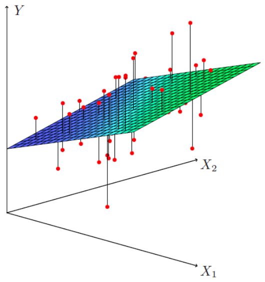
Regresión Lineal Múltiple. Parte 1

Departamento de Estadística e I.O.
Universitat de València

Regresión Lineal Múltiple

Regresión lineal múltiple

- Cuando se dispone de varios predictores, se puede plantear una regresión lineal múltiple
- Y usar Mínimos cuadrados para estimar los coeficientes:
  - Valores ajustados: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
  - Residuos: $e_i = y_i - \hat{y}_i$
  - Minimizar: $\sum_i e_i^2$



Fuente: An Introduction to Statistical Learning, with applications in R. (James et al.)

Regresión con dos predictores: La predicción por mínimos cuadrados, sería el plano que minimiza las distancias verticales todas las observaciones (rojo).

Modelo lineal múltiple

- Asumiendo normalidad y homocedasticidad se tiene el modelo de regresión lineal múltiple normal homocedástico

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{con} \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Razones para ajustar un modelo múltiple

- Si el modelo tiene un fin predictivo: el objetivo es mejorar la predicción tanto como sea posible.
  - **Criterio**: Sólo incluir aquellas variables que aumentan la varianza explicada por el modelo u otra medida de bondad.
- Si modelo tiene un fin explicativo del efecto de un predictor de interés: el objetivo es controlar la posible confusión causada por factores no medidos. Los confusores pueden crear relaciones artefactuales entre el predictor y la respuesta!
  - **Criterio**: incluir todas aquellas variables relacionadas con la respuesta y el predictor que producen un cambio sensible ($10\,\%$ ó $20\,\%$) en el coeficiente de la variable de interés (confusores).
  - Cuidado con los efectos muy pequeños es fácil producir sobre ellos un cambio sensible
  - Una variable es confusora cuando estando relacionada con alguna variable independiente, a su vez afecta a la dependiente.

Confusores: un ejemplo

- Datos 'deportistas.csv', X = 'MCMagra' e Y = 'PrctGrasa'
  - Parece que la relación es decreciente

```
(Intercept) MCMagra
24.6246 -0.1714
```

  - Pero es creciente para los hombres

```
(Intercept) MCMagra
0.3433 0.1193
```

  - También es creciente para las mujeres

```
(Intercept) MCMagra
0.2846 0.3200
```

- Paradoja de Simpson: Hay un factor oculto: el género, que genera una asociación ficticia con sentido inverso al real.
  - Hay que incluir en el modelo los confusores, para identificar la verdadera relación

```
(Intercept) Genero MCMagra
7.7246 -12.2430 0.18443
```

Mejora en la predicción: El mismo ejemplo

- En el análisis anterior, la capacidad predictiva del modelo sin genero

```
> lm1 <- lm(PrctGrasa ~ MCMagra, data=deportistas)
> summary(lm1)$r.squared
[1] 0.1309357
> summary(lm1)$sigma
[1] 5.784788
```

- Es superada por la del modelo con género:

```
> lm2 <- lm(PrctGrasa ~ MCMagra+Genero, data=deportistas)
> summary(lm2)$r.squared
[1] 0.5493066
> summary(lm2)$sigma
[1] 4.176289
```

- El hiperplano de mejor ajuste es

$$\widehat{\texttt{PcrtGrasa}} = 7{,}72462 - 12{,}243 \cdot \texttt{Genero} + 0{,}18443 \cdot \texttt{MCMagra},$$

siendo `Genero`=1, si hombre.

