Universidad de Valladolid

Análisis de Supervivencia: Estudio acerca de cáncer de pulmón

Álvaro Berrío Galindo
Álberto Calvo Madurga

Métodos Estadísticas de Computación Intensiva

Curso 2018-2019

1. Objetivos

El trabajo tiene como objetivo aplicar algo de lo visto en las horas de teoría de la asignatura de Métodos Estadísticos de Computación Intensiva. Nosotros hemos optado por realizar un análisis de supervivencia ya que pensamos que es lo más cercano a lo que nosotros podamos realizar en un futuro y es más intuitivo y fácil de mostrar.

La primera tarea es la adquisición de los datos. De esta manera, hemos encontrado un dataset interesante de un repositorio que contiene información acerca de 228 enfermos de cáncer de pulmón recogido por el Grupo de Tratamiento de Cáncer Norte Central (NCCTG en inglés).

Es interesante hacer un primer análisis descriptivo de algunas de las covariables, como por ejemplo **sexo**, **ph.ecog** y **edad**. Para la última haremos una categorización para menores de 65, 65-74 y mayores de 74, que es una división estándar utilizada en estos tipos de análisis médicos. Además miraremos posibles correlaciones entre las covariables.

Después, mediante un modelo de regresión de Cox de riesgos proporcionales, vamos a analizar el conjunto de datos para concluir qué variables explicativas tienen influencia significativa a la hora de predecir una muerte por cáncer de pulmón y explicar de qué manera influyen.

Finalmente estudiaremos los residuos para ver el ajuste del modelo, puntos influyentes o si aparecen outliers.

2. Análisis de las variables

- **inst**: número de la institución de la que proviene el enfermo.
- **time**: días que ha sobrevivido, variable de interés, tiempo desde que se detecta el cáncer hasta que fallece.
- **status**: indica que si el individuo ha fallecido o si ha sido una censura (1: dead, 2: censured).
- **age**: edad a la que se detectó la enfermedad.
- **ph.ecog**: puntuación de ECOG (indica la capacidad del enfermo de poder llevar una vida diaria normal(0: good, 5: dead)).
- **ph.karno**: puntuación de Karnofsky dicha por un médico (lo mismo que la anterior pero 0 indica mal y 100 bien).
- **pat.karno**: puntuación de Karnofsky dicha por el paciente.
- **meal.cal**: calorías consumidas en las comidas.
- **wt.loss**: peso perdido en los últimos 6 meses.

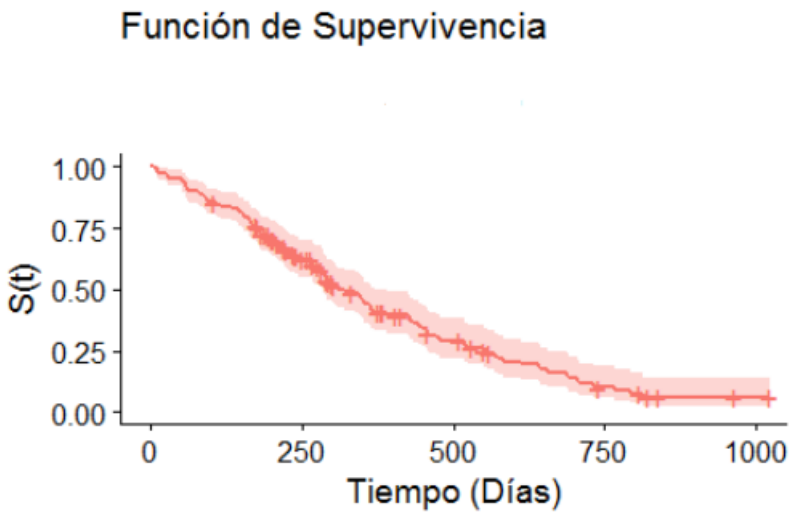
El objetivo de este estudio es ver cómo afectan las diferentes puntuaciones (ph.ecog, ph.karno, pat.karno) y el resto de variables como la edad en la supervivencia de enfermos de cáncer de pulmón.

