

Bioestadística (Master en Bioinformática)

Bloque: AMECP

Análisis de Modelos Estadísticos de Comparación y Predicción

Sesión: Análisis de regresión lineal

Manuel Franco
Dpto. Estadística e Investigación Operativa

Índice

1. Regresión lineal simple	1
1.1. Estimación y análisis del modelo	2
1.2. Predicción del modelo de regresión	5
1.3. Análisis de los residuos	6
1.4. Ajustes de regresión linealizables	8
1.5. Caso práctico	10
2. Regresión lineal múltiple	20
2.1. Interpretación y estimación del modelo	20
2.2. Análisis del modelo de regresión múltiple	21
2.3. Selección del modelo de regresión	23
2.4. Caso práctico	25

1. Regresión lineal simple

El Análisis de Regresión es una técnica estadística que se encarga de establecer la relación entre un conjunto de variables cuantitativas, disponibles en nuestro campo de estudio experimental a través de sus observaciones o datos registrados, analizando la relevancia de cada una de ellas en la relación con el objetivo de estimar o predecir una variable respuesta o dependiente de tipo continuo. En nuestro caso, representa un mecanismo para inferir o pronosticar una característica o variable de especial interés, con un elevado coste o riesgo asociado en su observación, mediante un modelo estadístico con una o varias características con bajos costes o riesgos asociados en su medición.

Un paso previo recomendable, para llevar a cabo un análisis de regresión, sería el estudio de dependencia, por ejemplo analizando el coeficiente de correlación visto en un tema anterior, dado que es lógico pensar que si no existe relación alguna entre las variables disponibles en el estudio, el modelo de regresión no será de utilidad. En este caso, la variable respuesta Y se representa a través de su media y las desviaciones sobre dicha media:

$$Y = E(Y) + \varepsilon$$

En cambio, si existe una gran relación de dependencia entre la variable respuesta y el resto de variables (llamadas independientes o predictoras) entonces el modelo de regresión que representa dicha relación

podría ser útil para predecir la variable respuesta. Así, se podría decir que esta metodología estadística consiste en la obtención de las leyes experimentales (modelo de regresión) que rigen la relación entre variables observables, y por consiguiente, estarán basadas en la experiencia (en las observaciones sobre los individuos) y sujetas a desviaciones con respecto al conjunto de toda la población

$$\text{"Respuesta = Modelo + Desviación"} \Leftrightarrow Y = E(Y|X) + \varepsilon,$$

es decir, el modelo es la representación del valor medio de la variable respuesta a través de la información del resto de variables observables.

En este contexto, nos centraremos en el análisis de regresión lineal, que ha tenido una gran importancia en las aplicaciones prácticas, y en general, puede considerarse como una aproximación en los casos de relación no lineal entre las variables en estudio. Así, comenzaremos con el modelo de regresión lineal simple, su aplicación a modelos no lineales, y por último, veremos la extensión al modelo de regresión lineal múltiple.

El análisis de regresión lineal simple pretende estimar y predecir una variable respuesta Y medida sobre los individuos de la población a través de una ecuación lineal de una variable predictora X (una recta)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

donde incluimos un término de desviación ε pues como hemos mencionado, no será una relación exacta, ya que estará determinada por las observaciones experimentales sobre los individuos objeto de estudio, debiéndose incluir este término que representa la desviación o error, y que engloba el aspecto experimental de la relación buscada y el error propio del modelo considerado al no contemplar todas los posibles variables o factores de riesgo sobre la variable Y .

Además, este modelo de regresión lineal se plantea como consecuencia de la relación entre ambas variables, respuesta Y y predictora X , detectada a través de las observaciones experimentales (también llamadas output e input), es decir, a través de los datos de la muestra (x_i, y_i) con $i = 1, \dots, n$, para lo que resulta de gran ayuda la representación de la nube de puntos o diagrama de dispersión (visto en un tema anterior) para describir el tipo de relación entre ambas características, dado que una gráfica de puntos con cierta tendencia lineal se interpreta como una "buena" relación lineal. Así, Y_i representa la variable Y medida sobre el individuo i observado en la población y para la que su valor observado ha sido y_i cuando la variable predictora ha tomado el valor $X = x_i$, y por tanto

$$Y_i = E(Y_i | X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{para } i = 1, \dots, n$$

es el modelo de regresión lineal simple, siendo $\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i$ el error o residuo aleatorio para cada individuo i , producido al ajustar la variable respuesta Y por la recta $\beta_0 + \beta_1 x_i$, siendo el valor experimental de este error $e_i = y_i - \beta_0 - \beta_1 x_i$ (ver Figura 1).

1.1. Estimación y análisis del modelo

1.1.1. Términos del modelo y condiciones iniciales

Como vemos en el modelo, hay una relación directa entre los coeficientes del modelo y la importancia o relevancia de la variable predictora en el modelo de regresión lineal en su relación con la respuesta, dado que

- β_1 representa el efecto medio (positivo o negativo) sobre la variable respuesta al aumentar la predictora en una unidad, $E(Y | X = x + 1) = E(Y | X = x) + \beta_1$, y se llama pendiente o coeficiente de regresión
- β_0 representa el valor medio de la variable respuesta cuando la predictora es cero, y se llama constante o intercepto: $E(Y | X = 0) = \beta_0$

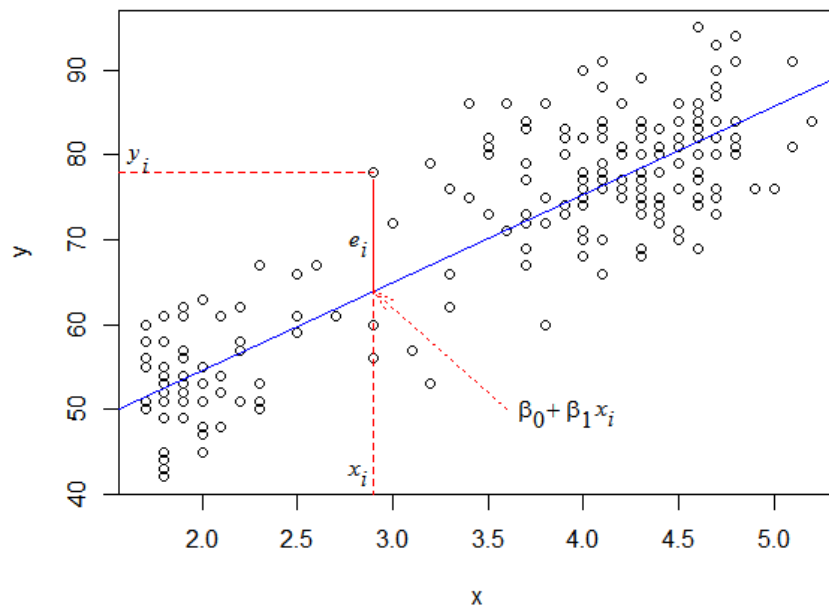


Figura 1: Representación de residuo en una nube de puntos

y evidentemente, si $\beta_1 = 0$ la variable predictora no interviene en el modelo.

Además, el desarrollo del análisis del modelo de regresión (la estimación del modelo, los contrastes de significación y la predicción de la respuesta) requiere de unas ciertas condiciones que deben mantenerse si el experimento se ha realizado en unas "buenas" condiciones que garantiza la aleatoriedad de los individuos observados y su representatividad en la población bajo estudio, llamadas condiciones iniciales:

- **linealidad:** existencia de la relación lineal entre las variables, es decir, error esperado o medio nulo
- **homogeneidad de varianzas:** igualdad de variabilidad o dispersión, es decir, observaciones bajo idénticas condiciones
- **incorrelación:** aleatoriedad en las observaciones o residuos, es decir, no presentan situaciones dinámicas o temporales dependientes
- **normalidad:** modelo de distribución supuesto sobre la respuesta o residuo para medir la incertidumbre

En toda aplicación del modelo de regresión, se deben analizar estas condiciones iniciales, lo que constituye la fase de análisis o diagnóstico de los residuos.

1.1.2. Recta de regresión: Modelo estimado

Una vez planteado el modelo de regresión lineal simple para predecir la variable respuesta a través de dicha ecuación lineal, se procede a la estimación de los coeficientes constante (intercepto) y pendiente del modelo para determinar, de entre todas las posibles rectas, la que mejor se ajusta en algún sentido a la información sobre ambas obtenida en la muestra, este modelo se llama recta de regresión

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \Leftrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \text{ para } i = 1, \dots, n$$

donde \hat{y} es el **modelo estimado** o **ajustado**, y para cada observación muestral i , \hat{y}_i son los valores ajustados o pronosticados a través del modelo para la variable respuesta, cuya observación registrada ha sido y_i , y por tanto, el error o residuo con el modelo ajustado es $\hat{e}_i = y_i - \hat{y}_i$.

Los valores estimados de los coeficientes del modelo, $\hat{\beta}_0$ y $\hat{\beta}_1$, se obtienen por el criterio de mínimos cuadrados (ver tema anterior), y coinciden con el procedimiento de máxima verosimilitud, bajo la condición de normalidad,

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \text{ y } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

siendo el error acumulado de los desajustes entre la recta de regresión y las observaciones de la respuesta (suma de cuadrados de los residuos o **mínimo error cuadrático**)

$$SS_R = \sum_{i=1}^n \hat{e}_i^2 = ns_y^2(1 - r_{xy}^2)$$

una medida de la variación del error experimental obtenida a través del modelo de regresión lineal, y depende del coeficiente de correlación lineal entre X e Y . Así, puede observarse que si la relación lineal es perfecta ($r_{xy} = -1$ ó 1) la suma de cuadrados del error es nula; sin embargo, si no hay relación entre ellas ($r_{xy} = 0$) entonces alcanza su máximo valor que representa toda la variabilidad o dispersión de la respuesta. Y por tanto, SS_R se interpreta como la variabilidad de la respuesta en el experimento que no queda explicada por el modelo de regresión.

1.1.3. Contrastes de significación

En la presentación de los resultados de un análisis de regresión lineal, resulta imprescindible la realización e interpretación de los contrastes de significación individual y conjunta de los términos incluidos en el modelo, es decir, analizar la importancia o relevancia individual de cada variable predictora o explicativa utilizada en el modelo para predecir la respuesta, y la relevancia global, que en este caso es equivalente a la individual al considerar una regresión simple (sólo una variable predictora). Esto también proporciona uno de los mecanismos que veremos para la selección del modelo de regresión en el caso de múltiples variables predictoras.

Los **contrastos individuales** que se realizan en la práctica son si el coeficiente constante o intercepto del modelo es cero frente a la alternativa de que no lo sea, y el contraste de hipótesis nula que el coeficiente pendiente de la recta sea cero frente a que no lo sea. Ambas pruebas se desarrollan a través de un estadístico t de Student con $n - 2$ grados de libertad, y con regiones de rechazo de dos colas. Por ejemplo, el contraste del intercepto $\beta_0 = 0$ permite discutir si es o no significativo dicho término en el modelo, decisión que adoptaremos a través del p -valor asociado a la muestra experimental disponible, y su valor próximo a cero mostrará su significación, es decir, su relevancia en el modelo.

En particular, cabe destacar el contraste de la pendiente $\beta_1 = 0$, ya que representa la importancia o relevancia directa entre la variable predictora y la respuesta, es decir, la utilidad de X para explicar a la variable Y a través de este modelo de regresión lineal. Un p -valor próximo a cero indicará que las observaciones están significativamente en contra de que $\beta_1 = 0$, y por tanto al rechazar esta hipótesis, se concluye que $\beta_1 \neq 0$ o equivalentemente, que X debe mantenerse en el modelo de regresión.

Observar que en este caso de regresión lineal simple, hemos comentado que la significación conjunta se reduce a la individual de la variable predictora. No obstante, comentamos brevemente como realizar el análisis de significación conjunto, llamado **contraste de regresión** pues analiza la importancia de la variables regresoras en el modelo, aunque en este caso sea equivalente al contraste anterior de $\beta_1 = 0$, servirá para ver su presentación en el caso múltiple. Así, este contraste de regresión se basa en la variabilidad o dispersión de la variable respuesta dada a través de la información recogida en las observaciones, y en su descomposición mediante la componente de variabilidad explicada por el

modelo de regresión estimado \hat{y} (suma de cuadrados de la regresión) y la variabilidad no explicada por el modelo (suma de cuadrados de los residuos),

$$SS_T = SS_E + SS_R \Leftrightarrow \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

lo que permite realizar el contraste de regresión mediante el estadístico F mostrado en la Tabla 1, que sigue una distribución F de Snedecor con $(1, n - 2)$ grados de libertad bajo la hipótesis nula del contraste de regresión.

Tabla 1: Tabla ANOVA del contraste de regresión

Fuente	S.Cuadrados	G.Libertad	M.Cuadrados	Estadístico	p-valor
Regresión:	SS_E	1	$MS_E = SS_E$	$F = \frac{MS_E}{MS_R}$	p
Residuos:	SS_R	$n - 2$	$MS_R = \frac{SS_R}{n-2}$		
Total:	SS_T	$n - 1$			

En esta Tabla 1 se observa que el estadístico F está basado en la tasa entre la variabilidad explicada y la no explicada por el modelo, de modo que a mayor valor de la componente explicada por el modelo respecto de la no explicada, más significativo o relevante será el modelo de regresión, y más pequeño será el p -valor correspondiente a esas observaciones.

1.1.4. Coeficiente de determinación

En relación a la descomposición de la variabilidad utilizada en el contraste de regresión, resulta inmediato definir y construir una medida de bondad del modelo de regresión, es decir, un indicador que permita cuantificar lo "bueno" que es el modelo de regresión para predecir la variable respuesta. Este indicador viene dado por el **Coeficiente de Determinación**

$$R^2 = \frac{SS_E}{SS_T}$$

y que representa la proporción de variabilidad explicada por el modelo de regresión con respecto a la total de la variable respuesta.

