Clase 4 - Módulo 2: Introducción a la analítica

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

¿Cómo mejorar los modelos lineales?

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

- $\epsilon_i \sim N(0, \sigma^2)$
- $\epsilon_1, \ldots, \epsilon_n$ son independientes.

Predictibilidad

- Cuando ajustamos modelos lineales y la relación entre Y y las covariables $X_1, X_2, \cdots, X_p$ es aproximadamente normal, el método de mínimos cuadrados funciona bien y los estimadores de los parámetros tienen poco sesgo.

Predictibilidad

- Cuando ajustamos modelos lineales y la relación entre Y y las covariables $X_1, X_2, \cdots, X_p$ es aproximadamente normal, el método de mínimos cuadrados funciona bien y los estimadores de los parámetros tienen poco sesgo.
- Si $n >> p$, es decir, el número de individuos es muy grande comparado con el número de covariables, entonces mínimos cuadrados también presenta estimadores con varianza pequeña y por tanto funcionará bien haciendo predicciones con observaciones nuevas.

Predictibilidad

- Cuando ajustamos modelos lineales y la relación entre Y y las covariables $X_1, X_2, \cdots, X_p$ es aproximadamente normal, el método de mínimos cuadrados funciona bien y los estimadores de los parámetros tienen poco sesgo.
- Si $n >> p$, es decir, el número de individuos es muy grande comparado con el número de covariables, entonces mínimos cuadrados también presenta estimadores con varianza pequeña y por tanto funcionará bien haciendo predicciones con observaciones nuevas.
- Si n no es mucho más grande que p entonces mínimos cuadrados tiende a sobreajustar el modelo y por tanto, las predicciones con nuevas observaciones no resultan ser buenas.

Predictibilidad

- Cuando ajustamos modelos lineales y la relación entre Y y las covariables $X_1, X_2, \cdots, X_p$ es aproximadamente normal, el método de mínimos cuadrados funciona bien y los estimadores de los parámetros tienen poco sesgo.

- Si $n >> p$, es decir, el número de individuos es muy grande comparado con el número de covariables, entonces mínimos cuadrados también presenta estimadores con varianza pequeña y por tanto funcionará bien haciendo predicciones con observaciones nuevas.

- Si n no es mucho más grande que p entonces mínimos cuadrados tiende a sobreajustar el modelo y por tanto, las predicciones con nuevas observaciones no resultan ser buenas.

- Si $p > n$ se puede llegar a que los estimadores de mínimos cuadrados no son únicos.

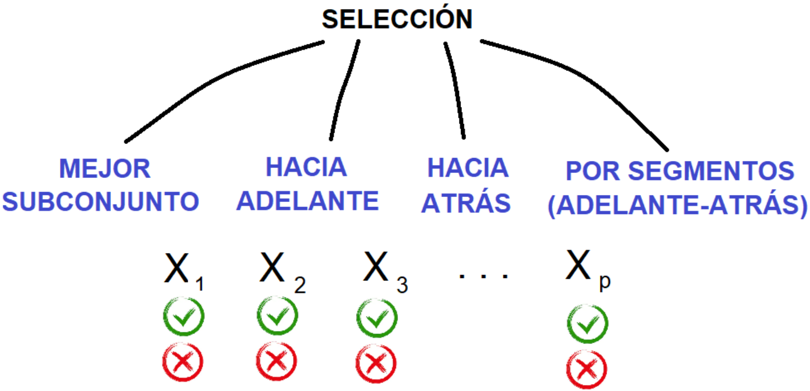
Interpretabilidad

- Cuando tenemos muchas covariables que no están asociadas con la variable respuesta, mínimos cuadrados tiende a mostrar que son significativas y no las elimina.

Interpretabilidad

- Cuando tenemos muchas covariables que no están asociadas con la variable respuesta, mínimos cuadrados tiende a mostrar que son significativas y no las elimina.
- Si se remueven las variables **irrelevantes**, haciendo que las estimaciones de los correspondientes coeficientes se "hagan" cero, se obtiene un modelo más fácil de interpretar.

- **Selección de subconjuntos de covariables:** Las covariables más relevantes para explicar la variabilidad de Y son seleccionadas.
- **Contracción (shrinkage):** Este método consiste en forzar que las estimaciones de los coeficientes correspondientes a covariables **irrelevantes** se "hagan" cero. Este proceso, también conocido como **regularización**, reduce la varianza de los estimadores de los coeficientes.
- **Reducción de dimensionalidad:** Consiste en proyectar las p covariables en un subespacio de dimensión reducida, M, con $M < p$. Esto produce M combinaciones lineales de las variables originales y dichas combinaciones se usan luego como predictores o covariables.