Preguntas abiertas importantes

- Tras un ajuste por mínimos cuadrados debemos cuestionarnos:
  - ¿El modelo tiene alguna utilidad para explicar o predecir la respuesta Y?
  - ¿Son todas las variables necesarias o basta con un subconjunto?
  - ¿Cuánto se ajusta el modelo a mis datos?
  - ¿Cuál es la precisión de nuestra estimación?¿Y de nuestras predicciones?
- Para responder a esas preguntas, debemos explorar las propiedades que se deducen del modelo lineal. En concreto, conocer el comportamiento en el muestreo de los estimadores de sus parámetros.

Regresión Lineal Múltiple

Notación matricial

- El modelo lineal, para una muestra de tamaño n, puede expresarse matricialmente:

$$Y = X\beta + \varepsilon$$

  - Donde $Y = (Y_1, \ldots, Y_n)'$ es el vector respuesta
  - $\beta = (\beta_0, \ldots, \beta_p)'$ es el vector de coeficientes
  - X, matriz de diseño, es una matriz $n \times (p+1)$, con el vector de unos, 1_n, en la primera columna y el vector asociado al predictor i en la columna $i+1$
  - $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ es el vector de errores
- Si se supone independencia de los datos, normalidad y homocedasticidad:

$$\varepsilon \sim N_n(0, \sigma^2 I_n)$$

  - Donde I_n es la matriz identidad $n \times n$

Estimadores mínimos cuadrados

- El vector de residuos, para un vector de coeficientes β, es
$$e = Y - X\beta$$
- Por tanto, la suma de cuadrados es
$$\sum_{i=1}^{n} e_i^2 = e'\,e = (Y - X\beta)'(Y - X\beta) = Y'Y - 2Y'X\beta + \beta'X'X\beta$$
- Ecuaciones normales. Derivando e igualando a 0, los estimadores mínimos cuadrados deben cumplir
$$X'Y = X'X\hat{\beta}$$
- Si la matriz X es de rango completo: $\hat{\beta} = (X'X)^{-1}X'Y$
  - Valores ajustados: $\hat{Y} = X\hat{\beta} = HY$, con $H = X(X'X)^{-1}X'$
  - Residuos: $e = Y - \hat{Y} = Y - HY = (I - H)Y$

Propiedades

Nota 1: $\mathrm{E}(AY+B) = A\,\mathrm{E}(Y)+B$, $\mathrm{V}(AY+B) = A\,\mathrm{V}(Y)A'$ y $\mathrm{Cov}(AY+B, CY+D) = A\,\mathrm{V}(Y)C'$

Nota 2: H e $I-H$ son matrices idempotentes y simétricas. Son matrices de proyección ortogonal, cumplen: $HX = X$, $(I-H)X = 0$ y $H(I-H) = 0$

- $\hat{\beta}$ es insesgado y con varianza $\sigma^2(X'X)^{-1}$

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathrm{E}(Y) = (X'X)^{-1}X'X\beta = \beta \\ \mathrm{V}(\hat{\beta}) &= \mathrm{V}((X'X)^{-1}X'Y) = (X'X)^{-1}X'\mathrm{V}(Y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}\end{aligned}$$

- Los residuos tienen media 0 y están correlados

$$\begin{aligned}\mathrm{E}(e) &= \mathrm{E}((I-H)Y) = (I-H)\mathrm{E}(Y) = (I-H)X\beta = 0 \\ \mathrm{V}(e) &= \mathrm{V}((I-H)Y) = (I-H)\mathrm{V}(Y)(I-H) = \sigma^2(I-H)\end{aligned}$$

- Los vectores $\hat{\beta}$ y e están incorrelados

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}, e) &= \mathrm{Cov}((X'X)^{-1}X'Y, (I-H)Y) = (X'X)^{-1}X'\mathrm{V}(Y)(I-H) = \\ &= \sigma^2(X'X)^{-1}X'(I-H) = 0\end{aligned}$$

Regresión Lineal Múltiple

Un ejemplo: Consumo de verduras

- En el estudio sobre consumo de verduras en niños (base de datos Pesos), la variable respuesta Peso se podría explicar utilizando las demás variables continuas de la base.

