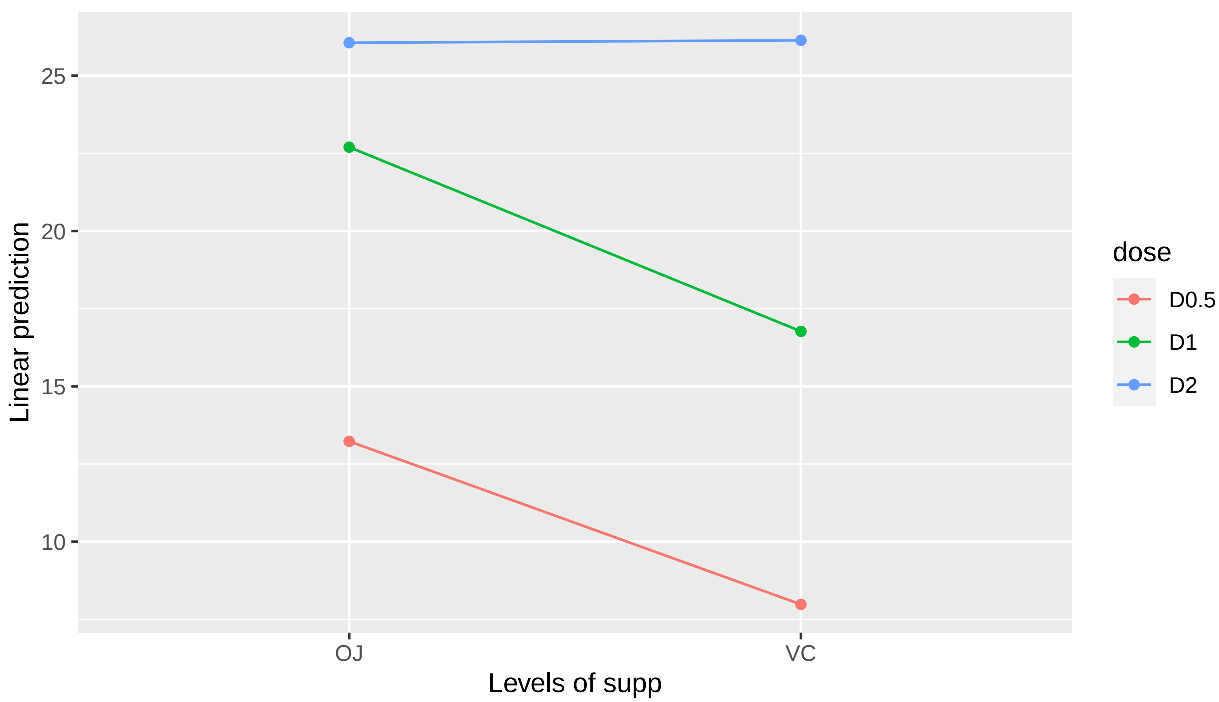
La tabla ANOVA con los resultados de estos contrastes para nuestro problema se muestra a continuación.

	Df	Sum Sq	Mean Sq	F value	P value
supp	1	205.350	205.35000	15.571979	0.0002312
dose	2	2426.434	1213.21717	91.999965	0.0000000
supp:dose	2	108.319	54.15950	4.106991	0.0218603
Residuals	54	712.106	13.18715		

- Como $F_A = \frac{MS_A}{MS_E} = 15.57 > F_{\alpha,(a-1),ab(n-1)} = F_{0.05,2,ab(n-1)}$ (o alternativamente, el P-valor es menor que el nivel α) rechazamos $H_0 : \tau_1 = \tau_2 = ... = \tau_a = 0$ y concluimos que el factor Cantidad de vitamina C afecta significativamente en la longitud de la célula encargada del crecimiento de los dientes de los animales.
- Como $F_B = \frac{MS_B}{MS_E} = 91.999965 > F_{\alpha,(b-1),ab(n-1)} = F_{0.05,1,ab(n-1)}$ (o alternativamente, el P-valor es menor que el nivel α) rechazamos $H_0 : \beta_1 = \beta_2 = ... = \beta_b = 0$ y concluimos que el factor método de suministro afecta significativamente en la longitud de la célula encargada del crecimiento de los dientes de los animales.
- Por último, como $F_{AB} = \frac{MS_{AB}}{MS_E} = 4.106991 > F_{\alpha,2,ab(n-1)} = F_{0.05,1,ab(n-1)}$ (o alternativamente, el P-valor es menor que el nivel α) rechazamos $H_0 : (\tau\beta)_{ij} = 0$ para todo ij y concluimos que la interacción entre los factores cantidad de vitamina C y método de suministro afecta significativamente en la longitud de la célula encargada del crecimiento de los dientes de los animales.

