

Taller de Consultoria - TP4

Gonzalo Barrera Borla

6/10/2019

Setup

Ejercicio 1

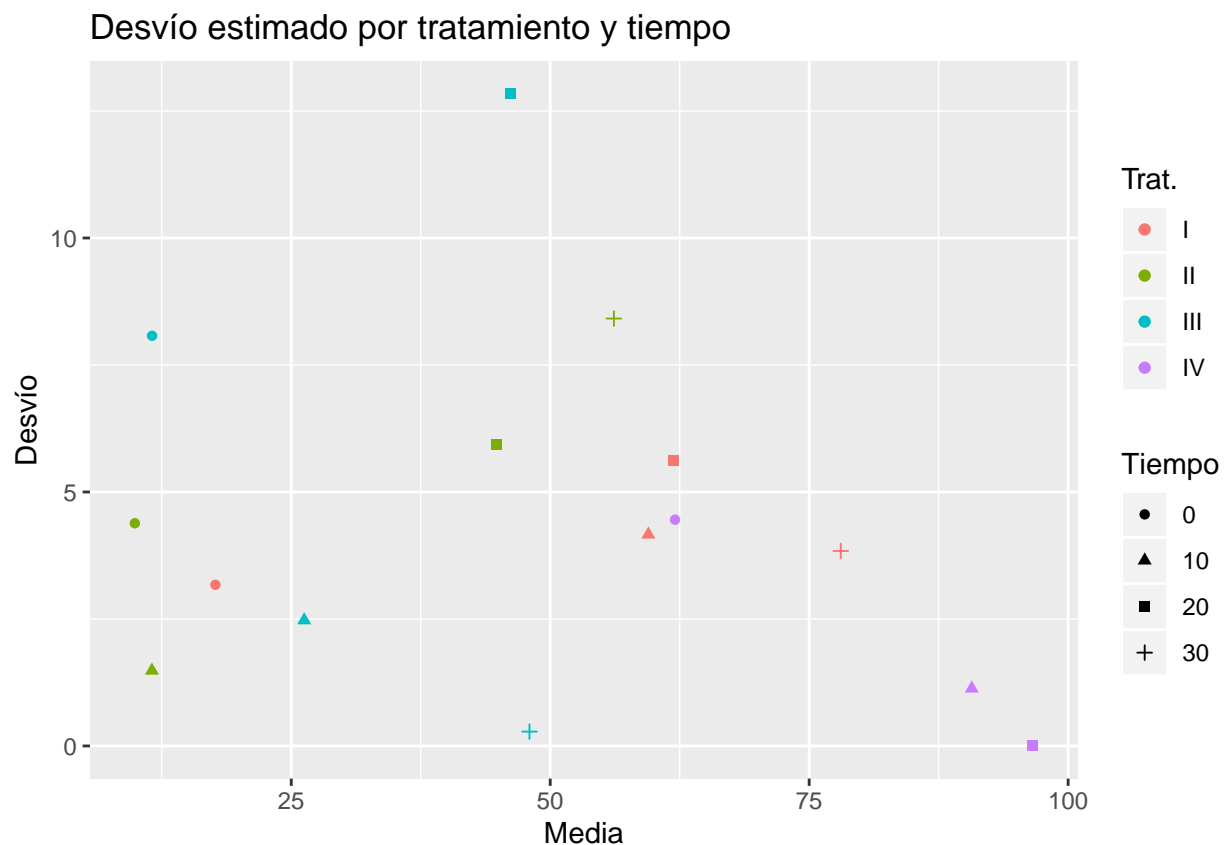
Se midió el daño al ADN (porcentaje de ADN que se separa durante la electroforesis) en raíces de habas sometidas a 4 tratamientos, aplicados durante diferentes tiempos. Interesa decidir si algún tratamiento es demostrablemente más dañino. Los “-” indican datos no medidos. El tiempo cero para cada uno de los tratamientos puede interpretarse como la cantidad de ADN que está inicialmente dañado, o sea que no es equivalente a “ningún tratamiento”. Notar que los resultados son porcentajes, lo que implica heteroscedasticidad.