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=Pesos)
coef(lm_pesos)
```

- El hiperplano de mejor ajuste es

```
(Intercept)        Edad      Altura    Verduras
13.83515376 1.87475191 -0.05688285 -8.71272835
```

Notad que los coeficientes obtenidos tienen el signo que se esperaba, salvo el de la altura.

¿Es útil alguno de los predictores?

- Sí alguno de los predictores es útil, debemos rechazar
  $$H_0 : \beta_1 = \ldots = \beta_p = 0$$
- Ejemplo: Con los datos `pesos.cvs`

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=pesos)

Residual standard error: 2.116 on 96 degrees of freedom
Multiple R-squared: 0.5841, Adjusted R-squared: 0.5711
F-statistic: 44.95 on 3 and 96 DF, p-value: < 2.2e-16
```

  - El estimador insesgado de σ es $\hat{\sigma} = 2{,}116$. Su cuadrado sería MS_e.
  - En este caso, $n = 100$ y $p = 3$, test F con 3 y 96 gl
  - P-valor muy pequeño: se rechaza H_0
- El modelo es explicativo. Algunos predictores son útiles.

Utilidad del modelo: test F Global

- El test F global puede plantearse de manera alternativa como una tabla ANOVA.

Fuente de variación	Suma de cuadrados	grados de libertad	cuadrados medios	F
Ajuste (Entre)	$SS_A = \sum_i (\hat{y}_i - \bar{y})^2$	p	$MS_A = \frac{SS_A}{p}$	$\frac{MS_A}{MS_e}$
Error (Dentro)	$SS_e = \sum_i (y_i - \hat{y}_i)^2$	$n-p-1$	$MS_e = \frac{SS_e}{n-p-1}$	
Total	$SS_y = \sum_i (y_i - \bar{y})^2$	$n-1$		

- Se cumple: $SS_y = SS_A + SS_e$.
- F sigue una distribución F de Snédecor: $F \sim F_{p,n-p-1}$.
- Contrastar: $H_0 : \beta_1 = \ldots = \beta_p = 0$. El rechazo de H_0 ($F > F^{\alpha}_{p,n-p-1}$) implica que el modelo es útil en alguna medida.
- Con un solo predictor, este test es equivalente al contraste sobre la pendiente de la recta

¿Son necesarios todos los predictores?

- El test t-Student sobre cada uno de los coeficientes permite eliminar predictores de uno en uno, ya que la distribución en el muestreo para cada $\hat{\beta}_i$ es en presencia del resto de predictores.
- Ejemplo: Con los datos `pesos.cvs`

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=pesos)

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)  13.83515   5.91272      2.340     0.02136 *
 Edad         1.87475    0.45813      4.092     8.9e-05 ***
 Altura       -0.05688   0.06352      -0.895    0.37278
 Verduras     -8.71273   2.97552      -2.928    0.00426 **
```

  - Altura no es significativo, podemos eliminarla.

¿Cuanto se ajusta el modelo a los datos?

- El coeficiente de determinación, R^2 es una medida de bondad de ajuste.

$$R^2 = \frac{SS_A}{SS_y} = 1 - \frac{SS_e}{SS_y}$$

  - Es el porcentaje de variabilidad explicado por el modelo
  - Permite comparar modelos con el mismo número de predictores
  - Aumenta sistemáticamente al incluir nuevos predictores

- Por ello, se introduce el R^2 ajustado, que penaliza por el nº de predictores:

$$R^2\text{ajustado} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

- Ejemplo: Con los datos `pesos.cvs`

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=pesos)
```

```
Residual standard error: 2.116 on 96 degrees of freedom
Multiple R-squared: 0.5841, Adjusted R-squared: 0.5711
F-statistic: 44.95 on 3 and 96 DF, p-value: < 2.2e-16
```

¿Qué confianza tenemos en el ajuste?

