线性回归分析 HW1

尹秋阳 2015011468

2017年3月5日

```
## Loading required package: ggplot2
```

Problem 1

a

不能，因为尽管知道了方差和均值，无法确定随机误差项的分布。也就无法确定分位数的情况。

事实上，不同分布对应的答案是不一样的。例如ϵ $N(200, 25)$ 时对应的0.683，
ϵ $u(200 - 5\sqrt{3}, 200 + 5\sqrt{3})$ 时，应对的概率是0.577

b

如果分布已知，可以求出相应概率，只需求出205和195的CDF对应的值相减即可。如下：

```
pnorm(205,mean = 200,sd = 5)-pnorm(195,mean = 200,sd = 5)
```

```
## [1] 0.6826895
```

Problem 2

a

应该属于observational data。

experimental data一般是指研究者运用控制变量法、对照组等等。通过当一个变量变化观察另一个变量的变化状态；

observational data没有那么多要求，观测值即可。

b

这简直是扯淡。

首先，这个研究者没有控制变量。本身影响cold的因素非常多。不同参与者的身体状况、地理环境都不同。

其次，即使控制了变量，顶多说变量之间有负相关性，不能推出因果性。

需要继续分析其他因素才能推出合理的结果。

c

气候因素、个人身体状况、当地地理环境等等。

d

首先建议研究者随机设立观测组和对照组，观测组和对照组需要尽量控制其他影响的因素。

其次两个组的人应该被要求进行不同的exercise锻炼，研究者随后观测cold和exercise的相关性。

Problem 3

a

```
data=read.table("CH01PR19_818607472.txt")
colnames(data)=c("Y","X")
bc=lm(data=data,Y~X)
(b0=bc$coefficients[1])
```

```
## (Intercept)
##    2.114049
```

```
(b1=bc$coefficients[2])
```

