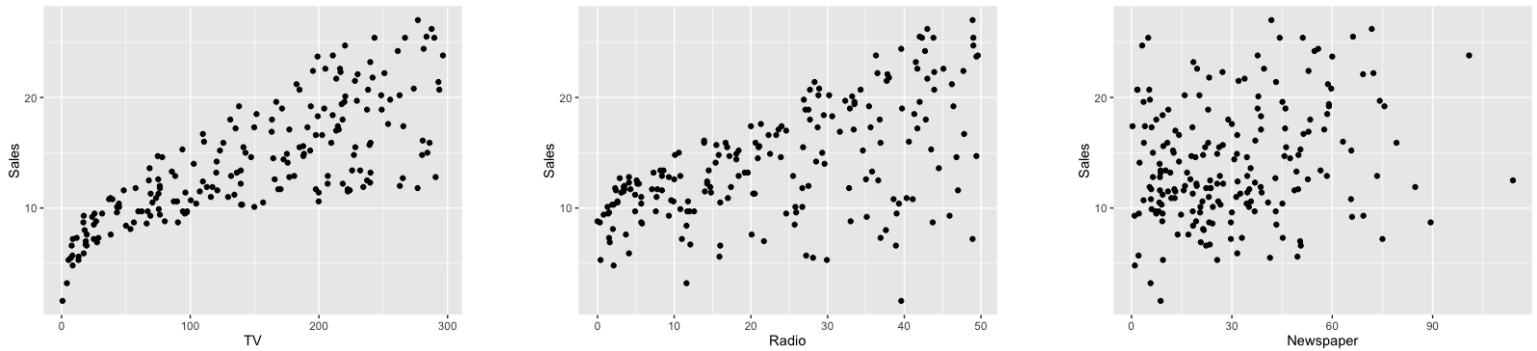
Regresión Lineal

Máster en Data Science y Big Data

Mario Encinar, PhD, MAPFRE
encinar@ucm.es

Ejemplo: Queremos aconsejar a una compañía sobre cómo mejorar las ventas de un determinado producto (en miles de unidades). Para conseguirlo, nos proporcionan un set de datos que contiene las ventas del producto en 200 mercados diferentes, junto con el presupuesto de publicidad en televisión, radio y periódicos en cada uno de tales mercados (en miles de dólares).



Naturalmente, nuestro cliente no tiene la capacidad de incrementar las ventas de su producto directamente, pero sí tiene la capacidad de decidir qué presupuesto dedicar a cada uno de los canales de publicidad.

Para responder al cliente, lo ideal sería encontrar una función f de modo que

$$\texttt{Sales} = f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

y, si P es el presupuesto total del que dispone nuestro cliente, la cuestión se reduciría a buscar el máximo de `Sales` sobre el conjunto

$$\{\texttt{TV} \geq 0, \texttt{Radio} \geq 0, \texttt{Newspaper} \geq 0, \texttt{TV} + \texttt{Radio} + \texttt{Newspaper} \leq P\}.$$

El problema es que encontrar una tal f no es fácil y, en la mayoría de los casos, es imposible, por tres motivos fundamentales:

- Los datos disponibles estarán, generalmente, afectados de ruido.
- Sería de extrañar que no hubiese ninguna otra variable relevante en el estudio.
- Incluso si no se dan los dos anteriores, los datos siempre serán insuficientes para determinar f con seguridad.

En este escenario, a las variables de presupuesto se les llama **variables de entrada (*inputs*, predictores, *features*, variables explicativas o variables independientes)**, y suelen denotarse usando la letra X, con algún subíndice para distinguir a unas de otras.

Por su parte, a la variable dependiente `Sales` se le llama **variable de salida, respuesta o variable objetivo**, y suele denotarse por la letra Y.

Más generalmente, supondremos que se observa una respuesta **cuantitativa** Y y p predictores diferentes $X_1, X_2, \ldots, X_p$. Suponemos que existe una cierta relación entre $X = (X_1, X_2, \ldots, X_p)$ e Y que es de la forma

$$Y = f(X) + \epsilon,$$

donde f es una función fija (pero desconocida) y ϵ es un término de **error** que es independiente de X y que tiene **media** 0.

Por otro lado, f representa la información **sistemática** que aporta X sobre Y.

Un **problema de regresión** consiste en estimar una función hipótesis h que aproxime a f de la mejor forma posible, al menos, en los datos disponibles.

Existen diversos motivos por los que uno querría estimar dicha hipótesis, que principalmente pueden clasificarse en relativos a la **realización de predicciones** y a los relativos a **inferencia de relaciones**.

