

title: “statLearn Lab6(a) Model Selection”

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com> (<http://rmarkdown.rstudio.com>).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

```
library(ISLR)
dim(Hitters)
```

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

```
names(Hitters)
```

```
## [1] "AtBat" "Hits" "HmRun" "Runs" "RBI"
## [6] "Walks" "Years" "CAtBat" "CHits" "CHmRun"
## [11] "CRuns" "CRBI" "CWalks" "League" "Division"
## [16] "PutOuts" "Assists" "Errors" "Salary" "NewLeague"
```

```
summary(Hitters)
```

```
##           AtBat           Hits           HmRun           Runs
## Min.      : 16.0   Min.      : 1   Min.      : 0.00   Min.      : 0.00
## 1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   1st Qu.: 30.25
## Median :379.5   Median : 96   Median : 8.00   Median : 48.00
## Mean     :380.9   Mean     :101   Mean     :10.77   Mean     : 50.91
## 3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00   3rd Qu.: 69.00
## Max.     :687.0   Max.     :238   Max.     :40.00   Max.     :130.00
##
##           RBI           Walks           Years           CAtBat
## Min.      : 0.00   Min.      : 0.00   Min.      : 1.000   Min.      : 19.0
## 1st Qu.: 28.00   1st Qu.: 22.00   1st Qu.: 4.000   1st Qu.: 816.8
## Median : 44.00   Median : 35.00   Median : 6.000   Median : 1928.0
## Mean     : 48.03   Mean     : 38.74   Mean     : 7.444   Mean     : 2648.7
## 3rd Qu.: 64.75   3rd Qu.: 53.00   3rd Qu.:11.000   3rd Qu.: 3924.2
## Max.     :121.00   Max.     :105.00   Max.     :24.000   Max.     :14053.0
##
##           CHits           CHmRun           CRuns           CRBI
## Min.      : 4.0   Min.      : 0.00   Min.      : 1.0   Min.      : 0.00
## 1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.: 88.75
## Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
## Mean     : 717.6   Mean     : 69.49   Mean     : 358.8   Mean     : 330.12
## 3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
## Max.     :4256.0   Max.     :548.00   Max.     :2165.0   Max.     :1659.00
##
##           CWalks           League Division PutOuts           Assists
## Min.      : 0.00   A:175   E:157   Min.      : 0.0   Min.      : 0.0
## 1st Qu.: 67.25   N:147   W:165   1st Qu.: 109.2   1st Qu.: 7.0
## Median : 170.50           Median : 212.0   Median : 39.5
## Mean     : 260.24           Mean     : 288.9   Mean     :106.9
## 3rd Qu.: 339.25           3rd Qu.: 325.0   3rd Qu.:166.0
## Max.     :1566.00           Max.     :1378.0   Max.     :492.0
##
##           Errors           Salary           NewLeague
## Min.      : 0.00   Min.      : 67.5   A:176
## 1st Qu.: 3.00   1st Qu.: 190.0   N:146
## Median : 6.00   Median : 425.0
## Mean     : 8.04   Mean     : 535.9
## 3rd Qu.:11.00   3rd Qu.: 750.0
## Max.     :32.00   Max.     :2460.0
##
##           NA's           :59
```

Hitters dataset consists of Baseball stats, with Salary as response (Y). We are going to use these Hitters features to predict the player's Salary. There are some missing values in the dataset, and we need to remove those:

```
sum(is.na(Hitters$Salary))
```

```
## [1] 59
```

```
Hitters = na.omit(Hitters)
with(Hitters, sum(is.na(Salary)))
```

```
## [1] 0
```

```
dim(Hitters)
```

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

Best Subset Regression

Use the ‘leaps’ package to evaluate all the best-subset model of given size. For each subset size, there is a star that indicates variable in the model.

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
regfit.full = regsubsets(Salary~., data=Hitters)
summary(regfit.full)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = Hitters)
## 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 8
## Selection Algorithm: exhaustive
##
```

| | | AtBat | Hits | HmRun | Runs | RBI | Walks | Years | CAtBat | CHits | CHmRun | CRuns |
|------|-------|-------|------|-------|------|-----|-------|-------|--------|-------|--------|-------|
| ## 1 | (1) | " " | " " | " " | " " | " " | " " | " " | " " | " " | " " | " " |
| ## 2 | (1) | " " | " * | " " | " " | " " | " " | " " | " " | " " | " " | " " |
| ## 3 | (1) | " " | " * | " " | " " | " " | " " | " " | " " | " " | " " | " " |
| ## 4 | (1) | " " | " * | " " | " " | " " | " " | " " | " " | " " | " " | " " |
| ## 5 | (1) | " * | " * | " " | " " | " " | " " | " " | " " | " " | " " | " " |
| ## 6 | (1) | " * | " * | " " | " " | " " | " * | " " | " " | " " | " " | " " |
| ## 7 | (1) | " " | " * | " " | " " | " " | " * | " " | " * | " * | " " | " " |
| ## 8 | (1) | " * | " * | " " | " " | " " | " * | " " | " " | " * | " * | " * |

