ANALYSIS OF CUSTOMER SPENDING USING LINEAR REGRESSION

MSC. DATA SCIENCE AND ECONOMICS

Tommaso Pessina[*]
Department of Economics, Management and Quantitative Methods
University of Milan
Milan, Italy
tommaso.pessina@studenti.unimi.it

September 1, 2020

ABSTRACT

The aim of this paper is to analyze how the customer expenditure vary with respect to different factor, like age, gender and income, in order to make a model for predicting the customer's expenditure. The data are collected by an American survey run outside of a supermarket. The analysis is done by running some bivariate test, association and correlation test in order to do a complete linear regression for correctly understand the implication of some factors on customer spending. After that we will discuss the fitting of a regression tree for the same purpose.

***K*eywords** Statistical Learning · Linear Regression · Regression Tree · Supervised Learning · More

1 Introduction

Firstly we can say few words about the structure of the dataset. We have a column that identify anonimously the customer (CustomerID) by ascendent number. The second column identify the gender of the relative customer (Gender) and, as we can imagine, it can contain only two values: Male and Female. The third column contains the age of the customer (Age) in numeric format. Finally, we have the last two column that identify correspondingly Income and Spending score. The first one (income) it is expressed as numeric value with range 10-140, while the second (spending) it is expressed as numeric value with range 0-100. The first thing that it was done to the dataset is to rename the last two column in spending and income for simplicity. Now, given this data, there are many question that we can ask to ourself, for example:

- Does Female spend more than Male?
- Does younger people spend more than old people?
- Does higher income imply higher expenditure?

And basically this three questions rapresents our keypoints that we are going to analyse in this paper.

2 Main findings and statistical analysis

In this chapter we will discuss, for each key point, how the analysis has been articulated. Firstly, we run some bivariate test in order to know which factor really matter with respect to our response variable, that is spending. In this way we can know which variables should be included and which should not in our model.

[*] MSc. Data Science and Economics student | Bsc. Computer Science

Table 1: Spending by gender

t = 0.80488; p-value = 0.422	
Gender	Mean in group
Female	51.52679
Male	48.51136

Table 2: Spending and Age for Male

Pearson's Chi-squared test	
X-squared	2300.3
df	2236
p-value	0.1678

2.1 Bivariate test

What we want to know, at this point, is which variable is associated to our response variable: the customer spending. Firstly, we may ask ourself if mean spending differ by gender. We can easly answer to that by running a t-test with gender as explanatory variable and we will obtain the result shown in the Table 1. The t-test are useful to determine if two dataset are significantly different from each other. The the formula computed by R is something like that:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

From this we know, although there is little difference in mean, since the p-value is bigger than 0.05, we fail to reject the Null Hypothesis and we can say that there is no statistically significant difference between spending for Male and for Female.
Thereafter, we can search for some association between the response and the explanatory variable by running a set of Pearson's Chi-Squared association test. In the next subchapters we will discuss all of this Chi-square test.

2.1.1 Are spending and age associated?

Firstly, we divided the dataset for Male and for Female in order to analyze and answer in a more detailed manner to our question. The question that we want to analyze at this point is if Spending and Age are associated and, for answer to that, we run two separated Chi-Square test, one for each different gender.
We use the chi-square test because it compare observed data with data that we would expect to obtain, according to a specific hypothesis, and it is used when we have categorical variable. The formula that R compute in background is something like:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad in \quad which \quad O = Observed \quad frequence \quad E = Expexted \quad frequence$$

Note that the Null Hypothesis is that the two variable are associated (i.e. there is no statistical difference between the two variable, a "no difference" situation). So, if the p-value is very low (we consider as "threshold" value 0.05) we reject the Null Hypothesis (i.e. the two variable are not associated). Alternatively, we can focus on the X-squared value in which higher value means higher association between the two variable.
Now we can analyze the result:

- For Male, as we see in Table 2: since the p-value is 0.16 (that is bigger than 0.05) we fail to reject our Null Hypothesis, so Age and Spending for Male are associated.
- For Female, as we see in Table 3: since the p-value is 0.49 (that is bigger than 0.05) we fail to reject our Null Hypothesis, so Age and Spending for Female are associated.

For both cases, the high value of the X-square confirm that they are associated.