- Estimar $f(x) = x'\beta$ para un vector concreto de valores de los predictores x_0, entendido como punto de muestreo.
  - El estimador puntual es $x_0'\hat{\beta} \sim N(x_0'\beta, \sigma^2 x_0'(X'X)^{-1}x_0)$
  - Por tanto, un pivote para $f(x_0) = x_0'\beta$ es:

$$\frac{x_0'\beta - x_0'\hat{\beta}}{\hat{\sigma}\sqrt{x_0'(X'X)^{-1}x_0}}$$

- Un ejemplo:

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=pesos)
x0 <- data.frame(Edad=10, Altura=150, Verduras = 0.3)
predict(lm_pesos,newdata=x0,interval='confidence')
```

  - Estimación puntual, 21.4; IC95 %, (20.1, 22.8)

```
       fit      lwr      upr
1 21.43643 20.09378 22.77908
```

¿Qué confianza tenemos en la predicción?

- Predecir Y, para un nuevo individuo con predictores x_0
  - La predicción puntual es $\hat{y}_0 = x_0'\hat{\beta}$
  - Como $Y_0 - \hat{y} \sim N(0, \sigma^2(1 + x_0'(X'X)^{-1}x_0))$, un pivote para y_0 es:

$$\frac{y_0 - x_0'\hat{\beta}}{\hat{\sigma}\sqrt{1 + x_0'(X'X)^{-1}x_0}}$$

- Un ejemplo:

```
lm_pesos <- lm(Peso ~ Edad + Altura + Verduras, data=pesos)
x0 <- data.frame(Edad=10, Altura=150, Verduras=c(0.3,0.7))
predict(lm_pesos,newdata=x0,interval='prediction')
```

  - Las predicciones son

```
       fit      lwr      upr
1 21.43643 17.02669 25.84617
2 17.95134 12.88268 23.01999
```

Regresión Lineal Múltiple

Predictores categóricos

- Los factores también pueden ser Predictores
- Se necesitan códigos numéricos para las categorías
  - Si el predictor es dicotómico, la primera categoría se toma como referencia y se codifica como 0, la otra como 1
    - Como el género en los datos de deportistas

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.72462   1.94170   3.978 9.72e-05 ***
MCMagra      0.18443   0.03454   5.339 2.53e-07 ***
Generomale -12.24299   0.90078 -13.591  < 2e-16 ***
```

  - Con k categorías, se añaden $k-1$ predictores codificados como 0 ó 1, denominados variables dummy
    - No hay una representación única, pero siempre $k-1$ dummy's, o tendremos parámetros no identificables
  - Cuidado! no puede eliminarse ninguna de estas dummy's de forma aislada.

Análisis de la covarianza (ANCOVA)

- Con un predictor numérico y otro categórico, el ajuste obliga a rectas paralelas
  - En el ejemplo de los deportistas: para chicas y chicos $\hat{y} = 7{,}7 + 0{,}18x$, $\hat{y} = -4{,}5 + 0{,}18x$, respectivamente
- Para evitar esa restricción, introducir el producto de los dos predictores
  - En el ejemplo de los deportistas:

```
lm(PrctGrasa ~ MCMagra * Genero, data=deportistas)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.28459  3.30006  0.086 0.93137
MCMagra 0.31997 0.05965 5.364 2.26e-07 ***
Generomale 0.05873 4.53697 0.013 0.98969
MCMagra:Generomale -0.20065 0.07258 -2.765 0.00624 **
```

- Ese producto es la interacción, los predictores originales son los efectos principales
- El estudio de la interacción se conoce como ANCOVA

Efecto interacción

- Si existe interacción, el efecto de cada predictor depende de los niveles del otro. Los dos predictores interactúan
  - Así, el efecto sobre PrctGrasa de incrementar en una unidad MCMagra en una chica es 0.32, en un chico es 0.12
- La interacción también puede darse entre dos predictores numéricos, o dos categóricos
  - Ejemplo: datos Advertising.csv, comparad los dos modelos siguientes e interpretad los resultados

```
ml1 <- lm(sales ~ TV + radio, data= Advertising)
ml2 <- lm(sales ~ TV * radio, data= Advertising)
```