3. Censuras

Tenemos la variable **status** que es la que nos dice si una observación es una censura o no. No tenemos información acerca de qué tipo sea, pero lo más lógico es que sean censuras por la derecha que indiquen que el individuo no había muerto para la fecha en la que el estudio terminó, murió por otras causas o bien se perdió la observación.

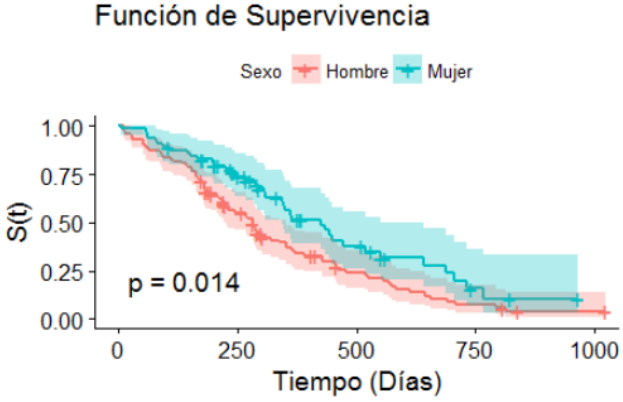
4. Análisis Descriptivo

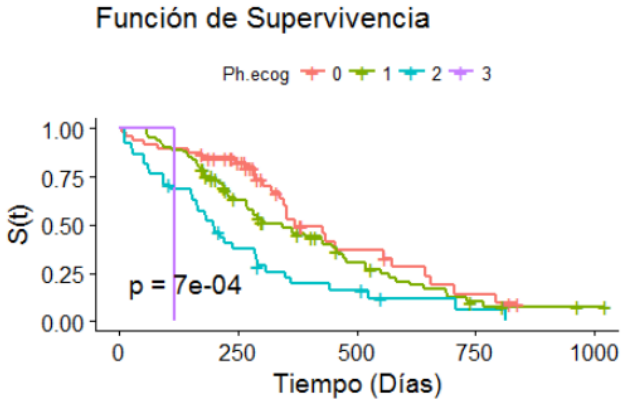
Lo primero que vamos a hacer es analizar la función de supervivencia general.



La función de supervivencia obtenida tiene una forma de descenso casi lineal, ya que no presenta ningún salto y parece tener una pendiente constante. Además una vez llegado a un tiempo mayor que 800 días, parece estabilizarse en torno al 0 por lo que a esa fecha, la mayoría han fallecido.

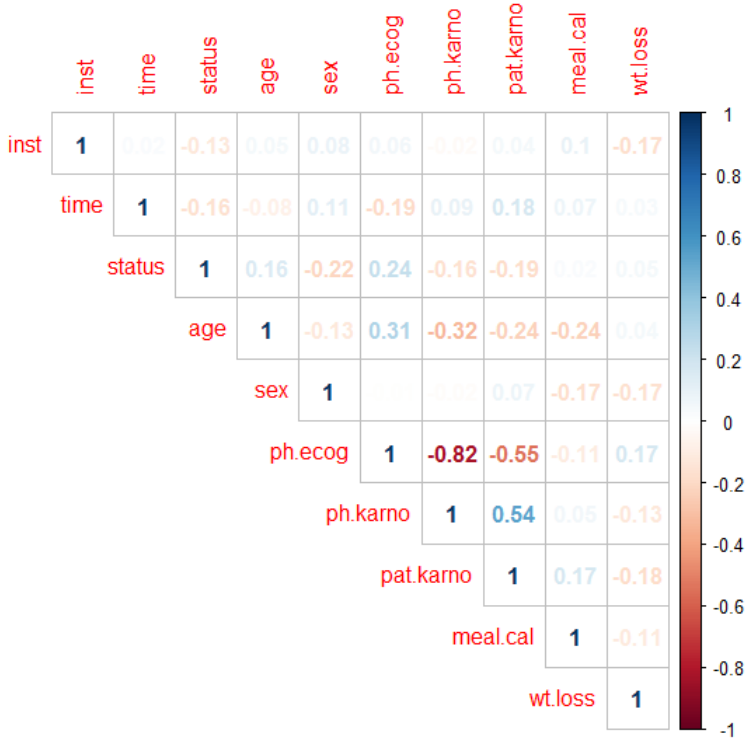
A continuación analizaremos las curvas para las variables **sexo**, que presenta dos niveles: 1 para hombre y 2 para mujer, y **ph.ecog** que en la descripción del dataset vemos que tiene un rango de valores de 0 a 5, que va de bueno a malo siendo el último nivel la muerte. Comentar que únicamente hay valores de 0 a 2 en el conjunto de los datos y un caso de nivel 3.