2.1.2 Does mean spending differ by gender?

Firstly, we can analyse a boxplot with Gender as explanatory variable (over spending) and we can see, as shown in Figure 1, that there is no statistically significant difference between the two group. In particular, Females have a very

Table 3: Spending and Age for Female

Pearson's Chi-squared test	
X-squared	2772.8
df	2772
p-value	0.4924

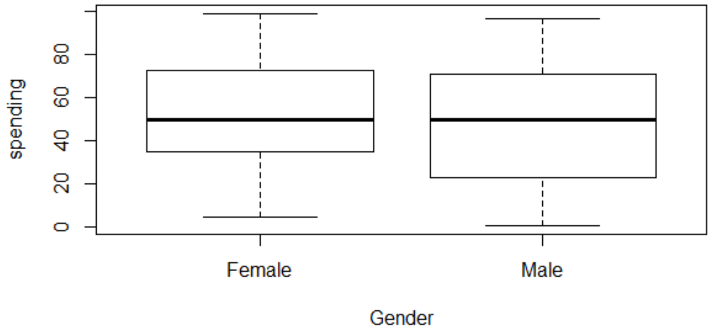

Figure 1: Mean Spending by Gender.

little bigger upper limit while Males have higher lower limit. The interval and the mean are basically the same.
As we can see in Table 4 there are no statistical significant difference in mean for the two group. Moreover, as we can see in Table 5, since the variance are a measure of dispersion and the F-statistic is a ratio of two quantities that are expected to be roughly equal with an F-value of approximately 1, a value of 0.671 confirm that there is no statistical significant difference. So, we fail to reject and we can say that mean spending does not differ by gender. Furthermore, since the Sum of Square of the residual are very high we can deduce that gender are not explicative in our model.

Table 4: Mean spending by gender

Gender	Mean	sd	n
Female	51.52679	24.11495	112
Male	48.51136	27.89677	88

Table 5: Spending and gender ANOVA

	df	Sum Sq	Mean Sq	F value	Pr(>F)
Gender	1	448	448.1	0.671	0.414
Residuals	198	132256	668.0		

2.1.3 Are spending and income associated?

Now, we can analyze if Spending are associated with Income in order to know if it is explicative in our model.
Here we use the Pearson's Chi-squared test again that is, as said, a statistical test applied to sets of categorical data to evaluate the strength of a relationship between two variable. We can consider our variable categorical because they are a sort of index for quantify the spending score or the income score and they can get a value from a fixed range.
We start by running the test separately for each gender, and we can see that:

- Male: as shown in Table 6, since the p-value is 0.044 that is less than 0.05, we reject the Null Hypothesis. So, for Male, Spending and Income are not associated; but this can be ambiguous since the X-square is pretty high;

- Female: as shown in Table 7, since the p-value is 0.16, we fail to reject the Null Hypothesis. So, Spending and Income for Female are associated.

In the light of the above, we can analyze the general case, for which we can run a Chi-square test as before and we will obtain a p-value of 0.33 that is bigger than 0.05, implying that Age and Income are not in general associated.

Table 6: Spending and Income for Male

Pearson's Chi-squared test	
X-squared	2404.5
df	2288
p-value	0.04423

Table 7: Spending and Income for Female

Pearson's Chi-squared test	
X-squared	3847.9
df	3762
p-value	0.1609

2.1.4 Correlation testing

The last thing that we can do now, is to run a correlation test in order to see (also graphically) which variable makes sense to include in our regression model.
The basic formula for correlation testing is the following:

$$Correlation = \frac{Cov(x,y)}{\sigma x * \sigma y}$$

Firstly, we have to convert Gender to numeric value. After that, we can run the test with: Age, Gender, Income and Spending. The graphical result can be seen in Figure 2.
Analyzing a little bit we can see that:

- Spending increase if income increase, although it has a complex shape. This suggest that it is necessary to do a deeper analysis maybe about different type of customer and this can be done with Clusterization (but this is not the aim of this paper);
- Spending decrease if age increase;

