Classification and Regression Trees (CART)

Regression Trees: Ozone data

This example is adapted from the book "Extending the Linear Model with R", by Julian J. Faraway. Here is a quote from that book describing the data:

> We apply the regression tree methodology to study the relationship between atmospheric ozone concentration and meteorology in the Los Angeles Basin in 1976. A number of cases with missing variables hve been removed for simplicity [but Evan notes that trees are among the regression and classification methods that are best able to handle missing data]. We wish to predict the ozone level from the other predictors.

The variables in the data set are as follows:

- `O3`: Ozone concentration (ppm) at Sandbug Air Force Base
- `vh`: Vandenburg 500 millibar height (inches)
- `wind`: wind speed (miles per hour)
- `humidity`: humidity (percent)
- `temp`: temperature (degrees C)
- `ibh`: inversion base height (feet)
- `dpg`: Daggett pressure gradient (mmhg)
- `ibt`: inversion base temperature (degrees F)
- `vis`: visibility (miles)
- `doy`: day of the year

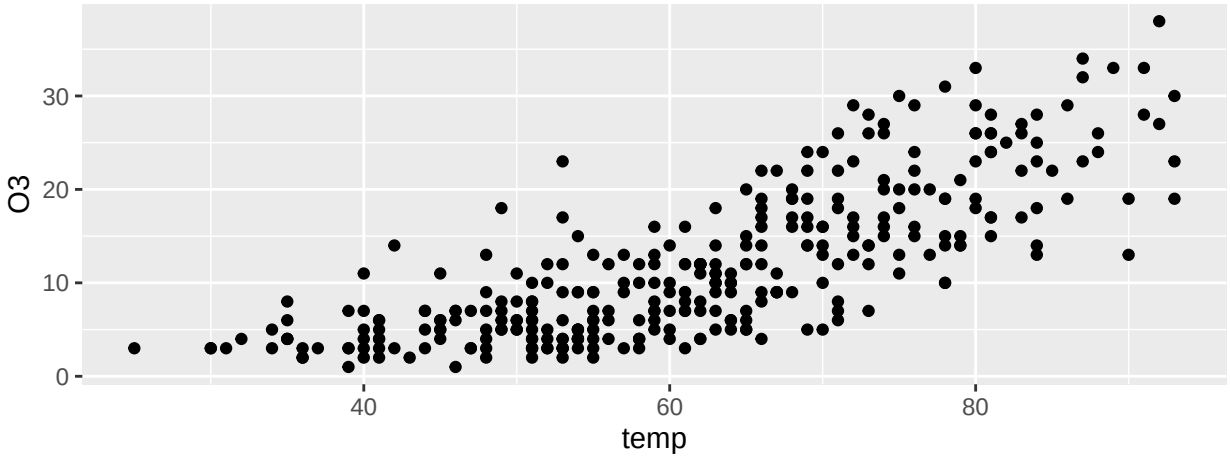
```
head(ozone)
```

```
##   O3   vh wind humidity temp  ibh dpg ibt vis doy
## 1  3 5710    4       28   40 2693 -25  87 250  33
## 2  5 5700    3       37   45  590 -24 128 100  34
## 3  5 5760    3       51   54 1450  25 139  60  35
## 4  6 5720    4       69   35 1568  15 121  60  36
## 5  4 5790    6       19   45 2631 -33 123 100  37
## 6  4 5790    3       25   55  554 -28 182 250  38
```

```
dim(ozone)
```

```
## [1] 330  10
```

```
ggplot(data = ozone, mapping = aes(x = temp, y = O3)) +
  geom_point()
```



