数据科学家直通车项目考试

Phase 1： Lending Club风险评估项目

Meina Wang

Q1 (10pts):

Monty hall problem: assume that a room is equipped with three doors. Behind two are goats, and behind the third is a shiny new car. You are asked to pick a door, and will win whatever is behind it. Let's say you pick door 1. Before the door is opened, however, someone who knows what's behind the doors (Monty Hall) opens one of the other two doors, revealing a goat, and asks you if you wish to change your selection to the third door (i.e., the door which neither you picked nor he opened). The Monty Hall problem is deciding whether you do. Please also provide an R implementation to justify your answer for this problem.

A1:

The player should switch to the other door to increase the probability of winning.

When the first selection was made, the probability of winning the car is $\frac{1}{3}$. So door 1 has a probability of $\frac{1}{3}$ having a car behind it.
After Monty Hall reveals the other door with a goat behind it, the probability of winning the car for door 1 stays the same $\frac{1}{3}$, while the door being revealed has 0 probability of having the car, and this adds to the probability of the other door having the car to be $\frac{1}{3} + \frac{1}{3} = \frac{2}{3}$.

So, if the player does not switch, there is $\frac{1}{3}$ chance of winning. If the player switches to the other door, now the probability of winning is $\frac{2}{3}$.

In [2]:

```
#Monty hall problem

doors <- c("1","2","3")

data = c()

for(i in 1:10000){
  prize <- sample(doors)[1]
  pick <- sample(doors)[1]
  open <- sample(doors[which(doors != pick & doors != prize)])[1]
  switch <- doors[which(doors != pick & doors != open)]

  if(pick == prize){
    data = c(data, "noswitch_win")}

  if(switch == prize){
    data = c(data, "switch_win")}
}

length(which(data == "noswitch_win"))
length(which(data == "switch_win"))
```

3266

6734

So the simulated result matches with our previous calculation.

Q2 (5pts):

On the average, how many times must a die be thrown until one gets a 6?

A2:

Assume the number of rolls until one 6 appears is X, then all the (X-1) rouns must have any numbers expect 6. This case follows the geometric distribution. The probability of having one 6 at X round is $(\frac{5}{6})^{X-1}\frac{1}{6}$.

The expectation of X is $E = \sum_{n=1}^{\infty} nP(X=n)$. Plug in the above equation for $P(X=n)$ yields,

$E = \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty} n(\frac{5}{6})^{n-1}\frac{1}{6} = \frac{1}{6}\frac{6}{5}\sum_{n=1}^{\infty} n(\frac{5}{6})^{n} = \frac{1}{6}\frac{6}{5}\frac{\frac{5}{6}}{(1-\frac{5}{6})^2} = 6.$

Therefore, the expected number of rolls until one gets a 6 is 6.

Q3 (10pts):

The game of craps, played with two dice, is one of America's fastest and most popular games. Calculating the odds associated with it is an instructive exercise. The rules are these. Only totals for the two dice count. The player throws the dice and wins at once if the total for the first throw is 7 or 11, loses at once if it is 2, 3 or 12. any other throw is called his “point”. If the first throw is a point, the player throws the dice repeatedly until he either wins by throwing his point again or loses by throwing 7. What is the player's chance to win?

A3:

First, calculate the probability of winning by getting 7 or 11 at first throw.
7 comes from 1+6, 2+5, 3+4, 4+3, 5+2, 6+1. So $P(7) = \frac{6}{36} = \frac{1}{6}$
11 comes from 5+6, 6+5. So $P(11) = \frac{2}{36} = \frac{1}{18}$
Adding $P(7)$ and $P(11)$ yields the probability of winning at first throw, which is $\frac{2}{9}$.

Second, calculate the probability of losing by getting 2, 3 or 12 at first throw.
2 comes from 1+1. So $P(7) = \frac{1}{36}$
3 comes from 1+2, 2+1. So $P(11) = \frac{2}{36} = \frac{1}{18}$
12 comes from 6+6. So $P(11) = \frac{1}{36}$
Adding $P(2)$, $P(3)$ and $P(12)$ yields the probability of losing at first throw, which is $\frac{1}{9}$.