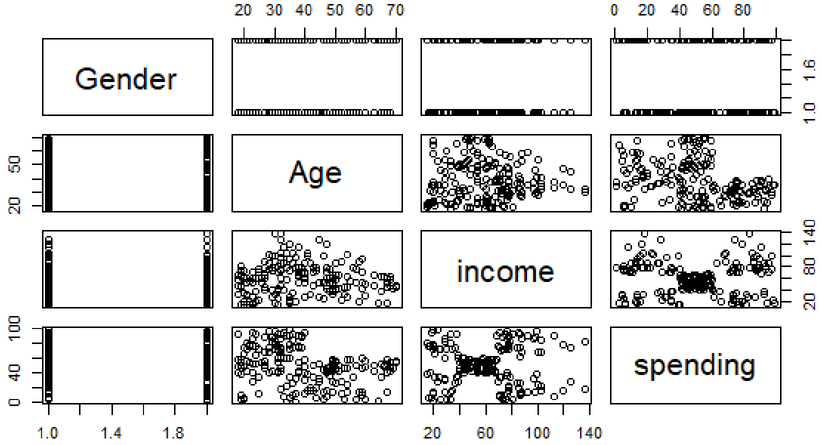

Figure 2: Mean Spending by Gender.

3 Linear Model

Now we can move on to analyze the linear model. Actually the only two variable that seems sensed to introduce are: Age and Income.
Here, firstly, we think about Multiple Linear regression, a method that is used to test how well multiple variables predict the variable of interest. By going deeper with our analysis, it will turns out that only one variable matter, so we should use a Simple Linear Regression model, that is used to test how well a varible predicts another variable.
In the next subchapter we will discuss separately the two linear model, since Income (as we seen previously) has ambiguous effect on spending.

3.1 Linear model for Age

Firstly, we can create a linear regression model with Age and Gender as variable. The formula of our linear model will be the following:

$$y = \beta_0 + \beta_1 x_i + \beta_2 \gamma_i + \epsilon$$

Where:

- y is the dependent variable;
- β_0 is the intercept, the predicted value of y when the x is 0;
- β_1 is the a regression coefficient (i.e. how much we expect y to change as x_i increases). In which x_i is a dummy variable (called independent variable) that can assume the value 0 if the individual i is Female and 1 otherwise;
- β_2 is the other regression coefficient (i.e. how much we expect y to change as γ_i increases). In which γ_i is a dummy (independent) variable for the age of the individual i;
- ϵ is the random error component of the estimation, or how much variation there is in our estimate of the regression coefficient.

From Table 8 we can see that only Age really matter in order to predict the spending of a customer. This because the p-value (of Age) is 2.85e-06 that is less than 0 (although we consider as threshold value 0.05 but 0 is even less) we can reject the Null Hypothesis and conclude that Age has a statistically significant effect on spending. So far nothing new, because we had previously seen that Gender is not statistical significant for our purpose.

Table 8: Linear model for spending

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	76.3981	6.9875	10.934	< 2e-16 ***
Gender	-1.9892	3.4973	-0.0.569	0.57
Age	-0.6006	0.1246	-4.821	2.85e-06 ***
Signif. codes: 0 '***', Adjusted R-squared: 0.09949				

Furthermore, we can analyze the Variance Inflation Factor in order to check if the introduction of this two variable is explanatory each other; we see that the value are near one so also we not have a problem of multi collinearity.
Afterwards, we run two different linear model using both Gender as reference. What we can see here is that, in both case, only Age has a positive effect in the regression over spending.

3.2 Linear model for income

From what previously analyzed we know that Income by itself does not have an effect on spending but, in the bivariate test, we seen different result by gender. In particular seems that income are associated only with Female's spending. As said, this should be analysed in more detail to be sure of its implications.
To cover any doubts we can anyway setup a linear model for income (over spending), with the following formula:

$$y = \beta_0 + \beta_1 \lambda_i + \beta_2 \gamma_i + \epsilon$$

Where:

- y is the dependent variable;
- β_0 is the intercept;
- β_1 is the regression coefficient in which λ_i is a dummy variable for the income of the individual i;
- β_2 is the other regression coefficient in which γ_i is a dummy variable for the age of the individual i;
- ϵ is the error of the estimation.

As we might expect, income is not relevant in predicting the customer spending, as shown in Table 9. This because the p-value (for income) is 0.931 that is bigger than 0.05 (our threshold value) and so we fail to reject the Null Hypothesis and we can conclude that income has not a statistically significant effect on spending.

