线性回归分析 HW2

尹秋阳 2015011468

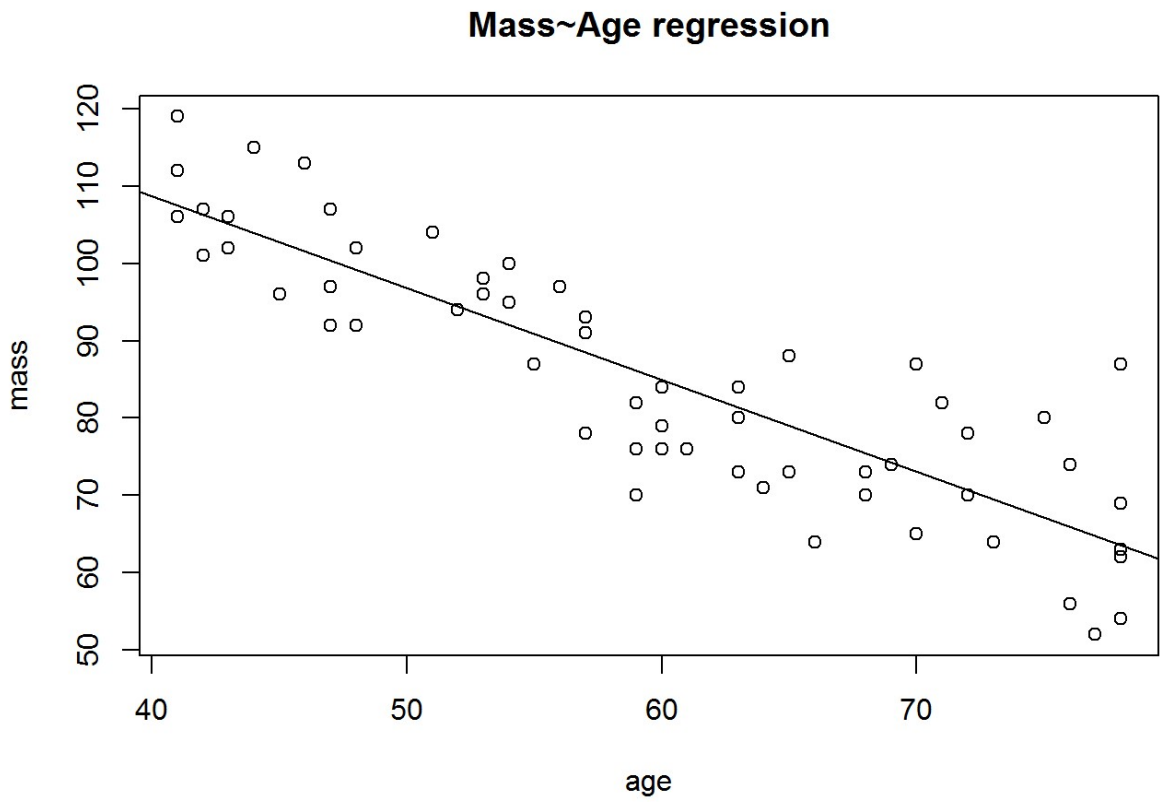
2017年3月15日

Problem 1

a

- Obtain the estimated function.Plot the estimated regression function and the data
- A good fit?
- Does your plot support the anticipation?

```
data_1=read.table("CH01PR27_967407278.txt",header = T)
x=data_1$age
y=data_1$mass
fit=lm(y~x)
plot(x,y,xlab = "age",ylab = "mass",main = "Mass~Age regression")
abline(fit)
```



基本上还算个不错的fit

的确从来看的确随着年龄的增长muscle mass会下降。

b

- a point estimate of the difference in the mean muscle mass differing in age by one year
- a point estimate of the mean muscle mass aged X=60 years
- the value of the residual for the eighth case
- a point estimate of σ^2

```
b0=fit$coefficients[1]
(b1=fit$coefficients[2])
```

```
##         x
## -1.189996
```

b1 就是斜率，也就是年龄上升一岁，肌肉质量平均下降的重量(-1.19)

```
b1*60+b0
```

```
##        x
## 84.94683
```