Para la variable **sexo**, la curva asociada a la mujer se mantiene por encima que la de el hombre. Además, rechazamos que las distribuciones para los dos factores sean iguales (pvalor de 0.014). En el caso de la variable **ph.ecog** no presentamos los intervalos de confianza ya que para el valor 3 hay solo una observación. El pvalor asociado al test de hipótesis es prácticamente 0, por lo que rechazamos que se comporten de la misma manera. Se ve que a mayor nivel de la variable menor supervivencia.

También nos parece interesante analizar las correlaciones de las variables para observar posible multicolinealidad, es decir, que toda la variabilidad que podría explicar una variable haya sido ya explicada por otra con la que esté fuertemente correlada.



Podemos destacar cómo están fuertemente correladas las variables **ph.ecog** y **ph.karno**, ya que ambas representan la calidad de vida de los pacientes. Es de destacar que la correlación es negativa porque en la primera variable el valor mínimo (el 0) es el mejor de cara a la supervivencia y por el contrario, en la segunda es el valor máximo (el 100). También con **pat.karno** presentan correlaciones significativas y esto se tendrá en cuenta a la hora de ver la significancia de las variables al ajustar el modelo de Cox.

5. Modelo de Cox

Vamos a realizar el modelo de Cox con el objetivo de descubrir qué variables influyen en el análisis de manera significativa. Para ello, vamos a usar el dataset completo eliminando los individuos que presentan NAN's. De esta forma, se reduce el número de observaciones de 228 a 167.

Primeramente, tenemos que validar la hipótesis de Riesgos Proporcionales.

	rho	chisq	p
inst	-0.0697	0.721	0.3958
age	0.0869	1.033	0.3094
sex	0.1250	1.777	0.1825
ph.ecog	0.0633	0.559	0.4547
ph.karno	0.1913	3.360	0.0668
pat.karno	0.0428	0.268	0.6049
meal.cal	0.1572	3.633	0.0566
wt.loss	0.0532	0.484	0.4868
GLOBAL	NA	11.804	0.1602

Los datos cumplen la hipótesis de riesgos proporcionales ya que vemos que no rechazamos el pvalor global ni el de las covariables. Esto nos permite continuar con el modelo de Cox aunque deberíamos tener en cuenta para el posterior análisis que el pvalor no es sumamente grande como para hacer una afirmación rotunda.

El test proporcionado en el modelo de Cox evalúa la hipótesis nula de que los distintos coeficientes de las covariables sean 0. Además, proporciona tests para comprobar la significancia de estas variables en el modelo.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
inst	-3.037e-02	9.701e-01	1.312e-02	-2.315	0.020619	*
age	1.281e-02	1.013e+00	1.194e-02	1.073	0.283403	
sex	-5.666e-01	5.674e-01	2.014e-01	-2.814	0.004890	**
ph.ecog	9.074e-01	2.478e+00	2.386e-01	3.803	0.000143	***
ph.karno	2.658e-02	1.027e+00	1.163e-02	2.286	0.022231	*
pat.karno	-1.091e-02	9.891e-01	8.141e-03	-1.340	0.180160	
meal.cal	2.602e-06	1.000e+00	2.677e-04	0.010	0.992244	
wt.loss	-1.671e-02	9.834e-01	7.911e-03	-2.112	0.034647	*

Se aprecia que las variables **inst**, **sex**, **ph.ecog**, **ph.karno** y **wt.loss** son las significativas, el resto no lo son ya que proporcionan un p-valor demasiado alto que indican que no se rechaza la hipótesis de que sean 0. El p-valor del modelo global es 0 ya que sí existen covariables significativas.