Table 9: Linear model for income

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.347852	6.552966	11.193	< 2e-16 ***
income	0.05749	0.066197	0.087	0.931
Age	-0.604787	0.124465	-4.859	2.4e-06 ***
Signif. codes: 0 '***', Adjusted R-squared: 0.09805				

4 Step-wise regression model

In this chapter we will discuss the stepwise regression, a method by which the choice of predictive variable is tasked to an automatic process, which can give suggestion about our way.
Firstly we can see in Table 10 what the full model returns.

Table 10: Linear model for income

(intercept)	CustomerID	GenderMale	Age	income
73.786431	-0.005187	-2.011315	-0.600690	0.019099

Then, in Table 11 we can see what the step model returns as predictive variable for the regression model.

Table 11: Linear model for income

(intercept)	Age
73.7012	-0.6049

The model returns only age because we will prefer the model with lower AIC (Akaike Information Criteria) value. AIC is similar to the Adjusted R-squares and it also penalises for adding more variables to our model (we prefer parsimonious model).
So, this will confirm what we argue in the previous chapter.

5 Regression tree

Finally, since we know that only Age is relevant in the prediction of customer spending, we can fit a Tree in order to predict the mean spending knowing the Age of the customer.
We can simply grow a Tree, using the method "anova" (ANalysis Of VAriance), and obtain the Tree shown in Figure 3 below.

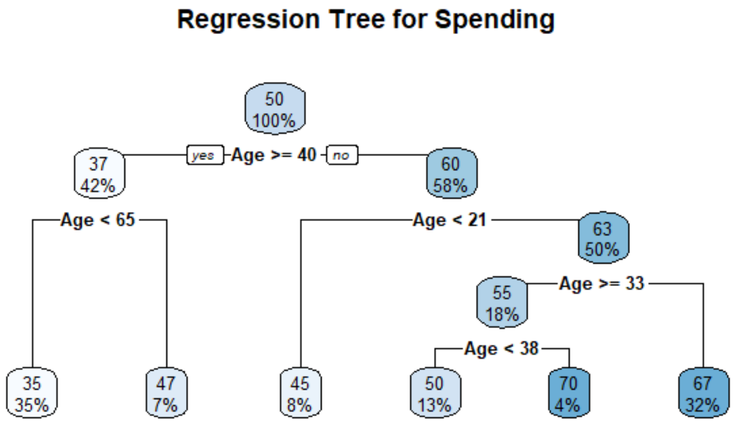

Figure 3: Regression tree for spending

Now, we can think if it necessary to prune the Tree. For doing so, we can analyse what shown in Table 12. Actually, what really matters, is to minimise the cross-validation error that here is expressed as xerror [2]. We can see that the xerror is minimised at 0.87462, that implies 1 split. This is also evident from the Figure 4.

Table 12: Linear model for spending

	CP	nsplit	rel error	xerror	xstd
1	0.196034	0	1.00000	1.00824	0.076984
2	0.034945	1	0.80397	0.87462	0.079903
3	0.022580	2	0.76902	0.88095	0.081583
4	0.021225	3	0.74644	0.90927	0.083336
5	0.012450	4	0.72521	0.89436	0.082878
6	0.010000	5	0.71277	0.92271	0.086340

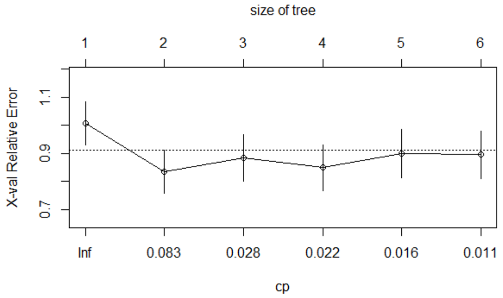

Figure 4: Cross-validation error for the library rpart

We run several time the fitting of the tree because, as you can see, seems that six is very close to the threshold and it continuously jump up, near and down. As matter of fact, according to the Figure 4, on average six is above the cutting line.
So, we choose not to prune the regression tree.

[2] The result comes from the R library "rpart"

6 Further analysis

As further analysis we can think, as said before, on clustering the different type of customer, but this will be unsupervised learning and this is not the aim this paper.
In addition to the above, it may be convenient to include other factor in the regression tree (e.g. income).