Comenzamos por analizar la relación entre media y varianza para cada tiempo y tratamiento. Como se ve a continuación, los resultados no son muy alentadores:

trt	t	mean	sd
III	20	46.183	12.835
II	30	56.150	8.415
III	0	11.567	8.075
II	20	44.800	5.940
I	20	61.925	5.625
IV	0	62.050	4.455
II	0	9.900	4.384
I	10	59.475	4.163
I	30	78.050	3.839
I	0	17.675	3.172
III	10	26.250	2.475
II	10	11.550	1.485
IV	10	90.700	1.131
III	30	48.000	0.283
IV	20	96.600	0.000

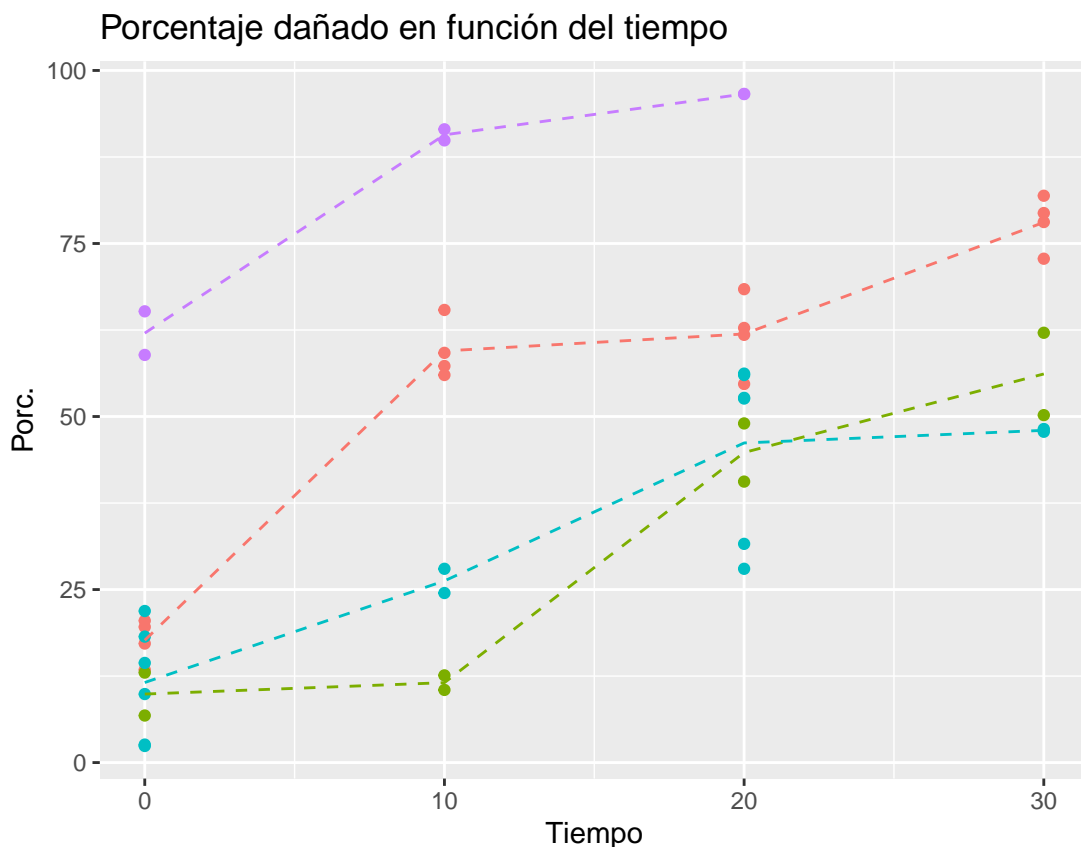
En primer lugar, para el tratamiento IV y $t = 20$, la varianza estimada es exactamente 0, por lo que la razón entre la máxima y mínima varianza para todas las combinaciones de bloque y tratamiento está indefinida, y la regla heurística que plantea Seber [1977, p.195] y utilizaremos en el TP3 no se puede usar. Aún descartando esta última observación, para el tratamiento III la varianza estimada es 12.84 en $t = 20$ y sólo 0.28 en $t = 30$, que arroja una altísima razón de 43.35.