Para ayudar en la interpretación del experimento, es útil construir un gráfico de las respuestas medias de los tratamientos. Dicho gráfico se muestra a continuación.





La significatividad de la interacción se indica por la falta de paralelismo en las rectas.

A través de estos gráficos podemos observar que, en general, cuanto mayor es la cantidad de vitamina C que se suministra al animal, mayor es la longitud de la célula encargada del crecimiento de los dientes. También podemos ver que el método de suministro mediante zumo de naranja afecta en una mayor longitud de la célula sobretodo para dosis pequeñas de vitamina C (0.5 mg y 1mg), mientras que si la dosis es la mas grande (2mg) parece que el método de suministro es indiferente.

Mas tarde utilizaremos métodos para contrastar diferencias entre las medias.

Un contraste de la F se utiliza para calcular la variabilidad de la variable respuesta capturada por el modelo.

R-squared	Adjusted R-squared	Standard error	Statistic	P value	Df
0.7937246	0.774625	3.631411	41.55718	0	6

El P-valor es muy pequeño. La interpretación de este contraste es que al menos uno de los tres términos en el modelo es significativo. También tenemos que el coeficiente de variacion es $R^2 = 0.7937246$ lo cual indica que sobre un 80 por ciento de la variablilidad de la variable longitud de la célula es explicada por los dos factores y la interacción.

En la siguiente seccion, discutimos el uso de los residuos y gráficos de los residuos para la comprobaciòn de las hipótesis de modelo.