SELECCIÓN
MEJOR SUBCONJUNTO
HACIA ADELANTE
HACIA ATRÁS
POR SEGMENTOS (ADELANTE-ATRÁS)
X_1 X_2 X_3 . . . X_p

Selección de subconjuntos:

Algoritmo - Selección del mejor Subconjunto:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

Algoritmo - Selección del mejor Subconjunto:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

2. Para $k = 1, 2, \ldots, p$:

- Ajuste todos los $\binom{p}{k}$ modelos que contienen exactamente k variables explicativas.
- Seleccione el **mejor** modelo entre todos los $\binom{p}{k}$ posibles, y llamelo $\mathcal{M}_k$. En este contexto, el **mejor** es aquel que tiene el menor RSS (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.

Algoritmo - Selección del mejor Subconjunto:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

2. Para $k = 1, 2, \ldots, p$:

   - Ajuste todos los $\binom{p}{k}$ modelos que contienen exactamente k variables explicativas.
   - Seleccione el **mejor** modelo entre todos los $\binom{p}{k}$ posibles, y llamelo $\mathcal{M}_k$. En este contexto, el **mejor** es aquel que tiene el menor RSS (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.

3. Seleccione el **mejor** entre todos los $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando: C_p (cross validated prediction error), AIC, BIC o el R^2 ajustado.


SELECCIÓN
MEJOR SUBCONJUNTO
HACIA ADELANTE
HACIA ATRÁS
POR SEGMENTOS (ADELANTE-ATRÁS)

Algoritmo - Selección hacia adelante:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

Algoritmo - Selección hacia adelante:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

2. Para $k = 0, 2, \ldots, (p-1)$:

- Ajuste todos los $(p-k)$ modelos que adicionan un predictor adicional en el modelo $\mathcal{M}_k$.
- Seleccione el **mejor** modelo entre todos los $(p-k)$ posibles, y llamelo $\mathcal{M}_{k+1}$. En este contexto, el **mejor** es aquel que tiene el menor *RSS* (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.

Algoritmo - Selección hacia adelante:

1. Denote por $\mathcal{M}_0$ al modelo de solo intercepto, es decir, sin variables explicativas.

2. Para $k = 0, 2, \ldots, (p-1)$:

    - Ajuste todos los $(p-k)$ modelos que adicionan un predictor adicional en el modelo $\mathcal{M}_k$.
    - Seleccione el **mejor** modelo entre todos los $(p-k)$ posibles, y llamelo $\mathcal{M}_{k+1}$. En este contexto, el **mejor** es aquel que tiene el menor *RSS* (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.

3. Seleccione el **mejor** entre todos los $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando: C_p (cross validated prediction error), AIC, BIC o el R^2 ajustado.

Selección hacia atrás:

Algoritmo - Selección hacia atrás:

1. Denote por $\mathcal{M}_p$ al modelo con todas las variables.

Algoritmo - Selección hacia atrás:

1. Denote por $\mathcal{M}_p$ al modelo con todas las variables.
2. Para $k = p, (p-1), \ldots, 1$:

- Ajuste todos los k modelos que eliminan un predictor (dejando los demás) en el modelo $\mathcal{M}_k$.
- Seleccione el **mejor** modelo entre todos los k posibles, y llamelo $\mathcal{M}_{k-1}$. En este contexto, el **mejor** es aquel que tiene el menor RSS (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.

Algoritmo - Selección hacia atrás:

1. Denote por $\mathcal{M}_p$ al modelo con todas las variables.
2. Para $k = p, (p-1), \ldots, 1$:
   - Ajuste todos los k modelos que eliminan un predictor (dejando los demás) en el modelo $\mathcal{M}_k$.
   - Seleccione el **mejor** modelo entre todos los k posibles, y llamelo $\mathcal{M}_{k-1}$. En este contexto, el **mejor** es aquel que tiene el menor RSS (Suma Cuadrática de los residuales), o lo que es equivalente, el que tenga el mayor R^2.
3. Seleccione el **mejor** entre todos los $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ usando: C_p (cross validated prediction error), AIC, BIC o el R^2 ajustado.