Third, calculate the probability of getting a point, which is $1 - P(winning) - P(losing) = 1 - \frac{2}{9} - \frac{1}{9} = \frac{2}{3}$, and it comes from getting the numbers 4, 5, 6, 8, 9, 10.
Similarly, calculate $P(4)$ with probability of getting 1+3, 2+2, 3+1. $P(4) = \frac{1}{12}$.
Calculate $P(5)$ with probability of getting 1+4, 2+3, 3+2, 4+1. $P(5) = \frac{1}{9}$
Calculate $P(6)$ with probability of getting 1+5, 2+4, 3+3, 4+2, 5+1. $P(6) = \frac{5}{36}$
Calculate $P(8)$ with probability of getting 2+6, 3+5, 4+4, 5+3, 6+2. $P(8) = \frac{5}{36}$
Calculate $P(9)$ with probability of getting 3+6, 4+5, 5+4, 6+3. $P(9) = \frac{1}{9}$
Calculate $P(10)$ with probability of getting 4+6, 5+5, 6+4. $P(10) = \frac{1}{12}$

Since based on the rule of this game, if the first throw is a point, then the player throws the dice repeatedly until he either wins by throwing his point again or loses by throwing 7. We will calculate the probability of winning given each point.

If the point is 4, then the probability of winning now is $P(4) = \frac{1}{12}$, and the probability of losing is $P(7) = \frac{1}{6}$, so the probability of the game continues is $(1 - P(winning) - P(losing) = \frac{3}{4})$. Therefore, base on the geometric distribution, the probability of winning given point 4 is
$\frac{1}{12} + \frac{3}{4}\frac{1}{12} + \frac{3}{4}^2\frac{1}{12} + \ldots = \frac{1}{12}(\frac{1}{1-\frac{3}{4}}) = \frac{1}{12}\frac{1}{3}$. Point 10 has the same probability of winning as point 4.

Similarly, calculate the probability of winning if the point is 5. $P(5) = \frac{1}{9}$, $P(7) = \frac{1}{6}$, so the probability of the game continues is $(1 - P(winning) - P(losing) = \frac{13}{18})$. Therefore, the probability of winning given point 5 is $\frac{1}{9} + \frac{13}{18}\frac{1}{9} + \frac{13}{18}^2\frac{1}{9} + \ldots = \frac{1}{9}(\frac{1}{1-\frac{13}{18}}) = \frac{1}{9}\frac{2}{5}$. Point 9 has the same probability of winning as point 5.

Next, calculate the probability of winning if the point is 6. $P(6) = \frac{5}{36}$, $P(7) = \frac{1}{6}$, so the probability of the game continues is $(1 - P(winning) - P(losing) = \frac{25}{36})$. Therefore, the probability of winning given point 5 is $\frac{5}{36} + \frac{25}{36}\frac{5}{36} + \frac{25}{36}^2\frac{5}{36} + \ldots = \frac{5}{36}(\frac{1}{1-\frac{25}{36}}) = \frac{5}{36}\frac{5}{11}$. Point 8 has the same probability of winning as point 6.

Based on the above calculation, the overall winning probability therefore is,
$P(getting 7 or 11 at first throw) + P(getting point again) = \frac{2}{9} + 2 * \frac{1}{12}\frac{1}{3} + 2 * \frac{1}{9}\frac{2}{5} + 2 * \frac{5}{36}\frac{5}{11} \approx 0.493$

So, there is ~49.3% chance that the player wins.

Q4 (10pts):

Explain the functionality of a regularizer. Let $y = \sum_{i=1}^{d} \alpha_i x_i$, please provide the mathematical forms for: “Lasso”, “Ridge”, “Elastic Net” and explain characteristics of each regularizer.

A4:

The use of a regularizer is to penalize the loss function by adding a complexity term that would give a bigger loss for more complex models.

For the function $y = \sum_{i=1}^{d} \alpha_i x_i$, its loss function therefore is $L = \sum_{i=1}^{d} (\alpha_i x_i - y_i)^2$

- Ridge (L2 norm): add the penalty term of $\lambda \sum_{i=1}^{p} \alpha_i^2$ to the original loss function, where λ is learning rate, and p is the total number of parameters.
  Ridge regression shrinks the feature parameters, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the trained model.
- LASSO (L1 norm): add the penalty term of $\lambda \sum_{i=1}^{p} \mid \alpha_i \mid$ to the original loss function.
  LASSO (least absolute shrinkage and selection operator) method not only punishes high values of the feature coefficients but actually setting them to zero if they are not relevant. Therefore, the model might end up with fewer features, and it can be used for feature selection.
- Elastic Net: add the penalty term of $\lambda_2 \sum_{i=1}^{p} \alpha_i^2 + \lambda_1 \sum_{i=1}^{p} \mid \alpha_i \mid$ to the original loss function.
  Elastic Net linearly combines the L1 and L2 penalties of the lasso and ridge methods, where L1 penalty helps generating a sparse model, and L2 penalty overcomes a strict selection.