这是60岁的muscle的估计值,约为84.94

```
fit$residuals[8]
```

```
##        8
## 4.443252
```

第8个样例的残差，约为4.443

```
m.df=fit$df.residual
SSE=sum(fit$residuals^2)
(o2=SSE/m.df)
```

```
## [1] 66.80082
```

这是方差的估计值，其中SSE是手算的,方差约为66.8

Problem 2

a

- Conduct a test to decide whether there is a negative linear assocciation
- Type I error at 0.05
- State the alternatives, decision rule and conclusion
- P-value

We let

$$H_0 : \beta_1 \geq 0$$
$$H_1 : \beta_1 < 0$$

We use the following statics:

$$\frac{b_1}{s(b_1)} \sim t_{n-2}$$

where $s(b_1)$ stands for $ $ i.e. $\sqrt{\frac{SSE}{S_{xx}(n-2)}}$

Then We do the testing:

```
sxx=sum((x-mean(x))^2)
s.b1=sqrt(o2/sxx)
(T=(b1-0)/s.b1)
```

```
##         x
## -13.19326
```

```
qt(0.05,df = m.df)
```

```
## [1] -1.671553
```

```
pt(T,df = m.df)
```

```
##            x
## 2.061993e-19
```

As you can see above, P-value is really low, (about 2e-19), so we just deny the Null hypothesis

b

No.

整个实验的对象应该是指成年人，不能肆意扩展到儿童。事实上，常识上可以知道成长期的肌肉质量肯定是随着年龄增长而增长的。那需要另外的实验和检验。

c

```
confint(fit)[2,]
```

```
##     2.5 %    97.5 %
## -1.370545 -1.009446
```

其次再说明置信区间的获得不需要特定的X。

由于实际上置信区间的获取如下：

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$$

其中$s(b_1) = \frac{MSE}{\sum (X - \bar{X})}$ 与具体的X无关

所以整个置信区间的获取与特定点X无关，即不需要specific point

Problem 3

a

- Calculate the power at $\beta = -1$
- generate the power function

Since it is one-sided test. We slightly change the original code.

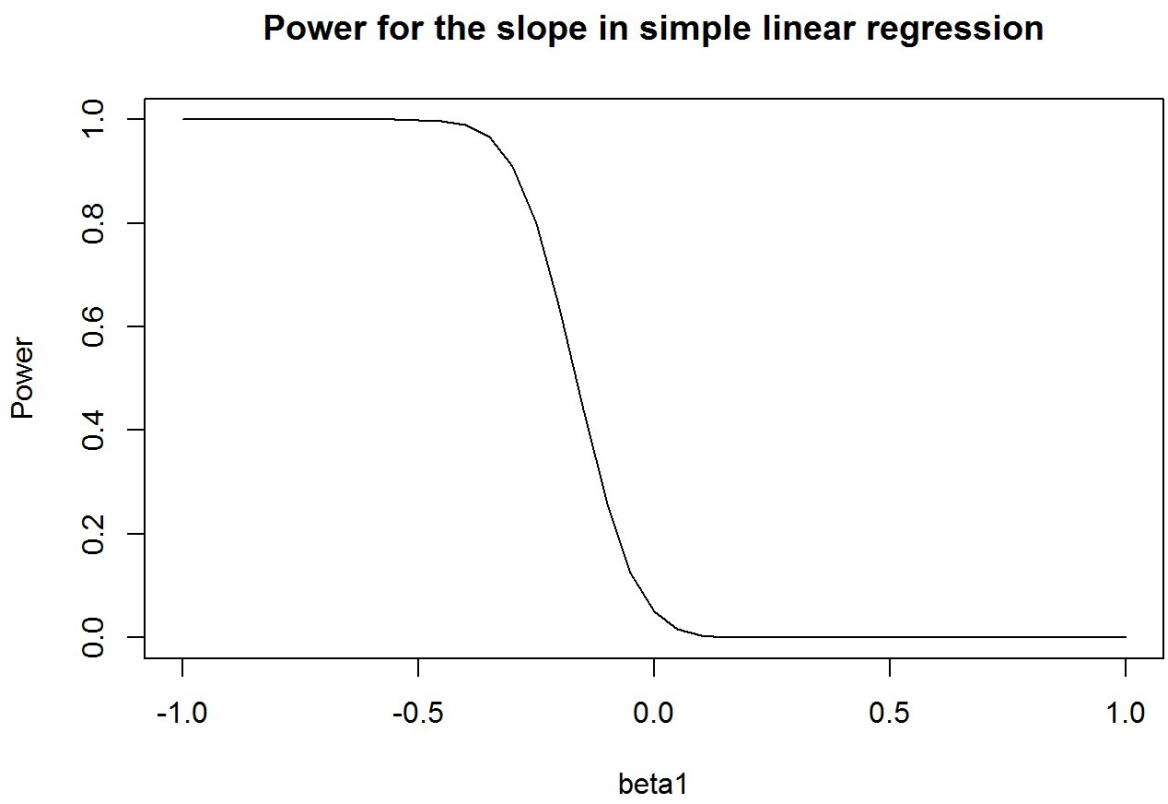
```
find_power <- function(n, sig2, ssx, beta1, alpha){
  sig2b1 = sig2/ssx
  df = n-2
  delta = beta1/sqrt(sig2b1)
  tstar = qt(alpha, df)
  power = pt(tstar,df,delta)
  return(power)
}
```

```
source("power.r")
n = length(x)
sig2 = 81
ssx = sxx
beta1 = -1
alpha = 0.05
find_power(n, sig2, ssx, beta1, alpha)
```

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

So we conclude that the power is near to 1. the power function is ploted below:

```
beta1 = seq(-1, 1, by=0.05)
power_vec = rep(0, length(beta1))
for( i in 1:length(beta1)){
  power_vec[i] = find_power(n, sig2, ssx, beta1[i], alpha)
}
plot(power_vec ~ beta1, type="l", main="Power for the slope in simple linea
r regression", ylab="Power")
```



It is reasonable. For it is a one-sided test. When β_1 is greater than zero, the power of which should be less than 0.05(Type I error), and in the case power(0)=0.05

b

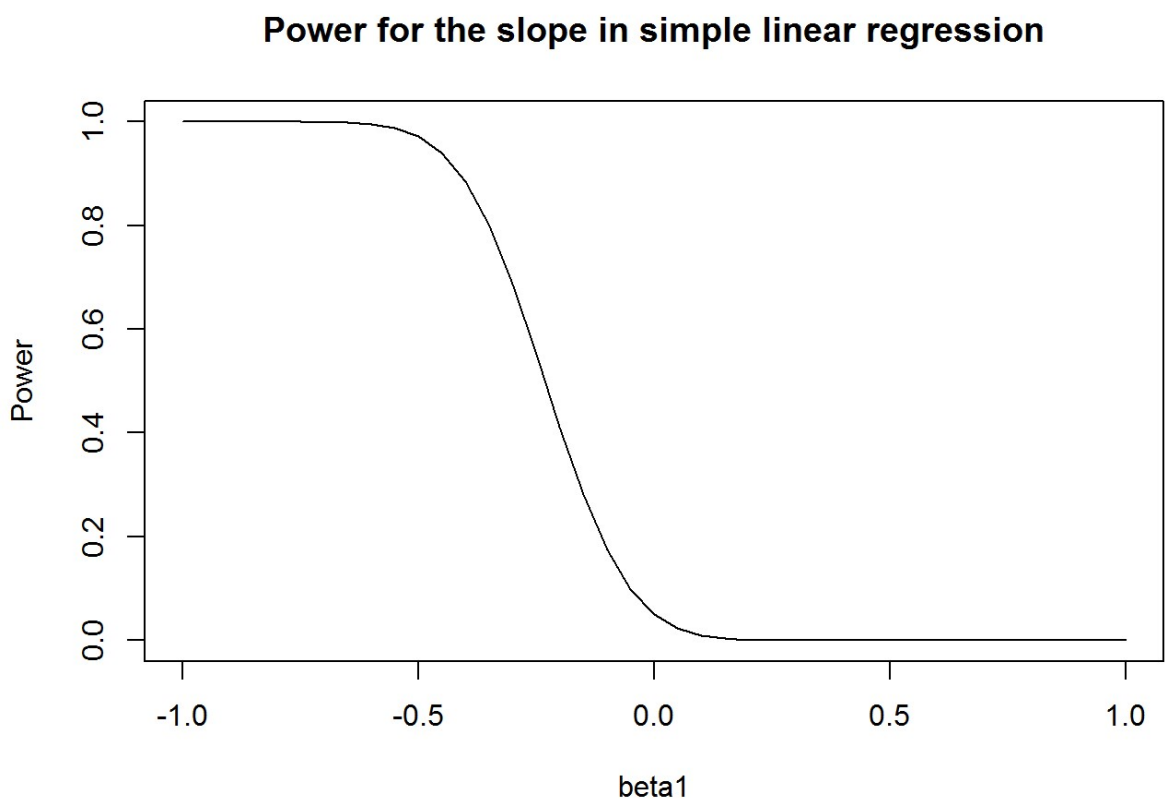
We first do the random sampling (using r sample fumction) and so the same things as part a

```
my_sample=NULL
for(i in seq(0,45,by = 15)){
  my_sample=c(my_sample,sample(i:(i+14),8))
}
x_new=x[my_sample]
n = length(x_new)
sig2 = 81
ssx = sum((x_new-mean(x_new))^2)
beta1 = -1
alpha = 0.05
find_power(n, sig2, ssx, beta1, alpha)
```

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

So we conclude that the power is still near to 1.

```
beta1 = seq(-1, 1, by=0.05)
power_vec = rep(0, length(beta1))
for( i in 1:length(beta1)){
  power_vec[i] = find_power(n, sig2, ssx, beta1[i], alpha)
}
plot(power_vec ~ beta1, type="l", main="Power for the slope in simple linea
r regression", ylab="Power")
```



C

We can conclude that when sample size is greater, for the same β , the power of which tends to be greater as well. So sample size really matters in power testing.

To conclude the pros and cons, greater sample size:

- tends to have greater power
- sometimes it is good, but sometimes it is over-powered
- cost money and efforts to collect data

while smaller size:

- tends to have smaller power
- easy to collect data

D

- test the power when $\beta = -0.02$

```
n = length(x)
sig2 = 81
ssx = sxx
beta1 = -0.2
alpha = 0.05
find_power(n, sig2, ssx, beta1, alpha)
```

```
## [1] 0.6350378
```

It is not properly powered.

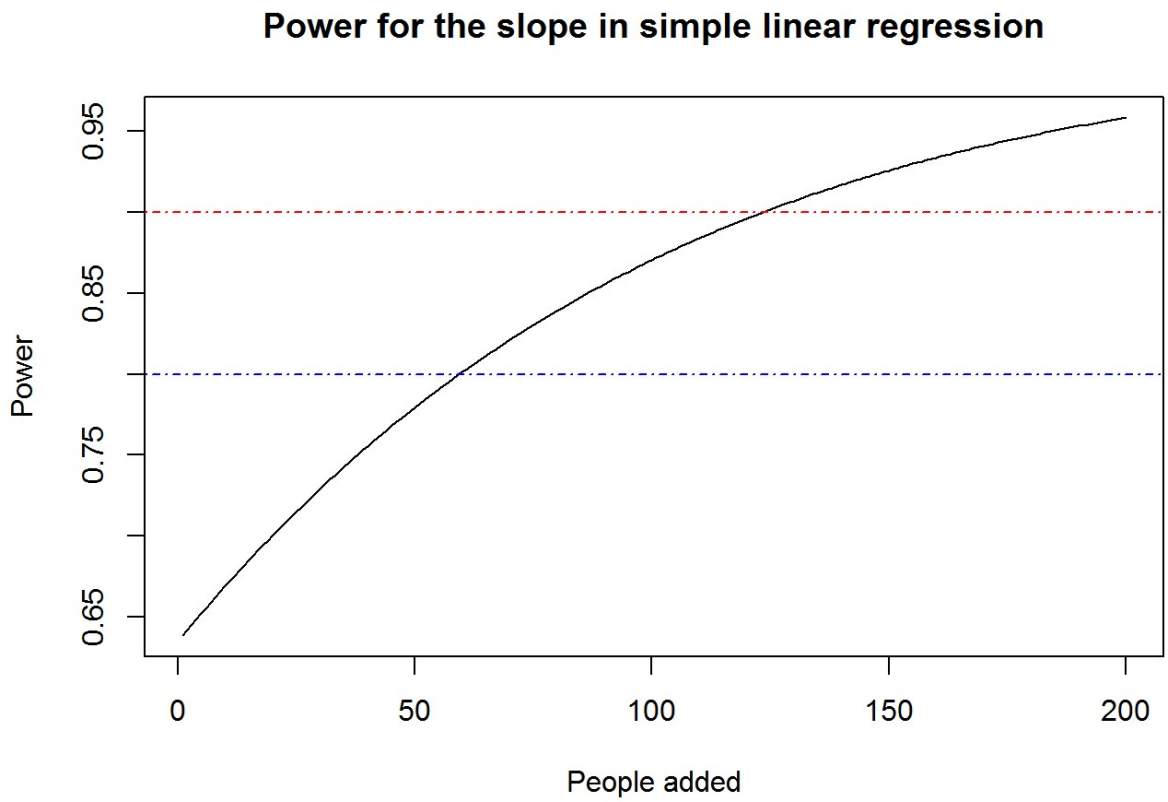
E

We should enlarge the sample size, or broaden the scope of X.

Assume that X is totally random. The following code is to decide the proper sample size.

I use R to test much sample is more proper to test as the level of -0.02:

```
people.max=200
power_vec = rep(0, people.max)
for (i in 1:people.max){
  add_people=i;
  n=60+i
  sig=81
  ssx=sxx+75*add_people #average Sxx plus is 75 for the interval of 30
  beta1 = -0.2
  alpha = 0.05
  power_vec[i]=find_power(n, sig2, ssx, beta1, alpha)
}
plot(y=power_vec,x=1:people.max, type="l", main="Power for the slope in simp