- También puede darse entre más de dos predictores
  - Interacciones de orden dos son las que relacionan a dos predictores
  - Las de orden mayor son más difíciles de interpretar, suelen utilizarse menos

Regresión Lineal Múltiple

Equivalencia con el test t y ANOVA

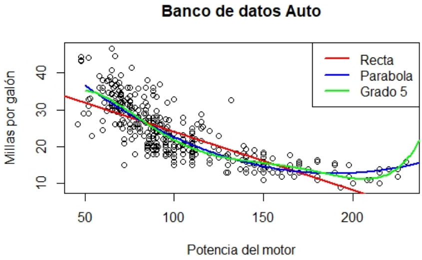
- Modelos ANOVA: todos los predictores categóricos
  - Test t-Student para una y dos muestras, ANOVA de una vía

```
lm(EDATDIAG  CHD, data = diabetes)
F-statistic: 4.801 on 1 and 147 DF, p-value: 0.03002


t.test(EDATDIAG  CHD, alternative='two.sided', conf.level=.95,
var.equal=T, data=diabetes)
t = -2.1911, df = 147, p-value = 0.03002
lm(EDATDIAG  ECG, data = diabetes)
F-statistic: 3.884 on 2 and 146 DF, p-value: 0.02273


mod.anova<-aov(EDATDIAG  ECG, data=diabetes)
          Df Sum Sq Mean Sq F value Pr(>F)
ECG        2   911   455.6   3.884   0.0227
Residuals 146 17129 117.3
```

Regresión polinómica

- La regresión polinómica es una forma de extender el modelo lineal a situaciones de no linealidad
  - Con los datos Auto del paquete ISLR,

```
lm(mpg ~ horsepower + I(horsepower**2), data=Auto)
```



- `lm(mpg ~ poly(horsepower,5), data=Auto)` utiliza polinomios ortogonales, que facilitan los cálculos pero dificultan la interpretación de los coeficientes

Regresión Lineal Múltiple

Comparación de modelos

- Ejemplo: Datos `Credit` del paquete `ISLR` ¿Cuál de los dos modelos siguientes parece más adecuado?

```
ajuste1 <- lm(Balance  Age + Rating , data=Credit)
ajuste2 <- lm(Balance  Age + Rating + Limit , data=Credit)
```

  - El coeficiente de determinación, R^2, del segundo es mayor
  - Pero el coeficiente de `Limit` podría ser cero
- Al añadir un nuevo predictor, R^2 aumenta, aunque el predictor no sea informativo
  - R^2 es útil para comparar modelos con el mismo número de predictores. No en otros casos
  - Notad que maximizar R^2 es equivalente a minimizar la suma de cuadrados de los residuos, pues $R^2 = 1 - SS_e/SS_y$
- R^2 ajustado ya penaliza por el número de predictores

Comparación de modelos

- Para la comparación de modelos conviene distinguir si son anidados (todos los predictores de un modelo están incluidos en el otro modelo)
- Para modelos No anidados: Criterio de información de Akaike (función `AIC`)
  - El mejor modelo es el de mínimo Akaike.
  - Existen criterios similares (llamados de 'información') que emplean otras penalizaciones (p.e. BIC)
- Para modelos Anidados: pueden usarse también Test F parcial, o test LRV (función `anova`)

  H_0 : Todos los coeficientes eliminados del modelo completo son cero

Otros criterios de información

- Criterio de información de Akaike. Bajo normalidad:

$$AIC = 2\text{ num.par} - 2\text{ logLik} = 2\,(p+1) + n\ln(2\pi \cdot SS_e/n) + n$$