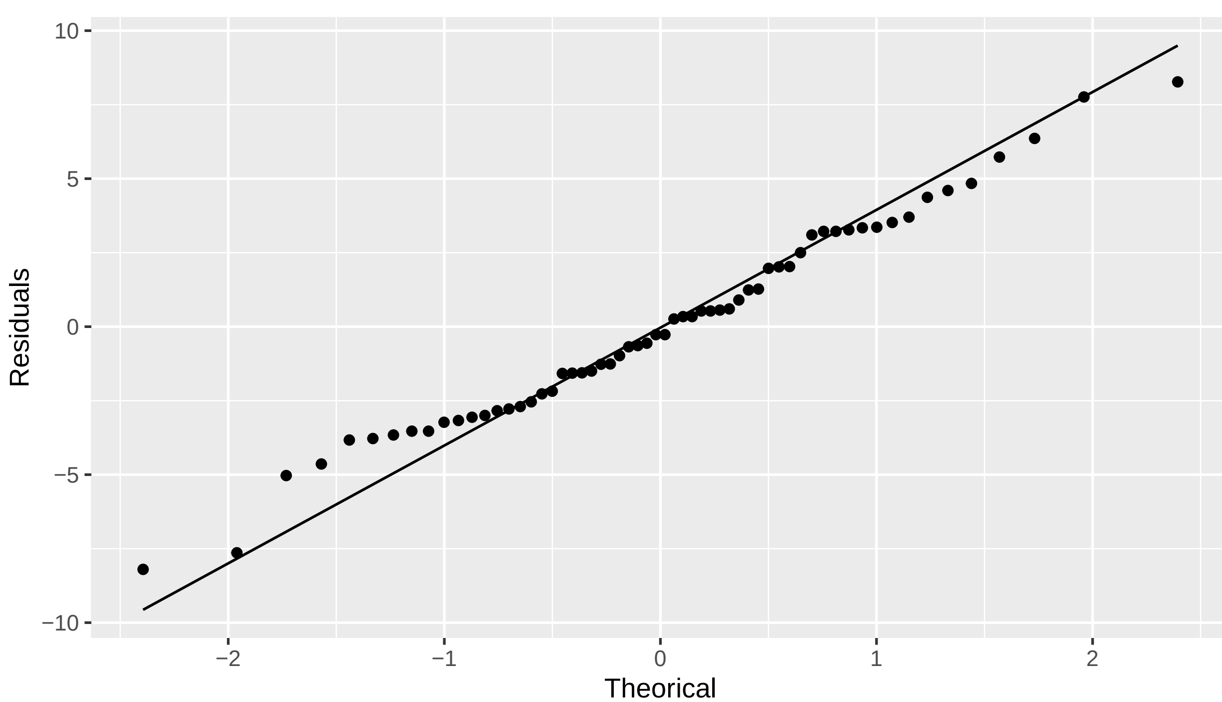
0.4 Hipótesis del modelo

El incumplimiento de los supuestos básicos puede ser investigado fácilmente examinando los **residuos**. Los residuos para un diseño factorial de dos factores son $e_{ijk} = y_{ijk} - \hat{y}_{ijk} = y_{ijk} - \overline{y}_{ij.}$ ya que los valores predichos son $\hat{y}_{ijk} = \overline{y}_{ij.}$ (la media de las observaciones en la celda ijesima).

0.4.1 La hipótesis de normalidad

Para comprobar la hipótesis de normalidad podemos ver el histograma de los residuos. Si se verifica la hipótesis de que los residuos son $NID(0, \sigma^2)$ este gráfico deberia parecerse al de una muestra de una distribución normal centrada en cero. Desafortunadamente, con muestras pequeñas, a menudo se produce una fluctuación considerable en la forma del histograma, por lo que una desviación moderada de la normalidad no implica necesariamente una violación grave de los supuestos. Una desviación mas fuerte es potencialmente grave y requiere un análisis mas detallado.

Un procedimiento muy útil es contruir un gráfico de probabilidad normal, normalmente llamado qqplot. Si los errores se distribuyen segun una distribución normal, este gráfico debera dibujar una línea recta. Cuando visualizamos esta recta, se suele hacer mas énfasis en los valores centrales que en los extremos.



La impresion general despues de analizar este gráfico es que los errores siguen, aproximadamente, una distribución normal.

Podemos usar el test de **Shapiro-Wilk** para comprobar la normalidad de los errores. En este caso la hipótesis nula es que los errores siguen una distribución normal.

Los resultados de este test se muestran en la siguiente tabla.