Pasando a analizar las covariables significativas, hay que fijarse en las variables **coef** y **exp(coef)**. Por ejemplo, para la variable **sex**, el signo negativo de coef indica que a valores mayores de estas covariables, menor es el riesgo de muerte. La variable **sex** disminuye el riesgo de muerte, es decir, las mujeres (que se codifican con el número 2) tienen menor riesgo de morir que los hombres (que se codifican con el número 1). Para saber en qué medida afecta la covariable hay que mirar la variable **exp(coef)**, ser mujer disminuye el riesgo de muerte por un factor de 0.57, es decir, las mujeres tienen algo más de la mitad de riesgo de morir que los hombres.

Podríamos querer efectuar la regresión solo con las variables consideradas significativas para ver la diferencia. Para ello, volvemos a realizar el test de riesgos proporcionales y así confirmar que se pueda realizar el modelo de Cox sobre las variables que hemos considerado significativas.

	rho	chisq	p
inst	-0.0831	0.9720	0.3242
sex	0.0993	1.1098	0.2921
ph.ecog	0.0263	0.0926	0.7609
ph.karno	0.1818	3.0106	0.0827
wt.loss	0.0503	0.4122	0.5208
GLOBAL	NA	7.7289	0.1718

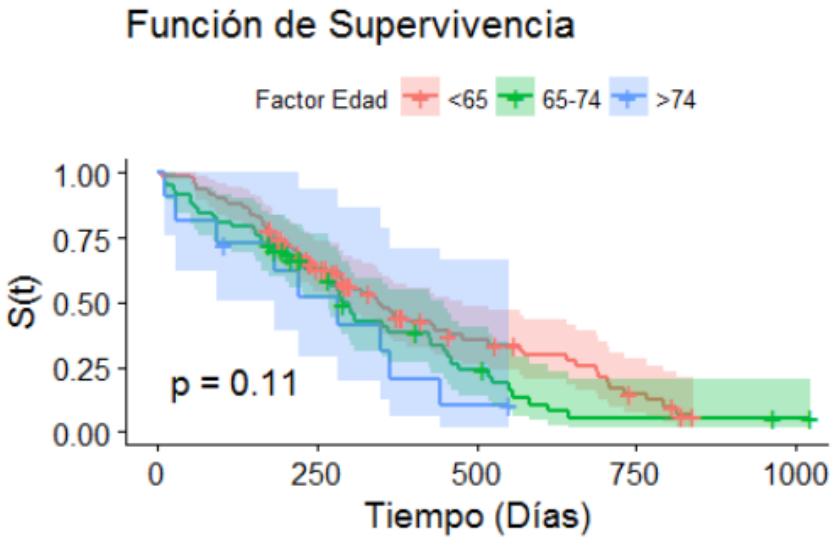
Se vuelve a cumplir la hipótesis ya que todos los p-valores tienen un valor suficientemente elevado. Además apenas han cambiado los coeficientes para las diferentes covariables. Para los residuos utilizaremos el modelo completo.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
inst	-0.030042	0.970404	0.012931	-2.323	0.02016	*
sex	-0.571959	0.564419	0.198865	-2.876	0.00403	**
ph.ecog	0.993224	2.699926	0.232115	4.279	1.88e-05	***
ph.karno	0.021492	1.021725	0.011222	1.915	0.05547	.
wt.loss	-0.014800	0.985309	0.007664	-1.931	0.05348	.

Los resultados obtenidos nos aportan como retroalimentación que todas las covariables que hemos seleccionado en primera estancia siguen siendo significativas aunque algunas en mayor medida que otras, por ejemplo **ph.karno** y **wt.loss** no son prácticamente significativas en este caso a nivel 0.05. En el caso de **pat.karno**, que no ha sido incluída en este modelo, ya habíamos comentado que estaba fuertemente correlada con **ph.ecog** y **ph.karno** y es probable que con esas variables se haya explicado toda la variabilidad posible en el modelo.

También hemos probado si existía alguna interacción que fuese significativa y pudiese ser metida en el modelo pero no hemos encontrado ninguna cuyo p-valor sea lo suficientemente grande.