Seleccionando el “mejor” modelo

- Las medidas R^2 y RSS son buenas medidas de evaluación del modelo en el proceso de **entrenamiento**, sin embargo, no son buenas para evaluar el **error de prueba**.
- Se hace nesesario entonces planter una método diferente para seleccionar el “mejor” modelo. Se proponen dos alternativas:

Indirecta usando medidas que cuantifican el sobreajuste.

Directa utilizando el método del conjunto de validación o validación cruzada.

- Cada uno de los $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ modelos tiene asociada una suma cuadrática de los residuales RSS.
- El modelo completo, es decir, el modelo con todas las p covariables tiene un estimador insesgado para σ^2 dado por:

$$\widehat{\sigma}^2 = \left[\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \ldots - \widehat{\beta}_p X_{ip}\right)^2\right] \Big/ (n-p-1)$$

- La suma cuadrática total está dada por

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Basados en las consideraciones anteriores, se definen las medidas de sobreajuste, para un modelo con d covariables, de la siguiente manera:

- $C_p = \frac{1}{n}\left(RSS + 2d\widehat{\sigma}^2\right)$
- $AIC = \frac{1}{n\widehat{\sigma}^2}\left(RSS + 2d\widehat{\sigma}^2\right)$
- $BIC = \frac{1}{n\widehat{\sigma}^2}\left(RSS + \ln(n)d\widehat{\sigma}^2\right)$
- $R^2_{Ajustado} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

Con cada una el mejor modelo es el de menor valor,

excepto con el $R^2_{Ajustado}$

Trabajando con R: Mejor subconjunto

```r
require(ISLR)
```

```
## Loading required package: ISLR
```

```r
dim(Hitters)
```

```
## [1] 322  20
```

```r
names(Hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Wa
##  [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CR
## [13] "CWalks"    "League"    "Division"  "PutOuts"   "Assists"   "Er
## [19] "Salary"    "NewLeague"
```

```r
Hitters<-na.omit(Hitters) # Eliminando NA's
dim(Hitters)
```

```
## [1] 263  20
```

```r
sum(is.na(Hitters))
```

```r
require(leaps)
```

```
## Loading required package: leaps
```

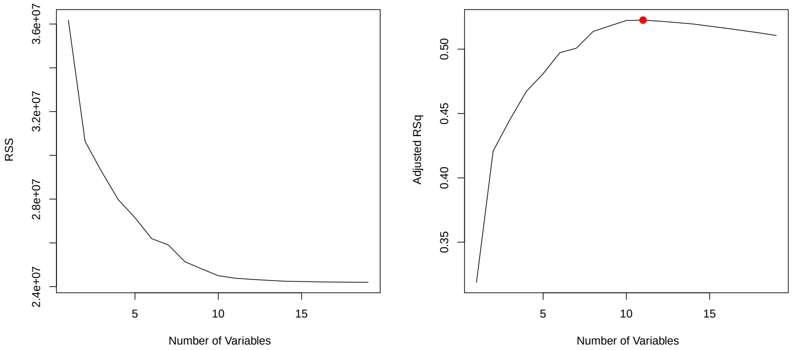
```r
regfit.full<-regsubsets(Salary~., data=Hitters, nvmax=19)
reg.summary<-summary(regfit.full)
```

```r
names(reg.summary)
```