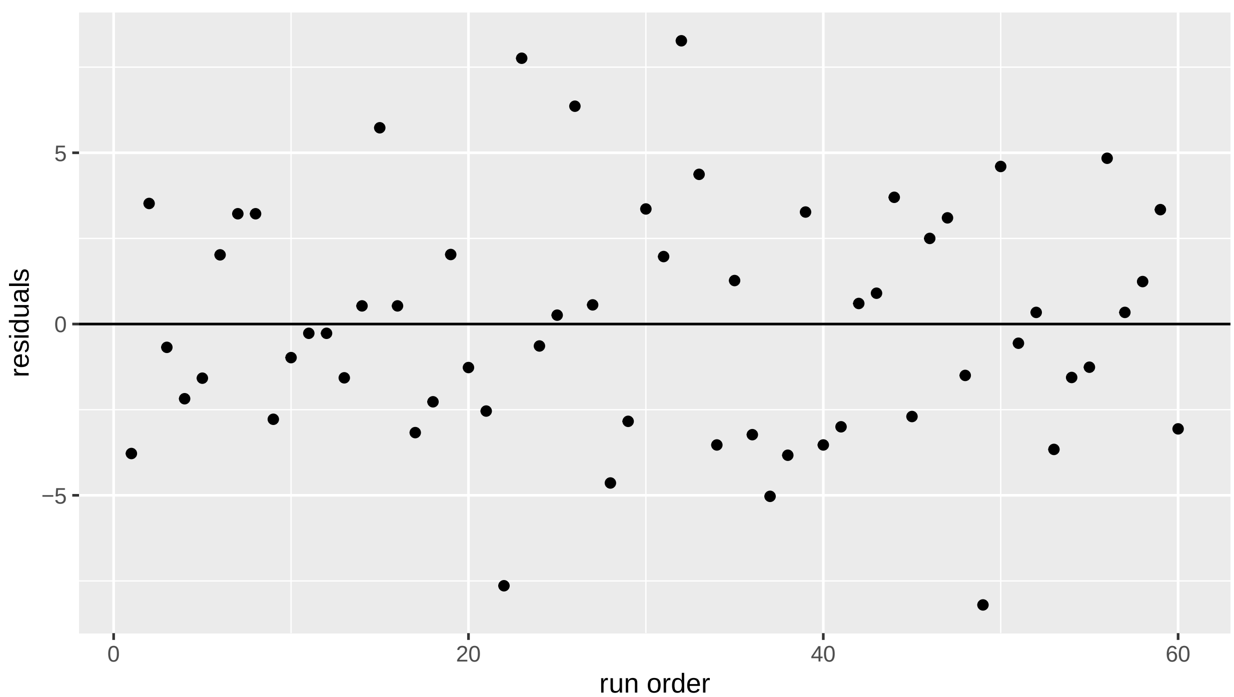
	Statistic	P value
	0.9849884	0.6694242

Como el P-valor es mayor que el nivel α, no podemos rechazar la hipótesis nula de que los errores siguen una distribución normal. Esta es la misma conclusión que obtuvimos tras analizar el gráfico de probabilidad normal de los residuos.

0.4.2 Independencia de los errores

Un gráfico de los residuos frente al orden temporal en el que se recogieron los datos puede ayudar a detectar una fuerte correlación de los residuos. Este gràfico solo es apropiado si conocemos el orden en el cual los datos fueron recogidos. Si hay alguna tendencia entre residuos positivos y negativos puede indicar una correlación positiva. Esto podria implicar que la hipótesis de independecia de los errores no se verifica.

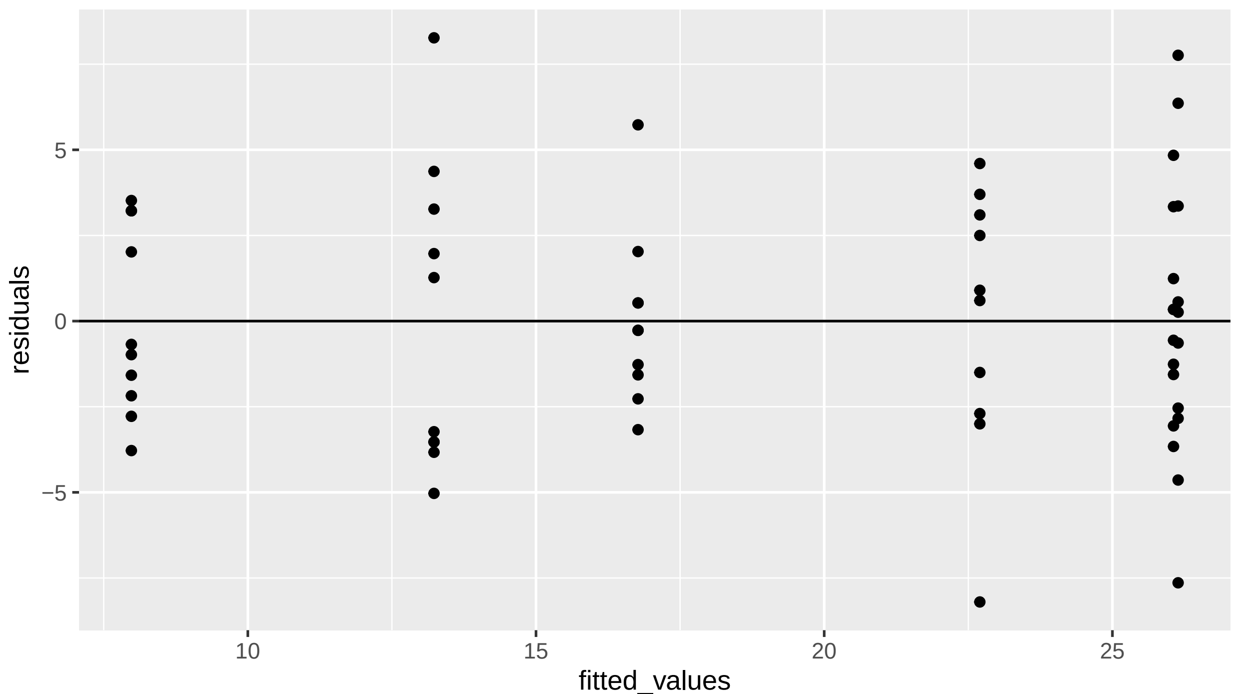
EL gráfico de los residuos frente al orden temporal, suponiendo que el orden en el cual se recogieron los datos es el orden en el que vienen los datos, se muestra en la siguiente figura.