Evidentemente, dado que R^2 es un cociente de términos positivos, su mínimo valor es 0 y la proximidad de R^2 a cero indica que el modelo no es bueno (explica poco o nada de la respuesta). Además, su mejor valor es 1 y la proximidad de R^2 a uno indica que el modelo explica gran parte de la variabilidad de la respuesta. En general, esta medida de bondad del modelo suele presentarse en términos porcentuales. Y en el caso de regresión lineal simple, puede comprobarse que el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación.

En la práctica, es frecuente incluir una alternativa del coeficiente de determinación, llamado coeficiente de determinación ajustado, $R_a^2 = 1 - \frac{n-1}{n-2}(1 - R^2)$, con el fin de tener en cuenta la dimensión del modelo de regresión analizado, y cuya interpretación es la misma que el coeficiente de determinación.

1.2. Predicción del modelo de regresión

Una vez hemos especificado, estimado y contrastado un modelo de regresión, el principal objetivo del mismo es utilizarlo en la estimación y predicción de la variable respuesta Y .

En este caso, a través de la recta de regresión (modelo ajustado) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, obtenemos la estimación puntual para el valor medio de la respuesta cuando la predictora $X = x$.

No obstante, la estimación de la media de la respuesta $\mu_{Y_x} = E(Y|X = x) = \beta_0 + \beta_1 x$ puede realizarse mediante un intervalo de valores factibles con un nivel de confianza $1 - \alpha$ prefijado, a través de las distribuciones muestrales de los estadísticos que intervienen en este modelo ajustado. Observar que en realidad, se trata de una colección de intervalos de confianza para cada valor x de la variable predictora en el que se estima μ_{Y_x} mediante $\hat{\beta}_0 + \hat{\beta}_1 x$, por lo que se le conoce como banda de confianza, su amplitud varía con respecto al valor de x y está centrada sobre la recta de regresión, siendo menor la amplitud cuando x está próximo al centro de la masa de información observada entre las dos variables, es decir, si $x \rightarrow \bar{x}$ hay mayor precisión en la estimación, y la amplitud aumenta al distanciarse x de dicho centro, es decir, hay menor precisión.

Esta observación deja al descubierto el problema de extrapolación, el error que se cometería si el modelo de regresión ajustado a través de las observaciones que contienen la información disponible entre las variables, se utiliza fuera del rango de valores observado, es decir, para valores x en los que no se dispone de información alguna entre las variables, incluso podría ser inadecuado un modelo de regresión lineal fuera de dicho rango.

Por otro lado, también puede desarrollarse la predicción del valor de la variable a través de la aproximación proporcionada por el modelo ajustado $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ para cada valor $X = x$. En este caso, se observa que disponemos de la misma herramienta (estadístico) para estimar la media y el valor de la respuesta, siendo lógico pensar que habrá mayor incertidumbre a la hora de aproximar la variable (tiene mayor variabilidad) que en la aproximación de la media, es decir, se pierde precisión en el proceso de predecir la respuesta, lo que significa que los intervalos de predicción estarán centrados sobre la recta de regresión pero con una mayor amplitud, constituyendo una banda de predicción que contiene a la banda de confianza.

1.3. Análisis de los residuos

El análisis de los residuos es fundamental en el estudio del modelo de regresión lineal, dado que todo su desarrollo parte de la suposición de unas condiciones iniciales que deben ser comprobadas, y de esto se encarga el análisis de los residuos. No obstante, las hipótesis iniciales pueden resultar imposibles de contrastar en muchos casos prácticos, debido a la falta de información muestral suficiente, por lo que en estas situaciones conviene disponer de un mecanismo que permita al menos detectar las posibles desviaciones con respecto a estas hipótesis iniciales. Observar que este apartado está dentro del análisis del modelo de regresión lineal simple, por lo que nos centraremos en la interpretación del análisis de los residuos para este modelo; sin embargo, los elementos del análisis de residuos gráficos y analíticos son generales, es decir, no están restringidos a un número concreto de variables predictoras, dado que se desarrolla a partir de los propios residuos (desajustes entre respuesta y modelo) y los ajustes (modelo), independientemente de los términos que intervengan en el modelo.

Así, en la práctica es habitual el uso de las gráficas residuales, para realizar un análisis descriptivo gráfico de los residuos obtenidos a través del modelo de regresión ajustado (recta de regresión), teniendo en cuenta que no es determinante y se requiere de los contrastes para tomar las decisiones sobre las condiciones con garantías (con un nivel de significación).

El gráfico de residuos frente los valores ajustados representa la nube de puntos de los residuos, y es utilizado para describir la linealidad y homogeneidad de varianzas, detectando las tendencias no centradas de los residuos y las diferencias de dispersión a lo largo de la nube de puntos. El gráfico de los residuos ordenados representa las secuencias de signos entre los residuos de forma temporal, y se utiliza para detectar las dependencias que pudieran existir. También se utiliza la representación de los residuos en un papel probabilístico normal, pero esta representación puede acompañarse del contraste correspondiente. En la Figura 2 se muestra un ejemplo de situaciones extremas de estas gráficas.

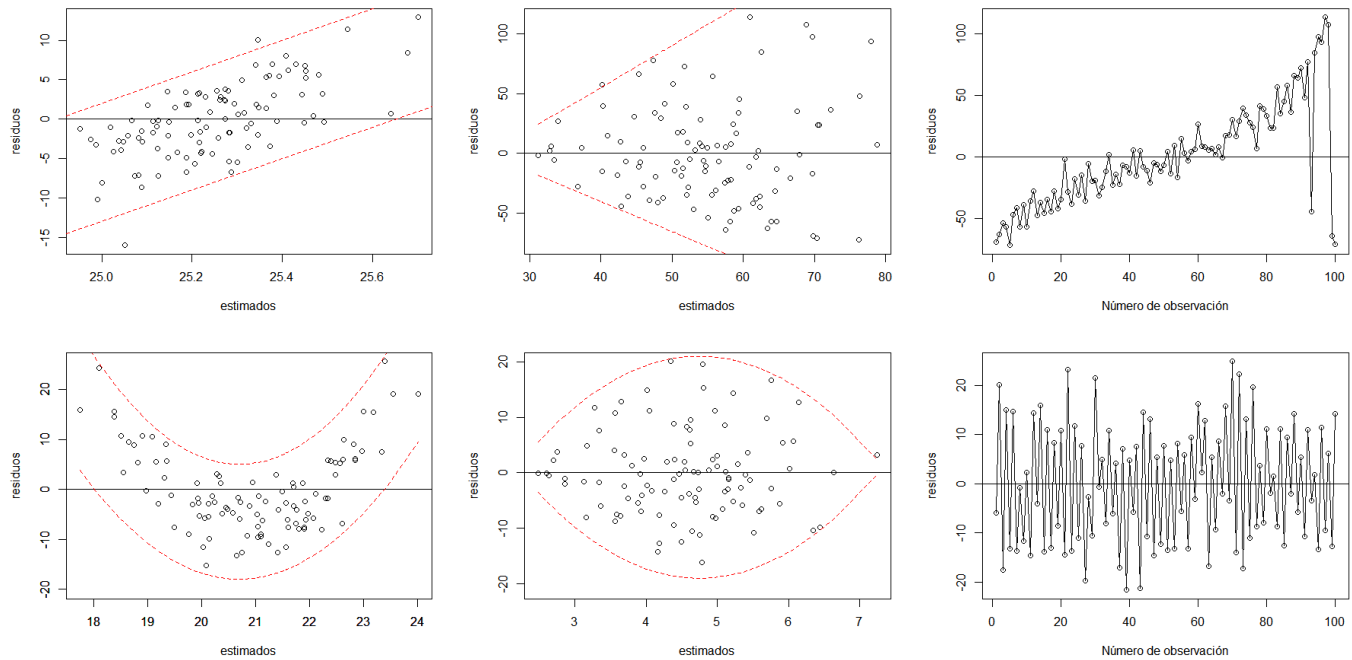


Figura 2: Situaciones del análisis gráfico de los residuos

1.3.1. Contrastes de los residuos

Respecto de los contrastes de hipótesis para estas condiciones, el **test de linealidad** o falta de ajuste contrasta la hipótesis nula de linealidad, que el modelo de regresión lineal representa bien a la variable respuesta, es decir, que el error medio es nulo, lo que equivale a que la media de la variable respuesta es el modelo de regresión, $\mu_{Y_x} = E(Y|x) = \beta_0 + \beta_1 x$. Este contraste se desarrolla a través de la descomposición del error o residuo en las componentes de error debido al desajuste del modelo (falta de ajuste) y del error debido a la aleatoriedad en toda experimentación (error puro), el estadístico del contraste se basa en el cociente de estos dos tipos de errores, y se suele presentar incluido en la Tabla 1 del contraste de regresión. La particularidad de este contraste de ajuste consiste en que para poder obtener las dos componentes de error, se necesita que exista al menos una observación "repetida" en el sentido de que al menos dos individuos de la muestra observada registren la misma medida de la variable predictora, y por tanto se disponga de varios valores de la respuesta para un mismo valor de la predictora.

El **contraste de homogeneidad de varianzas** es algo más restrictivo, dado que para realizar una comparación de las varianzas debe ser posible estimarlas en cada grupo formado por los diferentes valores de la variable predictora, es decir, requiere que las observaciones contengan más de un valor de la respuesta por cada uno de los diferentes valores de la predictora. En este caso, se puede realizar el contraste de hipótesis nula que las varianzas son iguales, $\sigma_1^2 = \dots = \sigma_k^2$, existiendo diversas alternativas, entre las que destacan el test de Bartlett bajo la condición de normalidad, y el test de **Levene** bajo desviaciones de la normalidad.

En relación a la condición de **incorrelación**, el estadístico más frecuente es el de Durbin-Watson, basado en las autocorrelaciones entre residuos adyacentes, y bajo la hipótesis nula de incorrelación, el estadístico D tiene una distribución simétrica centrada en el punto 2 y acotada en el intervalo $(0, 4)$. Así, el valor del estadístico D próximo a los extremos del intervalo indica una tendencia de autocorrelación (positiva o negativa, según la asimetría) y un valor próximo a 2 no detectaría una falta de incorrelación. También pueden utilizarse otros estadísticos de autocorrelación menos habituales en análisis de regresión, como el de Ljung-Box, que incluye las posibles autocorrelaciones entre residuos

no adyacentes.

Por último, la condición de **normalidad** puede contrastarse con las técnicas utilizadas en los temas anteriores.

1.3.2. Estadísticos de influencia

El diagnóstico del modelo se completa con el análisis de las observaciones influyentes en el modelo de regresión, que consiste en determinar o identificar que observaciones tienen mayor influencia en el modelo de regresión estimado, y por tanto merecen una comprobación y discusión. En este sentido, los estadísticos de influencia más frecuentes utilizan las medidas de apalancamiento o *leverages* de cada observación, que reflejan el distanciamiento de cada observación con respecto al nivel medio de las restantes, siendo su suma el número de términos en el modelo. Así, un valor de *leverage* próximo a uno representa una observación con un fuerte efecto palanca, desplazando el centro de masa de información contenida en las observaciones. Por ejemplo, en el caso de regresión lineal simple, los *leverages* $h_i = \frac{1}{n} \left(1 + \frac{(\bar{x} - x_i)^2}{s_x^2} \right)$ suman $k = 2$, y una observación se considera influyente cuando su *leverage* es superior a $3k/n$ (en este caso $6/n$).

A partir de estos niveles de apalancamiento o *leverage*, se definen los diferentes estadísticos de influencia, como el estadístico de Cook que detecta las observaciones influyentes en el modelo de regresión, es decir, en el conjunto de los coeficientes estimados en el modelo, y su representación gráfica ayuda a detectar y discutir las observaciones más influyentes en el modelo (mayor a 1). También pueden utilizarse los residuos estandarizados para detectar las observaciones atípicas y las desviaciones estandarizadas de los ajustes para cada observación. En general, las diferentes medidas de influencia pretenden identificar las observaciones experimentales diferenciadas del resto y que por tanto su influencia puede provocar errores en los contrastes de significación.

1.4. Ajustes de regresión linealizables

En el modelo de regresión lineal simple, se determina el mejor modelo que explica la relación entre dos variables con el objetivo de predecir o pronosticar una variable respuesta Y que puede ser de mayor coste, riesgo o exigir una prueba invasiva, a través de una variable predictora X más fácil, económica e incluso no invasiva. En el análisis de esta recta de regresión se incluye la estimación, pruebas de significación y diagnóstico de los residuos.

En este punto cabe destacar que en la práctica pueden resultar significativos los contrastes individuales y de regresión del modelo, y sin embargo no serlo las pruebas de diagnóstico de los residuos, por ejemplo el contraste de linealidad, como en el caso anterior. Esto significa que la variable predictora X se considera relevante para explicar la variable respuesta a través de un modelo de regresión, pero no lo suficiente para que la recta de regresión se ajuste bien y sea útil para la predicción. En este sentido, puede plantearse dos situaciones, por un lado que X es significativa pero no lo suficiente por sí sola para predecir la variable respuesta, siendo necesario mejorar el modelo de regresión con otras variables predictoras (modelo de regresión lineal múltiple), o por otro lado, que X es significativa pero no linealmente, es decir, sería necesario analizar un modelo de regresión no lineal que represente mejor la relación existente entre ambas variables con el objetivo predecir la variable respuesta (modelo de regresión no lineal).

En esta última dirección, comentaremos brevemente algunas formas habituales de "linealizar" el modelo de regresión mediante transformaciones adecuadas para desarrollar el análisis de regresión lineal sobre estas transformaciones, permitiendo aplicar un modelo linealizable para mejorar el ajuste de forma sencilla frente a otras técnicas más complejas de regresión no lineal.

1.4.1. Modelo de regresión cuadrático y cúbico

El modelo de regresión de partida está formado por la recta $y = \beta_0 + \beta_1 x$, es decir, un polinomio de grado 1, por lo que una forma natural de extender el modelo para mejorar el ajuste es considerar un modelo polinómico de mayor grado.

Así, el modelo de regresión cuadrático viene dado por el polinomio de grado dos:

$$E(Y | X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

cuyos coeficientes se determinan a través de la información muestral (x_i, y_i) con $i = 1, \dots, n$, como en el caso de regresión simple, obteniéndose el modelo cuadrático ajustado $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$ para cada observación $i = 1, \dots, n$.