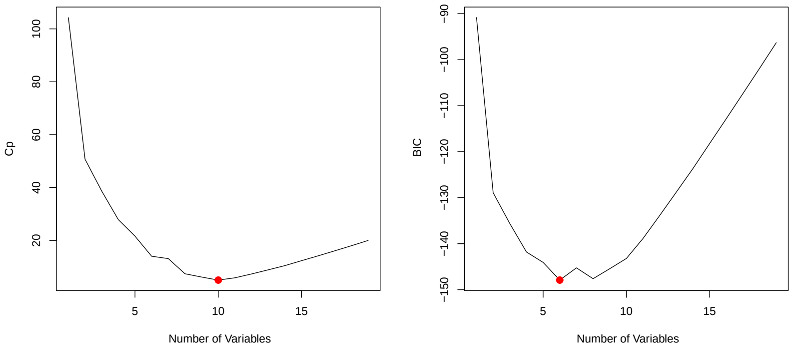
```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

Trabajando con R: Mejor subconjunto

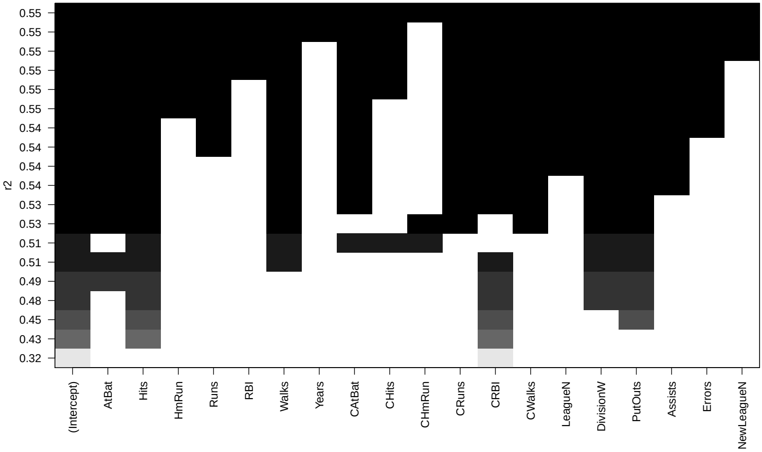
```
par(mfrow =c(1,2))
plot(reg.summary$rss ,xlab=" Number of Variables",ylab=" RSS",type="l")
plot(reg.summary$adjr2 ,xlab =" Number of Variables",ylab=" Adjusted RSq"
     ,type="l")
a1<-which.max (reg.summary$adjr2)
points (a1, reg.summary$adjr2[a1], col ="red",cex =2, pch =20)
```



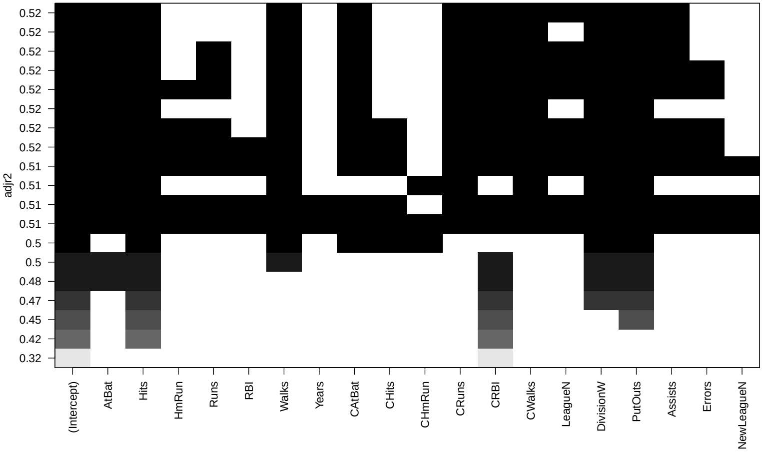
```
par(mfrow =c(1,2))
plot(reg.summary$cp ,xlab =" Number of Variables", ylab="Cp", type="l")
a2<-which.min(reg.summary$cp)
points (a2, reg.summary$cp [a2], col ="red",cex =2, pch =20)
a3<-which.min(reg.summary$bic)
plot(reg.summary$bic ,xlab=" Number of Variables",ylab=" BIC",type="l")
points (a3, reg.summary$bic [a3], col =" red",cex =2, pch =20)
```