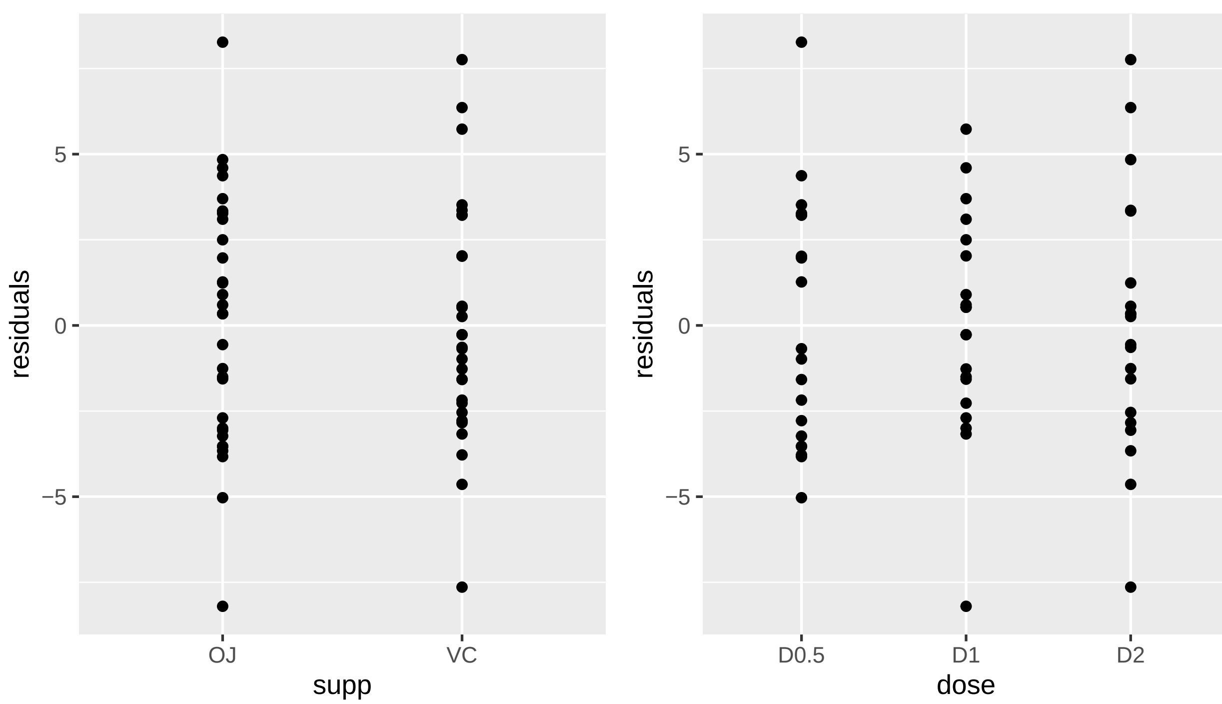
Observando este gráfico no hay razon para sospechar ninguna violación de independencia o de varianza constante de los errores.