El estudio de este modelo de regresión cuadrático se desarrolla de forma similar al caso de regresión lineal simple, analizando los contrastes de significación y el diagnóstico del modelo.

Análogamente, en el caso cúbico, se analiza el ajuste a través del polinomio de grado tres:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + \hat{\beta}_3 x_i^3.$$

En general, obsérvese que el modelo de regresión polinómico puede interpretarse como un falso o ficticio modelo de regresión lineal múltiple, donde las sucesivas variables predictoras del modelo son las potencias de la variable predictora original.

1.4.2. Modelo de regresión multiplicativo o potencial

Un modelo común de regresión linealizable es el modelo de regresión multiplicativo o potencial, dado por

$$E(Y | X = x) = \beta_0 x^{\beta_1},$$

fácilmente linealizable mediante la transformación logarítmica, $\log(y) = \log(\beta_0) + \beta_1 \log(x)$, es decir, $y^* = \beta_0^* + \beta_1 x^*$ con $y^* = \log(y)$, $x^* = \log(x)$ y $\beta_0^* = \log(\beta_0)$. Por tanto, para cada $i = 1, \dots, n$, los valores ajustados a través de la regresión potencial pueden expresarse como

$$\hat{y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1 x_i^* \Leftrightarrow \hat{y}_i = \hat{\beta}_0 x_i^{\hat{\beta}_1}$$

donde $\hat{\beta}_0 = \exp(\hat{\beta}_0^*)$.

1.4.3. Modelo de regresión exponencial

Para el modelo de regresión exponencial existen diversas representaciones equivalentes,

$$E(Y | X = x) = \exp(\beta_0 + \beta_1 x) \Leftrightarrow E(Y | X = x) = \beta_0^* \beta_1^{*x}$$

siendo $\beta_0^* = e^{\beta_0}$ y $\beta_1^* = e^{\beta_1}$. Como en el caso de regresión potencial, utilizando la función logarítmica se obtiene su expresión linealizada $y^* = \log(y) = \beta_0 + \beta_1 x$, cuyos coeficientes se determinan a través de los valores muestrales (x_i, y_i) , como en los casos anteriores, obteniéndose el modelo de regresión exponencial ajustado

$$\hat{y}_i = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \quad \text{para } i = 1, \dots, n.$$

1.5. Caso práctico

Veamos un caso práctico de aplicación del análisis de regresión lineal a través del programa estadístico R. En este estudio se analiza un conjunto de variables de tipo socioeconómico, antropométrico y gestacional en cien parejas sanas y sus recién nacidos ($n = 100$). El conjunto de datos observados se puede encontrar en Carrasco y Hernán (1993), y el objetivo es analizar la relación de las diferentes características y factores que intervienen en el estudio y su relevancia para predecir a través de un modelo el peso del recién nacido.

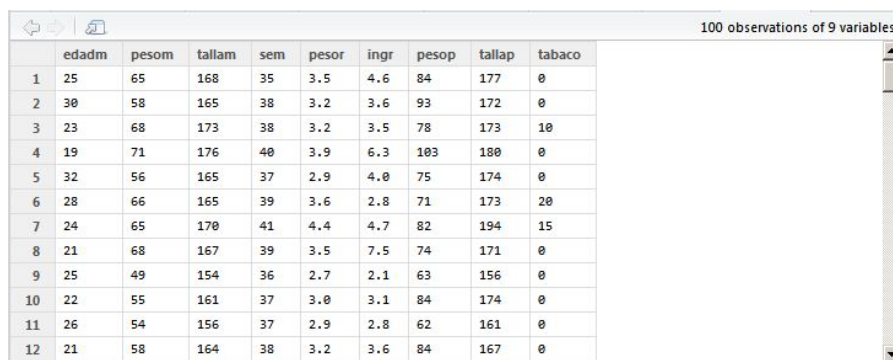
Los datos de este estudio se pueden cargar en R, por ejemplo desde el fichero *Bioestadística-BloqueAMECP.RData*, mediante la opción de abrir ficheros de la pestaña *Workspace* de RStudio, y cargando la base de datos en *R-Consola* mediante el comando:

```
> attach(peso)
```

las variables que contiene se describen en la Tabla 2, y las observaciones se muestran en la tabla de datos (Figura 3).

Tabla 2: Descripción de las variables del archivo *peso*

Nombre	Descripción
edadm	Edad de la madre en años en el momento del parto
pesom	Peso de la madre en kg
tallam	Altura de la madre en cm
sem	Tiempo de gestación en semanas
pesor	Peso del recién nacido en kg
ingr	Ingresos familiares en millones de ptas/año
pesop	Peso del padre en kg
tallap	Altura del padre en cm
tabaco	Número de cigarrillos al día de la madre



	edadm	pesom	tallam	sem	pesor	ingr	pesop	tallap	tabaco
1	25	65	168	35	3.5	4.6	84	177	0
2	30	58	165	38	3.2	3.6	93	172	0
3	23	68	173	38	3.2	3.5	78	173	10
4	19	71	176	40	3.9	6.3	103	180	0
5	32	56	165	37	2.9	4.0	75	174	0
6	28	66	165	39	3.6	2.8	71	173	20
7	24	65	170	41	4.4	4.7	82	194	15
8	21	68	167	39	3.5	7.5	74	171	0
9	25	49	154	36	2.7	2.1	63	156	0
10	22	55	161	37	3.0	3.1	84	174	0
11	26	54	156	37	2.9	2.8	62	161	0
12	21	58	164	38	3.2	3.6	84	167	0

Figura 3: Datos del archivo *peso*

1.5.1. Caso práctico de regresión simple

En este primer caso práctico, analizaremos el modelo de regresión lineal simple para la variable respuesta *pesor* del peso del recién nacido a través del tiempo de gestación en semanas *sem*, para mostrar el procedimiento y los resultados que presenta el programa. Para ello, una vez abierto el fichero de datos *peso* en R, y antes de iniciar el análisis, podemos realizar la representación gráfica de la nube de puntos o diagrama de dispersión, para describir la posible tendencia lineal entre las dos variables, mediante el siguiente código:

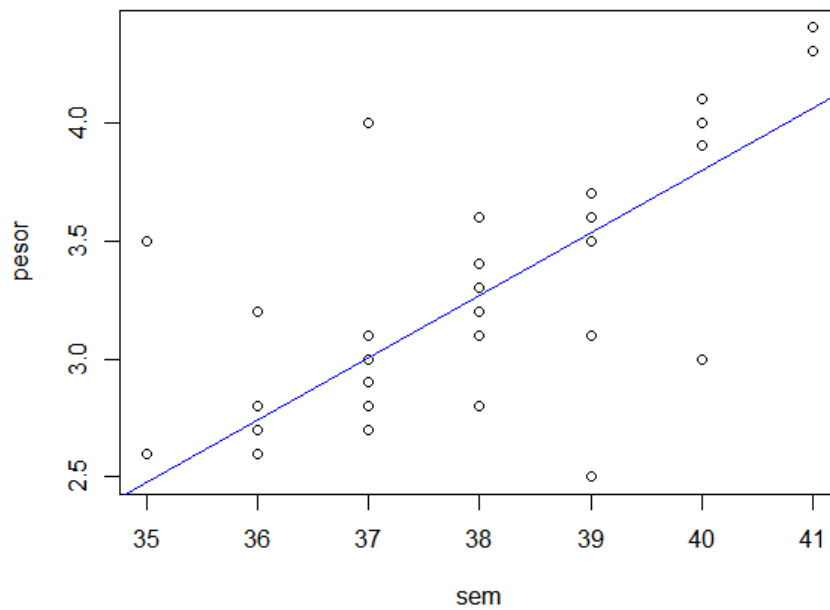


Figura 4: Diagrama de dispersión entre *pesor* y *sem*

```
> plot(sem, pesor)
```

En la representación gráfica obtenida (Figura 4), se observa una posible alineación de la nube de puntos en sentido creciente, que comprobaremos en el análisis de la recta de regresión.

Además, la función *lm* (*linear model*) de R nos permite estimar un modelo de regresión lineal usando la estructura $lm(y \sim x)$. Así, podemos incorporar al diagrama de dispersión la línea ajustada usando la función *abline* de la siguiente manera:

```
> fit <- lm(pesor ~ sem)
> abline(fit, col="blue")
```

como puede verse en la Figura 4, mostrándonos los coeficientes de la recta al indicar en R que muestre la variable *fit* donde hemos almacenado el resultado del ajuste:

```
> fit

Call:
lm(formula = pesor ~ sem)

Coefficients:
(Intercept)      sem
    -6.7845     0.2646
```

Para obtener los resultados del análisis de regresión lineal, utilizaremos la función *summary* sobre la variable donde hemos guardado el resultado del ajuste, obteniendo información sobre los residuos estimados, los contrastes de los coeficientes, el coeficiente determinación, etc., es decir,

```
> summary(fit)

Call:
lm(formula = pesor ~ sem)
```

```

Residuals:
    Min       1Q   Median       3Q      Max
-1.03369 -0.06913 -0.03369  0.07359  1.02458

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.78447     0.76207   -8.903 2.91e-14 ***
              sem      0.26457     0.02001   13.220 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2448 on 98 degrees of freedom
Multiple R-squared:  0.6407, Adjusted R-squared:  0.6371
F-statistic: 174.8 on 1 and 98 DF,  p-value: < 2.2e-16

```

Obsérvese que el contraste de regresión está incluido en los resultados anteriores, y también puede presentarse de la forma habitual mediante la función *anova*:

```

> anova(fit)
Analysis of Variance Table

Response: pesor
      Df Sum Sq Mean Sq F value    Pr(>F)
sem     1 10.4743  10.4743   174.77 < 2.2e-16 ***
Residuals 98   5.8732   0.0599
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

A partir de los resultados anteriores, la recta de regresión obtenida para estas observaciones entre el peso del recién nacido y las semanas de gestación es

$$pesor = -6.78 + 0.264sem$$

donde se observa la relación positiva entre ambas variables, y la importancia de las semanas de gestación en el peso que, por término medio, aumentará en 0.264 gramos a la semana, teniendo en cuenta el rango de valores de utilidad del modelo, dado que no puede extrapolarse donde no se dispone de información entre estas características. Asimismo, se obtiene que ambos términos (constante y predictora) son muy significativos en el modelo ($p \simeq 0$); idéntica conclusión se obtiene en el contraste de regresión a partir del estadístico F . El coeficiente de determinación indica que el 64.1 % de la variabilidad del peso de un recién nacido queda explicada a través de la recta de regresión con el tiempo en semanas de gestación.

Además de los contrastes de los coeficientes del modelo, R permite calcular los intervalos de confianzas para los coeficientes con la función *confint*, utilizando por defecto el nivel 95 % mediante el argumento *level=0.95*:

```

> confint(fit)
              2.5 %      97.5 %
(Intercept) -8.2967730 -5.2721658
              sem      0.2248541  0.3042825

```

Asimismo, cuando ejecutamos la función *lm* en R para realizar un ajuste, se calculan diferentes términos del análisis que podemos ver con el comando *attributes*:

```

> attributes(fit)

```

```

$names
[1] "coefficients" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "xlevels" "call" "terms" "model"

$class
[1] "lm"

```

Además, puede usarse el operador `$` para acceder a la información contenida en cada atributo, incluso algunos pueden ser guardados e incluidos como nuevas columnas en la base de datos de trabajo (*data.frame*), entre los que destacamos los residuos y ajustes o valores ajustados, necesarios para los contrastes de las condiciones iniciales. Por ejemplo, para incluir los residuos y ajustes en la base de datos *peso*, se puede proceder como sigue:

```

> peso$residuals <- fit$residuals
> peso$fitted.values <- fit$fitted.values

```

Otra forma de manejar estos resultados en el análisis del modelo, es mediante funciones de R, como las siguientes para los coeficientes, su matriz de varianza-covarianza, los ajustes y los residuos:

```

coef(fit)
vcov(fit)
fitted(fit)
residuals(fit)

```

Por ejemplo, para visualizar la aproximación a la normalidad de los residuos mediante el gráfico probabilístico de normalidad y el histograma, como se muestra en la Figura 5, ejecutaríamos las siguientes funciones en R:

```

> par(mfrow=c(1,2))
> qqnorm(residuals(fit));qqline(residuals(fit), col="blue")
> hist(residuals(fit), col="orange")

```

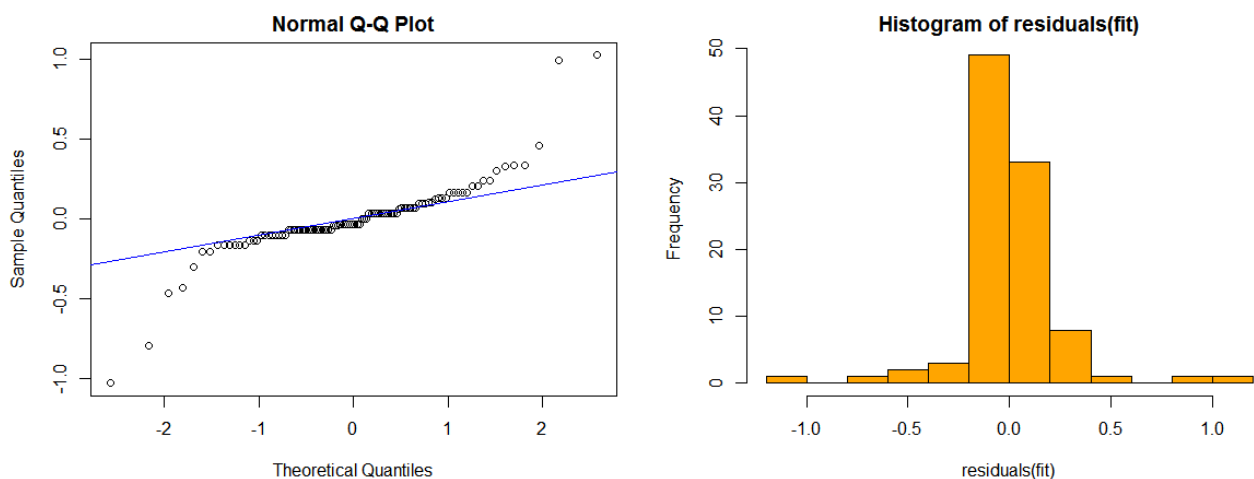


Figura 5: Gráficos de normalidad para los residuos

En cualquier caso, R permite comprobar directamente el diagnóstico gráfico de los residuos del modelo de regresión lineal, es decir, la aproximación a la normalidad, a la igualdad de varianzas (homoscedasticidad), el ajuste (linealidad) e identificar posibles observaciones influyentes:

```

> plot(fit)

```

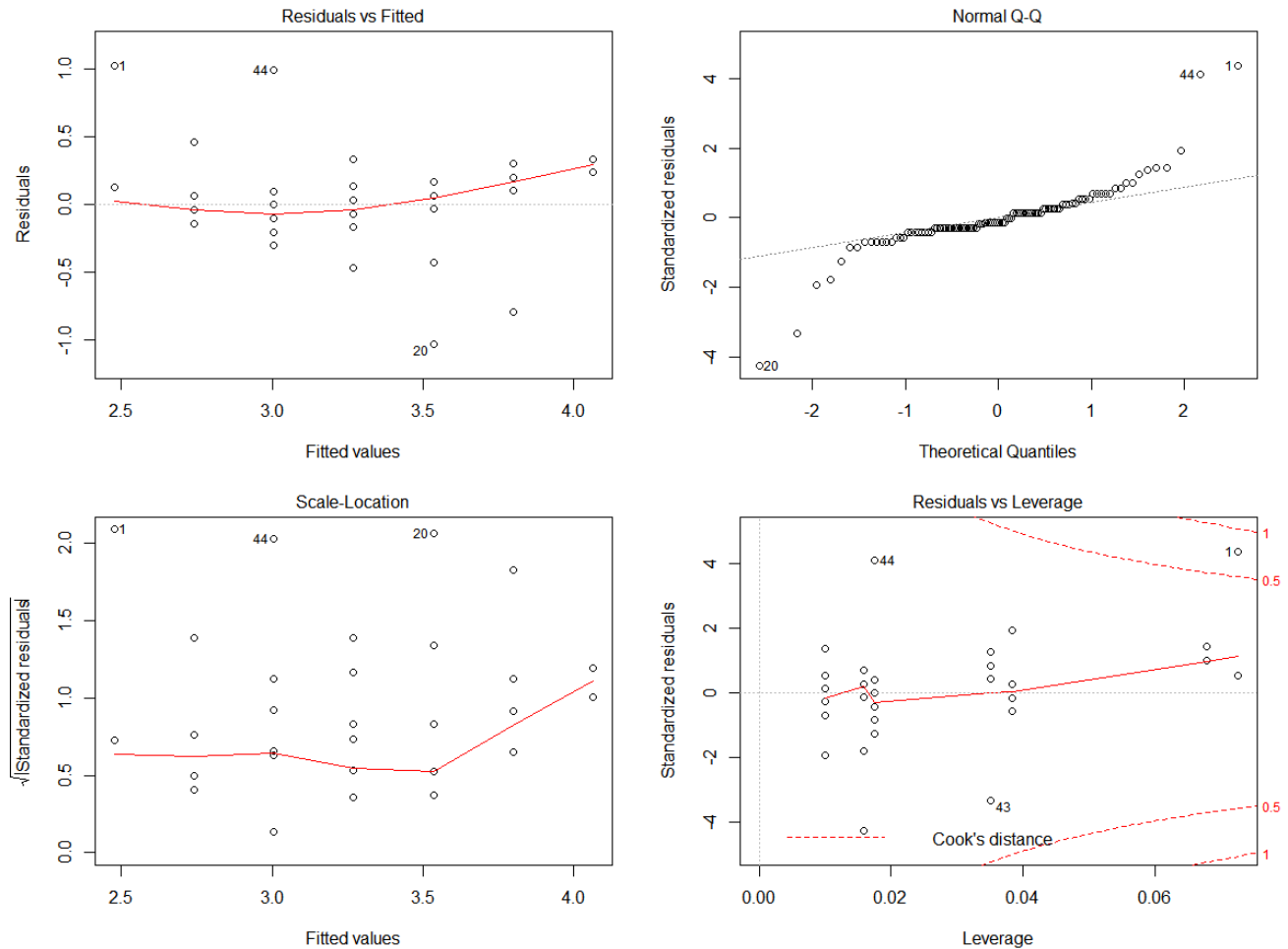


Figura 6: Gráficas para el diagnóstico del modelo ajustado

obteniéndose las gráficas de los residuos de la Figura 6.

- (1) *Valores ajustados versus residuos*: Nube de puntos de los residuos respecto de las predicciones para detectar la falta de linealidad. No se aprecia una falta de linealidad evidente. Obsérvese que las observaciones con los residuos más extremos son: 1, 20 y 44.
- (2) *QQ-plot de los residuos estandarizados*: Representación de cuantiles de los errores del ajuste respecto de la normalidad. Las posibles desviaciones a la normal se concentran en los extremos correspondientes a las observaciones: 1, 20 y 44.
- (3) *Valores ajustados versus residuos estandarizados*: Nube de puntos de los residuos estandarizados para detectar falta de homoscedasticidad. No se aprecia diferencias en la dispersión de los errores respecto de los ajustes. Las dispersiones más extremas corresponden a las observaciones: 1, 20 y 44.
- (4) *Leverage versus residuos estandarizados*: Gráfico de los residuos respecto de los leverages o apalancamientos de cada observación, para detectar las observaciones de mayor influencia a través del estadístico de distancias de Cook. En este caso las más influyentes en el ajuste son marcadas en la gráfica: 1, 43 y 44, aunque todas tiene valores de Cook inferiores a 1.
- (5) *Residuos en orden muestral*: Gráfico de residuos respecto del índice muestral para detectar la falta de independencia mediante rachas de dependencias de signos entre residuos consecutivos. La representación mostrada en la Figura 7 se realiza con las instrucciones:

```
> plot(residuals,type="o")
> abline(h=0)
```

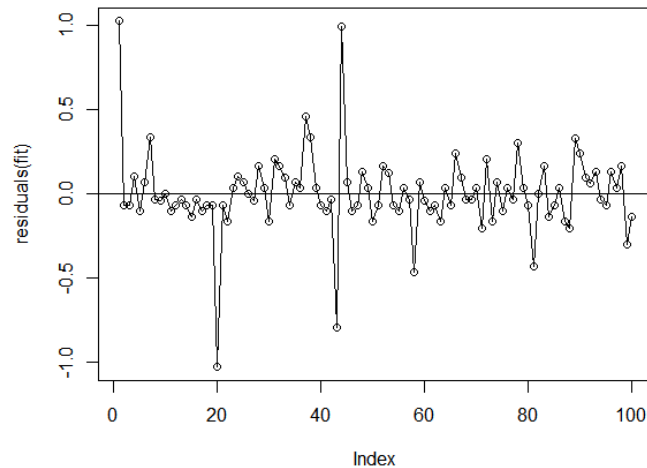


Figura 7: Series temporales de residuos

No obstante, deben realizarse los contrastes sobre estos residuos para garantizar la fiabilidad de las conclusiones, y sólo nos limitaremos a interpretar estos gráficos cuando no sea posible contrastar una condición inicial por falta de información experimental.

Así, el contraste de linealidad o falta de ajuste puede realizarse mediante la función *anova* comparando el modelo de regresión anterior *fit* con el modelo que considera la variable predictora como factor, en este caso *lm(pesor ~ factor(sem))*:

```
> falta <- lm( pesor ~ factor(sem))
> anova(fit,falta)
Analysis of Variance Table

Model 1: pesor ~ sem
Model 2: pesor ~ factor(sem)
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     98  5.8732
2     93  4.8020  5    1.0712 4.1491 0.001904 **
```

No obstante, la librería *EnvStats* de R nos proporciona otra forma de ejecutar el contraste de ajuste, presentando el error puro y la falta de ajuste, e incluyendo el test en la tabla ANOVA de regresión utilizando la función *anovaPE(fit)*.

Observamos que el *p*-valor del contraste de ajuste es aproximadamente 0.002, que nos lleva a la decisión de rechazar la condición de linealidad, al detectar significativamente un desajuste en la recta de regresión. Aunque este desajuste puede deberse a las observaciones influyentes identificadas en la Figuras 6 y 7, no pueden ignorarse dichas observaciones para mejorar el ajuste, a no ser que se disponga de información suficiente para considerarlas erróneas, por lo que en la práctica esta falta de linealidad nos conducirá a plantear una regresión linealizable o múltiple para mejorar el ajuste.

Para discutir la condición de incorrelación, es habitual utilizar el test de Durbin-Watson, para lo que necesitamos cargar en R la librería *lmtest*, obteniéndose en este caso que 1.95642 no resulta significativamente alejado del 2 para rechazar la incorrelación, como se deduce del *p*-valor 0.4179:

```
> library(lmtest)
> dwtest(fit)
```

Durbin-Watson test

```
data: fit
DW = 1.9574, p-value = 0.4179
alternative hypothesis: true autocorrelation is greater than 0
```

Ejercicio 1.1 Realizar el test de Ljung-Box para analizar la incorrelación de los residuos a través de la función `Box.test`, incluyendo la autocorrelación hasta el grado lag e indicando el número de términos en el modelo para la corrección de los grados de libertad del estadístico `fitdf`:

```
Box.test(residuals(fit), type="Ljung-Box", lag=4, fitdf=2)
```

Respecto de las condiciones de igualdad de varianzas, aplicamos los test de Bartlett y de Levene, y para la normalidad el test de Shapiro-Wilk, como sigue:

```
> bartlett.test(residuals(fit),fitted(fit))

Bartlett test of homogeneity of variances

data: residuals(fit) and fitted(fit)
Bartlett's K-squared = 34.9203, df = 6, p-value = 4.465e-06

> leveneTest(residuals,fitted(fit))
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  6  1.6966  0.1305
      93

Mensajes de aviso perdidos
In leveneTest.default(residuals, fitted(fit)) :
  fitted(fit) coerced to factor.

> shapiro.test(residuals)

Shapiro-Wilk normality test

data: residuals
W = 0.8061, p-value = 3.838e-10
```

El estadístico de Bartlett indica que existe una diferencia significativa entre las varianzas. No obstante, este estadístico es muy sensible a desviaciones de la normalidad, condición que es rechazada por el test de Shapiro-Wilk ($p \simeq 0$). En esta situación, es más robusto el test de Levene, que en este caso proporciona un p -valor 0.130 y no detecta ninguna diferencia significativa entre las varianzas. Recordar que los tests de igualdad de varianzas requieren que la varianza sea estimable en cada grupo, sin embargo, si no hay información suficiente en un grupo para estimar la varianza en dicho grupo es usual omitir en estas pruebas la comparación de aquellos casos de las predictoras en las que no hay información suficiente.

Por último, el objetivo final del análisis de regresión es aplicar el modelo ajustado para predecir la variable respuesta, siempre que se cumplan las condiciones iniciales y proporcione un ajuste adecuado. A modo ilustrativo, indicamos el código para obtener las predicciones junto con sus correspondientes intervalos de confianza para los pesos medios e intervalos de predicción para los pesos, utilizando la función `predict` y un nivel de confianza del 99% en ambos casos, en el primero para un caso (38 semanas) y en el segundo para dos casos (38 y 39 semanas):


```
> predict(fit, interval="confidence", level=0.99, data.frame(sem=38))
      fit      lwr      upr
1 3.269126 3.204739 3.333513

> predict(fit, interval="prediction", level=0.99, data.frame(sem=c(38,39)))
      fit      lwr      upr
1 3.269126 2.622817 3.915435
2 3.533694 2.885506 4.181882
```

y se observa en el primer caso (38 semanas), se estima con mayor precisión la media que una observación, dado que utilizando el mismo nivel de confianza, el intervalo de predicción tiene mayor amplitud que el de confianza. La Figura 8 representa las bandas de confianza y predicción con un nivel de confianza del 95 % para el peso correspondientes a la colección de intervalos obtenidos al recorrer el conjunto de observaciones, y que, por ejemplo, puede realizarse mediante las instrucciones de R recogidas en la función *bandasIC*:

```
> bandasIC(fit,0.95)
```

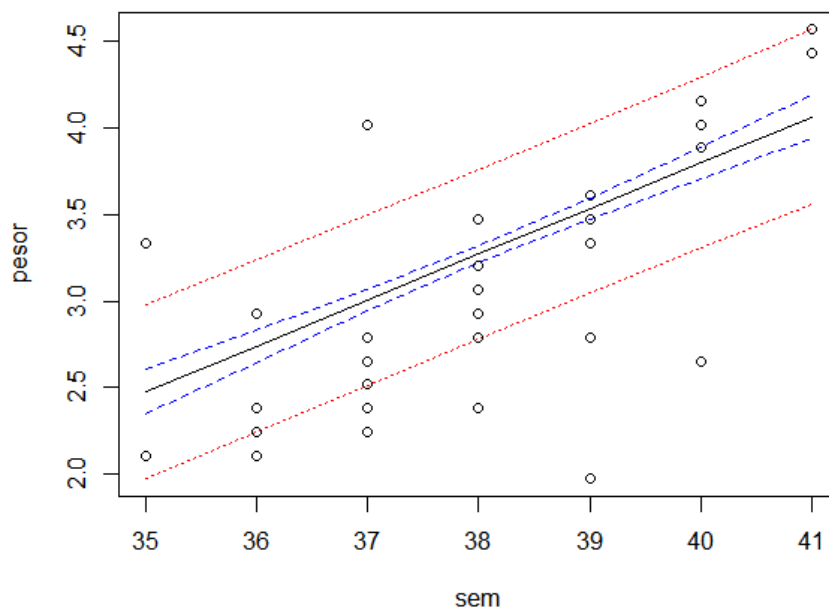


Figura 8: Bandas de confianza y de predicción para el *pesor*

En resumen, a partir del análisis anterior, el tiempo de gestación en semanas es significativo para analizar el peso del recién nacido, pero no lo suficiente para proporcionar buenas estimaciones de este peso, y además se han detectado desviaciones de las condiciones iniciales que también harían desestimar la efectividad del modelo de regresión lineal simple estimado. Estas desviaciones pueden deberse a la no linealidad de la relación entre ambas o la influencia de otras variables o factores predictores que no se han incluido en este modelo.

1.5.2. Caso práctico de regresión linealizable

En el caso anterior, el tiempo de gestación en semanas ha resultado significativo para analizar el peso del recién nacido, sin embargo la recta de regresión ajustada no satisfacía todas las condiciones para estimar adecuadamente el peso. En este caso práctico, aplicamos el modelo de regresión cuadrático para

la variable respuesta *pesor* del peso del recién nacido a través del tiempo de gestación en semanas *sem*, como ejemplo de ajuste de un modelo linealizable mediante el programa estadístico R, para obtener un modelo de regresión mejor entre ambas variables *pesor* y *sem*.

Cabe señalar que el proceso es análogo al caso del modelo lineal simple, ya que se realiza con la misma función *lm*, y que la diferencia estriba en la expresión de la fórmula del argumento de la misma. Así, en el caso de que nos dispongamos a ajustar nuestra nube de puntos a un polinomio de segundo grado en *sem*, $y = a + b \cdot sem + c \cdot sem^2$ (modelo de regresión cuadrático), procederemos como sigue:

```
> fitseg <- lm(pesor ~ sem + I(sem^2), data=peso)
```

en donde la expresión $I(sem^2)$ indica la inclusión de sem^2 en el modelo de regresión estimado. Y al igual que en el caso anterior, los resultados del análisis del modelo cuadrático se obtienen a partir del comando *summary*:

```
> summary(fitseg)
```

Call:

```
lm(formula = pesor ~ sem + I(sem^2), data = peso)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.00607	-0.09416	-0.00607	0.08894	1.00584

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.87903	14.61516	3.413	0.000939 ***
sem	-2.71216	0.76710	-3.536	0.000626 ***
I(sem^2)	0.03905	0.01006	3.882	0.000189 ***

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

Residual standard error: 0.2289 on 97 degrees of freedom

Multiple R-squared: 0.689, Adjusted R-squared: 0.6826

F-statistic: 107.5 on 2 and 97 DF, p-value: < 2.2e-16

por lo que el modelo de regresión cuadrático estimado para el peso de un recién nacido es

$$pesor = 49.879 - 2.712sem + 0.03905sem^2$$

y cuyo coeficiente de determinación ha aumentado ligeramente hasta el 68.9 % al incluir el término de segundo grado en el modelo. En la Figura 9, se representa este modelo de regresión cuadrático sobre la nube de puntos, como ejemplo ilustrativo de los modelos linealizables, incluyendo sus correspondientes bandas de confianza y de predicción al 95 %.

Obsérvese que el análisis del modelo cuadrático (en general, de un modelo linealizable) se realiza de forma similar al anterior, indicando que muestre los resultados del ajuste (modelo estimado, contrastes de coeficientes, contraste de regresión, coeficiente de determinación, ...), las gráficas de los residuos y guardando los residuos y ajustes para realizar los contrastes de las condiciones iniciales.

De forma similar al análisis del modelo de regresión cuadrático, el estudio de un modelo de regresión cúbico se realiza incluyendo dicho término en el comando de ajuste lineal:

```
> fitcub <- lm(pesor ~ sem + I(sem^2) + I(sem^3), data=peso)
```

Y análogamente, se llevan a cabo los ajustes a cualquier modelo linealizable indicando en la expresión los términos de las transformaciones adecuadas, por ejemplo el modelo exponencial para el peso del recién nacido a través del tiempo de gestación se desarrolla mediante la orden:

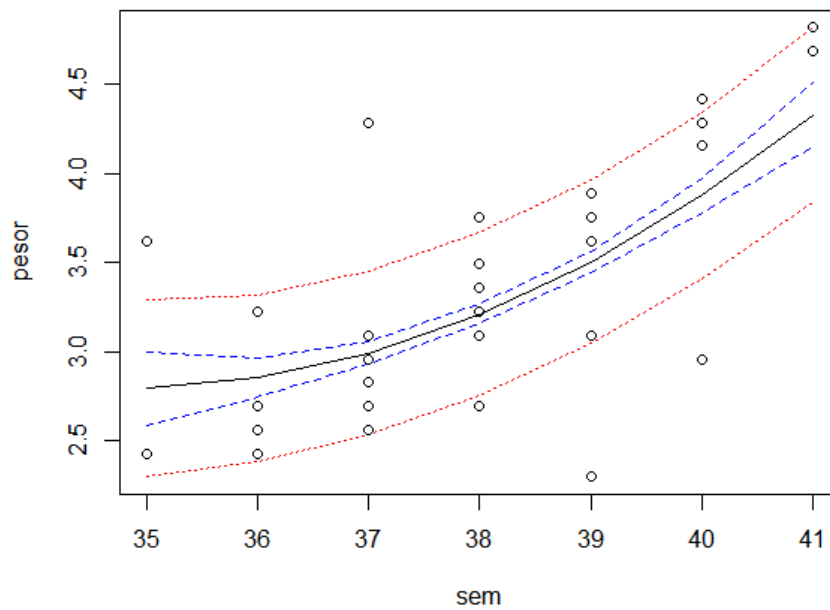


Figura 9: Regresión cuadrática, bandas de confianza y de predicción

```
> fitexp <- lm(log(pesor) ~ sem, data=peso)
```

Ejercicio 1.2 Realizar el análisis del modelo de regresión lineal cuadrático para el peso de un recién nacido a través del tiempo de gestación.

- (1) Presentar la tabla ANOVA para el contraste de regresión:

```
anova(fitseg)
```

- (2) Obtener los intervalos de confianza para los coeficientes del modelo:

```
confint(fitseg)
```

- (3) Añadir los valores ajustados y los errores del ajuste en la base de datos:

```
peso$resid2 <- fitseg$residuals
peso$estim2 <- fitseg$fitted.values
```

- (4) Representar los gráficos de los residuos e interpretarlos para el diagnóstico del modelo:

```
plot(fitseg)
plot(resid2, type="o"); abline(h=0)
```

- (5) Realizar e interpretar los contrastes de hipótesis del diagnóstico del modelo ajustado:

```
anovaPE(fitseg)
dwtest(fitseg)
bartlett.test(resid2, estim2)
leveneTest(resid2, estim2)
shapiro.test(resid2)
```

Ejercicio 1.3 Realizar el análisis del modelo de regresión exponencial para el peso de un recién nacido a través del tiempo de gestación, siguiendo los pasos del ejercicio anterior.

- (1) Presentar la tabla ANOVA para el contraste de regresión:

```
anova(fitexp)
```

(2) Obtener los intervalos de confianza para los coeficientes del modelo:

```
confint(fitexp)
```

(3) Añadir los valores ajustados y los errores del ajuste en la base de datos:

```
peso$residEx <- fitexp$residuals
peso$estimEx <- fitexp$fitted.values
```

(4) Representar los gráficos de los residuos e interpretarlos para el diagnóstico del modelo:

```
plot(fitexp)
plot(residEx, type="o"); abline(h=0)
```

(5) Realizar e interpretar los contrastes de hipótesis del diagnóstico del modelo ajustado:

```
anovaPE(fitexp)
dwtest(fitexp)
bartlett.test(residEx, estimEx)
leveneTest(residEx, estimEx)
shapiro.test(residEx)
```

2. Regresión lineal múltiple

Como hemos comentado, el análisis del modelo de regresión lineal múltiple se plantea como la técnica estadística que modeliza la relación entre una variable respuesta Y y un conjunto de variables predictoras (X_1, \dots, X_k) , con el fin de pronosticar los valores de la respuesta a través de este modelo. Por un lado, puede considerarse como una ampliación del modelo de regresión lineal simple, que mejora el ajuste de un modelo inicial incorporando más información experimental contenida en las variables que se incluyen en el modelo de regresión, y por otra parte, como el modelo de regresión de partida incluyendo todas las variables predictoras posibles que se consideren relevantes para analizar la respuesta. Ambos procedimientos constituyen los elementos principales de la selección del mejor modelo de regresión lineal que serán comentados posteriormente.

En este contexto, el modelo de regresión lineal múltiple, para cada observación i de la muestra, viene dado por

$$Y_i = E(Y_i | (X_1, \dots, X_k) = (x_{1i}, \dots, x_{ki})) + \varepsilon_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

para $i = 1, \dots, n$, siendo $\beta_0, \beta_1, \dots, \beta_k$ los coeficientes del modelo y ε_i el residuo al ajustar cada Y_i mediante este modelo de regresión cuando $(X_1, \dots, X_k) = (x_{1i}, \dots, x_{ki})$, cuyo valor experimental es $e_i = y_i - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$. Por conveniencia, denotaremos por $X_0 = 1$ la componente asociada al coeficiente constante β_0 , esta constante o intercepto β_0 representa el valor medio de la respuesta cuando todas las predictoras son cero, y su papel en el modelo no es representar la relación entre las predictoras y la respuesta, sino mejorar el ajuste del modelo de regresión a las observaciones muestrales.

2.1. Interpretación y estimación del modelo

Al igual que en el caso de la regresión simple, se observa una relación directa entre los coeficientes del modelo y la importancia de su correspondiente variable predictora. Evidentemente, si un coeficiente $\beta_j = 0$, su variable predictora asociada X_j no interviene en el modelo con $j = 0, 1, \dots, k$. Asimismo, el coeficiente pendiente β_j representa el efecto medio (positivo o negativo) sobre la variable respuesta al aumentar la predictora X_j en una unidad cuando se mantienen constantes las restantes predictoras. Por lo que tendrá interés en el estudio de la significación individual de dicha predictora.

Para desarrollar el análisis de regresión múltiple se requieren las condiciones iniciales indicadas en el caso simple: linealidad (error medio nulo), homogeneidad de varianzas (igual dispersión), incorrelación (aleatoriedad) y normalidad. Condiciones que se deben analizar a posteriori en el análisis de los residuos, la única diferencia con el caso simple está en la representación del modelo y sus ajustes, siendo el concepto de los residuos el mismo.

En este caso, el criterio de mínimos cuadrados y el de máxima verosimilitud bajo normalidad, permiten obtener el modelo de regresión lineal múltiple estimado

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k \Leftrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki} \text{ para } i = 1, \dots, n$$

donde \hat{y}_i es el valor ajustado para cada observación i , siendo su error o residuo $\hat{e}_i = y_i - \hat{y}_i$.

Las estimaciones de los coeficientes del modelo pueden representarse en forma vectorial $\hat{\beta}^t = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ a través del vector aleatorio respuesta \vec{Y} correspondiente a sus valores muestrales, $\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \vec{Y}$, donde cada fila de la matriz $\mathbf{X} = (\vec{X}_0, \vec{X}_1, \dots, \vec{X}_k)$, llamada matriz del diseño, corresponde a cada conjunto de valores de las variables predictoras. El error acumulado de los desajustes entre el modelo de regresión lineal múltiple estimado y la respuesta,

$$SS_R = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

proporciona una medida de la variabilidad del error experimental (variación no explicada por el modelo de regresión).

2.2. Análisis del modelo de regresión múltiple

2.2.1. Contrastes de significación individual

Los contrastes de significación individual de los términos incluidos en el modelo, se realizan igual que en el caso simple, la única diferencia está en el número de estos términos, es decir, el número de variables predictoras más el término constante, de modo que los estadísticos de estos contrastes siguen distribuciones muestrales t de Student con $n - k - 1$ grados de libertad, y tienen regiones de rechazo de dos colas.

Por ejemplo, el contraste de hipótesis nula $\beta_i = 0$ frente a la alternativa $H_1 : \beta_i \neq 0$, contrasta la importancia o relevancia individual de la variable predictora X_i en el modelo, es decir, la utilidad de X_i para explicar a la variable Y a través de este modelo de regresión lineal. Un p -valor próximo a cero indicará que las observaciones están significativamente en contra de que $\beta_i = 0$, y por tanto al rechazar esta hipótesis, se concluye que $\beta_i \neq 0$, o equivalentemente, que X_i debe mantenerse en el modelo de regresión. Sin embargo, cuando no puede rechazarse esta hipótesis con la información disponible en la muestra, se concluye que X_i no es significativa, y por tanto, podría reducirse este término del modelo de regresión, dado que su posible relación lineal con la respuesta no es relevante en el modelo, resultando un modelo más sencillo con el resto de predictoras.

2.2.2. Contraste de regresión

En este caso, el contraste de significación global de todas las variables predictoras del modelo no se reduce a un contraste individual de un coeficiente como sucede en el caso simple, dado que el contraste de regresión incluye al conjunto de coeficientes $(\beta_1, \beta_2, \dots, \beta_k)$. Sin embargo, su desarrollo basado en la descomposición de la variabilidad es similar al utilizado en el caso simple, ya que se mantiene la descomposición de la variabilidad de la respuesta a través de la componente de variabilidad explicada

por el modelo de regresión y la variabilidad no explicada por el modelo

$$SS_T = SS_E + SS_R \Leftrightarrow \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

la única diferencia está en el modelo con el que se obtienen los ajustes \hat{y}_i , y por tanto, en los grados de libertad de la distribución en el muestreo del estadístico del contraste que se representa en la Tabla 3; su aplicación e interpretación del p -valor es idéntica a la usada en el caso simple.

Fuente	S. Cuadrados	G.Libertad	Media Cuadrados	Estadístico	p -valor
Regresión:	SS_E	k	$MS_E = \frac{SS_E}{k}$	$F = \frac{MS_E}{MS_R}$	p
Residuos:	SS_R	$n - k - 1$	$MS_R = \frac{SS_R}{n-k-1}$		
Total:	SS_T	$n - 1$			

Tabla 3: Tabla ANOVA del contraste de regresión múltiple

Observar que, en este caso, la estimación de la variabilidad viene dada por el estadístico media de cuadrados de los residuos de la Tabla 3, MS_R , donde interviene de forma explícita el número de predictoras k , y denotándose por $S = \sqrt{MS_R}$ a la estimación de la desviación típica.