```
##
```

| | | CRBI | CWalks | LeagueN | DivisionW | PutOuts | Assists | Errors | NewLeagueN |
|------|-------|------|--------|---------|-----------|---------|---------|--------|------------|
| ## 1 | (1) | " * | " " | " " | " " | " " | " " | " " | " " |
| ## 2 | (1) | " * | " " | " " | " " | " " | " " | " " | " " |
| ## 3 | (1) | " * | " " | " " | " " | " * | " " | " " | " " |
| ## 4 | (1) | " * | " " | " " | " * | " * | " " | " " | " " |
| ## 5 | (1) | " * | " " | " " | " * | " * | " " | " " | " " |
| ## 6 | (1) | " * | " " | " " | " * | " * | " " | " " | " " |
| ## 7 | (1) | " " | " " | " " | " * | " * | " " | " " | " " |
| ## 8 | (1) | " " | " * | " " | " * | " * | " " | " " | " " |

The varbles are not nested, by default up to subsets up to size 8. To get full subsets up to size 19, use the “nvmax=” command:

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

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

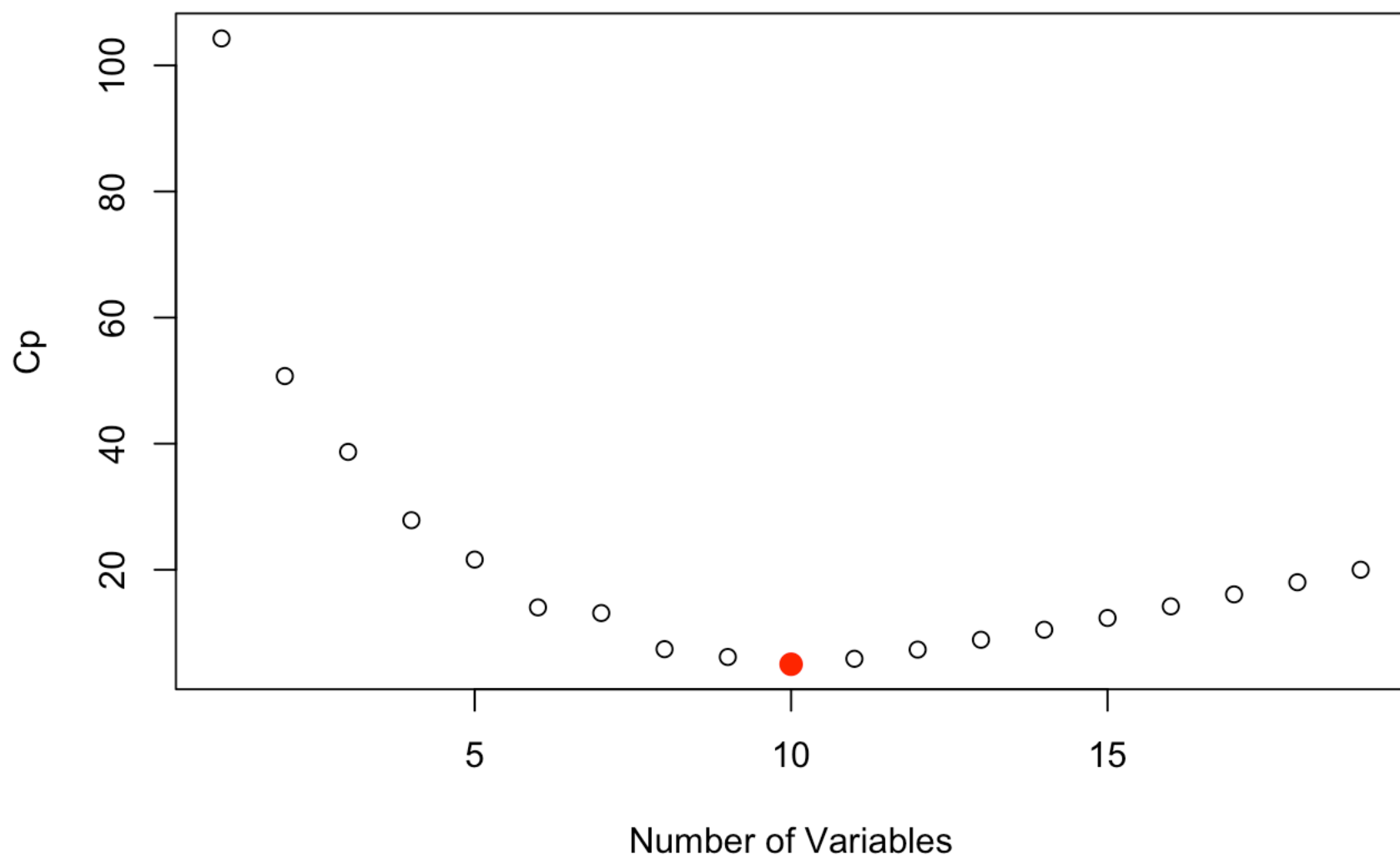
Returns information need to select the best model (rsq, rss, cp, bic) Plot cp component, estimate of prediction error, against # variables

Model with 10 predictors seems to have the lowest estimated error And we can verify this calling which.min() function for Cp