0.4.3 Varianza constante u homocedasticidad

Si el modelo es correcto y las hipótesis se cumplen, los residuos no deberian estar relacionados con ninguna otra variable, incluyendo la variable respuesta. Una simple comprobación se puede hacer graficando los residuos frente a los valores predichos $\hat{y}_{ij.}$ ($\hat{y}_{ij} = \overline{y}_{ij.}$). Este gráfico no deberia motrar ningun patron.



Las siguientes gráficas muestran los residuos frente a los factores tipo de suministro y cantidad de vitamina C, respectivamente.



Aunque los gráficos de los residuos se usan a menudo para diagnosticar la desigualdad de las varianzas, existen diferentes test estadísticos para comprobar esto. Estos test contrastan la hipótesis nula de que las varianzas son iguales en los diferentes grupos de los tratamientos.

Un test muy usado para contrastar la igualdad de las varianzas es el **test de Barlett**. Este procedimiento hace uso de un estadístico cuya distribución es aproximadamente una chi-cuadrado.

Ya que el test de Barlett es sensible a la hipótesis de normalidad vamos a utilizar otro método alternativo para contrastar las hipótesis del modelo de que la varianza de los residuos es constante. El **test modificado de Levene** es un procedimiento robusto a desviaciones de la normalidad que también nos permite contrastar la igualdad de las varianzas. Este test utiliza la desviación estándar de las observaciones y_{ij} en cada tratamiento a traves de la mediana del tratamiento.

Los resultados de este test se muestran en las siguiente tabla.

	Df	Statistic	P value
	5	1.708578	0.1483606

El P-valor es mayor que el nivel $\alpha = 0.05$, por tanto, no podemos rechazar la hipótesis nula (de que las varianzas son iguales) y podemos asumir la homogeneidad de las varianzas en los diferentes grupos.

0.5 Comparaciones múltiples

Cuando los resultados del modelo ANOVA indican hay diferencias significativas entre los tratamientos o que hay presencia de interacción entre los factores, es interesante hacer comparaciones entre las medias de los factores para descubrir las diferencias especificas.

Hay que tener en cuenta que, cuando la interacción entre los factores es significativa, las comparaciones de las medias de un factor pueden ser "escondidas" por la interacción. Un enfoque para esta situación es fijar un factor a un nivel específico y aplicar el test para contrastar las medias de los niveles del otro factor. Podríamos comparar un total de $ab = 6$ celdas (combinaciónes de los niveles de los factores). En este análisis, hay diferencias de las medias tanto para los efectos principales como para la interacción, por tanto, habría un total de 15 comparaciones para todos los pares de medias posibles.

Para ilustrar esto, supongamos un ejemplo en el que estamos interesados en hacer comparaciones entre las medias del factor cantidad de vitamina C (tres niveles). Como hay interacción entre los factores, hacemos esta comparación para un solo método de suministro (zumo de naranja). Así, queremos contrastar todas las comparaciones de medias por pares del factor cantidad de vitamina C, es decir, $H_0 : \tau_i = \tau_j$ frente a $H_1 : \tau_i \neq \tau_j$ para todo $i \neq j$ $i, j = 1, 2, 3$, para un nivel fijado del factor método de suministro (digamos β_1).

El test de **Tukey** es utilizado para encontrar las medias por pares que difieren significativamente y también es capaz de construir intervalos de confianza para todas las comparaciones por pares de medias.