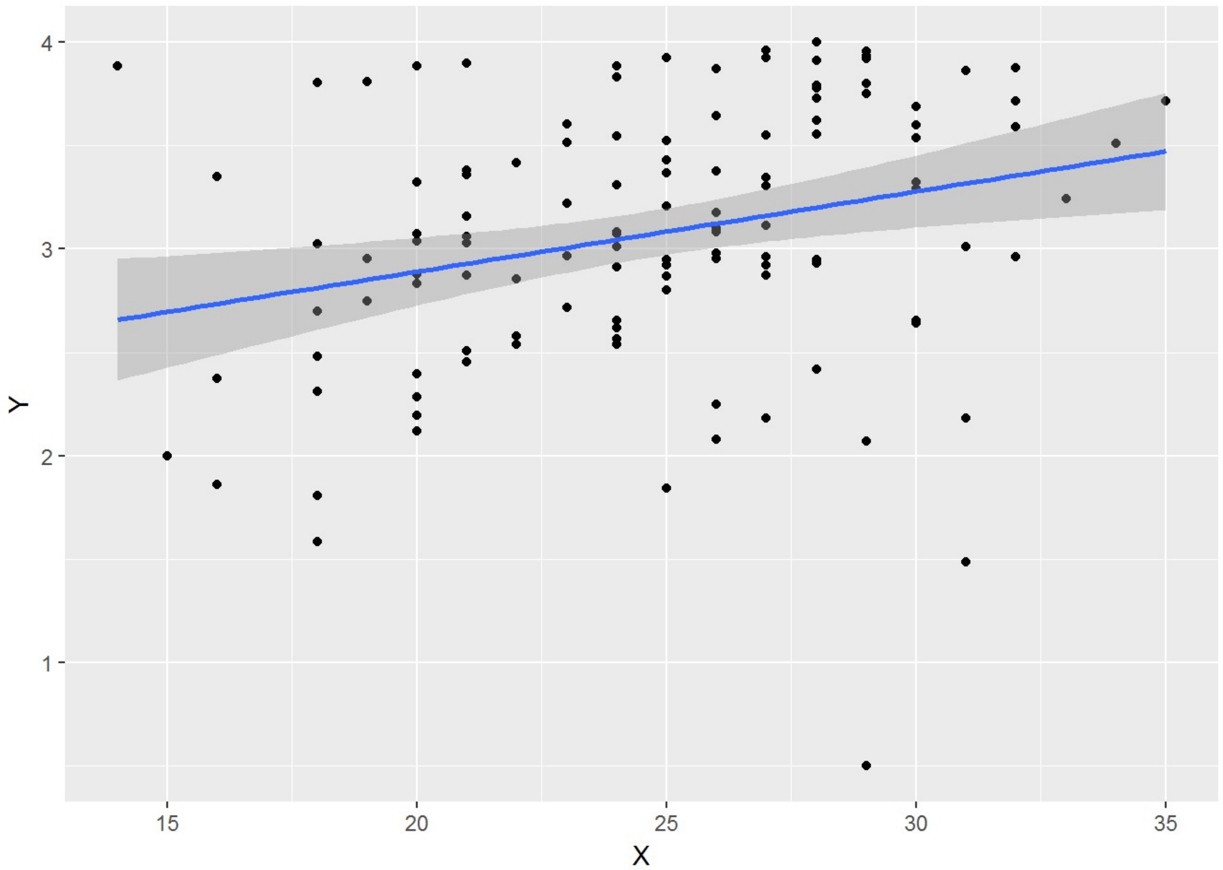
```
##          X
## 0.03882713
```

即回归方程为

$$Y = 0.03882713X + 2.114049$$

b

```
ggplot(data,aes(x=X,y=Y))+geom_point()+geom_smooth(method = "lm")
```



从图中可以看出，有大量的点落在置信区间外面

所以好像不是很适合线性模型,不fit well

c

```
(b1*30+b0)
```

```
##         X
## 3.278863
```

d

对应的是回归所得的斜率

```
(b1)
```

```
##          X
## 0.03882713
```

Problem 4

a

```
sum(bc$residuals)
```

```
## [1] -2.942091e-15
```

可以看到小于e-06，即残差之和为0

b

使用MSE作为σ^2 的估计$\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{118}$$

```
(o2=sum(bc$residuals^2)/118)
```

```
## [1] 0.3882848
```

```
(o=sqrt(o2))
```

```
## [1] 0.623125
```

σ 的单位与Y 相同，为GPA的单位

Problem 5

a

我们定义Q 为目标函数

$$Q = \sum (Y_i - \beta_1 X_i)^2$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum (X_i(Y_i - \beta_1 X_i)) = 0$$

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

b

$$L(\beta_1) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_1 X_i)^2)$$
$$\frac{d(logL(\beta_1))}{d\beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i(Y_i - \beta_1 X_i)) = 0$$
$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$

可以看到与最小二乘法所得结果相同

c

$$\hat{\beta}_1 = \frac{\sum X_i Y_i}{\sum X_i^2}$$
$$E(\hat{\beta}_1) = E(\frac{\sum X_i Y_i}{\sum X_i^2}) = E(\frac{\sum X_i(\beta_i X_i + \epsilon_i)}{\sum X_i^2})$$
$$= \beta_i + E(\frac{\sum X_i \epsilon_i}{\sum X_i^2}) = \beta_i + \frac{1}{\sum X_i^2}\sum X_i E(\epsilon_i)$$
$$= \beta_i$$

Problem 6

$$H_0 =" topicalpain\text{没有延长康复时间}"$$
$$H_1 =" topicalpain\text{延长了康复时间}"$$

```
data=c(128,135,121,142,126,151,123)
x=mean(data)
n=length(data)
sx=sqrt(var(data))
T=(x-123.7)/(sx/sqrt(n))
(p=1-pt(T,n-1))
```

```
## [1] 0.04201765
```

可以看到p-value小于5%，也就是说显著程度为5%和10%时，都拒绝原假设，即的确延长了康复时间