```
plot(reg.summary$cp, xlab="Number of Variables", ylab="Cp")
which.min(reg.summary$cp)
```

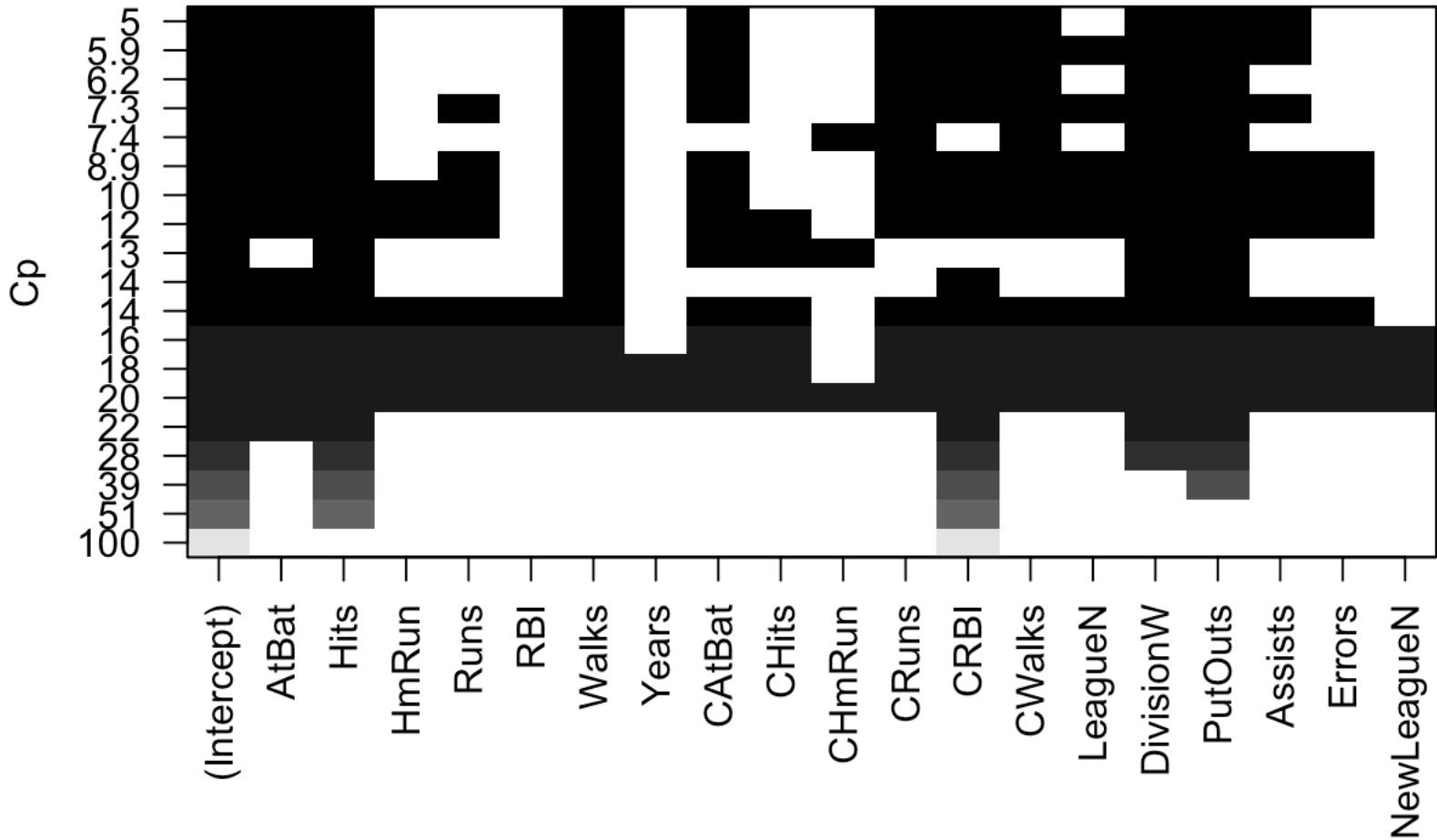
```
## [1] 10
```