Aún peor, cuando graficamos la varianza estimada en función de la media estimada, no se evidencia ninguna relación, por lo que ni siquiera sabríamos cómo morigerar la heteroscedasticidad existente:



Por el momento, nos bastará con haber notado esta complicación, y avanzaremos con el análisis.

¿Qué modelo puede llegar a representar adecuadamente estos datos? Para contestarnos, comenzamos graficando el porcentaje de ADN dañado en función de t para cada observación, coloreada según el tratamiento. De fondo, agregamos en línea punteada el “perfil” del porcentaje promedio de ADN dañado por tratamiento y tiempo,



para distinguir tendencias:

Pareciera que al menos el tratamiento IV se distingue desde el comienzo de los demás, y que el porcentaje de daño es siempre creciente en t , a una tasa constante o *tal vez* decreciente.

Aunque esté medida en pocos valores únicos, la covariable t (tiempo) es cuantitativa, y por ende entendemos que es más razonable plantear lo que en la literatura se conoce como un *análisis de la covarianza* (ANCOVA) de un factor (el tratamiento) con 4 niveles, antes que un modelo de análisis de la covarianza de 2 factores.

Combinando estas observaciones, planteamos como modelo de referencia - (de referencia) un ANOVA de 2 factores, - (y como candidatos) 6 modelos “ANCOVA” con un factor (**tratamiento**) sobre - polinomios de grado 1, 2 y 3 en t , - con y sin interacciones

Elegiremos polinomios “crudos” en lugar de los ortogonales por defecto de R para facilitar la comprensión de los coeficientes resultantes.

Si uno ajusta los modelos sobre los datos completos, no importa la métrica elegida, los modelos con más parámetros (ANOVA 2 factores y el cúbico multiplicativo) tienden a ser los que mejor ajustan: con sólo 46 datos, ajustar 16 coeficientes es casi una garantía de éxito.

Para evitar este sesgo hacia modelos más complejos, utilizamos el muy recomendable paquete **caret**, que nos permite ajustar modelos con distintas técnicas de resampleo. Para este ejercicio, optamos por el clásico split entre datos de entrenamiento (70%) y teste (30%), y repetimos el proceso 100 veces. A continuación, incluimos el error cuadrático medio estimado para cada modelo, y su desvío estándar estimado, para poder seleccionar el “mejor” modelo según la ya discutida regla de un desvío estándar:

Modelo	ECM	Desvio
$p \sim \text{factor}(t) * \text{trt}$	9.12	2.80
$p \sim \text{poly}(t, 2, \text{raw} = \text{TRUE}) + \text{trt}$	9.83	1.36
$p \sim \text{poly}(t, 1, \text{raw} = \text{TRUE}) + \text{trt}$	9.88	1.23
$p \sim \text{poly}(t, 3, \text{raw} = \text{TRUE}) + \text{trt}$	10.07	1.72

Modelo	ECM	Desvio
$p \sim \text{poly}(t, 2, \text{raw} = \text{TRUE}) * \text{trt}$	10.45	4.17
$p \sim \text{poly}(t, 1, \text{raw} = \text{TRUE}) * \text{trt}$	10.68	1.92
$p \sim \text{poly}(t, 3, \text{raw} = \text{TRUE}) * \text{trt}$	14.32	11.41

Como es de esperar, el modelo cúbico multiplicativo no resiste la validación cruzada. Sin embargo, el modelo que mejor ajusta, aún cuando usando técnicas de validación cruzada, sigue siendo el ANOVA de 2 factores, que cuenta con la mayor libertad posible para adecuarse a los datos (tiene la misma cantidad de coeficientes que el cúbico multiplicativo, pero sin correlación alguna entre las covariables predictoras, que sí tienen t, t^2, t^3). Sin embargo, se observa que el desvío estimado para el modelo ANOVA de 2 factores es bastante alto, lo suficiente como para que la regla de 1 SD contenga cómodamente a todos los modelos salvo, justamente el cúbico multiplicativo. En este contexto, nos inclinamos por el modelo más sencillo posible, que es el lineal aditivo. Ajustamos entonces

$$p_{ij} = \mu_i + \beta \times t + \epsilon_{ij} \quad \text{con } \epsilon_{ij} \sim N(0, \sigma^2), i \in \text{I, II, III, IV}, 1 \leq j \leq n_i$$

donde μ_i es la ordenada correspondiente al tratamiento i , t es el tiempo de medición y n_i la cantidad de observaciones del i -ésimo tratamiento, *a sabiendas* de que no hemos podido garantizar todavía la homocedasticidad de los ϵ_{ij} . Los principales estadísticos del modelo, efectos principales y coeficientes del modelo *ajustado con todos los datos* resultan:

R ² aj.	F obs.	P-valor
0.879	82.9	4.36e-19

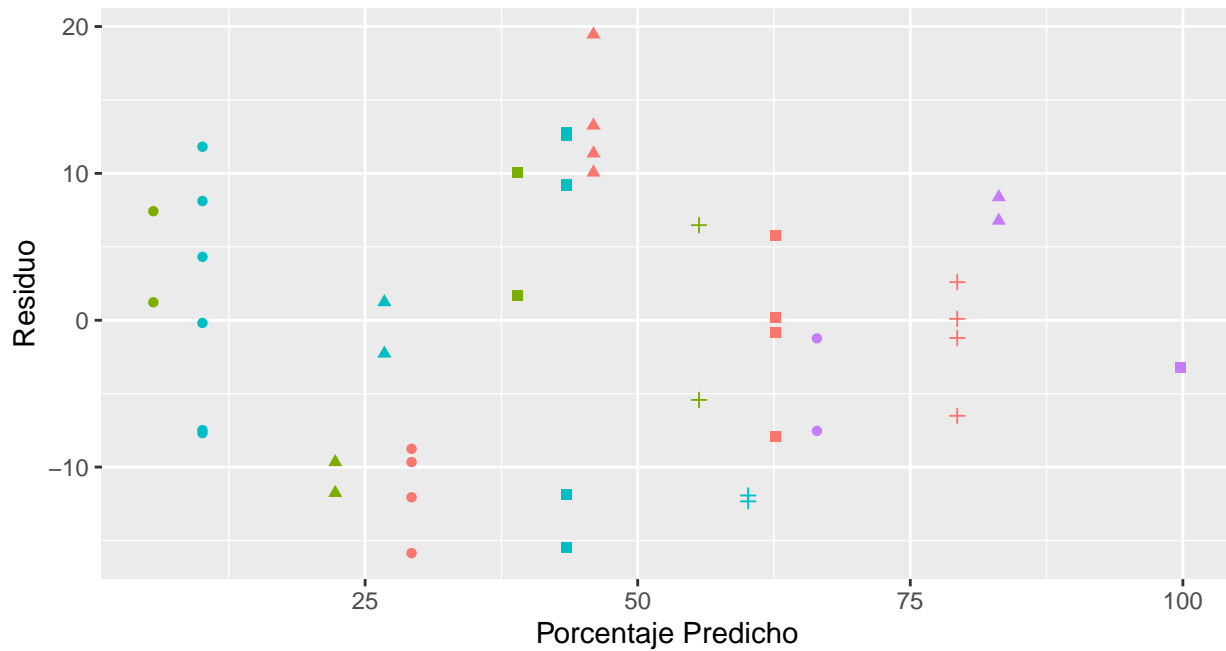
Coefs	GL	C. Medios	F obs.	P-valor
trt	3	55.5	55.5	1.71e-14
t	1	165	165	5.86e-16
Residuals	41	NA	NA	NA

Coef.	Estimado	P-valor
(Intercept)	29.3	5.57e-12
trtII	-23.7	8.79e-07
trtIII	-19.2	1.13e-06
trtIV	37.2	4.41e-10
t	1.67	5.86e-16

Podemos observar que tanto (i) el modelo completo, (ii) los efectos principales del tratamiento y el tiempo y (iii) cada coeficiente individual son significativos con bajísimo p-valor. Revisamos los residuos (absolutos y estudentizados) de las predicciones, para estimar si hemos capturado satisfactoriamente la estructura de los datos, y no observamos nada particularmente alarmante:

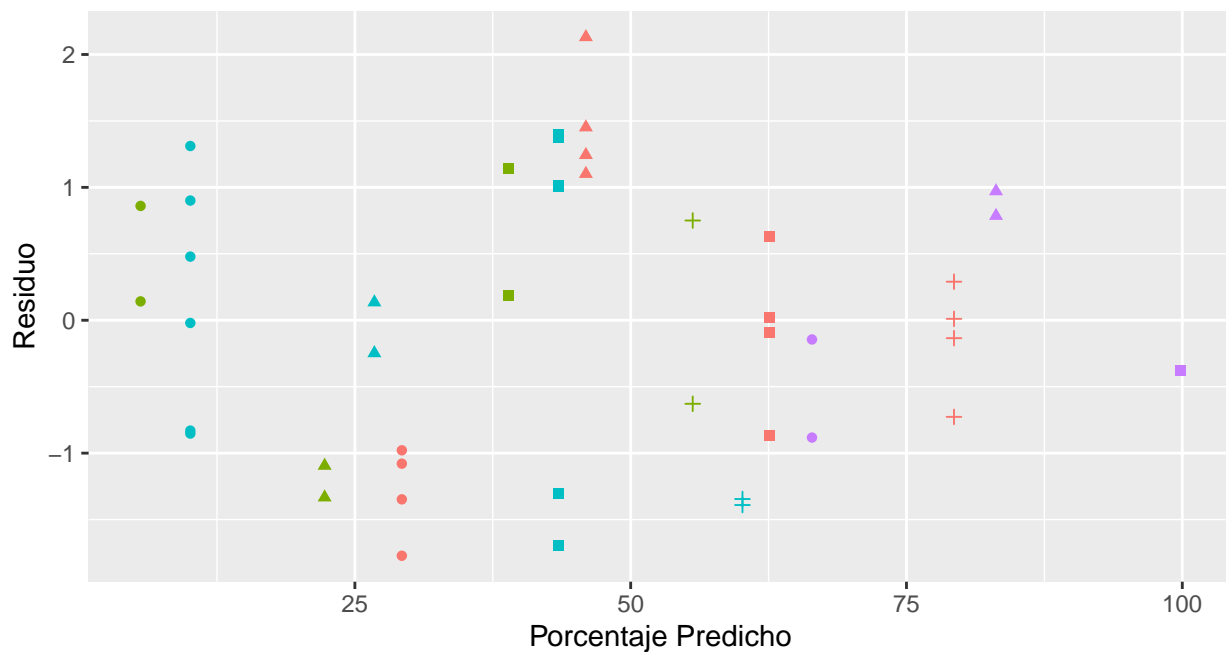
Residuos en función del valor predicho

No se evidencia estructura aparente



Residuos estudentizados en función del valor predicho

Ningún residuo tiene valor absoluto mayor a 2.5



A pesar de la evidente heterocedasticidad original de las observaciones, hemos obtenido un ajuste bastante

satisfactorio para los datos, con una directa interpretación física. Sabemos que cuando asumimos erróneamente una matriz de covarianza $\sigma^2 I_n$ en lugar de la verdadera V , los estimadores de los parámetros del modelo $\hat{\beta}$ son insesgados, pero no los de menor varianza. Así y todo, hemos obtenido unos estimadores indudablemente distintos a cero, con lo cual podemos en principio confiar en ellos, ya que de usar la verdadera matriz de covarianza su significación debería mejorar.

Finalmente, podemos asegurar que

- el porcentaje inicial de ADN dañado varía con el método,
- pero el avance del daño en el tiempo no varía con el método.

Los modelos ajustados son:

$$\begin{aligned} p_{I,j} &= 29.3 + 1.67 \times t + \epsilon_{I,j} \\ p_{II,j} &= 5.6 + 1.67 \times t + \epsilon_{II,j} \\ p_{III,j} &= 10.1 + 1.67 \times t + \epsilon_{III,j} \\ p_{IV,j} &= 66.5 + 1.67 \times t + \epsilon_{IV,j} \end{aligned}$$

Y los tratamientos, del más al menos dañino, son $IV > I > III > II$. Concluimos con un gráfico de las funciones ajustadas sobre los datos originales:

