post02-zhiheng-xu

ZHIHENG XU

2017/12/3

More about data visualization in R

Introduction

As a student in statistic major, we often need to analyze the relationship between a dependent variable and multiple independent variables (aka “linear regression”). Therefore, creating different plots are the most straightforward way to help us visualize the relationship. From previous lectures , we have learned some basic functions in ggplot2, such as making a scatter plot or a bar plot. In this post, I will show you how to create other useful plots to visualize data when we conduct a linear regression.

Import dataset

```
body <- read.csv("~/stat133/stat133-hws-fall17/post02/data/bodyfat.csv")
head(body)
```

```
##   Density bodyfat Age Weight Height Neck Chest Abdomen   Hip Thigh Knee
## 1  1.0708    12.3  23 154.25  67.75 36.2  93.1    85.2  94.5  59.0 37.3
## 2  1.0853     6.1  22 173.25  72.25 38.5  93.6    83.0  98.7  58.7 37.3
## 3  1.0414    25.3  22 154.00  66.25 34.0  95.8    87.9  99.2  59.6 38.9
## 4  1.0751    10.4  26 184.75  72.25 37.4 101.8    86.4 101.2  60.1 37.3
## 5  1.0340    28.7  24 184.25  71.25 34.4  97.3   100.0 101.9  63.2 42.2
## 6  1.0502    20.9  24 210.25  74.75 39.0 104.5    94.4 107.8  66.0 42.0
##   Ankle Biceps Forearm Wrist
## 1  21.9   32.0    27.4  17.1
## 2  23.4   30.5    28.9  18.2
## 3  24.0   28.8    25.2  16.6
## 4  22.8   32.4    29.4  18.2
## 5  24.0   32.2    27.7  17.7
## 6  25.6   35.7    30.6  18.8
```

This dataset is from my stat151A class. And this post is based on the lab1 notes from my stat151A class. To reproduce my plot and results, you can also find my dataset here(https://www.rdocumentation.org/packages/mfp/versions/1.5.2/topics/bodyfat). This dataset is “a data frame containing the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men”. And the variables are

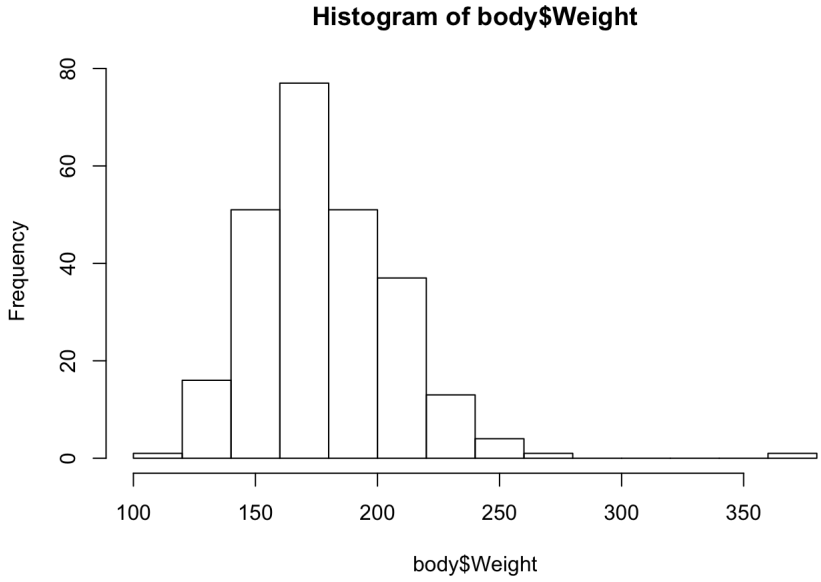
- Density determined from underwater weighing (g/cm^3)
- Percent body fat from Siri’s (1956) equation
- Age (years)
- Weight (lbs)
- Height (inches)
- Neck circumference (cm)
- Chest circumference (cm)
- Abdomen 2 circumference (cm)
- Hip circumference (cm)
- Thigh circumference (cm)
- Knee circumference (cm)
- Ankle circumference (cm)
- Biceps (extended) circumference (cm)
- Forearm circumference (cm)
- Wrist circumference (cm)

Create plots

In this plot, I’ll show you how to study the relationship between each variables in this dataset. From the lecture, we know that the easiest way to visualize the distribution of data is create a histogram .For example,

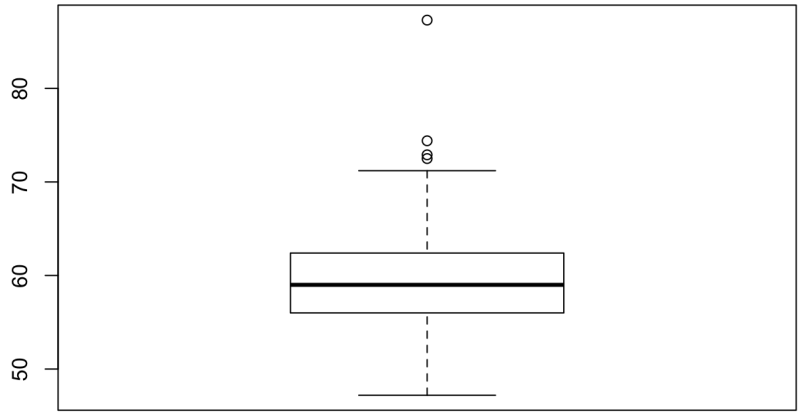
```
hist(body$Knee)
```


```
hist(body$Weight)
```

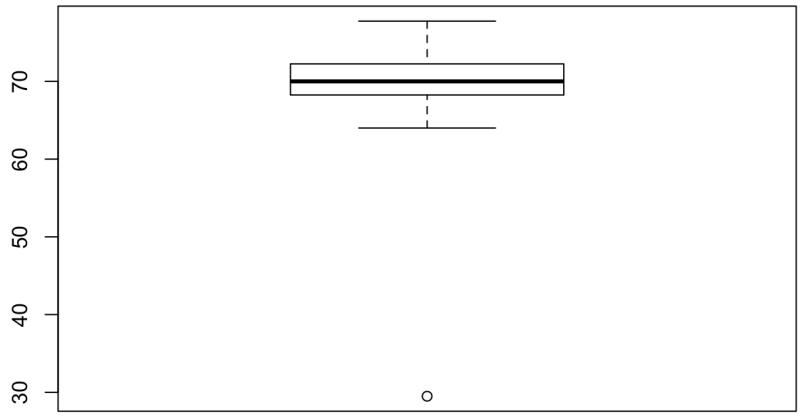


Or by creating a box plot.

```
boxplot(body$Thigh)
```



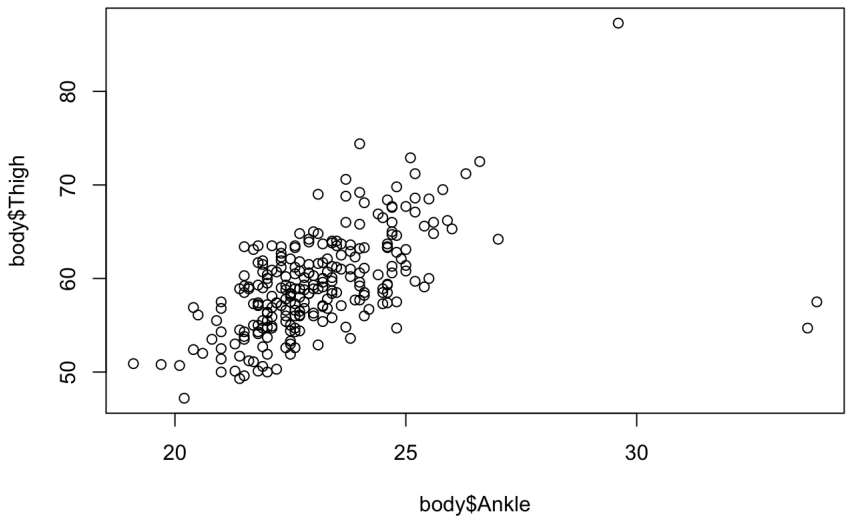

```
boxplot(body$Height)
```



From boxplot, we can see the distribution of data(including its mean and its spread), and also if there is an outlier in our data.