```
plot(regfit.full, scale ="r2")
```

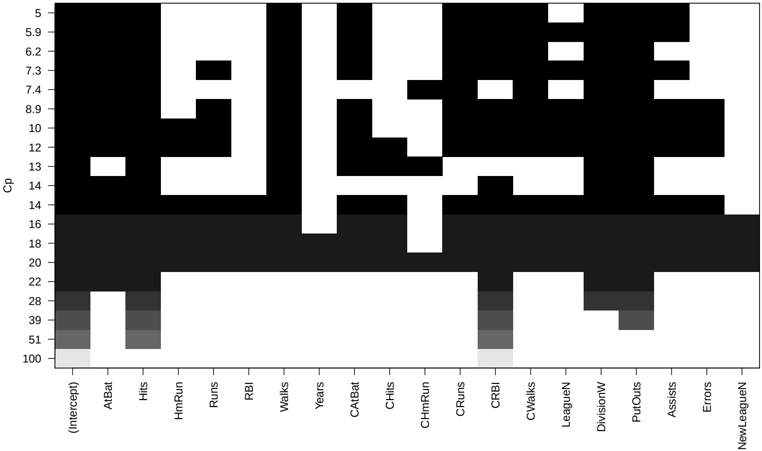


```r
plot(regfit.full, scale ="adjr2")
```



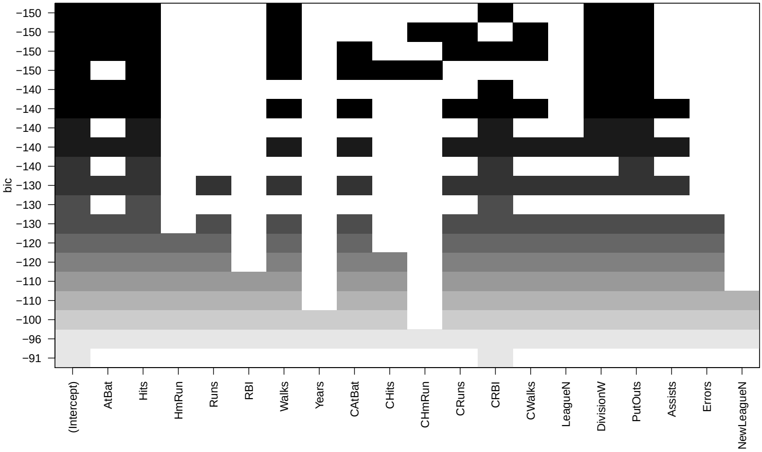
Trabajando con R: Mejor subconjunto

```
plot(regfit.full, scale ="Cp")
```



Trabajando con R: Mejor subconjunto

```r
plot(regfit.full, scale ="bic")
```

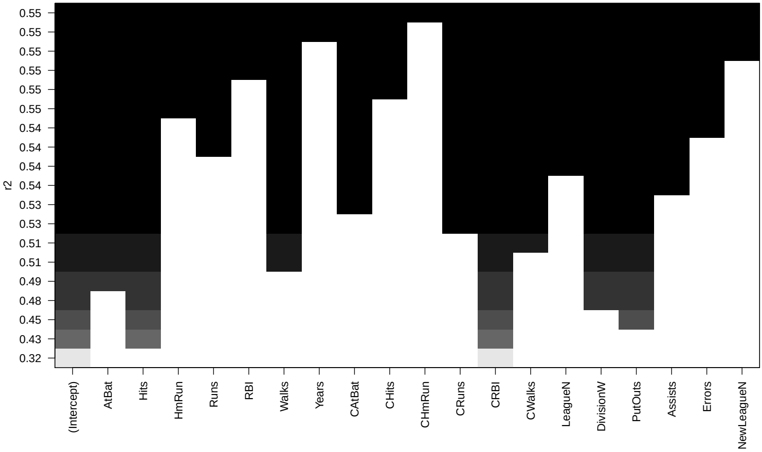


Trabajando con R: Selección hacia adelante

```r
regfit.fwd<-regsubsets(Salary~.,data=Hitters,
                         nvmax=19, method = "forward")
summary(regfit.fwd)
```

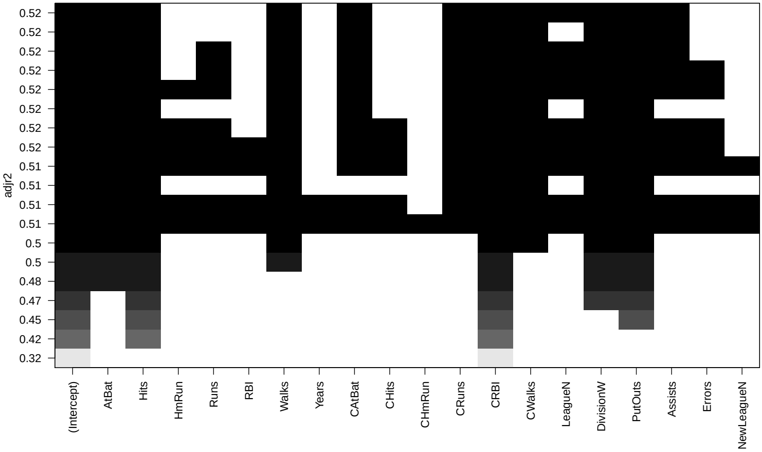
```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 19, method = "forward")
## 19 Variables  (and intercept)
##            Forced in Forced out
## AtBat          FALSE      FALSE
## Hits           FALSE      FALSE
## HmRun          FALSE      FALSE
## Runs           FALSE      FALSE
## RBI            FALSE      FALSE
## Walks          FALSE      FALSE
## Years          FALSE      FALSE
## CAtBat         FALSE      FALSE
## CHits          FALSE      FALSE
## CHmRun         FALSE      FALSE
## CRuns          FALSE      FALSE
## CRBI           FALSE      FALSE
## CWalks         FALSE      FALSE
## LeagueN        FALSE      FALSE
## DivisionW      FALSE      FALSE
## PutOuts        FALSE      FALSE
## Assists        FALSE      FALSE
## Errors         FALSE      FALSE
## NewLeagueN     FALSE      FALSE
## 1 subsets of each size up to 19
## Selection Algorithm: forward
##           AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
```