En multitud de situaciones, un cierto conjunto de entradas X está disponible pero no disponemos de las respuestas Y (o no se pueden obtener de una forma sencilla). Si h está construida de forma *razonable*, $h(X)$ será un estimador insesgado de Y y, por tanto, tiene sentido **predecir** Y mediante

$$\hat{Y} = h(X).$$

La bondad de tal estimación puede cuantificarse mediante

$$\begin{aligned} \mathrm{E}[(Y - \hat{Y})^2] &= \mathrm{E}[(f(X) + \epsilon - h(X))^2] \\ &= \mathrm{E}[(f(X) - h(X))^2] + \mathrm{V}(\epsilon) \end{aligned}$$

Una mejor elección de h puede disminuir el término $\mathrm{E}[(f(X) - h(X))^2]$ (**error reducible**) pero, en principio, no hay nada que podamos hacer con $\mathrm{V}(\epsilon)$.

En este caso, la forma de h no nos importa y, de hecho, con frecuencia se emplea como una *caja negra*. Lo único que importa es hacer que el error sea el mínimo posible.

En otro tipo de situaciones estamos interesados en **entender** de qué forma se ve afectada Y por los cambios de X. Concretamente, uno puede querer responder preguntas del tipo:

- ¿Qué predictores tienen relación con la variable respuesta?
- ¿Qué tipo de relación hay entre la variable respuesta y cada uno de los predictores?
- ¿Se puede afirmar que el crecimiento de Y con respecto a X_1 es lineal, cuadrático o de orden superior?

Para ello, uno necesita conocer la forma de h (y, naturalmente, que el error reducible no sea demasiado grande).

Volviendo al ejemplo del *dataset* `Advertising`, para sugerir un plan de marketing a nuestro cliente, podríamos plantearnos responder a las siguientes preguntas:

- ¿Hay relación entre el presupuesto en publicidad y las ventas? Si es que la hay, ¿cómo de fuerte es? ¿Es lineal?
- ¿Qué canales de publicidad contribuyen a las ventas?
- ¿Con cuánta precisión podemos estimar el efecto de cada canal de publicidad en las ventas?
- ¿Con cuánta precisión podemos estimar ventas futuras en base al presupuesto en publicidad?
- ¿Existe interacción entre los distintos canales de publicidad?

El **modelo de regresión lineal** puede ayudarnos a responder a cada una de estas preguntas.

El **modelo de regresión lineal simple** es la forma más sencilla de predecir una respuesta cuantitativa Y en base a una variable predictora X, y consiste en asumir que existe, aproximadamente, una relación lineal entre ambas variables

$$Y \approx \beta_0 + \beta_1 X.$$

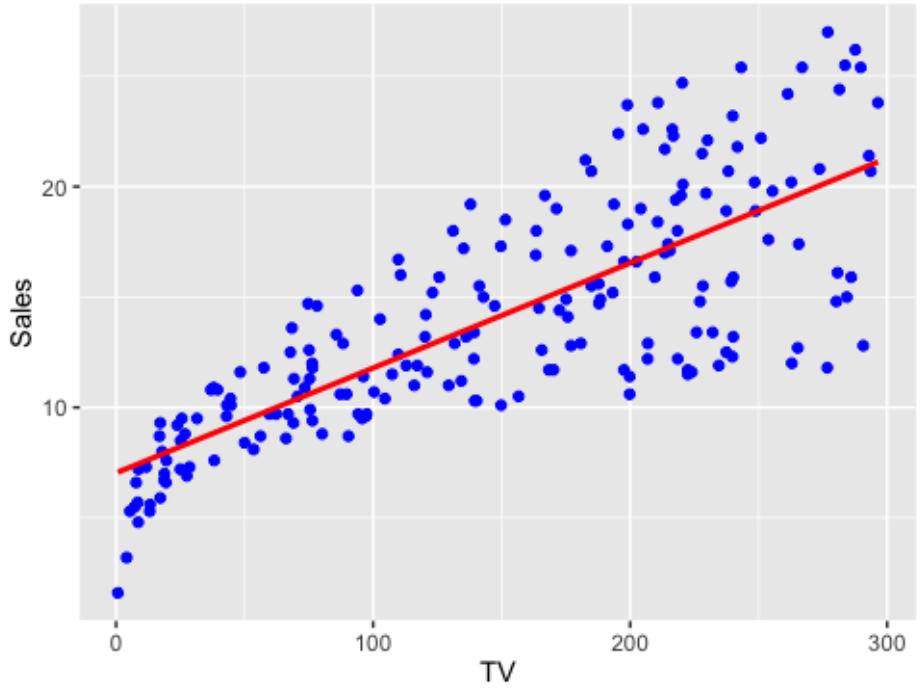
Una vez se han utilizado los datos disponibles para producir estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de los coeficientes, éstas se emplean para predecir valores futuros de la variable respuesta mediante

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

El primer paso, por tanto, es estimar $\hat{\beta}_0$ y $\hat{\beta}_1$.