le linear regression", ylab="Power",xlab="People added")
abline(h=0.9,col="red",lty=4)
abline(h=0.8,col="blue",lty=4)
```



Here I draw two lines, the first one of which is 0.9, the other one is 0.8. As Mr.Zhu has said in the class, the power between these two is considered good enough.

as you can infer from the plot.About 50 to 110 more people are expected to take part in the test.

F

I have already comment some in the part C

- greater sample size tends to have greater power
- for some minor change at the judging boarder, we need more samples to improve power

Problem 4 and 5

We have already derived the expression of $b1$

$$b1 = \beta_1 + \sum_{i=1}^{n} \frac{X_i - \bar{X}}{S_{xx}} \epsilon_i$$

Then we continue to observe the properties of $b0$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Expectations(Problem 4):

$$E(b_0) = E(\bar{Y}) - \bar{X} E(b_1)$$

While

$$E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i))$$
$$= \beta_0 + \beta_1 \bar{X}$$
$$E(b1) = \beta_1$$

Then

$$E(b_0) = \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X}$$
$$= \beta_0$$

Variance(Problem 5)

We can know from 2.31 that $\bar{Y}$ and b_0 are independent. Then

$$var(b_0) = var(\bar{Y}) + \bar{X}^2 var(b_1)$$

while

$$var(\bar{Y}) = \frac{1}{n^2}\sum var(Y_i) = \frac{1}{n}\sigma^2$$
$$var(b_1) = \frac{\sigma^2}{S_{xx}}$$

Then

$$var(b_0) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}})$$

where S_{xx} stands for $\sum(X_i - \bar{X})^2$

In the case o 2.29, just let X_h as 0 . Then we can get the variance of b_0

$$\frac{b_1}{s(b_1)} \sim t_{n-2}$$

where $s(b_1)$ stands for $ $ i.e. $\sqrt{\frac{SSE}{S_{xx}(n-2)}}$