2.2.3. Coeficiente de determinación

Recordamos que el coeficiente de determinación representa la proporción de variabilidad explicada por el modelo de regresión con respecto a la variabilidad total de la respuesta $R^2 = \frac{SS_E}{SS_T}$, e indica la medida de bondad del ajuste del modelo de regresión para representar y predecir la respuesta. Aunque este coeficiente de determinación del modelo de regresión toma sus valores entre $[0, 1]$, es habitual expresarlo e interpretarlo en términos porcentuales.

Además, el coeficiente de determinación ajustado, $R_a^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$, proporciona una medida alternativa de la bondad del modelo de regresión incluyendo de forma explícita el número de predictoras, siendo su interpretación similar al coeficiente de determinación.

2.2.4. Multicolinealidad

El proceso de análisis de los residuos del modelo de regresión múltiple se realiza del mismo modo que en el caso de regresión simple, tanto en la aplicación del estudio gráfico como del estudio analítico de los residuos del modelo de regresión múltiple, dado que el diagnóstico del modelo de regresión se realiza a posteriori sobre los residuos ajustados $\hat{e}_i = y_i - \hat{y}_i$, la única diferencia entre los residuos del caso simple y múltiple está en el modelo para obtener los valores ajustados \hat{y}_i con menor o mayor información muestral debida a la cantidad de predictores.

No obstante, en el caso de un modelo de regresión lineal con múltiples variables predictoras aparece un nuevo problema, conocido como problema de colinealidad o multicolinealidad, que debe ser tenido en cuenta para cuantificar la relevancia de cada variable predictora y su influencia en la utilidad global del modelo de regresión utilizado para pronosticar la respuesta.

El concepto de multicolinealidad se refiere a la existencia de relación de dependencia entre las variables predictoras. Por ejemplo, una predictora que resulta significativa en el modelo puede estar fuertemente relacionada con algunas o todas las restantes predictoras, de forma que su relevancia sobre la variable respuesta se reduce por dicha relación, es decir, el conjunto de predictoras con las que está relacionada proporciona la mayor parte de la variabilidad explicada por ésta sobre la respuesta, y desde un punto de vista práctico, dicha variable predictora no aporta información significativa en el modelo de regresión con respecto a la dada por el resto de predictoras.

Además, la multicolinealidad afecta tanto a la estimación de los coeficientes del modelo como a la variación de estos estimadores, dando lugar a un modelo de regresión estimado que puede ser ineficaz para predecir la variable respuesta, esto es el problema de multicolinealidad y por ello la importancia de su estudio. Evidentemente, la presencia de este tipo de variables predictoras en el modelo de regresión múltiple indica un alto grado de relación entre las columnas de la matriz de diseño \mathbf{X} , que se traduce en que el determinante de $\mathbf{X}^t\mathbf{X}$ está próximo a cero, y por tanto, la matrix inversa que define el vector de estimadores de los coeficientes y a su matriz de covarianzas tendrá valores elevados. En concreto, el problema de multicolinealidad implica una gran variación en la estimación de los coeficientes y del modelo de regresión estimado, lo que provoca imprecisión en el modelo, y consecuentemente, en las predicciones que se pretenden obtener a través de él.

Para analizar este problema de multicolinealidad se puede utilizar el estadístico *Factor de Incremento de la Varianza* para cada variable predictora, $FIV(X_j)$, basado en la matriz de correlaciones entre el conjunto de variables predictoras (X_1, \dots, X_k) , siendo su mínimo valor 1 (indica que la predictora X_j está incorrelada con el resto), y a mayor valor de $FIV(X_j)$ mayor grado de dependencia con las restantes predictoras, es decir, mayor grado de multicolinealidad, considerándose que un valor mayor a 10 indica una variable predictora con alta multicolinealidad. Este factor de incremento de la varianza identifica a las variables causantes del problema de multicolinealidad, y podría utilizarse como método de selección de variables predictoras en el modelo reduciendo una a una las de mayor multicolinealidad.

En situaciones de multicolinealidad, otras técnicas alternativas de estimación del modelo están disponibles en la librería *lmtest*, o también se puede utilizar el análisis de componentes principales de las variables predictoras, reduciendo la cantidad de predictores en el modelo de regresión para intentar evitar la multicolinealidad.

2.3. Selección del modelo de regresión

Veamos brevemente las técnicas utilizadas para seleccionar las mejores variables predictoras para formar el modelo de regresión lineal múltiple que proporcionará las predicciones de la variable respuesta.

2.3.1. Selección por etapas

Estos métodos se basan en la significación de las desviaciones producidas al incluir o reducir una determinada predictora en el modelo, tomando como referencia en estas desviaciones alguno de los elementos característicos del modelo de regresión (coeficiente de determinación, variabilidad no explicada, estadístico F o su p -valor).

En este contexto, comentamos el funcionamiento de estos procesos de selección utilizando su relevancia en el modelo mediante la significación individual de las predictoras en cada paso.

Selección Forward

El proceso de selección *Forward* consiste en incorporar por orden de menor p -valor las variables predictoras en el modelo hasta que las restantes predictoras por incluir no sean significativas.

Etapas 1. Iniciar con el modelo constante. Fijar un nivel de significación para entrar, α_{entrar} .

Etapas 2. Para todas las variables predictoras que no están en el modelo, obtener los modelos de regresión estimados añadiendo cada una de las predictoras y sus contrastes de significación. Elegir la predictora más significativa (menor p -valor $< \alpha_{\text{entrar}}$).

Etapas 3. Repetir la Etapa 2 hasta que ninguna variable predictora pueda añadirse al modelo.

Observar que el valor α_{entrar} para seleccionar una variable predictora no debe ser muy restrictivo, para que las variables no seleccionadas no tengan efectos relevantes sobre la significación de las predictoras incluidas en el modelo de regresión seleccionado.

Selección Backward

El proceso de selección *Backward* consiste en reducir por orden de mayor p -valor las variables predictoras del modelo hasta que las predictoras que permanecen en el modelo de regresión sean todas significativas.

Etapas 1. Iniciar con el modelo de regresión lineal con todas las variables predictoras. Fijar un nivel de significación para salir, α_{salir} .

Etapas 2. Elegir la predictora menos significativa del modelo (mayor p -valor $> \alpha_{\text{salir}}$). Obtener el modelo de regresión estimado al eliminar dicha predictora y los contrastes de significación de los términos que se mantienen.

Etapas 3. Repetir la Etapa 2 hasta que ninguna variable predictora del modelo pueda eliminarse.

Observar que el valor α_{salir} para remover una variable predictora no debe ser muy restrictivo, para que las variables eliminadas en el proceso no tengan efectos relevantes sobre la significación de las variables predictoras del modelo de regresión seleccionado.

Selección Stepwise

El proceso de selección *Stepwise* consiste en una combinación de los dos procesos de selección anteriores, alternándose los pasos de incluir una predictora en el modelo (*forward*) y de eliminar una predictora ya incluida en el modelo de regresión (*backward*).

Etapas 1. Iniciar con el modelo constante. Fijar los niveles de significación para entrar y salir, $\alpha_{\text{entrar}} \leq \alpha_{\text{salir}}$.

Etapas 2. Para todas las variables predictoras que no están en el modelo, obtener los modelos de regresión estimados añadiendo cada una de las predictoras y sus contrastes de significación. Elegir la predictora más significativa (menor p -valor $< \alpha_{\text{entrar}}$).

Etapas 3. Elegir la predictora menos significativa del modelo (mayor p -valor $> \alpha_{\text{salir}}$). Obtener el modelo de regresión estimado al eliminar dicha predictora y los contrastes de significación de los términos que se mantienen.

Etapas 4. Volver a la Etapa 2 hasta que ninguna variable predictora pueda añadirse al modelo o eliminarse del modelo en la Etapa 3.

Observar que la desigualdad, $\alpha_{\text{entrar}} \leq \alpha_{\text{salir}}$, para introducir y remover una variable predictora pretende evitar la formación de bucles, dado que la introducción en el modelo de la mejor predictora de las disponibles modifica la significación de las variables previamente incluidas, lo que puede provocar que alguna de estas predictoras del modelo deje de ser significativa.

Análogamente, se desarrollan estos métodos por etapas a través de los estadísticos R^2 y F , introduciendo en el modelo la predictora que provoca mayor aumento en el valor del estadístico R^2 (F) o eliminando la que produce menor reducción en el valor de R^2 (F). Por el contrario, utilizando el estadístico SS_R o MS_R , se introduce la predictora que provoca mayor reducción en SS_R o se elimina la que produce menor aumento en SS_R .

No obstante, estos procesos de selección por etapas pueden tener algunos inconvenientes, como sobrevalorar la importancia de las variables predictoras que permanecen en el modelo seleccionado y la pérdida del modelo óptimo considerando el objetivo final de la regresión para predecir la respuesta, y por tanto, deberá utilizarse algún procedimiento de evaluación del modelo seleccionado.

2.3.2. Criterios de evaluación del modelo

Criterios de información

En el análisis y la modelización de datos es habitual el uso de medidas para evaluar el ajuste del modelo estimado a las observaciones experimentales a través de los valores de la verosimilitud del modelo ajustado L y el número k de parámetros estimados. En particular, el criterio de información de Akaike viene dado por $AIC(k) = -2\log(L) + 2k$, y el criterio de información bayesiana de Schwarz por $BIC(k) = -2\log(L) + 2\log(n)$.

Estos criterios de información proporcionan procedimientos alternativos en la selección del modelo de regresión, dado que sus valores disminuyen al mejorar el ajuste, al igual que la medida de variabilidad del error. Por tanto, en una etapa de selección del modelo de regresión con i variables predictoras ($k = i + 2$ incluyendo el coeficiente constante y la varianza), se introduce en el modelo la predictora que provoca mayor reducción en el valor de AIC (BIC) o se elimina la que produce menor aumento en AIC (BIC).

Criterio de Mallows

Por último, observar que un proceso de selección del modelo de regresión lineal múltiple conlleva la reducción del número de variables predictoras, surgiendo la necesidad de medir la adecuación del modelo de regresión seleccionado.

El estadístico C_p de Mallows proporciona un indicador de idoneidad del número de predictoras seleccionadas basado en la variabilidad no explicada por el modelo de regresión con p predictoras seleccionadas (SS_{R^*}) y la variabilidad del modelo de regresión completo con las k predictoras (SS_R), y definido por $C_p = \frac{SS_{R^*}}{MS_R} + 2(p + 1) - n$, siendo su valor esperado el número de términos del modelo de regresión seleccionado, $E(C_p) = p + 1$.

En la práctica es habitual la aplicación del criterio de Mallows, que considera adecuado un modelo de regresión seleccionado con p predictoras para predecir la respuesta si su estimación de C_p es inferior al número de términos del modelo completo, $C_p \leq k + 1$.

2.4. Caso práctico

En primer lugar, obsérvese que el análisis de regresión múltiple se realiza de forma análoga al modelo de regresión lineal simple, como hemos comentado, la diferencia se reduce al problema de visualización de la tendencia lineal de la nube de puntos en una dimensión superior (incluso con dos predictoras) y en la obtención de las predicciones o ajustes teniendo en cuenta las múltiples predictoras.

Conviene recordar que la regresión simple, a través de las semanas de gestación, incumplía algunas condiciones iniciales, como la falta de linealidad, inconvenientes que no se solventaban con la regresión cuadrática, por lo que cabe analizar un modelo de regresión múltiple para mejorar el ajuste y que cumpla dichas condiciones iniciales. En este caso, consideramos el modelo de regresión múltiple con todas las predictoras disponibles para analizar la mejor relación lineal con el peso de un recién nacido, para ello utilizamos la función $lm(y \sim \text{modelo})$, donde los términos que forman el modelo se añaden separados por el signo $+$, es decir,

```
> fitm <- lm(pesor ~ edadm + pesom + tallam + sem + ingr + pesop + tallap + tabaco)
```

o bien se sustituye el modelo por un punto para indicar en R todas las variables de un conjunto de datos,

```
> fitm <- lm(pesor ~ . , data=peso)
```

obteniendo el siguiente resumen de resultados del análisis de regresión:

```
> summary(fitm)

Call:
lm(formula = pesor ~ edadm + pesom + tallam + sem + ingr + pesop +
    tallap + tabaco)

Residuals:
    Min       1Q   Median       3Q      Max
-0.41525 -0.09132 -0.03330  0.08703  0.49142

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.347375    0.837815  -3.995 0.000131 ***
edadm        -0.001316    0.003209  -0.410 0.682635
pesom         0.018968    0.004271   4.441 2.51e-05 ***
tallam       -0.008320    0.005233  -1.590 0.115335
sem           0.130900    0.020071   6.522 3.82e-09 ***
ingr          0.058564    0.016931   3.459 0.000827 ***
pesop         0.007625    0.002225   3.427 0.000918 ***
tallap        0.006002    0.002884   2.081 0.040223 *
tabaco        0.003890    0.002327   1.672 0.098012 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1701 on 91 degrees of freedom
Multiple R-squared:  0.8389, Adjusted R-squared:  0.8248
F-statistic: 59.25 on 8 and 91 DF,  p-value: < 2.2e-16
```

Como hemos visto anteriormente, otras funciones que nos proporcionan información para el análisis del modelo de regresión, así como para realizar los gráficos de los residuos, son: `coef(fitm)` (coeficientes estimados del modelo), `confint(fitm, level=0.95)` (intervalos de confianza para los coeficientes del modelo), `coeftest(fitm)` (contrastes individuales de los coeficientes), `fitted(fitm)` (valores estimados o ajustados), `residuals(fitm)` (errores del ajuste), `anova(fitm)` (tabla ANOVA de la regresión), `vcov(fitm)` (matriz de covarianzas de los coeficientes), `influence(fitm)` (influencia de las observaciones).