Finalmente nos ha sorprendido que la variable **edad** no sea significativa en ninguno de los dos modelos, por tanto planteamos una discretización de esta para observar posibles resultados que nos contradigan esto. Utilizaremos una división estándar en el ámbito médico que es la correspondiente a menores de 65 años, de 65 a 74 y mayores de 74. Una vez hecho esto analizamos la gráfica de su función de supervivencia.

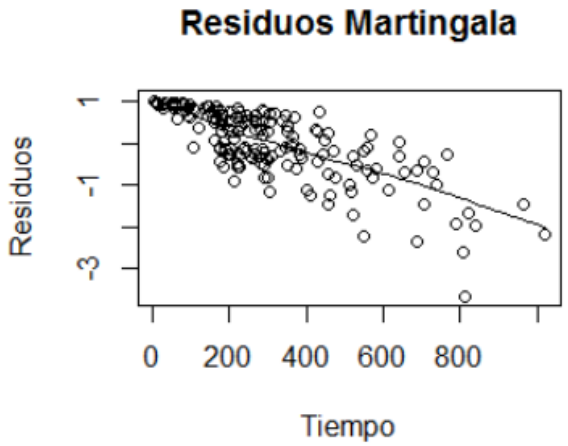


No podemos rechazar que los grupos que hemos elaborado sean significativamente diferentes por tanto confirmamos que la **edad** no influye en nuestros modelos.

6. Análisis de los Residuos

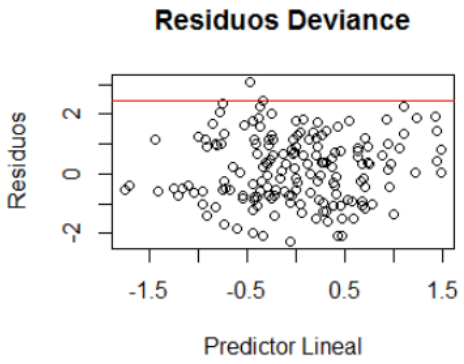
Dado el poco tiempo que tuvimos en clase para explicar los distintos tipos de residuos que podemos obtener para analizar el ajuste del modelo, vamos a profundizar más en este tema.

- **Residuos Martingala**: Asociados a la descomposición de Doob (valor Esperado - valor Observado). Son una transformación de los residuos de Cox-Snell en el intervalo (-∞,1]. Son útiles para valorar la forma funcional de las variables explicativas.



Este gráfico es más complicado de analizar que si los transformáramos para analizarlos mejor como vamos a hacer en el siguiente punto.