Supongamos que $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ son n pares observados de las variables (X, Y). Determinar $\hat{\beta}_0$ y $\hat{\beta}_1$ consiste en determinar, en algún sentido, una recta *que pase muy cerca* de todos los puntos (x_i, y_i), es decir

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ para todo } i.$$



La manera estándar de conseguir estas estimaciones es mediante **mínimos cuadrados ordinarios**. Para $i = 1, 2, \ldots, n$, definimos el i-ésimo **residuo** del modelo mediante

$$e_i = y_i - (\beta_0 + \beta_1 x_i),$$

es decir, el error (con signo) que se produce al predecir y_i mediante $\beta_0 + \beta_1 x_i$, y el **error cuadrático medio** mediante

$$\texttt{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2.$$

Es razonable elegir $\hat{\beta}_0$ y $\hat{\beta}_1$ de forma que **minimicen** `MSE` y, de esta forma, se obtiene

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

con lo que el modelo final es

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Para medir la precisión de $\hat{\beta}_0$ y $\hat{\beta}_1$ como estimaciones de los coeficientes *reales*, es necesario asumir que existe una relación real entre X e Y de la forma

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \quad \epsilon \perp\!\!\!\perp X,$$

expresión a la que nos referimos como **recta de regresión poblacional**. Se puede demostrar (con alguna hipótesis extra) que los errores estándar asociados a $\hat{\beta}_0$ y $\hat{\beta}_1$ vienen dados por

$$\texttt{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \texttt{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Aquí, aunque σ es desconocida, suele estimarse mediante el **error residual estándar**

$$\texttt{RSE} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n e_i^2}.$$

Las expresiones anteriores permiten calcular intervalos de confianza y realizar **contrastes de hipótesis relativos a los coeficientes del modelo**, ya que

$$\frac{\hat{\beta}_j - \beta_j}{\texttt{SE}(\hat{\beta}_j)} \sim t_{n-2}, j = 0, 1.$$

La aplicación más habitual de este resultado consiste en contrastar si la **respuesta** está **significativamente influida por alguna variable** (ya sea el término independiente o el predictor), mediante un contraste de hipótesis nula

$$H_0 : \beta_j = 0.$$

La medida absoluta más habitual para estimar la precisión del modelo es **la raíz cuadrada del error cuadrático medio**

$$\texttt{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2},$$

que, para valores *grandes* de n, es muy cercano al `RSE`. Esta magnitud mide, en unidades de la variable respuesta, el error medio que se produce al aproximar valores reales y_i por sus predicciones $\hat{y}_i$.

Para obtener una medición independiente de unidades, se emplea el **estadístico** R^2, que viene dado por

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y que representa la proporción de variabilidad de la variable respuesta que está explicada por el modelo.

Observación: Uno podría estar tentado de resolver esta última cuestión por medio del coeficiente de correlación

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

De hecho, es fácil ver que, en este caso, $R^2 = \rho^2$, pero esto no es así en el caso del modelo de regresión lineal múltiple.

El modelo de regresión lineal simple sólo tiene en cuenta una variable predictora. En general, ¿cómo podemos extender nuestro análisis para utilizar la información de varios predictores?

Una posibilidad es ajustar un modelo lineal simple

$$y = \hat{\beta}_{j0} + \hat{\beta}_{j1} x_j$$

que relacione Y con cada variable predictora X_j, y eso puede ser útil hasta cierto punto.

La mejor posibilidad es utilizar la misma estrategia que en el caso univariante, pero considerar todos los predictores al mismo tiempo. Así, asumimos que existe una **relación lineal** entre $X_1, X_2, \ldots, X_p$ e Y de la forma

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

Dadas n observaciones

$$(x_{i1}, x_{i2}, \ldots, x_{ip}; y_i),$$

definimos el i-ésimo **residuo** del modelo mediante

$$e_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \cdots + \beta_p x_{ip})$$

y elegimos $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ de modo que el **error cuadrático medio**

$$\texttt{MSE} = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

sea **mínimo**.