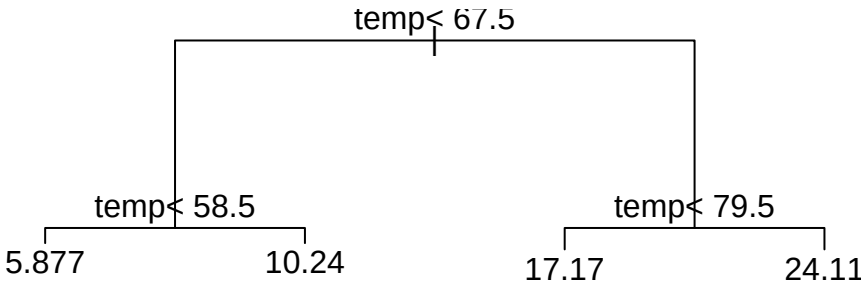
Regression tree with 1 explanatory variable:

Fit the model

```
tree_fit <- train(
  form = O3 ~ temp,
  data = ozone,
  method = "rpart",
  trControl = trainControl(method = "none"),
  tuneGrid = data.frame(cp = 0.01)
)
```

Display the estimated tree

```
plot(tree_fit$finalModel, margin = 0.1)
text(tree_fit$finalModel)
```



What's the predicted ozone level for a day with a temperature of 75 degrees?

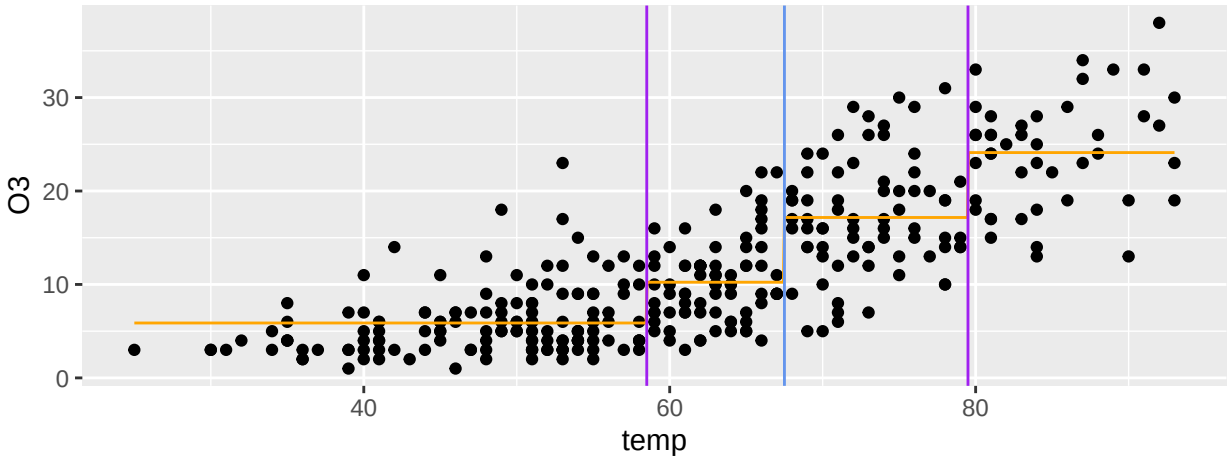
```
predict(tree_fit, newdata = data.frame(temp = 75))
```

```
##        1
## 17.16667
```

A picture of predicted ozone level as a function of temperature:

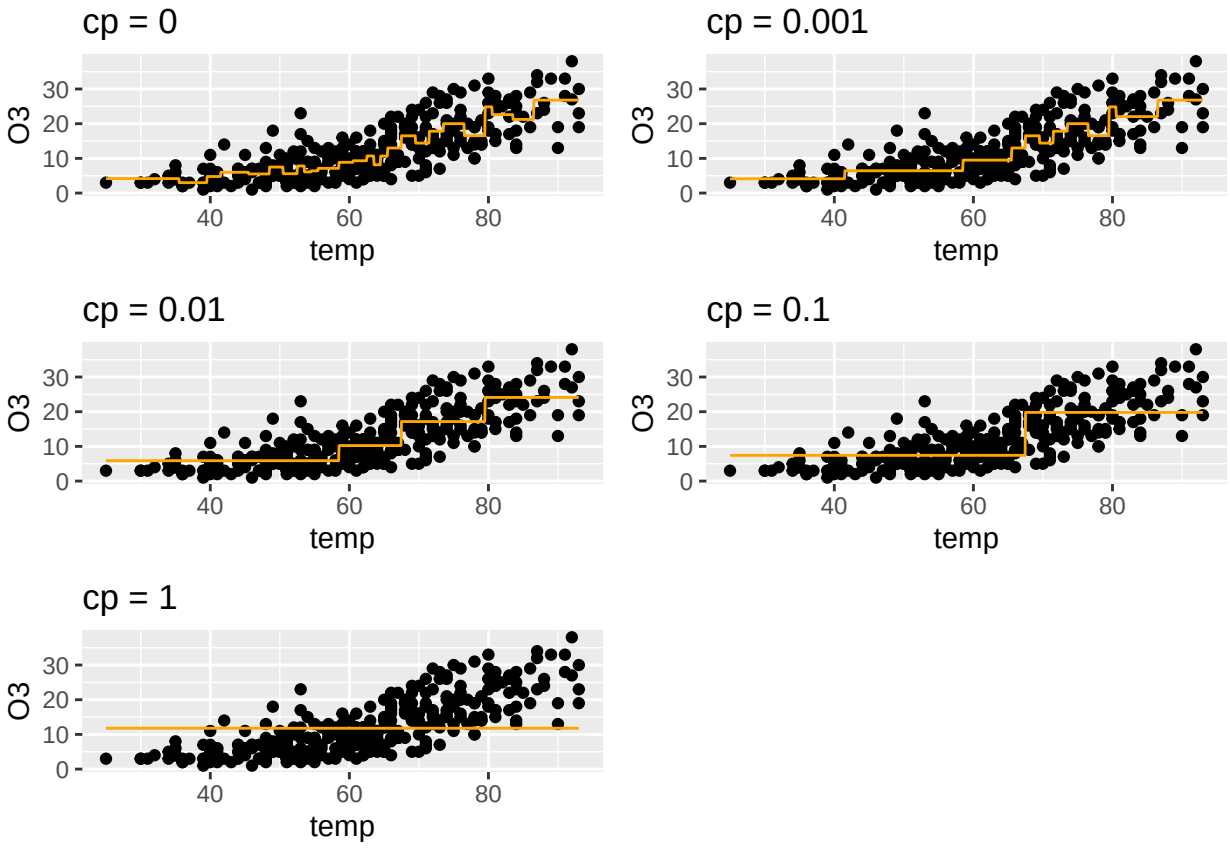
```
predict_tree <- function(x) {
  predict(tree_fit, newdata = data.frame(temp = x))
}

ggplot(data = ozone, mapping = aes(x = temp, y = O3)) +
  geom_point() +
  stat_function(fun = predict_tree, n = 1001, color = "orange") +
  geom_vline(xintercept = 67.5, color = "cornflowerblue") +
  geom_vline(xintercept = c(58.5, 79.5), color = "purple")
```



Effect of penalty parameter λ on estimation

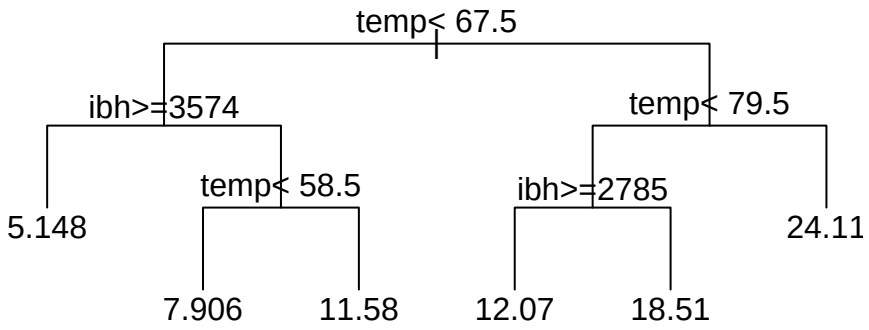
- In `rpart`, λ is denoted by `cp`, for complexity parameter