Al igual que en el análisis de la recta de regresión y de los modelos linealizables, se utilizan las gráficas de los residuos para describir su comportamiento e intentar detectar descriptivamente las desviaciones de las condiciones iniciales. Obsérvese que en el caso múltiple resulta más habitual el uso del análisis gráfico de los residuos, debido a la falta de información muestral suficiente para realizar los contrastes del diagnóstico en la mayor parte de los casos prácticos, como por ejemplo la linealidad e igualdad de varianzas que requieren más de una observación de la respuesta con los mismos valores de las variables predictoras. En este caso, ejecutando las siguientes instrucciones se obtienen las gráficas de la Figura 10 para el análisis de los residuos y medidas de influencia de las observaciones sobre el modelo estimado:

```
> par(mfrow=c(1,2))
> plot(fitm)
> plot(residuals(fitm),type="o"); abline(h=0)
```

Asimismo, cabe señalar que la principal diferencia del diagnóstico del modelo de regresión múltiple respecto del caso simple es la posible existencia del multicolinealidad entre las variables predictoras, lo que puede dar lugar a la acumulación de imprecisiones en la inferencia e invalidar las predicciones del modelo. La librería *car* en R, entre otras, permite el cálculo del factor de incremento de la varianza de cada variable predictora como sigue:

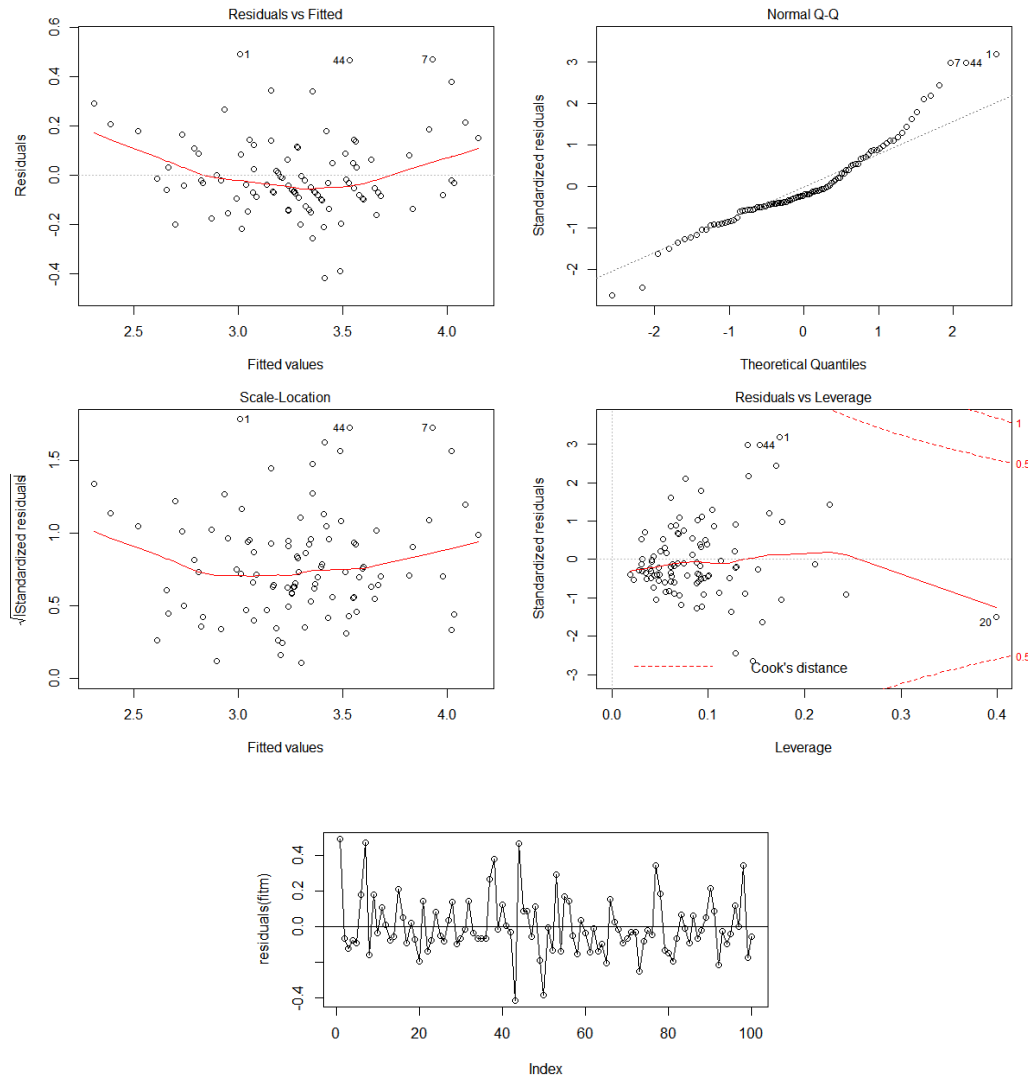


Figura 10: Gráficos de los residuos para el ajuste múltiple

```
> vif(fitm)
    edadm    pesom    tallam      sem    ingr    pesop    tallap    tabaco
1.049433 3.641424 3.912403 2.083636 1.679311 2.006822 2.111176 1.244632
```

siendo la altura de la madre (*tallam*) la variable predictora con mayor *FIV*, es decir, la variable predictora más relacionada con el resto y que provocaría mayor imprecisión en la estimación de los coeficientes del modelo; no obstante, se mantiene en un valor moderado de colinealidad.

Otro tópico que adquiere relevancia en el análisis de regresión múltiple, con respecto al simple, es el análisis y selección de las variables predictoras del modelo. En este caso práctico, a partir de los resultados de los contrastes de significación individual y conjunta de las predictoras en el modelo de regresión múltiple, se obtiene que, entre otras, la edad de la madre no resulta significativa en el modelo de regresión para predecir el peso del recién nacido ($p = 0.683$), siendo por tanto una de las candidatas a reducirse del modelo si aplicamos un método de selección de predictoras, dado que manteniendo el resto de variables predictoras constantes, aumentar en un año la edad de la madre supondría una reducción en media de 1.3 gramos en el peso del recién nacido, un efecto irrelevante sobre nuestra variable respuesta. Sin embargo, una variable predictora relevante en el modelo es el peso de la madre ($p \simeq 0$) y su aumento en un kg supondría un aumento medio de 18.9 g en el recién nacido manteniendo fijas el resto de variables.

En este sentido, alguna predictora podría eliminarse del modelo de regresión para trabajar en

la práctica con un modelo más manejable (con menos variables predictoras) sin perder de forma significativa la parte de la variabilidad explicada por el modelo. Esto plantea dos cuestiones, la primera comparar dos modelos de regresión para decidir si hay o no diferencias significativas entre ambos, y en segundo lugar, los métodos de selección de variables para un modelo de regresión a partir de las variables predictoras más relevantes en el estudio, en este caso, del peso de un recién nacido.

Por un lado, para realizar la comparación entre dos modelos de regresión en R, se utiliza la función *anova* sobre los dos modelos de regresión ajustado. Por ejemplo, si llamamos *fit0* al modelo de regresión constante, es decir, todas las variables predictoras con coeficientes nulos, y el modelo de regresión completo anterior *fitm* con todas las variables predictoras, evidentemente su comparación sería equivalente al contraste de regresión en el modelo completo:

```
> fit0 <- lm(pesor ~ 1, data=peso)
> summary(fit0)

Call:
lm(formula = pesor ~ 1, data = peso)

Residuals:
    Min       1Q   Median       3Q      Max
-0.785 -0.210 -0.085  0.215  1.115

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.28500     0.04064   80.84  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4064 on 99 degrees of freedom

> anova(fit0, fitm)
Analysis of Variance Table

Model 1: pesor ~ 1
Model 2: pesor ~ edadm + pesom + tallam + sem + ingr + pesop + tallap +
  tabaco
   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      99 16.3475
2      91  2.6329  8    13.715 59.253 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

y del mismo modo, con esta función *anova* puede desarrollarse el contraste de significación de cualquier subconjunto de variables predictoras del modelo de regresión.

No obstante, uno de los procesos de selección del modelo más sencillos para desarrollar en la práctica, es la selección por etapas *Backward* a través de los contrastes de significación individual de las predictoras. En este caso, partimos del modelo de regresión completo, *fitm* y fijamos un α_{salir} (por ejemplo, $\alpha_{salir} = 0.1$). En las sucesivas etapas se elimina la predictora menos significativa. Así, la primera candidata a ser suprimida del modelo es la predictora *edadm*, dado que su *p*-valor $p = 0.683 > 0.1$, es el menos significativo de todos, volvemos a estimar el modelo de regresión sin la edad de la madre y analizamos sus contrastes de significación.

Por ejemplo, llamando *fit7* al modelo que mantiene 7 predictoras y utilizando la librería *lmtest* se obtiene:

```
> library(lmtest)
> fit7 <- lm(pesor ~ . -edadm, data=peso)
> coeftest(fit7)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.3758737	0.8311462	-4.0617	0.0001023	***
pesom	0.0189313	0.0042506	4.4538	2.369e-05	***
tallam	-0.0082605	0.0052075	-1.5863	0.1161084	
sem	0.1301263	0.0198920	6.5416	3.374e-09	***
ingr	0.0593984	0.0167322	3.5500	0.0006093	***
pesop	0.0077083	0.0022054	3.4951	0.0007310	***
tallap	0.0060250	0.0028702	2.0992	0.0385421	*
tabaco	0.0040136	0.0022973	1.7471	0.0839564	.

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

donde se observa el efecto de la eliminación de la variable *edadm* en la significación del resto de predictoras, aunque está no era significativa. Ahora, la predictora menos significativa es la altura de la madre, *tallam*, correspondiente al mayor *p*-valor de las que permanecen en el modelo, $p = 0.116 > 0.1$, y repetimos el proceso con el modelo de regresión con 6 predictoras:

```
> fit6 <- lm(pesor ~ . -edadm-tallam, data=peso)
> coeftest(fit6)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.2039888	0.6519868	-6.4480	5.004e-09	***
pesom	0.0138662	0.0028284	4.9025	3.999e-06	***
sem	0.1283388	0.0200213	6.4101	5.945e-09	***
ingr	0.0577618	0.0168359	3.4309	0.0008991	***
pesop	0.0076412	0.0022229	3.4374	0.0008800	***
tallap	0.0051550	0.0028402	1.8150	0.0727414	.
tabaco	0.0036340	0.0023034	1.5777	0.1180268	

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

En estos resultados, se observa que al suprimir la altura de la madre, entre otros, se ha producido una pérdida en el nivel de significación de la variable *tabaco*, siendo la menos significativa del modelo, $p = 0.118 > 0.1$, y por tanto, analizamos el modelo de regresión con 5 predictoras:

```
> fit5 <- lm(pesor ~ . -edadm-tallam-tabaco, data=peso)
> coeftest(fit5)
```

t test of coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-4.4698295	0.6348071	-7.0412	3.095e-10	***
pesom	0.0144444	0.0028267	5.1100	1.687e-06	***
sem	0.1368218	0.0194380	7.0389	3.129e-10	***
ingr	0.0538785	0.0167864	3.2096	0.001819	**
pesop	0.0064475	0.0021067	3.0605	0.002881	**

```
tallap      0.0053703  0.0028593  1.8782  0.063455 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este último modelo de regresión, las variables predictoras que permanecen han aumentado su significación (reducido sus p -valores) siendo todos inferiores al $\alpha_{salir} = 0.1$ fijado al inicio. Por tanto, este método de selección concluye que estas son las cinco predictoras más relevantes para analizar el peso del recién nacido a través de un modelo de regresión lineal múltiple, siendo el modelo estimado:

$$pesor = -4.469 + 0.014pesom + 0.1368sem + 0.0538ingr + 0.006pesop + 0.005tallap$$

donde, entre otros aspectos, cabe señalar que el coeficiente de determinación del modelo completo, $R^2 = 0.8389$, sólo se ha reducido ligeramente a $R^2 = 0.8298$ en el modelo seleccionado, es decir, el modelo de regresión prácticamente no ha perdido variabilidad explicada aún habiéndose descartado la información de tres predictores.

Idénticos resultados se obtienen actualizando el modelo de regresión en cada paso al suprimir cada variable predictora para salir del modelo, como puede desarrollarse utilizando las siguientes instrucciones en R:

```
fitmult <- lm(pesor ~ . , data=peso)
fitmult <- update(fitmult, . ~ . -edadm)
coeftest(fitmult)
fitmult <- update(fitmult, . ~ . -tallam)
coeftest(fitmult)
fitmult <- update(fitmult, . ~ . -tabaco)
coeftest(fitmult)
```

o bien, al aplicar la función *fastbw* de selección *Backward* de predictoras que incluye la librería *rms*, utilizando previamente la función de mínimos cuadrados *ols* para ajustar el modelo de regresión, e indicando en los argumentos la regla de selección con el criterio de Akaike ("aic") o con los p -valores de los contrastes de significación ("p") junto con el límite para eliminar una predictora ($sls=\alpha_{salir}$), como se muestra a continuación:

```
> library(rms)
> fitrms <- ols(pesor ~ sem+pesom+pesop+ingr+tallap+tabaco+tallam+edadm)
> fastbw(fitrms, rule="p", sls=0.1)
```

	Deleted	Chi-Sq	d.f.	P	Residual	d.f.	P	AIC	R2
edadm	0.17	1	0.6817	0.17	1	0.6817	-1.83	0.839	
tallam	2.49	1	0.1143	2.66	2	0.2642	-1.34	0.834	
tabaco	2.51	1	0.1133	5.17	3	0.1599	-0.83	0.830	

Approximate Estimates after Deleting Factors

	Coef	S.E.	Wald Z	P
Intercept	-4.469829	0.627609	-7.122	1.064e-12
sem	0.136822	0.019218	7.120	1.082e-12
pesom	0.014444	0.002795	5.169	2.359e-07
pesop	0.006447	0.002083	3.096	1.964e-03
ingr	0.053879	0.016596	3.246	1.169e-03
tallap	0.005370	0.002827	1.900	5.747e-02

Factors in Final Model

```
[1] sem    pesom  pesop  ingr   tallap
```

Obsérvese que el modelo de regresión seleccionado debe analizarse como en los casos anteriores, incluyendo el análisis de las condiciones iniciales a través de los residuos, lo que permitirá medir y validar el ajuste del modelo con el propósito de pronosticar el peso de un recién nacido según los valores de estas cinco características, dejándose como ejercicio propuesto al lector para su aplicación práctica.