Vamos a asumir que la mejor estimación de la varianza del error es MS_E calculada en la tabla ANOVA, y ademas es la misma para todas las combinaciónes de los tratamientos. Este test rechaza H_0 (que las medias son iguales) si

$$|\overline{y}_{i.} - \overline{y}_{j.}| > T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

donde a es el número de niveles del factor y f es el número de grados de libertad asociado con el MS_E. La distribución del estadístico de recorrido estudentizado, q_α, se puede encontrar en tablas de muchos libros de estadística. Equivalentemente, se pueden contruir intervalos de confianza para todos los pares de medias

$$\overline{y}_{i.} - \overline{y}_{j.} - q_\alpha(a, f)\sqrt{\frac{MS_E}{n}} \leq \mu_i - \mu_j \leq \overline{y}_{i.} - \overline{y}_{j.} + q_\alpha(a, f)\sqrt{\frac{MS_E}{n}}$$

Hay que anotar que si tenemos un diseño no balanceado (el número de observaciones para cada nivel del factor es diferente) podríamos usar el método de **Tukey-Kramer** con una pequeña modificación en el cálculo de T_α, pero no es nuestro caso.

Los resultados del test de Tukey se muestran en la siguiente tabla.

	Difference	Lower ci	Uper ci	P value
VC:D0.5-OJ:D0.5	**-5.25**	**-10.0481**	**-0.4519**	**0.0243**
OJ:D1-OJ:D0.5	**9.47**	**4.6719**	**14.2681**	**0.0000**
VC:D1-OJ:D0.5	3.54	-1.2581	8.3381	0.2640
OJ:D2-OJ:D0.5	**12.83**	**8.0319**	**17.6281**	**0.0000**
VC:D2-OJ:D0.5	12.91	8.1119	17.7081	0.0000
OJ:D1-VC:D0.5	14.72	9.9219	19.5181	0.0000
VC:D1-VC:D0.5	**8.79**	**3.9919**	**13.5881**	**0.0000**
OJ:D2-VC:D0.5	18.08	13.2819	22.8781	0.0000
VC:D2-VC:D0.5	**18.16**	**13.3619**	**22.9581**	**0.0000**
VC:D1-OJ:D1	**-5.93**	**-10.7281**	**-1.1319**	**0.0074**
OJ:D2-OJ:D1	**3.36**	**-1.4381**	**8.1581**	**0.3187**
VC:D2-OJ:D1	3.44	-1.3581	8.2381	0.2936
OJ:D2-VC:D1	9.29	4.4919	14.0881	0.0000
VC:D2-VC:D1	**9.37**	**4.5719**	**14.1681**	**0.0000**
VC:D2-OJ:D2	**0.08**	**-4.7181**	**4.8781**	**1.0000**

En esta tabla se muestran los resultados de todas las comparaciones de medias por pares posibles.

- Las filas de la tabla marcadas en negrita muestran las comparaciones del ejemplo supuesto anteriormente. Podemos ver que no hay diferencias significativas entre las medias de los niveles 2mg-1mg del factor cantidad de vitamina C fijando el método de suministro zumo de naranja (el P-valor es mayor que el nivel α por lo que no podemos rechazar la hipótesis nula), mientras que si hay diferencias significativas entre las medias de los niveles 1mg-0.5mg y 2mg-0.5mg del factor cantidad de vitamina C fijando el método de suministro zumo de naranja (el P-valor es menor que el nivel α por lo que rechazamos la hipótesis nula).

- Las filas de la tabla marcadas en azul muestran las comparaciones del ejemplo supuesto anteriormente pero esta vez fijando el método de suministro ácido ascórbico. Podemos ver que hay diferencias significativas entre las medias de los niveles 2mg-0.5mg, 1mg-0.5mg y 2mg-1mg del factor cantidad de vitamina C fijando el método de suministro ácido ascórbico.

- Las filas de la tabla marcadas en gris muestras las comparaciones entre las medias de los dos niveles del factor método de suministro (zumo de naranja y ácido ascórbico.) fijando el nivel del factor cantidad de vitamina C a 0.5mg, 1mg, y 2mg, respectivamente. Podemos ver que hay diferencias significativas entre las medias de los dos niveles si fijamos la cantidad de vitamina C a 0.5mg o 1mg, mientras que no hay diferencias si fijamos la cantidad de vitamina C a 2mg.