Q5 (5pts):

What’s overfitting? How to detect it? How to resolve it?

A5:

Overfitting is when a model or an algorithm fits the data too well, and is too complicated, so that it starts to capture the noise of the data.

If a model has high variance, it won't generalize well on unseen data, and it's overfitting.

To resolve overfitting:

- Collect more training data if possible.
- Use a smaller set of features so the model is simpler.
- Use cross-validation on the training data to train the algorithm multiple times and pick parameter that minimize $J_{cv}(\theta)$, this could get a better estimation of the error on unseen data.
- Apply regularization to the model to penalize overfitting.

Q6 (5pts):

What is K-fold cross validation? What’s the advantage of cross validation, comparing to splitting data into train and test?

A6:

K-fold cross validation is a type of cross validation method. It randomly partitions the original training data set into K equal subsets, with each subset as a fold. For the k folds, a single fold is retained as the validation data for testing the model, and the remaining k – 1 folds are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k folds used exactly once as the validation data. The k results from the folds can then be averaged to produce a single estimation.

The advantage of cross validation comparing to splitting data into train and test is that all the observations are used for both training and validation, and each observation is used for validation exactly once. It also allows the test data to remain truly unseen.

Q7 (10pts):

Suppose you are shooting free throws and each shot has a 60% chance of going in (there is no "learning" effect and "depreciation" effect). Which of the following is the most likely scenario where you can win $1000:
(a) Make at least 2 out of 3,
(b) Make at least 4 out of 6, and
(c) Make at least 20 out of 30.
Explain why.

A7:

(a) Make at least 2 out of 3.
The probability of winning is the same across the three options, which is $\frac{2}{3}$. The difference is that the more trial one makes, the closer the actual probability will be to the expected probability. Also, the expected value of 60% is lower than the expect probability of winning ($\frac{2}{3}$). Therefore, the more shootings one takes, the actual probability will be closer to 60%, therefore less likely to reach or exceed the winning chance of $\frac{2}{3}$. So we need the variance to reach the chance of $\frac{2}{3}$. The more we play, the variance would reduce, and the actual probability will be closer to 60%.

Therefore, (a) has the best chance of winning comparing to the other options.

Q8 (10pts):

Suppose that we know a priori that the data points (X_i, Y_i) fit a straight line, except that there is a little error involved. That is to say, suppose that $X_1, \ldots, X_n$ are fixed and that we think of $Y_1, \ldots, Y_n$ as being random variables satisfying $Y_i = \lambda_1 X_i + \lambda_2 + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, where all the ϵ_i's are assumed to be independent. Please show that in this case, MLE is equivalent to minimizing squared error.

A8:

Based on the description, the likelihood function for the function $Y_i = \lambda_1 X_i + \lambda_2 + \epsilon_i$ is,

$$L(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} f(X_i; \lambda_1, \lambda_2, \epsilon)$$

. If the $\epsilon_i \sim N(0, \sigma^2)$, where all the ϵ_i's are assumed to be independent, then the likelihood function becomes,

$$L(Y_1, Y_2, \ldots, Y_n; \lambda_1, \lambda_2, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} exp(\frac{-1}{2\sigma^2}(\sum_{i=1}^{n}(Y_i - \lambda_1 X_i - \lambda_2)^2))$$

For a fixed positive σ value, maximize the likelihood function L is equivalent to minimize the term $\sum_{i=1}^{n}(Y_i - \lambda_1 X_i - \lambda_2)^2$ inside the likelihood function, which is to minimize the square error.

Q9 (5pts):

Logistic regression assumptions. Derive logistic regression likelihood and loss based these assumptions.

A9:

The assumptions for logistic regression are:

- the central assumption that $P(Y|X)$ can be approximated as a sigmoid function applied to a linear combination of input features. A sigmod function is defined as $\sigma(z) = \frac{1}{1+e^{-z}}$.
- binary logistic regression requires the dependent variable to be binary, and ordinal logistic regression requires the dependent variable to be ordinal.
- logistic regression requires the observations to be independent of each other, and there is little or no multicollinearity among the independent variables.
- logistic regression assumes linearity of independent variables and log odds. It requires that the independent variables are linearly related to the log odds.
- it requires quite large sample sizes.

