刘程华 2018011687 计91

1.The attached data set "data.csv" is a subset of the Ames House data set. The first row of this data set is the name of the variables, where "SalePrice" is the dependent variable (column 1) and "Lot_Area" is the independent variable (predictor; column 2). There are $n = 1598$ observations in total (rows 2-1599). Suppose we use a linear regression model to approximate the relationship between SalePrice and Lot_Area:

$$\text{SalePrice } = \beta_0 + \beta_1 \text{ Lot\_Area } + \epsilon$$

Suppose that the noise ϵ_i i.i.d. $\sim N\left(0, \sigma^2\right)$. Calculate the MLEs of $\beta_0, \beta_1, \sigma^2$

solve：

let$Y = SalePrice$ and $X = Lot\_Area$

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$$

$$\Rightarrow L\left(\beta_0, \beta_1, \sigma^2\right) = P\left(Y_1\right) P\left(Y_2\right) \ldots P\left(Y_n\right) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \sum \left(Y_i - \beta_0 - \beta_1 X_i\right)^2}$$

$$\Rightarrow \ell\left(\beta_0, \beta_1, \sigma^2\right) = \ln(L) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

To maxmize the ℓ

$$\Rightarrow \begin{cases} \frac{\ell\left(\beta_0, \beta_1, \sigma^2\right)}{\partial \hat{\beta}_0} & = 0 \\ \frac{\ell\left(\beta_0, \beta_1, \sigma^2\right)}{\partial \hat{\beta}_1} & = 0 \\ \frac{\ell\left(\beta_0, \beta_1, \sigma^2\right)}{\partial \sigma^2} & = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_0 = \bar{Y} - b_1 \bar{X} = 1.330 \times 10^5 \\ \hat{\beta}_1 = \frac{\sum_i \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_i \left(X_i - \bar{X}\right)^2} = 4.688 \\ \hat{\sigma}^2 = \frac{\sum_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{n} = \frac{1596}{1598} \times 74930^2 \end{cases}$$

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-370781  -46560  -18455   33684  418123

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.330e+05  3.795e+03   35.04   <2e-16 ***
X           4.688e+00  3.308e-01   14.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74930 on 1596 degrees of freedom
Multiple R-squared:  0.1118,   Adjusted R-squared:  0.1112
```

```
F-statistic: 200.8 on 1 and 1596 DF,  p-value: < 2.2e-16
```

Code

```
data=read.csv('/Users/liuchenghua/Downloads/data.csv',header = TRUE)
Y=data['SalePrice']
X=data['Lot_Area']
Y<-unlist(Y)
X<-unlist(X)
fit=lm(Y~X)
summary(fit)
```

2.Let $\boldsymbol{X}=(X_1,\cdots,X_n)$ be a random sample of size $n=15$ from a normal distribution $N\left(\mu,0.25^2\right)$, and the observed values are:

$$2.9, 2.8, 3.0, 2.8, 3.1, 2.7, 2.3, 2.8, 2.4, 2.8, 2.6, 2.6, 3.1, 3.2, 2.9$$

Construct a 95% confidence interval for μ

We know that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

```
> left
[1] 2.673485
> right
[1] 2.926515
```

Code

```
x=c(2.9,2.8,3.0,2.8,3.1,2.7,2.3,2.8,2.4,2.8,2.6,2.6,3.1,3.2,2.9)
sigma=0.25
alpha=0.05
len=length(x)
left=mean(x)-qnorm(1-alpha/2)*sigma/sqrt(len)
right=mean(x)-qnorm(alpha/2)*sigma/sqrt(len)
```