Para dar una expresión de $\hat{\beta}_j$, $j = 0, 1, \ldots, p$, escribimos

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}, \ \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}, \ \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

y entonces se busca resolver

$$\arg\min_{\hat{\beta}} \left\| \mathbf{y} - \mathbf{X}\hat{\beta} \right\|^2.$$

cuya **solución** es:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

La bondad de ajuste del modelo, al igual que en el caso univariante, se mide mediante el `RMSE`, el estadístico R^2 o, aún mejor, mediante el R^2 **ajustado**, que es una modificación del R^2 que tiene en cuenta el número de predictores empleado, y que viene dado por

$$R_a^2 = R^2 - (1 - R^2)\frac{p}{n - p - 1}.$$

Para contrastar si hay una verdadera relación lineal entre la variable respuesta y los predictores, se realiza el contraste de hipótesis nula

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

cuya hipótesis alternativa es

$$H_A : \text{al menos un } \beta_j \text{ es distinto de } 0.$$

El **estadístico** relevante en este caso es

$$F = \frac{n - p - 1}{p} \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

que, bajo la hipótesis nula, sigue una distribución F de **Snedecor** de parámetros $(p, n - p - 1)$.

Una vez que se sabe que, al menos, hay un predictor que influye en la variable respuesta, es natural preguntarse cuáles son aquellos predictores influyentes para ajustar un modelo que sólo tenga en cuenta a éstos. Podría parecer que tiene sentido realizar contrastes t individuales para cada predictor, pero esto puede llevar a errores si p es grande.

Idealmente, se deberían considerar todos los modelos lineales posibles de Y en función de subconjuntos de $X_1, X_2, \ldots, X_p$ para elegir el mejor de ellos, pero esto no es viable. Dos consideraciones:

- Existen diversas alternativas para comparar la bondad de ajuste de dos modelos lineales y elegir el mejor (`AIC`, `BIC`, C_p, R^2_a, análisis de residuos...).
- También hay diferentes estrategias para guiar la selección de variables (*forward selection*, *backward selection*, *mixed selection*...).

Asumiendo que la relación entre Y y X es aproximadamente lineal y que $n \gg p$, las predicciones obtenidas por el modelo de regresión lineal serán buenas. Cuando $n \approx p$ o $n < p$, esto es falso.

Por esto, se hace necesario construir modelos que empleen **pocas variables** para que el **error** en las predicciones sea **pequeño** y para que el modelo sea interpretable.

Para comparar la bondad de diferentes modelos, o para estimar el error de las predicciones arrojadas por un modelo, caben dos alternativas:

- *Ajustar* el error de entrenamiento para estimar el verdadero error de predicción.
- Emplear un conjunto de validación, que no intervenga en el proceso de entrenamiento del modelo, y calcular los errores de predicción sobre ese conjunto.

Naturalmente, incluir más variables en el proceso de ajuste del modelo hará que el RMSE o el R^2 (de entrenamiento) disminuyan, pero esto no quiere decir que el modelo vaya a ofrecer mejores predicciones.

Denotamos

$$\texttt{RSS} = n\texttt{MSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Para estimar el error de predicción, se emplean los siguientes estadísticos:

- El Criterio de Información de Akaike

$$\texttt{AIC} = \frac{1}{n\hat{\sigma}^2}(\texttt{RSS} + 2p\hat{\sigma}^2) + C \text{ (constante que depende de la muestra)}$$

- El Criterio de Información Bayesiana

$$\texttt{BIC} = \frac{1}{n}(\texttt{RSS} + p\hat{\sigma}^2 \log(n))$$

- El estadístico C_p de Mallows

$$C_p = \frac{1}{n}(\texttt{RSS} + 2p\hat{\sigma}^2)$$

- El R^2 ajustado

$$R_a^2 = R^2 - (1 - R^2)\frac{p}{n - p - 1}$$

El objetivo de la construcción del modelo debe ser minimizar cualquiera de los tres primeros o maximizar el cuarto.

Best subset selection:

1. Sea $\mathcal{M}_0$ el modelo *nulo*, es decir, el que no contiene predictores.
2. Para $k = 1, 2, \ldots, p$:
   1. Ajustar todos los $\binom{p}{k}$ modelos que contienen, exactamente, k predictores.
   2. Elegir $\mathcal{M}_k$, el mejor de ellos (es decir, el que tenga menor `RSS` o mayor R^2).
3. Elegir el mejor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ según usando `AIC`, `BIC`, C_p, R^2_a o error de validación.

¡Hay que probar 2^p modelos distintos!

Forward stepwise selection:

1. Sea $\mathcal{M}_0$ el modelo *nulo*, es decir, el que no contiene predictores.
2. Para $k = 0, 1, \ldots, p-1$:
   1. Ajustar todos los $p - k$ modelos que contienen, exactamente, un predictor añadido a $\mathcal{M}_k$.
   2. Elegir $\mathcal{M}_{k+1}$, el mejor de ellos (es decir, el que tenga menor `RSS` o mayor R^2).
3. Elegir el mejor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ según usando `AIC`, `BIC`, C_p, R^2_a o error de validación.