Existen varios prodedimientos alternativos al test de Tukey como el método de Bonferroni, el LSD de Fisher o el test de Duncan. En nuestro caso hemos aplicado varios métodos obteniendo resultados muy similares.

Para terminar esta seccion, debemos decir que, generalmente, los métodos de contrastes son utiles para comparaciones preplaneadas(especificadas antes de llevar a cabo el experimento y de examinar los datos). Esto es debido a que, si elegimos las comparaciones despues del realizar el experimento, probablemente contruiremos los test que corresponden a las diferencias de medias observadas mas grandes. Estas grandes diferencias pueden ser resultado de la presencia de efectos reales, o podrian ser resultado del error aleatorio. La idea de examinar los datos para seleccionar las comparaciones de gran ineteres es llamado **data snooping**. Existen métodos, como el método de Scheffe (que veremos a continuación), que posibilitan esta idea.

El método de **Scheffé** es utilizado para comparar cualquier posible contraste entre las medias de los tratamientos y también es capaz de contruir intervalos de confianza. Aunque el método de Scheffé se puede utilizar para muchas posibles comparaciones entre medias, cuando se estudian solo diferencias entre pares de medias resulta menos eficaz que los tets especificos para contrastar diferencias de pares de medias.

Para ilustrar este procedimiento, consideramos nuestro ejemplo y supongamos que el contraste de interés es, para el nivel fijado zumo de naranja del factor método de suministro β_1, $\Gamma_1 = \tau_1 + \tau_2 - \tau_3 = 0$.

El resultado de este contraste utilizando el método de Scheffé se muestra en la siguiente tabla.

	Difference	Lower ci	Uper ci	P value
OJ:D0.5,OJ:D1-OJ:D2	-8.095	-12.95289	-3.23711	7.09e-05

El P-valor del contraste Γ_1 es menor que el nivel $\alpha = 0.05$ por tanto el contraste $\Gamma_1 = \tau_1 + \tau_2 - \tau_3 = 0$ se rechaza y se concluye que, para el nivel fijado zumo de naranja del factor método de suministro, la media de las cantidades de vitamina C 0.5mg y 1mg como grupo difiere significativamente de la media de la cantidad de vitamina C 2mg.

0.6 Conclusiones

Para concluir, contestamos a las dos cuestiones que nos planteabamos en el análisis inical del problema:

- ¿ Que efectos tienen la cantidad de vitamina C y el método de suministro en la longitud de la célula encargada del crecimiento de los dientes de los animales?

Si se elige el ácido ascórbico como método de suministro, cuanto mayor sea la cantidad de vitamina C que se suministre al animal, mayor sera la longitud de la célula encargada del crecimiento de los dientes. Si se elige el zumo de naranja como método de suministro, lo anterior es cierto, salvo para las cantidades de vitamina C 1mg y 2mg donde no hay diferencias significativas.

Respecto al método de suministro, para dosis de 0.5mg y 1mg de vitamina C suministradas al animal, el método de suministro zumo de narnaja afecta en una mayor longitud de la célula encargada del crecimiento de los dientes. Para la dosis de vitamina C mas alta (2mg), no hay diferencias significativas en la longitud de la célula por el método de suministro.

- ¿ Que combinación de los factores produce una mayor longitud de la célula encargada del crecimiento de los dientes de los animales?

La cantidad de vitamina C mas pequeña (0.5mg) queda descartarda si se desea una mayor longitud de la célula, puesto que, independientemente del método suministrado, produce las longitudes mas bajas de estas células. La duda esta entre las dosis de vitamina C de 1mg y 2mg. Si se elige el método de suministro ácido ascórbico entonces hay diferencias significativas entre estas dos cantidades de vitamina C, procudiendo una mayor longitud de la célula la dosis de vitamina C de 2mg. Si en cambio se elige el método de suministro zumo de naranja entonces no hay diferencias significativas entre estas dos cantidades de vitamina C.

Así, las combinaciónes que producen una mayor longitud de la célula son: ácido ascórbico-2mg, zumo de naranja-2mg y zumo de naranja-1mg.