7 Conclusion

To conclude our analysis, we can summarise our main finding:

- Customer spending depend upon Age for both Male and Female;
- Mean spending does not differ by gender;
- Spending and income are associated only for Female, but this would require further analysis;
- From the correlation testing and the linear model comes up that only Age has a statistically significant effect on customer spending;
- We fit a regression tree and comes up that is not necessary to prune it;
- We discussed what can be analysed further.

We can also analyse a little bit our final linear model for predict the customer spending, that is:

$$y = \beta_0 + \gamma_i \beta_1 + \epsilon$$

where only the age of the i-th customer is included. The summary of this model is shown in Table 13.
Finally, in Figure 5 we can see graphically the linear regression model with Age and Spending for Male and Female.
And this conclude our analysis of customer spending.

Table 13: Linear model for income

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.7012	5.1238	14.384	< 2e-16 ***
Age	-0.6049	0.1241	-4.873	2.25e-06 ***
Signif. codes: 0 '***', Adjusted R-squared: 0.1026				

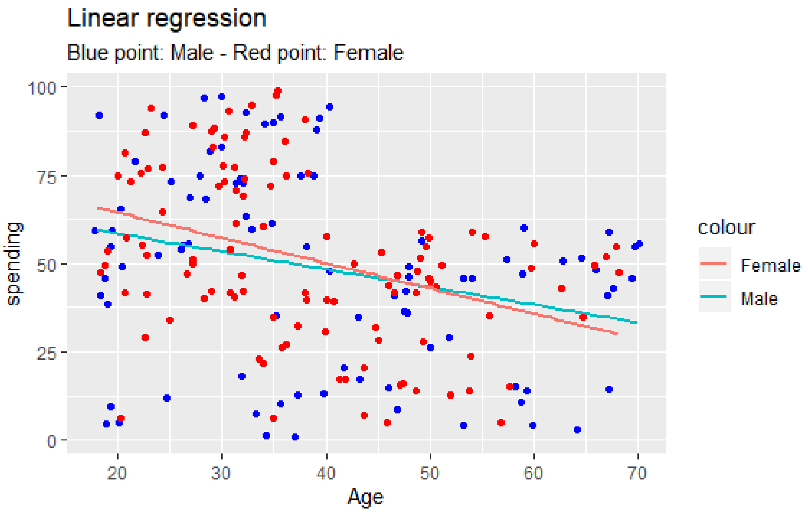

Figure 5: Linear model for Spending over Age

8 Code appendix

In this final chapter we will include all the R code (as appendix).

```
### Analysis of custmoer spending using linear regression

Dataset <- read.csv("Mall_Customers.csv")
colnames(Dataset)[4]<-"income"
colnames(Dataset)[5]<-"spending"
View(Dataset)

m = subset(Dataset, Gender=="Male")
f = subset(Dataset, Gender=="Female")
mod<-lm(m$spending~m$Age, data=m)
predicted_df <- data.frame(spendingPred = predict(mod, m), Age=m$Age)

library(ggplot2)

#plot the linear model
ggplot() +
  labs(x = "Age", y = "spending", title = "Linear␣regression",
       subtitle = "Blue␣point:␣Male␣-␣Red␣point:␣Female") +
  geom_jitter(aes(m$Age,m$spending), colour="blue") +
  geom_smooth(aes(m$Age,m$spending, color="Male"), method=lm, se=FALSE) +
  geom_jitter(aes(f$Age,f$spending), colour="red") +
  geom_smooth(aes(f$Age,f$spending, color="Female"), method=lm, se=FALSE)


## ->Bivariate tests<- ###

## Does mean spending differ by gender? YES BUT NOT SO MUCH (P-VALUE IS NOT VERY LOW)
t.test(spending~Gender, alternative='two.sided', conf.level=.95,
       var.equal=FALSE, data=Dataset)

## MALE Are Spending and Age associated? YES
# p-value is 0.16 (bigger than 0.05) we fail to reject
mytab <- xtabs(~spending+Age, data=m)
mytab
Test <- chisq.test(mytab, correct=FALSE)
Test

## FEMALE Are Spending and Age associated? YES
# p-value is 0.49 (bigger than 0.05) we fail to reject
mytab <- xtabs(~spending+Age, data=f)
mytab
Test <- chisq.test(mytab, correct=FALSE)
Test

## Does mean spending differ by gender? NO
boxplot(spending ~ Gender, data=Dataset) #no outlier, male larger lower limit
library(gplots)
plotmeans(spending ~ Gender, data=Dataset) # by mean famale little spend more
library(plyr)
ddply(Dataset,~Gender,summarise,mean=mean(spending),sd=sd(spending),
      n=length(spending))
summary(aov(spending ~ Gender, data=Dataset)) #p-value > 0.05 no difference

## MALE Are Spending and Income associated? NO
# p-value is 0.04 (less than 0.05) we reject (Age and Income are not associated)
```