```
plot(regfit.fwd, scale ="r2")
```

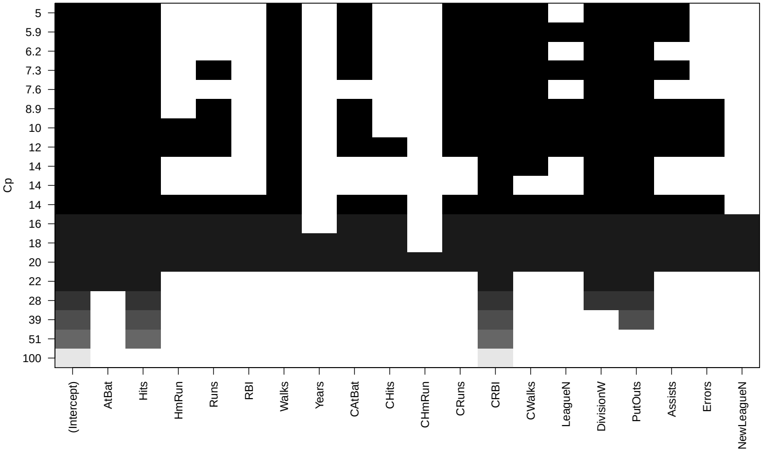


Trabajando con R: Selección hacia adelante

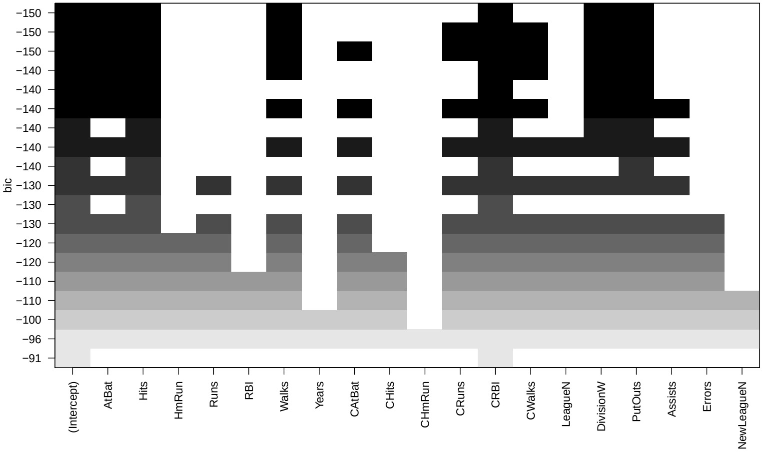
```r
plot(regfit.fwd, scale ="adjr2")
```



```
plot(regfit.fwd, scale ="Cp")
```



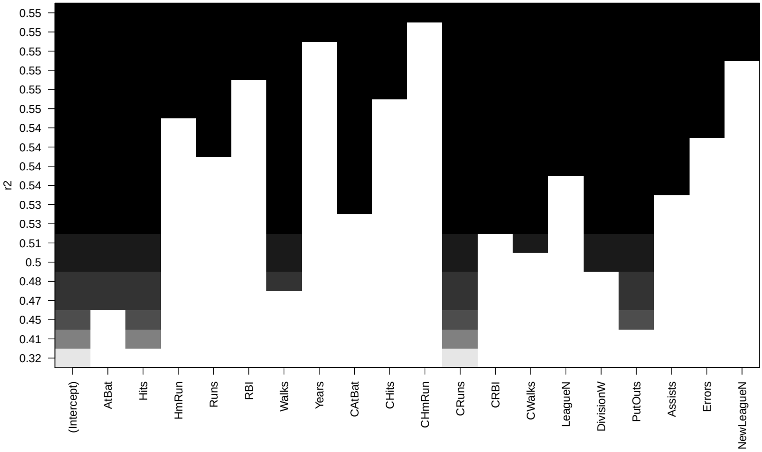
```r
plot(regfit.fwd, scale ="bic")
```