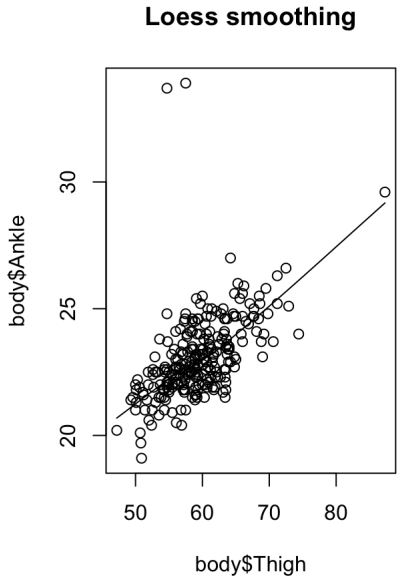
Histogram and boxplot is a good method to see the distribution of a single variable. But we sometimes need to study the relationship between two variables.For example, I want to know the relationship between Thigh and Ankle.

```
plot(body$Ankle,body$Thigh)
```



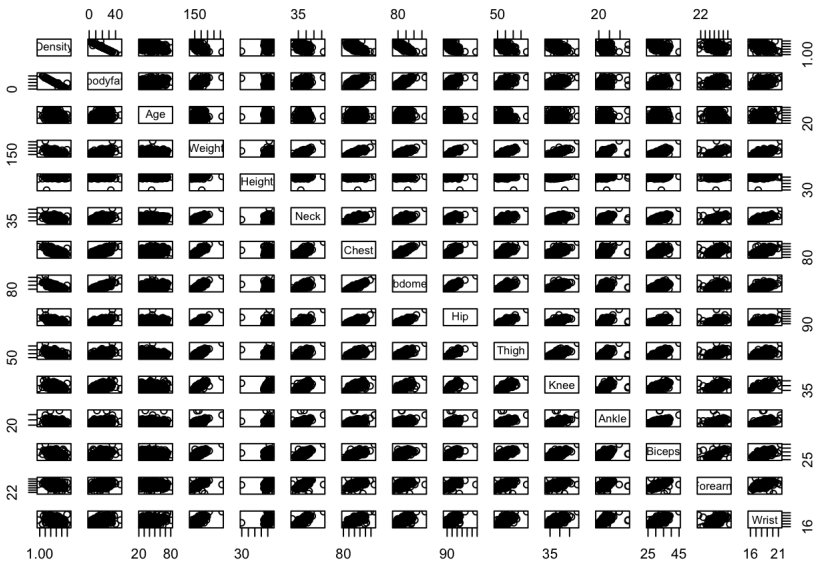
To better visualize the relationship between these variables, we can add a locally weighted regression line.

```
par(mfrow=c(1,2))
scatter.smooth(body$Thigh, body$Ankle)
title("Loess smoothing")
```



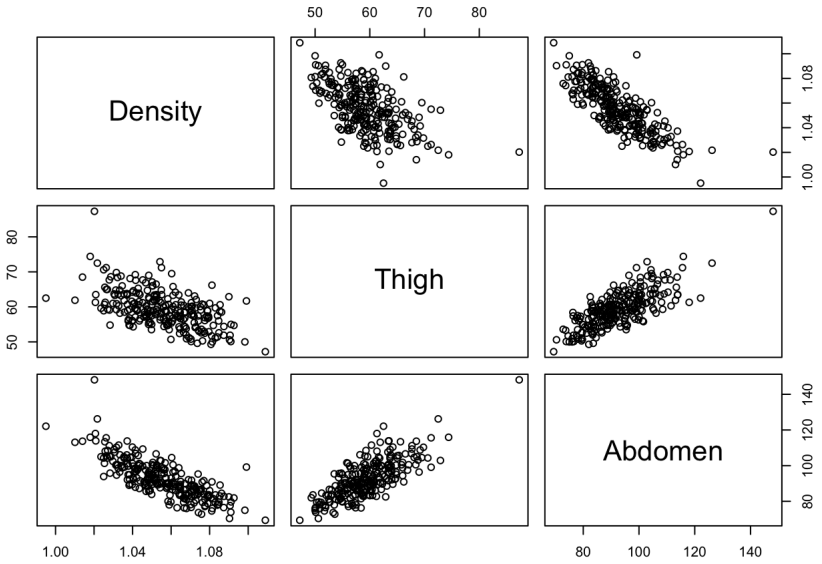
However, sometimes a dataset includes many variables, and it will be really time-consuming to draw the scatter plots one by one. Therefore, in this case, we can use pair() to see all the possible scatter plots.

```
pairs(body)
```



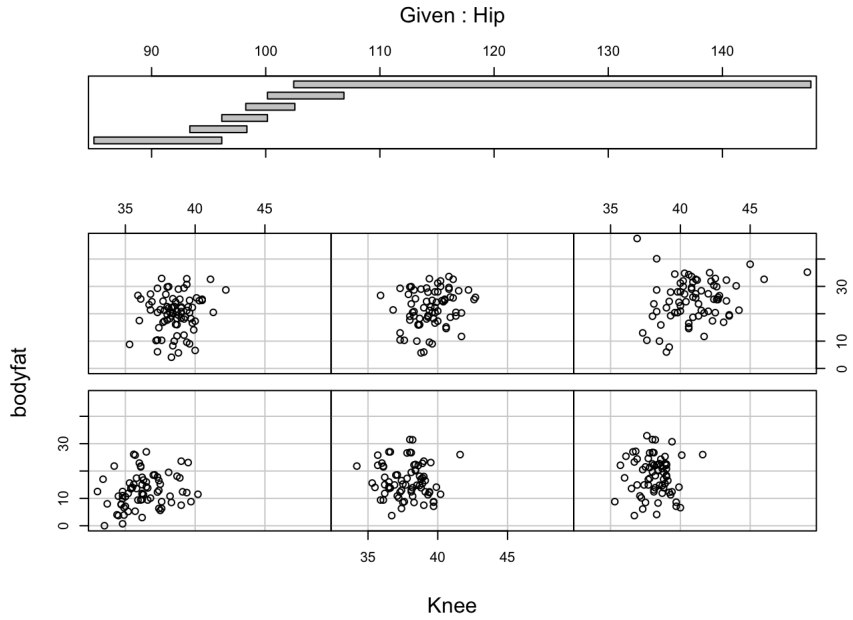
Moreover, we can also use pairs() to find the relationship on these two variable when there is a condition on a third variable. For example,

```
pairs(~Density + Thigh + Abdomen, data=body)
```



Lastly, I'll show to how to use coplot(). Coplot() is a better way to show the relationship between variables relation when given a condition. For example, i want to the relatioship between bodyfat and knee, when hip is given.

```
coplot(bodyfat ~ Knee | Hip, data=body)
```



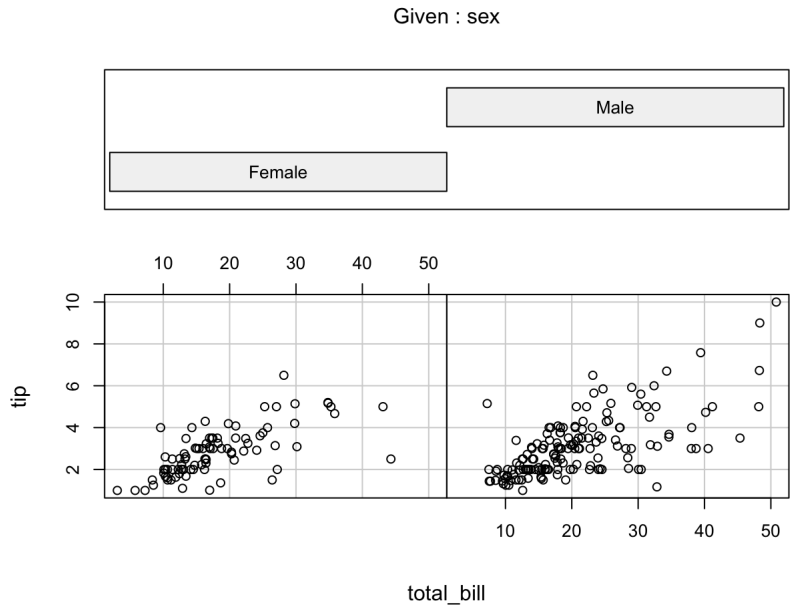
And actually, coplot() is extremely useful to study the conditional relationship when we have a cotagorical variable. Since bodyfat doesn't have catogorical variables, I will import another dataset to show you this.

```
data(tips, package="reshape2")
head(tips)
```

```
##   total_bill  tip    sex smoker day   time size
## 1      16.99 1.01 Female     No Sun Dinner    2
## 2      10.34 1.66   Male     No Sun Dinner    3
## 3      21.01 3.50   Male     No Sun Dinner    3
## 4      23.68 3.31   Male     No Sun Dinner    2
## 5      24.59 3.61 Female     No Sun Dinner    4
## 6      25.29 4.71   Male     No Sun Dinner    4
```