Backward stepwise selection:

1. Sea $\mathcal{M}_0$ el modelo *completo*, es decir, el que contiene todos los predictores.
2. Para $k = p, p-1, \ldots, 1$:
   1. Ajustar todos los k modelos que tienen los mismos predictores que $\mathcal{M}_k$ salvo uno.
   2. Elegir $\mathcal{M}_{k-1}$, el mejor de ellos (es decir, el que tenga menor `RSS` o mayor R^2).
3. Elegir el mejor modelo entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ según usando `AIC`, `BIC`, C_p, R^2_a o error de validación.

Existen otras estrategias de selección de variables: regularización, reducción de dimensiones...

Variables categóricas: Si una de las variables de nuestro set de datos es categórica, es necesario recodificarla en forma numérica para que el modelo de regresión lineal la acepte.

En general, no tiene sentido codificar una variable categórica como una única variable numérica, y por esto se utilizan variables *dummy*: si la variable categórica X toma v posibles valores $x_1, x_2, \ldots, x_v$, se sustituye por las $v - 1$ variables

$$X_{(k)} = \begin{cases} 1 \text{ si } X = x_k \\ 0 \text{ en otro caso} \end{cases}$$

Problemas potenciales:

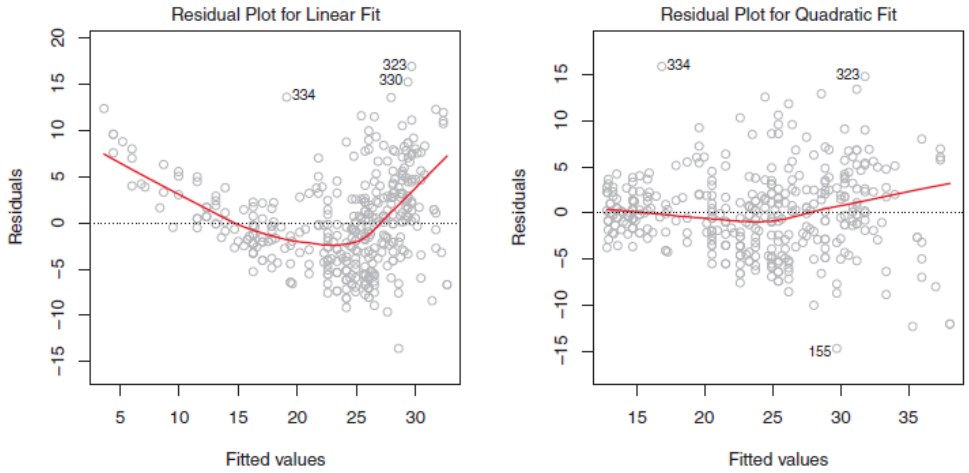
- Relaciones no lineales.
- Varianza no constante de los residuos.
- Outliers.
- High leverage points.
- Colinealidad.

Relaciones no lineales: El modelo lineal funciona razonablemente bien cuando existe una relación lineal entre los predictores y la variable respuesta, pero puede llevar a conclusiones erróneas en caso contrario.

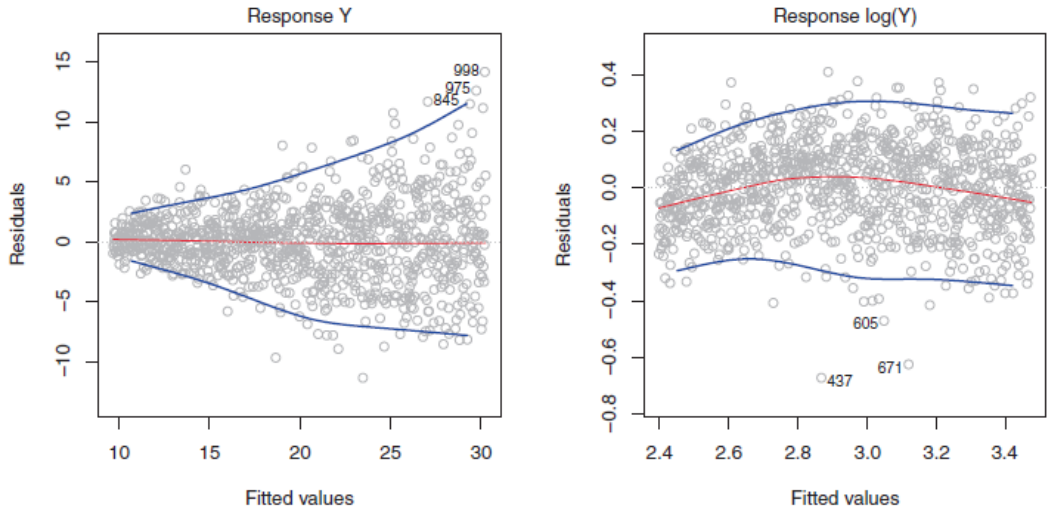
Para identificar este problema, es útil representar los residuos del modelo:

- e_i frente a x_i en el caso unidimensional.
- e_i frente a y_i o $\hat{y}_i$ en el caso multidimensional.

Idealmente, este tipo de diagramas no mostrarán ningún tipo de patrón. En caso de que lo haya, pueden emplearse transformaciones no lineales de los predictores para mejorar el modelo.

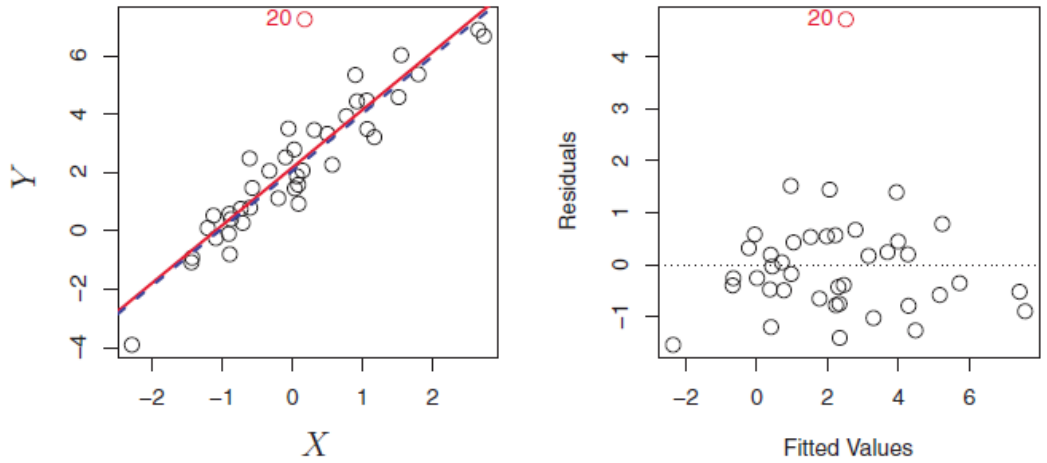


Varianza no constante de los residuos: Otra hipótesis importante del modelo de regresión lineal es que los residuos tienen varianza constante (homocedasticidad). Los diagramas anteriores también sirven para identificar este problema.



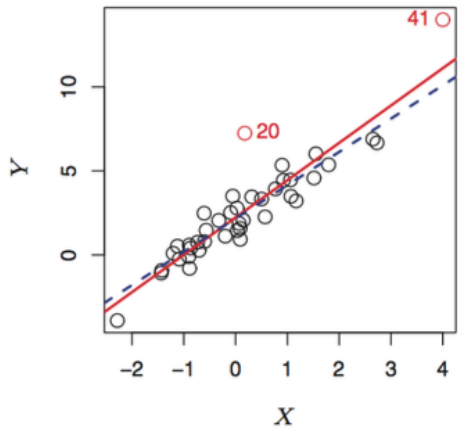
En caso de darse esta situación, puede resolverse aplicando una transformación lineal a la variable respuesta.

Outliers: Un outlier es un punto para el que el error e_i es anormalmente grande.



Típicamente, eliminar outliers en el ajuste del modelo no cambia gran cosa el modelo final, pero puede mejorar drásticamente la bondad de ajuste. En el ejemplo, el modelo con el outlier presenta un R^2 de 0,805, mientras que al eliminar el outlier se obtiene un R^2 de 0,892.

High leverage points: Un *high leverage point* (punto influyente) es, esencialmente, un outlier para las variables predictoras.



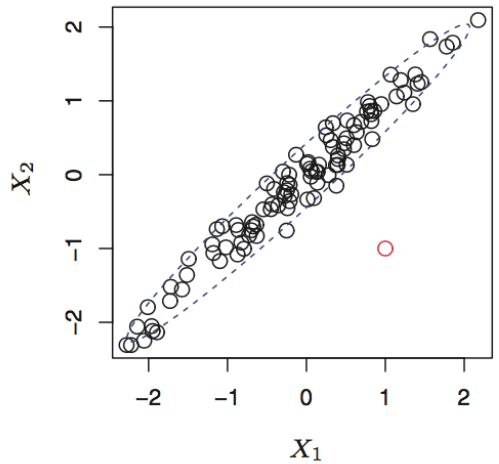
La presencia de high leverage points tiene un impacto reseñable sobre el modelo de regresión lineal estimado, y suele ser recomendable eliminarlos.