```r
regfit.bwd<-regsubsets(Salary~.,data=Hitters,
                          nvmax=19, method = "backward")
summary(regfit.bwd)
```

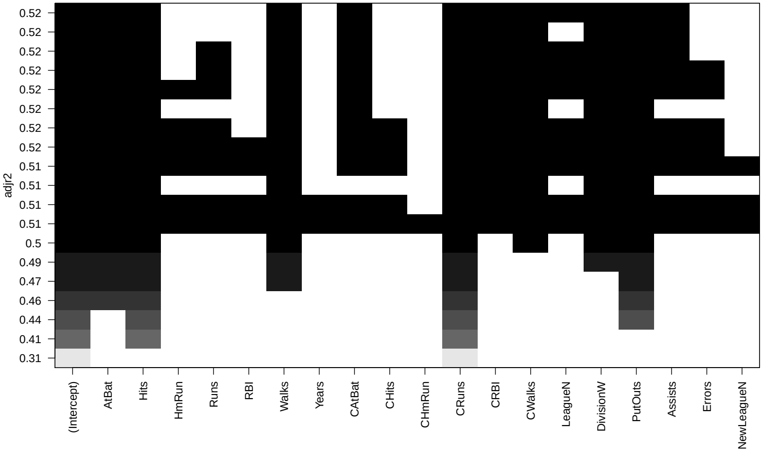
```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 19, method = "backward")
## 19 Variables  (and intercept)
##            Forced in Forced out
## AtBat          FALSE      FALSE
## Hits           FALSE      FALSE
## HmRun          FALSE      FALSE
## Runs           FALSE      FALSE
## RBI            FALSE      FALSE
## Walks          FALSE      FALSE
## Years          FALSE      FALSE
## CAtBat         FALSE      FALSE
## CHits          FALSE      FALSE
## CHmRun         FALSE      FALSE
## CRuns          FALSE      FALSE
## CRBI           FALSE      FALSE
## CWalks         FALSE      FALSE
## LeagueN        FALSE      FALSE
## DivisionW      FALSE      FALSE
## PutOuts        FALSE      FALSE
## Assists        FALSE      FALSE
## Errors         FALSE      FALSE
## NewLeagueN     FALSE      FALSE
## 1 subsets of each size up to 19
## Selection Algorithm: backward
##          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
```

```r
plot(regfit.bwd, scale ="r2")
```



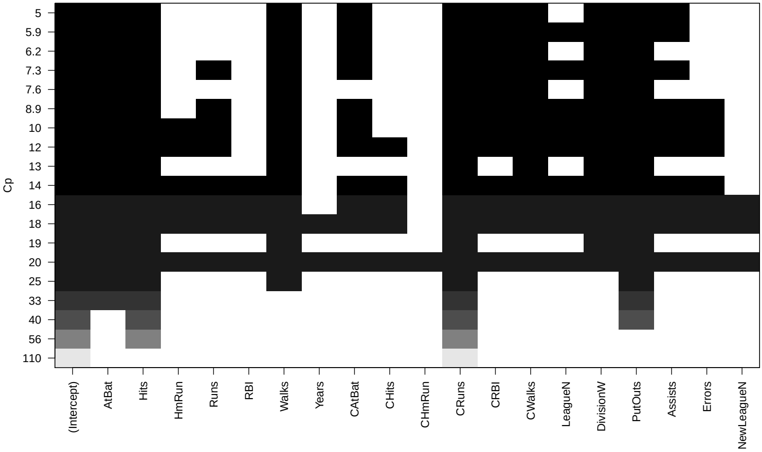
Trabajando con R: Selección hacia atrás

```
plot(regfit.bwd, scale ="adjr2")
```



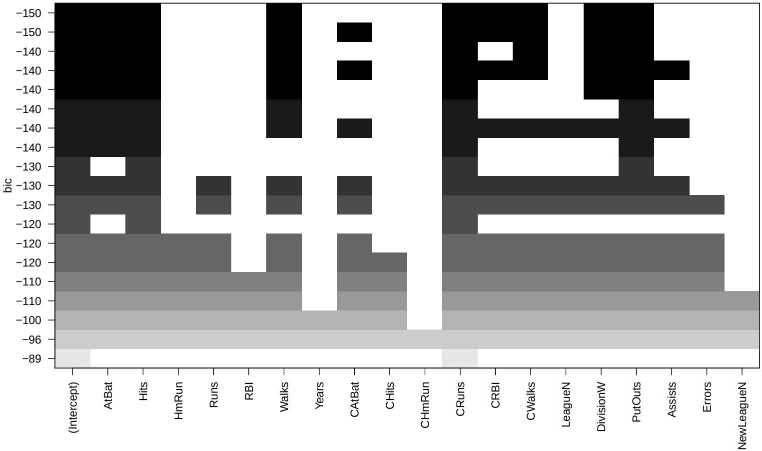
Trabajando con R: Selección hacia atrás