Asimismo, el proceso de selección de un modelo de regresión puede realizarse utilizando otros procedimientos y otras librerías de R. Por ejemplo, utilizando la librería *leaps*, la función *regsubsets* proporciona los modelos de regresión ordenados para cada cantidad de predictoras en el modelo estimado, considerando el coeficiente de determinación R^2 o el C_p de Mallows como medidas o criterios del orden entre los modelos con igual número de variables. En este caso, declaramos en el argumento que presente los dos mejores subconjuntos de variables para cada número de predictoras en el modelo de regresión:

```
> library("leaps")
> subconj <- regsubsets(pesor ~ ., nbest=2, data=peso)
> summary(subconj)
Subset selection object
Call: regsubsets.formula(pesor ~ ., nbest = 2, data = peso)
8 Variables (and intercept)
      Forced in Forced out
edadm      FALSE      FALSE
pesom      FALSE      FALSE
tallam      FALSE      FALSE
sem         FALSE      FALSE
ingr        FALSE      FALSE
pesop       FALSE      FALSE
tallap      FALSE      FALSE
tabaco      FALSE      FALSE
2 subsets of each size up to 8
Selection Algorithm: exhaustive
```

		edadm	pesom	tallam	sem	ingr	pesop	tallap	tabaco
1	(1)	" "	" "	" "	"*	" "	" "	" "	" "
1	(2)	" "	" "	" "	" "	" "	" "	"*	" "
2	(1)	" "	"*	" "	"*	" "	" "	" "	" "
2	(2)	" "	" "	" "	"*	" "	"*	" "	" "
3	(1)	" "	"*	" "	"*	" "	"*	" "	" "
3	(2)	" "	"*	" "	"*	"*	" "	" "	" "
4	(1)	" "	"*	" "	"*	"*	"*	" "	" "
4	(2)	" "	"*	" "	"*	"*	" "	"*	" "
5	(1)	" "	"*	" "	"*	"*	"*	"*	" "
5	(2)	" "	"*	" "	"*	"*	"*	" "	"*
6	(1)	" "	"*	" "	"*	"*	"*	"*	"*
6	(2)	" "	"*	"*	"*	"*	"*	"*	" "
7	(1)	" "	"*	"*	"*	"*	"*	"*	"*
7	(2)	"*	"*	" "	"*	"*	"*	"*	"*
8	(1)	"*	"*	"*	"*	"*	"*	"*	"*

En esta salida de resultados, se observa que aparecen ordenados por filas los mejores modelos de regresión para cada número de predictoras, siendo las variables incluidas en cada modelo las predictoras correspondientes a las columnas marcadas con "*" en dicha fila. Además, todos estos modelos pueden ordenarse con respecto a sus valores de C_p de Mallows, R^2 o R_a^2 , representando una tabla gráfica cuyas casillas de cada fila identifican cada modelo, por ejemplo con la instrucción `plot(subconj, scale="Cp")` o cambiando el valor del argumento `scale="r2"`.

No obstante, para facilitar la lectura e interpretación de estos estadísticos, repetimos la instrucción anterior para que calcule únicamente el mejor modelo de regresión para cada cantidad de predictoras ($nbest=1$), indicando que presente los valores de estos estadísticos para cada uno de los modelos y la representación de dichos valores según el número de variables predictoras en el modelo seleccionado (Figura 11):

```
> subconj <- regsubsets(pesor~., nbest=1, data=peso)
> valoresCp <- summary(subconj)$cp
> valoresR2 <- summary(subconj)$rsq
> valoresCp
[1] 106.998510 40.943888 20.539116 9.777641 8.168665 7.661769 7.168251
[8] 9.000000
> valoresR2
[1] 0.6407254 0.7611711 0.8008239 0.8234097 0.8297970 0.8342338 0.8386469
[8] 0.8389447
> valoresp <- subconj$first:subconj$last
> plot(valoresp, valoresCp, xlab="Nº coeficientes", ylab=expression(C[p]))
> abline(h=subconj$np, col="red", lty="dashed")
> plot(valoresp, valoresR2, xlab="Nº coeficientes", ylab=expression(R^2))
```

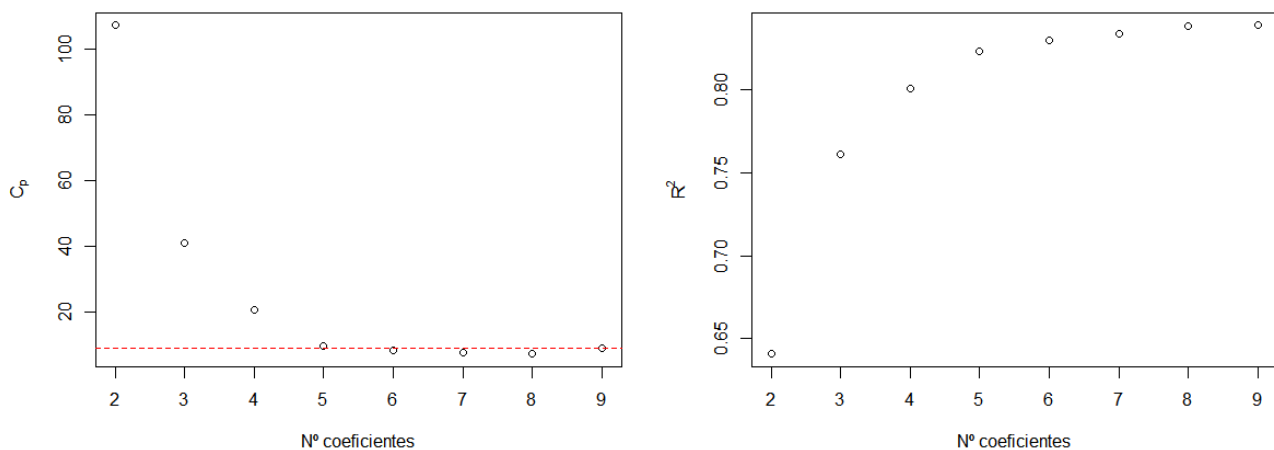


Figura 11: Gráficas para valores de C_p de Mallows y R^2

En esta Figura 11, se observa que en la reducción a 6 términos en el modelo, es decir, 5 variables predictoras más el intercepto del modelo, se obtiene un valor $C_5 = 8.168$ que se considera aceptable al ser inferior la número de términos del modelo completo, y su coeficiente de determinación $R^2 = 0.829$ se mantiene próximo a su valor máximo del modelo completo. Estos resultados, corroboran la selección de las cinco variables predictoras realizada en el procedimiento anterior.

Por otro lado, también puede aplicarse algoritmos automáticos de selección de predictoras, como el utilizado anteriormente con la función *fastbw* de la librería *rms*, que permiten elegir entre los diferentes métodos de selección por etapas, es decir, de introducción de predictoras *forward*, de reducción *backward* o por etapas entre ambos *both*. Para realizar dichos procedimientos en R, comprobamos que está cargada la librería *MASS*, y tendremos en cuenta que estos algoritmos de selección permiten introducir un modelo mínimo con un subconjunto de predictoras fijas en el modelo y un modelo máximo con el mayor conjunto de predictoras disponible. En nuestro caso, ya hemos establecido el modelo de regresión constante *fit0* y el modelo de regresión completo *fitm*, por lo que ejecutamos en R el comando *step* con los argumentos siguientes:

```
seleccionForw <- step(fit0, scope=list(lower=fit0, upper=fitm),
```



```
+ direction="forward", trace=0, criterion="AIC", k=2)
```

y aplicando las funciones *anova* y *summary*, presenta la tabla con la significación de las variables predictoras en el orden que han sido incluidas en el modelo, así como el resumen del ajuste del modelo de regresión seleccionado:

```
> anova(seleccionForw)
Analysis of Variance Table

Response: pesor
      Df Sum Sq Mean Sq  F value    Pr(>F)
sem      1 10.4743  10.4743 365.3276 < 2.2e-16 ***
pesom     1  1.9690   1.9690  68.6755 9.078e-13 ***
pesop     1  0.6482   0.6482  22.6092 7.317e-06 ***
ingr      1  0.3692   0.3692  12.8779 0.0005354 ***
tallap    1  0.1044   0.1044   3.6419 0.0594605 .
tabaco    1  0.0725   0.0725   2.5298 0.1151473
tallam    1  0.0721   0.0721   2.5163 0.1161084
Residuals 92  2.6377   0.0287
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(seleccionForw)

Call:
lm(formula = pesor ~ sem + pesom + pesop + ingr + tallap + tabaco + tallam)

Residuals:
      Min       1Q   Median       3Q      Max
-0.40604 -0.08702 -0.03087  0.08518  0.49152

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.375874    0.831146  -4.062 0.000102 ***
sem           0.130126    0.019892   6.542 3.37e-09 ***
pesom         0.018931    0.004251   4.454 2.37e-05 ***
pesop         0.007708    0.002205   3.495 0.000731 ***
ingr          0.059398    0.016732   3.550 0.000609 ***
tallap        0.006025    0.002870   2.099 0.038542 *
tabaco        0.004014    0.002297   1.747 0.083956 .
tallam       -0.008260    0.005207  -1.586 0.116108
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1693 on 92 degrees of freedom
Multiple R-squared:  0.8386, Adjusted R-squared:  0.8264
F-statistic: 68.31 on 7 and 92 DF,  p-value: < 2.2e-16
```

De forma similar a los métodos anteriores, se observa en la lista de variables predictoras seleccionadas en el modelo de regresión, que en cada paso incluye la predictora que proporciona mayor aportación en la suma de cuadrados explicada por el modelo, y por tanto, mayor reducción en la suma de cuadrados de los residuos. Además, en este caso el proceso finaliza con 7 variables predictoras en el modelo de regresión, incluyendo las variables *tallam* y *tabaco* cuyas aportaciones con respecto al

modelo con las cinco primeras predictoras no son significativas ($p > 0.115$), y por consiguiente, sólo elimina la edad de la madre del modelo final seleccionado. Obsérvese que este modelo de regresión con sólo una predictora menos también resulta aceptable según el criterio de Mallows, $C_p = 7.168$.

Análogamente, se obtiene conclusión al ejecutar los algoritmos de selección por etapas *backward* o *both*, que puede comprobarse como ejercicio utilizando los siguientes argumentos, respectivamente:

```
step(fitm, scope=list(lower=fit0, upper=fitm),
    + direction="backward", trace=0, criterion="AIC", k=2)
step(fit0, scope=list(lower=fit0, upper=fitm),
    + direction="both", trace=0, steps=100, k=2)
```

Otra forma de realizar estos algoritmos de selección del modelo de regresión es ejecutar la función *stepAIC*, también incluida en la librería *MASS*, y proporciona la misma salida que las instrucciones anteriores. A modo de ejemplo se muestra a continuación la tabla de significación de la función *anova* sobre el proceso de selección por etapas hacia ambos lados:

```
> seleccionAIC <- stepAIC(fit0, scope=list(lower=fit0, upper=fitm),
    + scale=0, trace=0, direction="both", steps=100, k=2)
> anova(seleccionAIC)
Analysis of Variance Table

Response: pesor
      Df Sum Sq Mean Sq  F value    Pr(>F)
sem      1 10.4743  10.4743 365.3276 < 2.2e-16 ***
pesom     1  1.9690   1.9690  68.6755 9.078e-13 ***
pesop     1  0.6482   0.6482  22.6092 7.317e-06 ***
ingr      1  0.3692   0.3692  12.8779 0.0005354 ***
tallap    1  0.1044   0.1044   3.6419 0.0594605 .
tabaco    1  0.0725   0.0725   2.5298 0.1151473
tallam    1  0.0721   0.0721   2.5163 0.1161084
Residuals 92  2.6377   0.0287
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ejercicio 2.1 Analizar del modelo de regresión para el peso de un recién nacido con las siete predictoras seleccionadas con el criterio de Akaike.

- (1) Presentar el resumen del ajuste del modelo de regresión y comentar los principales resultados.
- (2) Obtener e interpretar los intervalos de confianza para los coeficientes del modelo.
- (3) Representar los gráficos de los residuos e interpretarlos para el diagnóstico del modelo.
- (4) Realizar e interpretar los contrastes de hipótesis del diagnóstico del modelo ajustado.

Ejercicio 2.2 A partir de las variables predictoras semanas de gestación, peso de la madre, peso del padre, altura del padre e ingresos familiares que han sido seleccionadas en el proceso backwards con el criterio de niveles de significación:

- (1) Analizar el modelo de regresión para el peso de un recién nacido, siguiendo el esquema de análisis del ejercicio anterior.
- (2) Comparar la diferencia entre este modelo de regresión y el analizado en el ejercicio anterior.

Referencias

Bibliografía

- Draper, N.R.; Smith, H. (1998). Applied Regression Analysis, 3rd. John Wiley.
- Everitt, B.S.; Hothorn, T. (2010). A Handbook of Statistical Analysis Using R. Chapman Hall.
- Faraway, J. (2004). Linear Models with R. CRC Press.
- Faraway, J. (2005). Extending the Linear Model with R. CRC Press.
- García Pérez, A. (2008). Estadística aplicada con R. UNED.
- González Ortiz, F.J. (2007). Prácticas de Estadística con R (Parte I y Parte II). Universidad de Cantabria.
- Peña, D. (2002). Análisis de Datos Multivariantes. McGraw-Hill.

Recursos en Internet

- An Introduction to R: <http://cran.r-project.org/doc/manuals/R-intro.pdf>
- Comunidad R hispano: <http://www.r-es.org/>
- Curso introducción R: <http://www.uv.es/conesa/CursoR/cursor.html>
- icebreakeR: <http://www.ms.unimelb.edu.au/~andrewpr/r-users/icebreakeR.pdf>
- Introduction to Data Technologies: <http://www.stat.auckland.ac.nz/~paul/ItDT/itdt-2010-11-01.pdf>
- Practical Regression and ANOVA in R: on CRAN, Faraway, J.
- Quick-R: <http://www.statmethods.net/>
- RStudio: <http://rstudio.org/>