```
mytab <- xtabs(~m$spending+m$income, data=m)
mytab
Test <- chisq.test(mytab, correct=FALSE)
Test

## FEMALE Are Spending and Income assocciated? YES
# p-value is 0.16 (bigger than 0.05) we fail to reject
mytab <- xtabs(~f$spending+f$income, data=f)
mytab
Test <- chisq.test(mytab, correct=FALSE)
Test

##General case? NO
# Correlated p-value 0.33 not associated
mytab <- xtabs(~spending+income, data=Dataset)
mytab
Test <- chisq.test(mytab, correct=FALSE)
Test

### Correlation analysis
Dataset$Gender<-as.numeric(Dataset$Gender)
cor(Dataset[,c("Gender","Age","income","spending")])
pairs(~Gender+Age+income+spending ,data=Dataset)
#e.i. spending little increase if income increare
#     spending little decrease if age decrease

#seems sensed to include age and income

### -> Linear models for Spending <- ###

library(car)
Dataset<-na.omit(Dataset)
Dataset$spending<-as.numeric(Dataset$spending)
Dataset$Gender<-as.numeric(Dataset$Gender)
mod<-lm(Dataset$spending~Dataset$Gender+Dataset$Age, data=Dataset)
#Gender does not matter, Age matter
summary(mod)
par(mfrow=c(2,2))
plot(mod) #ok residual btw -2 2

vif(mod) # variance inflation factors
sqrt(vif(mod)) > 2 # no problem of collinearity

#first category as reference (Male in this case)
mod2<-lm(spending~Gender+Age+income, data=Dataset)
summary(mod2)

## use F as reference
library(tidyverse)
Dataset$Gender<-as.factor(Dataset$Gender)
Dataset <- Dataset %>% mutate(Gender = relevel(Gender, ref = "1")) #1 is Female
mod2<-lm(spending~Gender+Age+income, data=Dataset)
summary(mod2)

### model with all the variable
library(lubridate)
mod4<-lm(spending~., data=Dataset)
summary(mod4)
vif(mod4) # variance inflation factors
```

```
#### -> Check income#####

library(car)
Dataset<-na.omit(Dataset)
Dataset$income<-as.numeric(Dataset$income)
Dataset$Gender<-as.numeric(Dataset$Gender)
mod<-lm(Dataset$spending~Dataset$income+Age, data=Dataset)
summary(mod) #only age matters
par(mfrow=c(2,2))
plot(mod)

### model with all the variable
library(lubridate)
mod4<-lm(income~., data=Dataset) #customer are ordered by income asc
summary(mod4)

vif(mod4) # variance inflation factors
sqrt(vif(mod4)) > 2 # no problem of collinearity

### variable selection
library(MASS)
# Fit the full model
full.model <- lm(spending ~., data = Dataset)
# Stepwise regression model
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
full.model
step.model

boxplot(Age ~ Gender, data=Dataset)

boxplot(income ~ Gender, data=Dataset)

### -> CHECK WITH TREE <- ###

library(rpart)
library(rpart.plot)
library(partykit)

# grow tree
fit <- rpart(spending~Age,
             method="anova", data=Dataset)

fit1<-ctree(spending~Age,
            data=Dataset)

printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# create additional plots
par(mfrow=c(1,2)) # two plots on one page
rsq.rpart(fit) # visualize cross-validation results

# plot tree
rpart.plot(fit, uniform=TRUE, main="Regression␣Tree␣for␣Spending")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

```
# prune the tree
pfit<- prune(fit, cp=0.034945 ) # from cptable

# plot the pruned tree
rpart.plot(pfit, uniform=TRUE,
     main="Pruned␣Regression␣Tree␣for␣Spending")
text(pfit, use.n=TRUE, all=TRUE, cex=.8)
```