```
plot(regfit.bwd, scale ="Cp")
```



Trabajando con R: Selección hacia atrás

```r
plot(regfit.bwd, scale ="bic")
```



```
set.seed (1)
train<-sample (c(TRUE ,FALSE), nrow(Hitters),rep=TRUE)
test<-(!train )
regfit.best<-regsubsets(Salary~.,data=Hitters[train,],
                        nvmax =19)

# Creando una función para evaluar los modelos:
predict.regsubsets =function (object,newdata,y){
 form<-as.formula(object$call[[2]])
 mat<-model.matrix(form ,newdata)
 val.errors =rep(NA, (ncol(mat)-1))
 for(i in 1:length(val.errors)){
 coefi<-coef(object ,id=i)
 xvars<-names (coefi)
 pred<-mat[,xvars]%*%coefi
 val.errors [i]= mean((y-pred)^2)
 }
 val.errors
 }
```

Trabajando con R: Seleccionando el mejor modelo con CV

```
e1<-predict.regsubsets(regfit.best,Hitters[test,],
                        Hitters$Salary[test])

b1<-which.min(e1)

regfit.best<-regsubsets(Salary~.,data=Hitters ,nvmax =19)

e1
b1
coef(regfit.best,10)
```

Trabajando con R: Seleccionando el mejor modelo con k-fold

```
k<-10
set.seed (1)
folds<-sample (1:k,nrow(Hitters ),replace =TRUE)
cv.errors<-matrix (NA ,k,19,
                   dimnames =list(NULL,paste (1:19)))
```

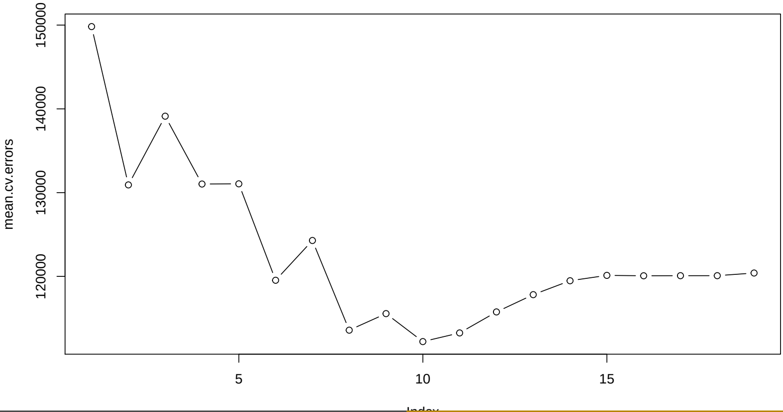
Trabajando con R: Seleccionando el mejor modelo con k-fold

```r
for(j in 1:k){
best.fit<-regsubsets(Salary~.,data=Hitters[folds!=j,],
nvmax=19)
pred<-predict.regsubsets(best.fit, Hitters[folds==j,],
                         Hitters$Salary[folds==j])
cv.errors[j,]<-pred
}
mean.cv.errors<-apply(cv.errors,2,mean)
mean.cv.errors
```

```
##        1        2        3        4        5        6        7        8
## 149821.1 130922.0 139127.0 131028.8 131050.2 119538.6 124286.1 113580.0
##        9       10       11       12       13       14       15       16
## 115556.5 112216.7 113251.2 115755.9 117820.8 119481.2 120121.6 120074.3
##       17       18       19
## 120084.8 120085.8 120403.5
```

Trabajando con R: Seleccionando el mejor modelo con k-fold

```
c1<-which.min(mean.cv.errors)
plot(mean.cv.errors,type="b")
```



Trabajando con R: Seleccionando el mejor modelo con k-fold

```r
reg.best<-regsubsets(Salary~.,data=Hitters, nvmax =19)
coef(reg.best,c1)
```

```
##   (Intercept)         AtBat          Hits         Walks        CAtBat 
##   162.5354420    -2.1686501     6.9180175     5.7732246    -0.1300798 
##          CRBI        CWalks     DivisionW       PutOuts       Assists 
##     0.7743122    -0.8308264  -112.3800575     0.2973726     0.2831680 
```

Utilizando los métodos de selección y validación cruzada, realice un análisis de la base de datos **`surgical`** del paquete **`olsrr`** donde seleccione las variables más importantes para explicar la variabilidad del tiempo de supervivencia.