For each data-point, there is a vector of features x_i, and an observed class y_i. The probability of that class was either p, if $y_i = 1$, or $1 - p$, if $y_i = 0$. The model assumes $P(Y = 1|X = x) = \sigma(z)$, where $z = \theta_0 + \sum_{i=1}^{m} \theta_i x_i$. The likelihood is then $L = \prod_{i=i}^{m} P(x_i)^{y_i}(1 - P(x_i))^{1-y_i}$. The negative log-likelihood therefore is $l = -\sum_{i=1}^{m} y_i log P(x_i) + (1 - y_i) log(1 - P(x_i))$. Notice that $-log(L) = l$.
Based on

$$P(Y = 1|X = x) = \sigma(f(x)) = \frac{1}{1 + e^{-f(x)}}$$

, and

$$P(Y = 0|X = x) = 1 - \sigma(f(x)) = \frac{1}{1 + e^{+f(x)}}$$

. Combining the two equations yield

$$P(y_i|x_i) = \frac{1}{1 + e^{-y_i f(x_i)}}$$

Assuming independence, the likelihood it the product of the $P(y_i|x_i)$, which is,

$$L = \prod_{i=i}^{m} \frac{1}{1 + e^{-y_i f(x_i)}}$$

. Take the log of the likelihood function yields,

$$l = \sum^{m} log(1 + e^{-y_i f(x_i)})$$

, which is the loss function.

Q10 (10pts):

You want to find someone whose birthday matches yours. What is the least number of strangers whose birthdays you need to ask about to have a 50-50 chance?

A10:

For this problem, we ingore the leap year, so that every 365 days have the equal probability of being a birthday. We generalize the problem as: given r people in a group, what is the probability that two share the same birthday. We can assume $r \leq 365$ since there will sure be two people share the same birthday if $r \geq 366$. We also assume each day within a year has equal probability of being a birthday. Define E as the event where at least two people share the same birthday. $P(E) = 1 - P(E^c)$ where %P(E^c)% is the probability that no people shares the same birthday. Total number of the event $E^c = 365 * 364 * 363... * (365 - r + 1) = \frac{365!}{(365-r)!}$. The total number of events for r people to have birthday is %365^r%.

Therefore, $P(E^c) = \frac{E^c}{365^r} = \frac{365!}{365^r(365-r)!}$, and $P(E) = 1 - P(E^c) = 1 - \frac{E^c}{365^r} = 1 - \frac{365!}{365^r(365-r)!}$

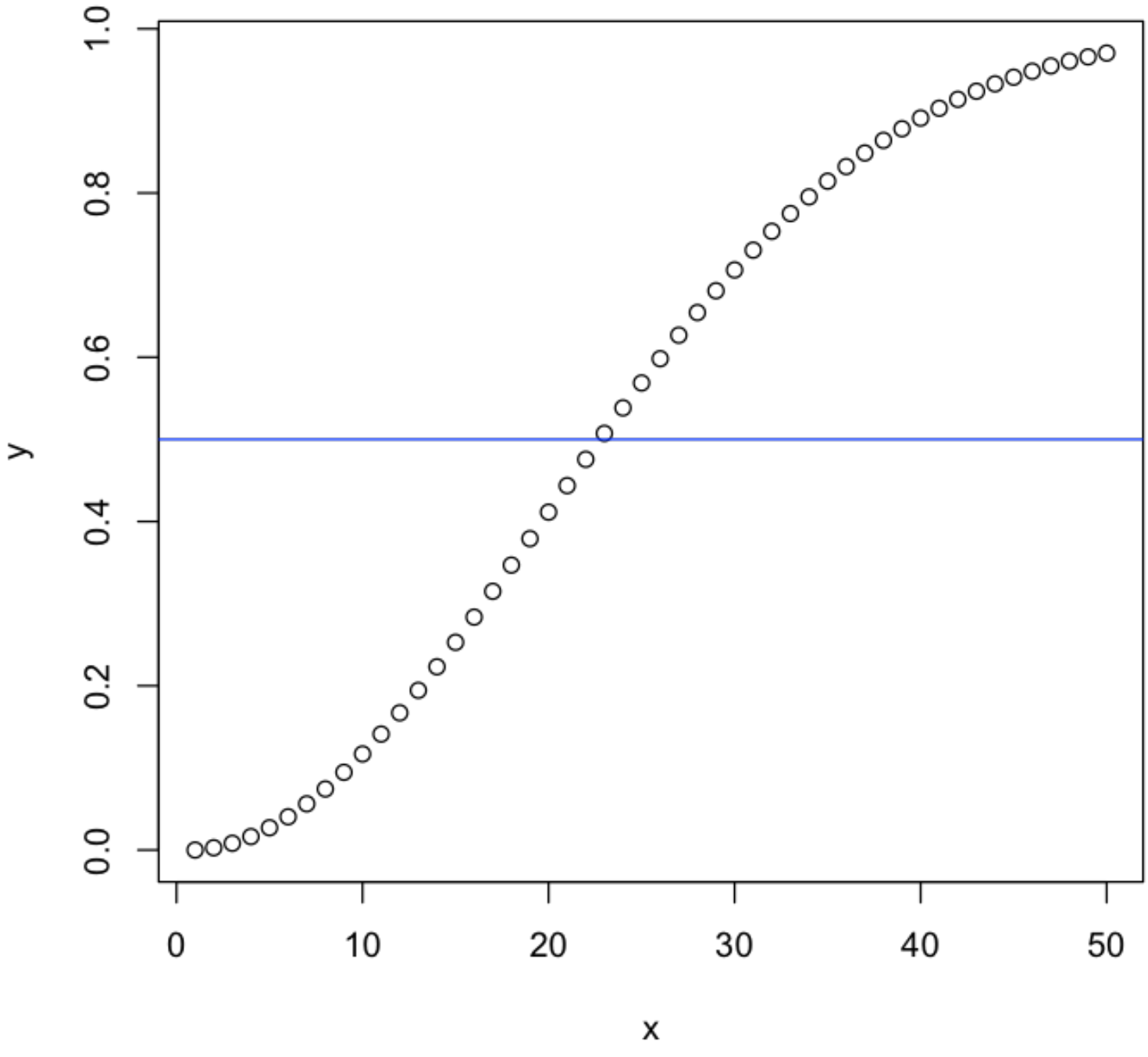
Based on the below plot, the least number of people required is 23.