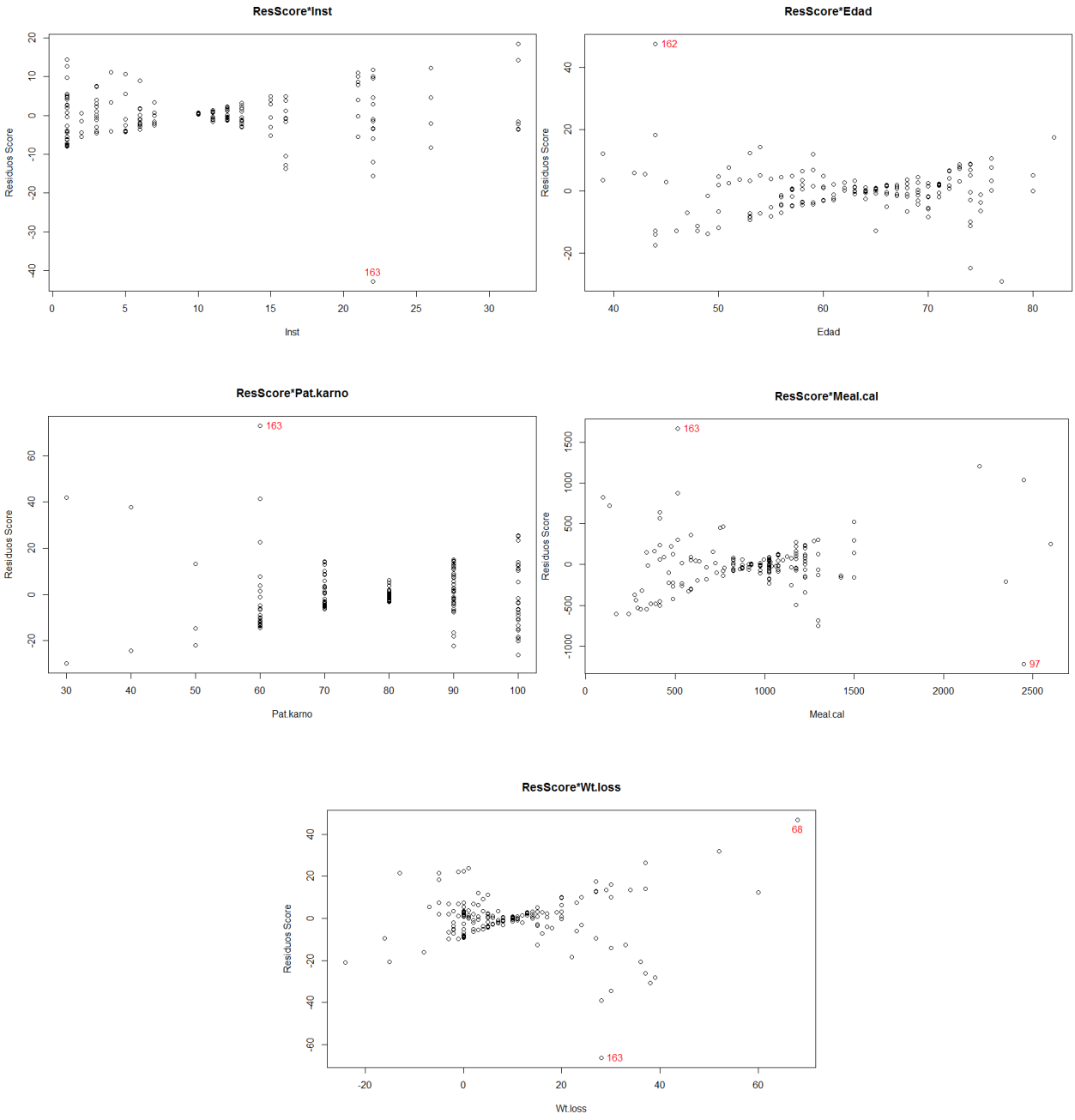
- **Residuos Deviance**: Transformación de los residuos martingala para hacerlos simétricos en torno al 0. Son útiles para el análisis de outliers. Debido a que la disposición se aproxima a una distribución Gaussiana podemos calificar de observaciones atípicas aquellas con valores fuera de [-2.5,2.5] aprox. Con un gráfico de residuos frente predichos podemos ver esto claramente.



	resids.dev	time	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
57	3.097894	5	0	100	80	338	5
111	2.467873	15	0	90	70	575	10

Las dos observaciones que podríamos catalogar como outliers son las que ocupan el 1er y 5o puesto en menores tiempos desde detección del cáncer hasta la muerte y sin embargo los valores de los indicadores de calidad de vida **ph.ecog** y **ph.karno** son muy buenos, por tanto efectivamente vemos que son dos observaciones atípicas.

- **Residuos Score**: Se calculan para cada individuo y variable explicativa y se interpreta como la diferencia media entre el valor de una covariable para este caso y el valor medio de esta covariable. Estos residuos estudian la influencia de observaciones en las variables "leverage"