Para determinar si una cierta observación j es de high leverage para la variable X, además de realizar un análisis gráfico, se calcula

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Es fácil ver que $\frac{1}{n} \leq h_j \leq 1$, y que la media de los h_j es $\frac{2}{n}$, así que cuando $h_j \gg \frac{2}{n}$, es razonable pensar que la observación j es de high leverage.

El problema es que, a veces, una observación no es de high leverage para ningún predictor en concreto, pero sí para el conjunto de todos los predictores.



Colinealidad: Se dice que en un problema de regresión existe colinealidad cuando dos o más predictores están estrechamente relacionados (de forma lineal) entre sí (es decir, cuando son *casi* linealmente dependientes).

La colinealidad de los predictores produce problemas porque:

- Introduce incertidumbre en las estimaciones de los coeficientes.
- Puede hacernos introducir variables irrelevantes en el modelo.

Para identificar variables colineales, se puede

- Buscar pares de variables (X_i, X_j) para las que $\rho(X_i, X_j)^2 \approx 1$.
- Calcular el *variance inflation factor*

$$\texttt{VIF}(X_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}.$$

Usualmente, un $\texttt{VIF} > 5$ sugiere que la variable en cuestión es colineal con el resto.

Referencias:

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Springer. ISBN: 978-1-4614-7138-7.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of Statistical Learning. Data Mining, Inference and Prediction.* Springer. ISBN: 978-0-387-84858-7.

Apéndice I: Propiedades de los Estimadores

- **Sesgo**: Diferencia entre la esperanza matemática del estimador y el valor numérico del parámetro que estima:

$$E(T) - \theta$$

  Un estimador cuyo sesgo es nulo se llama insesgado o centrado.

- **Eficiencia**: se dice que un estimador es más eficiente o más preciso que otro estimador, si la varianza del primero es menor que la del segundo. Por ejemplo, si $\hat{\theta}_1$ y $\hat{\theta}_2$ son ambos estimadores de θ y

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$$

  se dice que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$. Un estimador es más eficiente (más preciso), por tanto, cuanto menor es su varianza.

- **Consistencia**: a medida que el tamaño de la muestra crece, el valor del estimador tiende a ser el valor del parámetro.

- **Robustez, Suficiencia, Invariancia**

Apéndice II: Relación Sesgo - Varianza

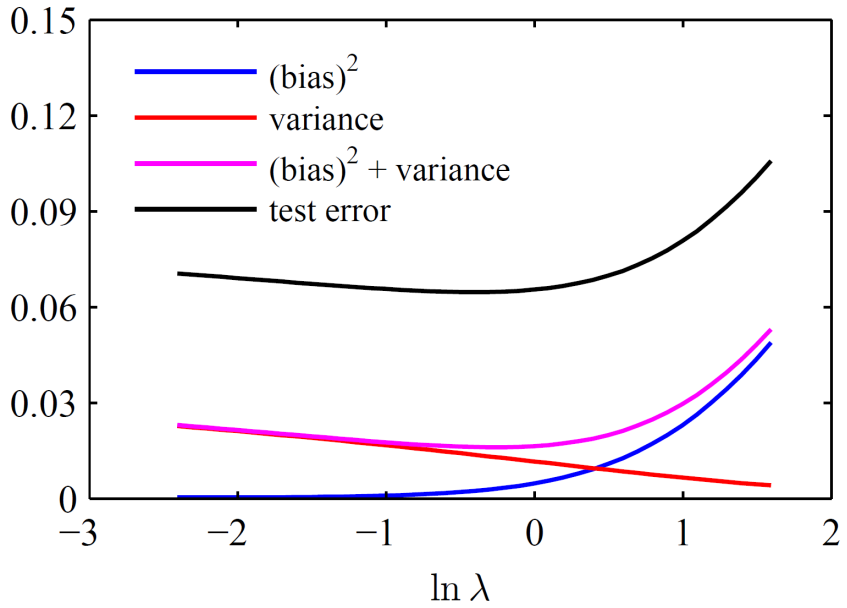
El error cuadrático medio (sin el término estocástico) se descompone a su vez en:

$$\mathrm{E}[(f(X) - h(X))^2] = \mathrm{E}[(h(X) - \bar{h}(X))^2] + (f(X) - \bar{h}(X))^2$$

El primer término es la varianza de la hipótesis h en X, y depende fuertemente de la muestra X.