Regression tree with 2 explanatory variables:

```
tree_fit <- train(
  form = O3 ~ temp + ibh,
  data = ozone,
  method = "rpart",
  trControl = trainControl(method = "none"),
  tuneGrid = data.frame(cp = 0.01)
)
```

```
plot(tree_fit$finalModel, margin = 0.1, uniform = TRUE)
text(tree_fit$finalModel)
```

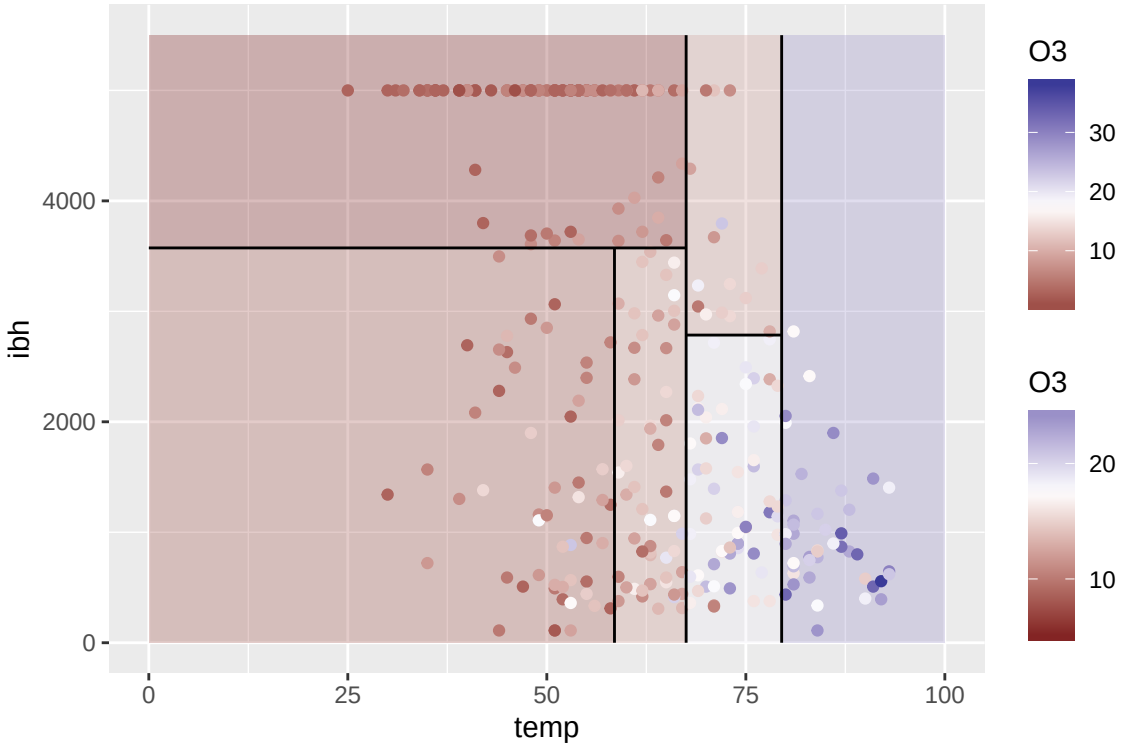


What's the predicted ozone level for a day with a temperature of 75 degrees and an inversion base height of 2000 feet?

```
test_data <- data.frame(
  temp = 75, ibh = 2000
)

predict(tree_fit, newdata = test_data)
```

```
##        1
## 18.50877
```