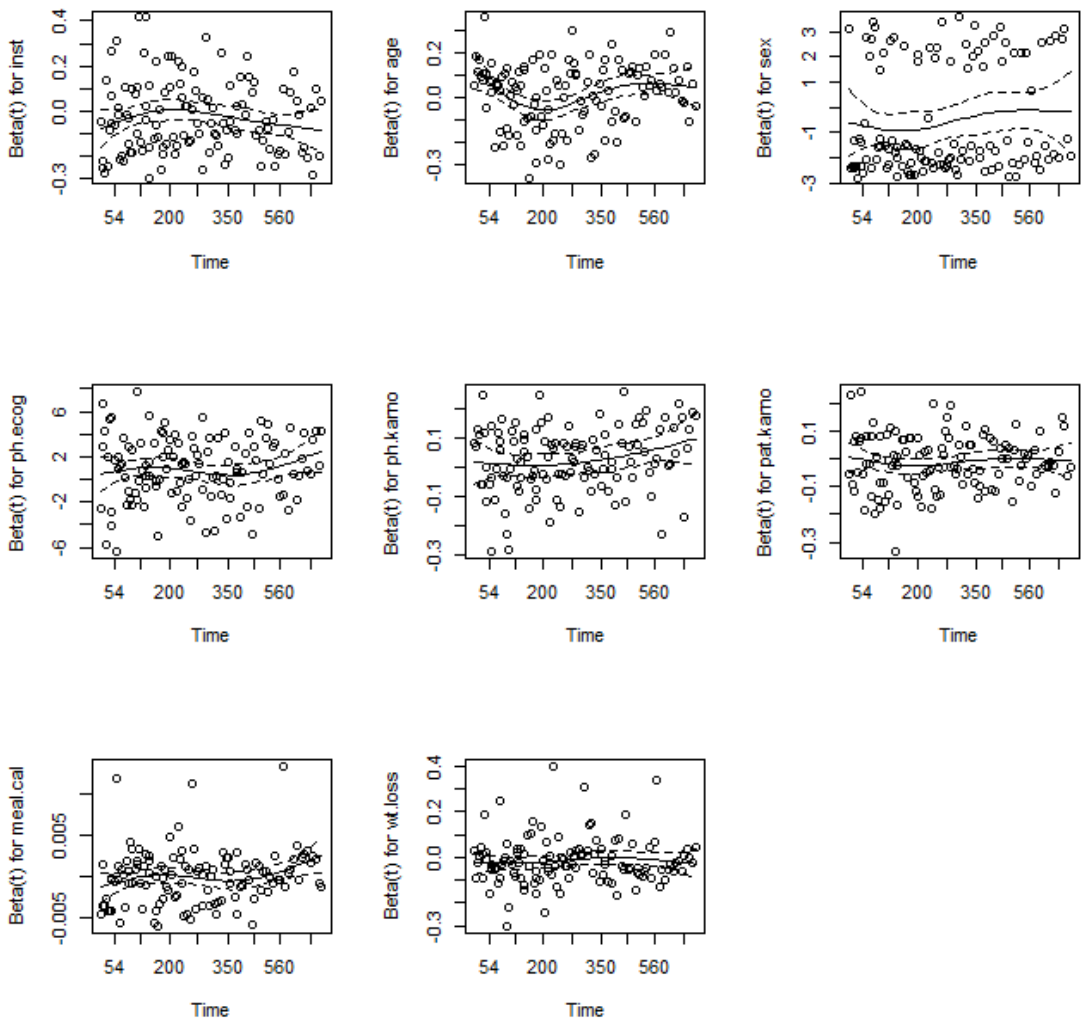
Vemos que destacan diversas observaciones por su influencia:

- Obs 163: Esta observación es realmente importante ya que es influyente para las variables Inst,Pat.karno,Meal.cal y Wt.loss
- Obs 162: Observación influyente para la variable Edad
- Obs 97: Observación influyente para la variable Meal.cal

- Obs 68: Observación influyente para la variable Wt.loss

- **Residuos Schoenfeld**:Son los más efectivos en cuanto a detectar anomalías para cada una de las variables que intervienen en el modelo sugiriendo posibles transformaciones para los datos. También se calculan para cada individuo y variable.

  Para facilitar la interpretación de estos gráficos se superpone una curva de ajuste alisada por splines junto con dos líneas adicionales +-2 error estándar. Si la hipótesis de riesgos proporcionales se cumple, los residuos deberían agruparse de forma aleatoria a ambos lados del valor 0 y la curva ajustada debería ser próxima a una recta.



Por ejemplo para las variables **wt.loss** y **pat.karmo** es donde más se nota que son prácticamente rectas. Sin embargo la variable edad, que no era significativa, es la que más claramente vemos que no cumple esto.

La hipótesis de riesgos proporcionales no había sido rechazada, aunque tampoco con mucha firmeza y esto se refleja observando los resultados del análisis de los residuos que no son tan idóneos como deseábamos.