In [49]:

```
x <- c(1:50)
y = 1 - exp(lfactorial(365) - lfactorial(365 - x))*(365^-x)

plot(x,y)
abline(h=0.5, col="blue")
value = x[which(y > 0.5)]
value
```

```
23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42
43  44  45  46  47  48  49  50
```



Q11 (10pts):

Let X denote the IQ of a randomly selected adult. Assume that X is normally distributed with unknown mean μ and (a strangely known) standard deviation of 16. Let sample size n = 64. And, while setting the probability of committing a Type I error to α = 0.05, test the null hypothesis H 0 : μ = 100 against the alternative hypothesis that H A : μ > 100. What is the power of the hypothesis test when μ = 108?

A11:

From looking up the z-table, for type I error of $\alpha = 0.05$, the z threshold value is 1.64.
Based on description $\sigma = 16$, $n = 64$, and $\mu = 100$, using the the equation $z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, we have

$$\bar{X} = \mu + z\frac{\sigma}{\sqrt{n}} = 106.56$$

The power of the hypothesis test when μ = 108 is

$$P = P(\bar{X} \geq 106.56 and \mu = 108) = P(z \geq \frac{106.56-108}{\frac{16}{\sqrt{16}}} = P(z \geq 0.36) = 1 - P(z < -0.36))$$

From looking up the Cumulative Probabilities of the Standard Normal Distribution table,
$P(z < -0.36) = 0.35942$. Therefore $P = 1 - 0.35942 \approx 0.64$.
So, we have $\sim 64\%$ of chance rejecting the null hypothesis test when $\mu = 108$.

Q12 (10pts):

Basic R operations
(a) (5pts) x1 <- c(1, 4, 3, NA, 7), how to calculate the mean of x1, excluding NA?
(b) (5pts) Use boxplot detecting outliers from cars2, and explain the default criterion

#Inject outliers into data.
cars1 <- cars[1:30,] # original data
cars_outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218)) #introduce outliers.
cars2 <- rbind(cars1, cars_outliers) # data with outliers

A12:

(a) use na.rm=TRUE within mean.

In [6]:

```
# (a)
x1 <- c(1, 4, 3, NA, 7)
x1_mean <- mean(x1, na.rm=TRUE)
x1_mean
```

3.75

(b) The default criterion for selecting outlier with boxplot in R is data outside the range. Range value determines how far the plot whiskers extend out from the box. If range is positive, the whiskers extend to the most extreme data point which is no more than range times the interquartile range from the box. A value of zero causes the whiskers to extend to the data extremes. The default value for range is 1.5, with means any data point outside the 1.5 interquartile range would be considered as an outlier.

Red dots in the below plot are the detected outliers with the default range value of 1.5.

In [50]:

```
# (b)
cars1 <- cars[1:30, ] # original data
cars_outliers <- data.frame(speed=c(19,19,20,20,20), dist=c(190, 186, 210, 220, 218))
cars2 <- rbind(cars1, cars_outliers) # data with outliers
boxplot(cars2, outline=TRUE, outcol="red")
```