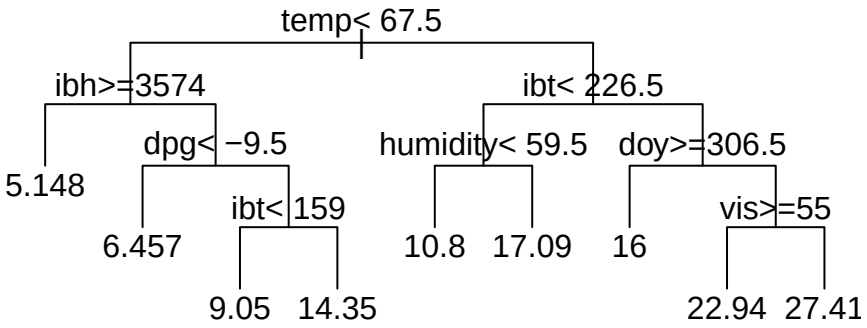


Note that the above is *not* a plot of decision boundaries for classification since this is a regression problem! The predictions are for a quantitative response.

More covariates:

```
tree_fit <- train(
  form = O3 ~ .,
  data = ozone,
  method = "rpart",
  trControl = trainControl(method = "none"),
  tuneGrid = data.frame(cp = 0.01)
)

# print picture of resulting tree
plot(tree_fit$finalModel, margin = 0.1, uniform = TRUE)
text(tree_fit$finalModel)
```



Classification Trees: Heart Disease data

We have data on 303 patients who presented with chest pain. The response variable is `AHD`, which is is "Yes" if an angiogrphic test indicates presence of heart disease, and "No" otherwise. There are 13 predictor variables which are a mix of quantitative and categorical variables.

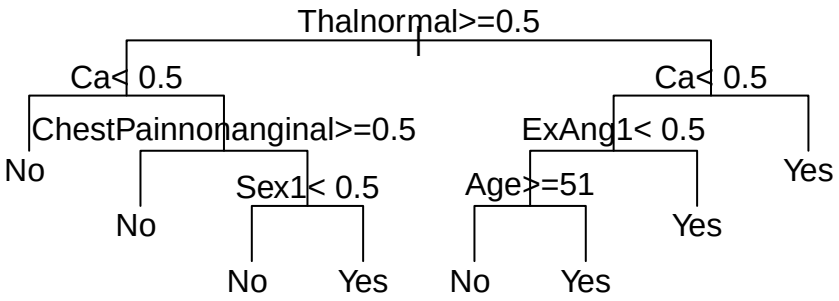
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
head(heart)
```

```
## # A tibble: 6 x 14
##     Age Sex   ChestPain RestBP  Chol Fbs   RestECG MaxHR ExAng Oldpeak
##   <dbl> <fct> <fct>      <dbl> <dbl> <fct> <fct>   <dbl> <fct>   <dbl>
## 1    63 1     typical      145   233 1     2         150 0         2.3
## 2    67 1     asymptom~    160   286 0     2         108 1         1.5
## 3    67 1     asymptom~    120   229 0     2         129 1         2.6
## 4    37 1     nonangin~    130   250 0     0         187 0         3.5
## 5    41 0     nontypic~    130   204 0     2         172 0         1.4
## 6    56 1     nontypic~    120   236 0     0         178 0         0.8
## # ... with 4 more variables: Slope <fct>, Ca <dbl>, Thal <fct>, AHD <fct>
```

```
tree_fit <- train(
  form = AHD ~ .,
  data = heart,
  method = "rpart",
  trControl = trainControl(method = "none"),
  tuneGrid = data.frame(cp = 0.01)
)

# print second picture of resulting tree
plot(tree_fit$finalModel, margin = 0.1, uniform = TRUE)
text(tree_fit$finalModel)
```




```
levels(heart$Thal)
```

```
## [1] "fixed"      "normal"     "reversable"
```

```
levels(heart$ChestPain)
```

```
## [1] "asymptomatic" "nonanginal"   "nontypical"   "typical"
```

What's the predicted class for someone with the following covariate values?

```
##   Age Sex ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca
## 1  55   1   typical    140  250   1       1   160     1     2.2     3  2
##     Thal
## 1 normal
```

```
predict(tree_fit, newdata = person_to_classify)
```

```
## [1] Yes
## Levels: No Yes
```