El segundo término es el cuadrado del sesgo (o error sistemático), que se asocia a la clase de hipótesis que estamos considerando.

Descomposición del Error



El eje x mide la **complejidad** de la hipótesis (que **crece de derecha a izquierda**).

- Las hipótesis más simples tienen usualmente mayor sesgo. Un sesgo alto puede conducir a que el modelo no capture las relaciones relevantes entre predictores y variable objetivo.
- Las hipótesis complejas tienen una varianza grande, la hipótesis es muy dependiente del conjunto de datos en el que se entrena. El modelo puede llegar a ajustar el ruido aleatorio, en lugar de las salidas propuestas (**sobreajuste**).

Apéndice III: Propiedades del estimador de mínimos cuadrados ordinarios (OLS)

Dado que $(\mathbf{X}^t\mathbf{X})\hat{\beta} = \mathbf{X}^t\mathbf{y}$ se tiene que $\mathbf{X}^t e = 0$. Los valores observados de los predictores y los residuos son ortogonales, y es consecuencia de la estimación por mínimos cuadrados. Las siguientes propiedades derivan todas de este resultado.

- **La suma de los residuos es cero.**
  La matriz $\mathbf{X}$ tiene como primera columna un vector de unos (relacionado con el término independiente) y de $\mathbf{X}^t e = 0$ se tiene que $\sum e_i = 0$.
- **La media de los residuos es cero.**
  Desde lo anterior, se tiene que $\bar{e} = \frac{\sum e_i}{n} = 0$.
- **Los valores observados de X están incorrelados con los residuos.**
  Si añadimos a $\mathbf{X}^t e = 0$ el resultado anterior se tiene que cada regresor tiene una correlación muestral nula con los residuos.

- **El hiperplano de regresión pasa por la media de los valores observados $(\mathbf{X}$ e $\mathbf{y}$.**
  Se sigue del hecho de que $\bar{e} = \bar{y} - \bar{\mathbf{X}}\hat{\beta} = 0$, es decir, $\bar{y} = \bar{\mathbf{X}}\hat{\beta}$
- **Los valores predichos de y están incorrelados con los residuos**
  Los valores predichos de $\mathbf{y}$ son tales que:

$$\hat{\mathbf{y}}^t e = (\mathbf{X}\beta)^t e = \beta^t \mathbf{X}^t e = 0$$

- **La media de los valores predichos $\mathbf{y}$ igualan siempre la media de las $\mathbf{y}$ observadas i.e. $\bar{\hat{\mathbf{y}}} = \bar{\mathbf{y}}$**

Estas propiedades siempre son ciertas, pero hay que tener cuidado de no inferir que alguna propiedad de los residuos la cumplen los errores.

Apéndice III: Hipótesis de Gauss-Markov

1. $\mathbf{y} = \mathbf{X}\beta + \epsilon$
   Asume que hay una relación lineal entre $\mathbf{X}$ e $\mathbf{y}$
2. $\mathbf{X}$ es una matriz de rango completo
   No hay multicolinealidad perfecta, todas las columnas de $\mathbf{X}$ son linealmente independientes.
3. $E[\epsilon \mid \mathbf{X}] = 0$
   Los errores promedian a cero dado cualquier valor de $\mathbf{X}$. Ninguna observación de las variables independientes incluye información acerca del valor del error. Esta hipótesis implica que $E[\mathbf{y}] = \mathbf{X}\beta$.
4. $E[\epsilon\epsilon^t \mid \mathbf{X}] = \sigma^2 I$
   Esto captura la hipótesis de homocedasticidad y no autocorrelación.
5. $\mathbf{X}$ puede ser determinista o aleatoria pero debe generarse por un mecanismo que no se relaciona con ϵ.
6. $\epsilon \mid \mathbf{X} \sim N(0, \sigma^2 I)$
   Esta hipótesis no se requiere para el Teorema de Gauss Markov pero se asume para hacer test de hipótesis. Se recurre al Teorema Central del Límite para justificarla.

Teorema de Gauss-Markov

Bajo estas condiciones el Teorema de Gauss-Markov asevera que no existe otro estimador lineal insesgado de los coeficientes β que tenga una varianza muestral inferior (BLUE, Best Linear Unbiased and Efficient Estimator) y se trata del estimador de mínimos cuadrados ordinarios (OLS):

$$\arg\min_{\hat{\beta}} \left\| \mathbf{y} - \mathbf{X}\hat{\beta} \right\|^2.$$

cuya **solución** es:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$