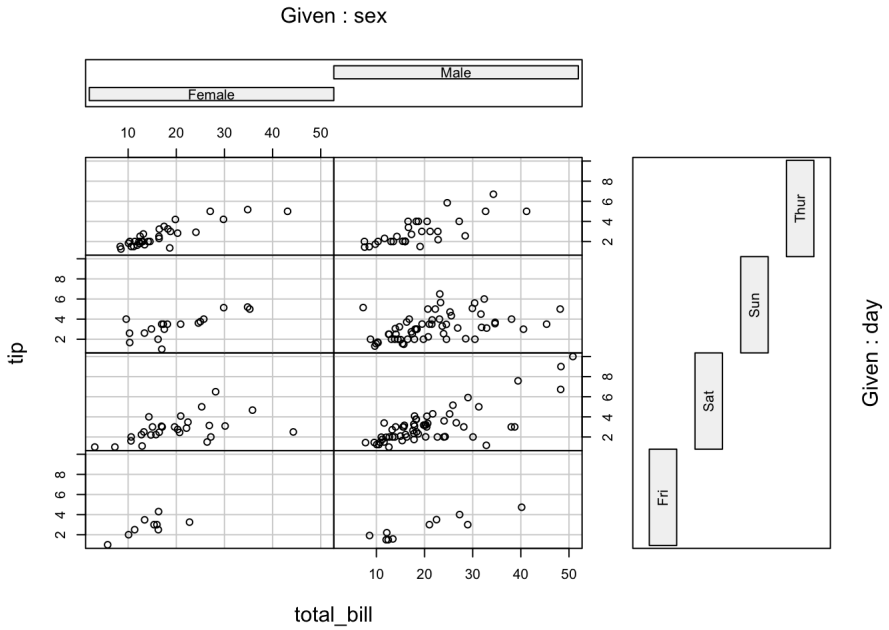
Suppose we want to know if gender has a impact on their tips, we can use coplot() to show this.

```
coplot(tip ~ total_bill | sex, data=tips)
```



From this coplot, we can see that males normally pay more tips than females. Moreover, we can use coplot() not only to study the relationship based one condition, but also on two or more conditions. For example, it we want to know if gender and day has a collective impact on the tips,

```
coplot(tip ~ total_bill | sex * day, data=tips)
```



From the plot, we can see that males tend to tip more on weekend, while females always tip less than males.

References

- https://www.rdocumentation.org/packages/mfp/versions/1.5.2/topics/bodyfat
- https://www.rdocumentation.org/packages/graphics/versions/3.4.0/topics/coplot
- https://www.statmethods.net/graphs/boxplot.html
- https://www.rdocumentation.org/packages/graphics/versions/3.4.0/topics/pairs
- https://www.datacamp.com/courses/data-visualization-in-r
- https://www.r-bloggers.com/7-visualizations-you-should-learn-in-r/
- https://www.dezyre.com/data-science-in-r-programming-tutorial/data-visualizations-tools-r