  - El mejor modelo es el que minimiza el AIC

- Criterio de información Bayesiano, BIC, similar al AIC

$$BIC = \ln(n)\text{ num.par} - 2\text{ logLik} = \ln(n)\;(p+1) + n\ln(2\pi \cdot SS_e/n) + n$$

  - El mejor modelo es el que minimiza BIC
  - BIC penaliza más que AIC ($\ln(n) > 2$ si $n > 3$)

- Criterio C_p de Mallow

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

  - $\hat{\sigma}^2$ es el estimador obtenido con todos los predictores. El modelo tiene buen ajuste si C_p está cerca de $p+1$

Regresión Lineal Múltiple

El mejor subconjunto de predictores

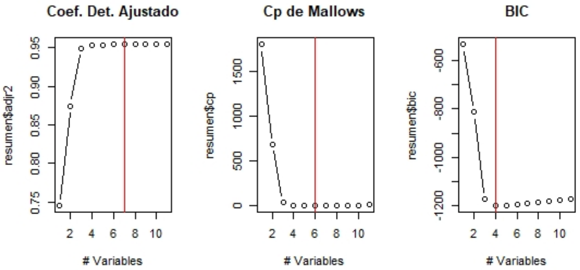
- Best subset selection. Si el número de predictores p no es muy grande
  - Para cada $q = 1, \ldots, p$, construir todos los modelos con q predictores y reservar el modelo de menor SS_e
  - Elegir, entre los reservados, el que optimize algún criterio: R^2 ajustado, AIC, BIC o C_p
- Un ejemplo: Datos `Credit` del paquete `ISLR`

```
library(leaps)
ajuste.todo <- regsubsets(Balance ~ . -ID , data=Credit,
              nvmax=11)
summary(ajuste.todo)
```

  - NOTA: En `leaps`, AIC y BIC se estiman de forma diferente
    - Primero se estima $\hat{\sigma}^2$ a partir del modelo completo
    - En cada modelo, se supone σ^2 conocida e igual a $\hat{\sigma}^2$
    - Así, AIC = C_p y BIC = $SS_e/\hat{\sigma}^2 + \ln(n)\ (p+1) - n$

Comparación de los mejores modelos

- Las gráficas resultantes son:



- R^2 ajustado es el criterio que menos penaliza
- C_p, igual a AIC, propone 6 predictores
- BIC es el que más penaliza, propone solo 4 predictores

Elección paso a paso

- Si p es grande, Best subset selection puede ser prohibitivo
  - El número de modelos es 2^p, 1024 si $p = 10$, por ejemplo
  - Una búsqueda exhaustiva puede producir sobreajuste, especialmente si p es grande
- Stepwise selection es una alternativa más eficiente
  - Forward Stepwise parte del modelo sin predictores
    - Busca el mejor modelo, añadiendo un solo predictor al modelo actual
    - Lo acepta, si es mejor que el actual, y continúa buscando
    - En otro caso, detiene el proceso
  - Backward Stepwise parte del modelo saturado
    - Busca el mejor modelo, eliminando un solo predictor del modelo actual
    - Lo acepta, si es mejor que el actual, y continúa buscando
    - En otro caso, detiene el proceso
  - Existen alternativas que mezclan backward y forward stepwise

Stepwise Selection: Un ejemplo

- El comando `step()` proporciona selección paso a paso
  - Este comando calcula AIC y BIC sin suponer σ^2 conocida
- Un ejemplo: Datos `Credit` del paquete `ISLR`
  - Opción por defecto, mezcla de backward y forward stepwise

```
Credit_sin_ID <- subset(Credit,select =
                         Income:Balance)
ajuste_todo <- lm(Balance ~ . , data=Credit_sin_ID)
traza <- step(ajuste_todo)
```

- Subcomando `direction` para backward o forward stepwise,
  - `direction = c("both", "backward", "forward")`
- Estos procedimientos, sobre todo si p es grande, pueden llegar a modelos distintos