```
points(10,reg.summary$cp[10],col="red",cex=2,pch=20)
```



plot method for 'regsubsets' object, displays a pattern picture with each predictor on X-axis, and values of Cp plotted on the Y axis Examine Coefficients for regfit, with 10 predictors

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



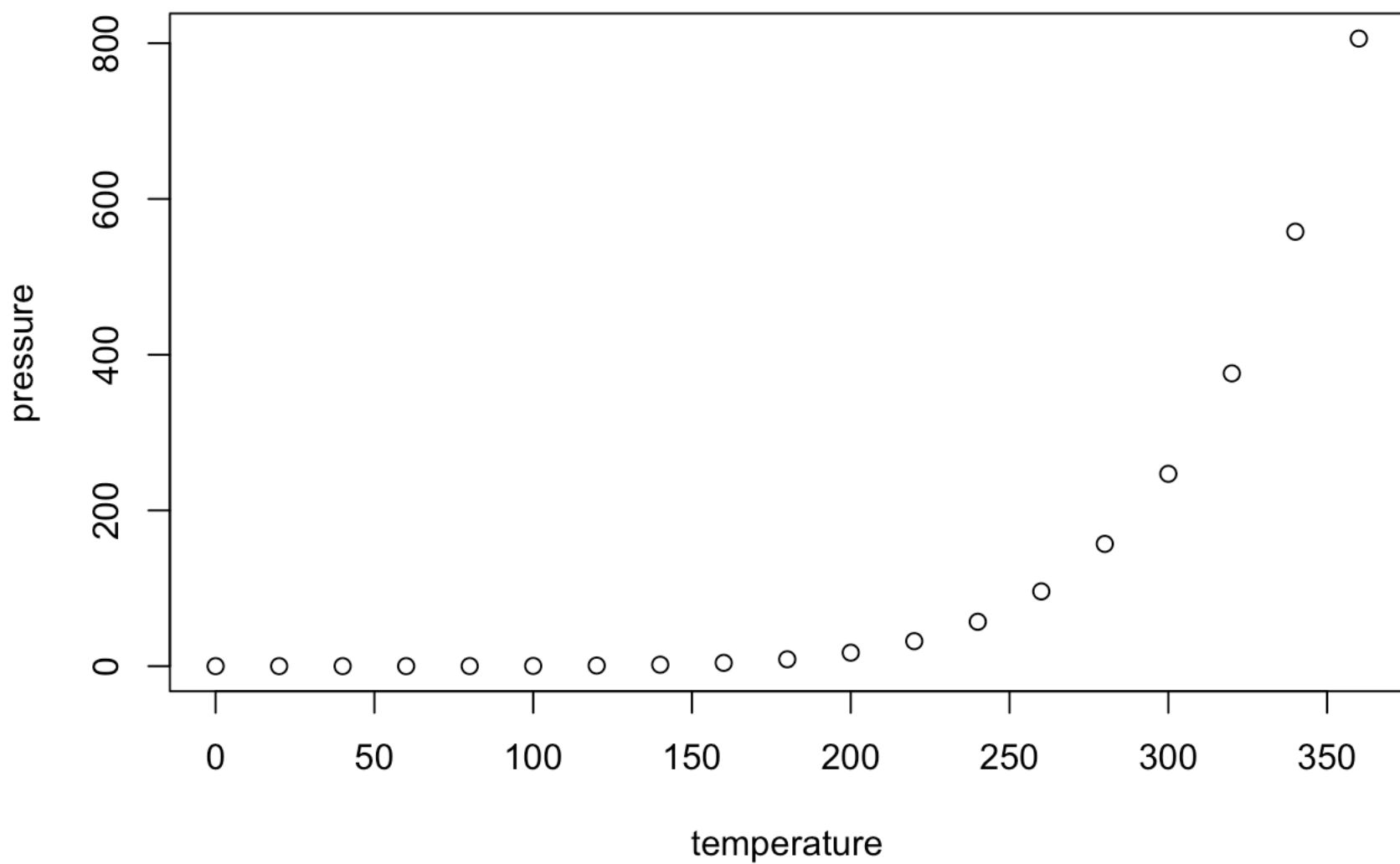
```
coef(regfit.full,10)
```

| | | | | | |
|----|-------------|------------|------------|--------------|------------|
| ## | (Intercept) | AtBat | Hits | Walks | CAtBat |
| ## | 162.5354420 | -2.1686501 | 6.9180175 | 5.7732246 | -0.1300798 |
| ## | CRuns | CRBI | CWalks | DivisionW | PutOuts |
| ## | 1.4082490 | 0.7743122 | -0.8308264 | -112.3800575 | 0.2973726 |
| ## | Assists | | | | |
| ## | 0.2831680 | | | | |

For printable summary of output in html, click ‘Knit HTML’ on toolbar

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.