线性回归分析 HW3

尹秋阳 2015011468

目录

1 Problem 1

1.1 a

- obtain 95 percent confidence interval for age 60
- interpret

```
m_data=read.table("CH01PR27_967407278.txt",header = T)
reg=lm(mass~age,data=m_data)
```

```
conf=predict(reg,se.fit = T,data.frame(age=60),interval = "confidence",level = 0.95)
(conf_ans=conf$fit)
```

```
##        fit      lwr      upr
## 1 84.94683 82.83471 87.05895
```

这说明对于回归结果来说，我们有 95% 的信心认为 Y(60) 属于 82.83 到 87.05895

1.2 b

- obtain 95 percent prediction interval for age 60
- interpret

```
pred=predict(reg,se.fit = T,data.frame(age=60),interval = "predict",level = 0.95)
(pre_ans=pred$fit)
```

```
##        fit      lwr     upr
## 1 84.94683 68.45067 101.443
```

这说明对于真实值来说，我们有 95% 的信心认为真实值 Y(60) 属于 62.45 到 101.443

1.3 c

- Boundary value for $X_h = 60$
- wider than confidence band ?

```
n = 60
alpha = 0.05
dfn = 2
dfd = n-2
w2 = 2 * qf(1-alpha, dfn, dfd)
w = sqrt(w2)
alphat = 2 * (1-pt(w, dfd))
conf_band=predict(reg,se.fit = T, newdata=data.frame(age=60), interval="confidence", le
(conf_band_ans=conf_band$fit)
```

```
##        fit      lwr      upr
## 1 84.94683 82.29593 87.59774
```

和 confidence interval 比较：

```
conf_ans
```

```
##        fit      lwr      upr
## 1 84.94683 82.83471 87.05895
```

可以看到 boundary value 的确比 confidence interval 宽
原因在于：对于 boundary band 的定义：

$$P[L(X) < \beta_0 + \beta_1 x < U(x)] = 1 - \alpha$$

for all the X

而 for one fixed x 没有 “所有 x” 的限制，要求宽松。

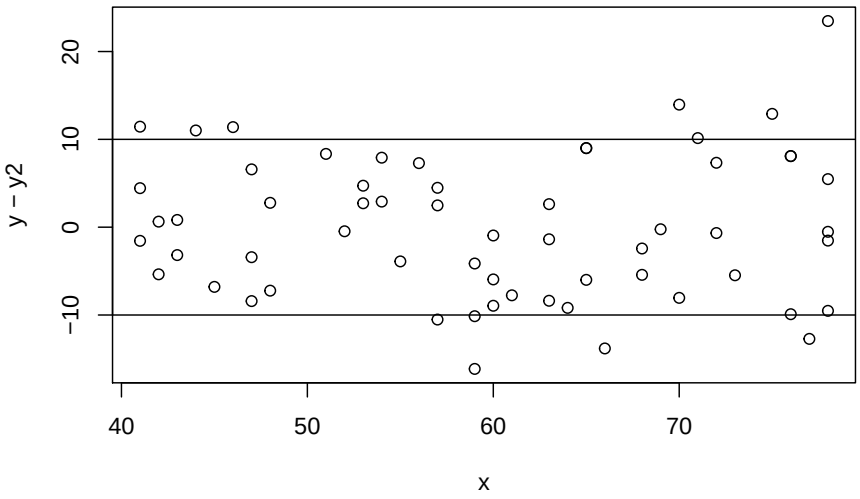
换句话说，在 confidence interval 连出来的带，所有 x 在其中的概率 $<1-\alpha$, 于是真实的置信带肯定宽于 confidence interval 连出来的带子

2 Problem 2

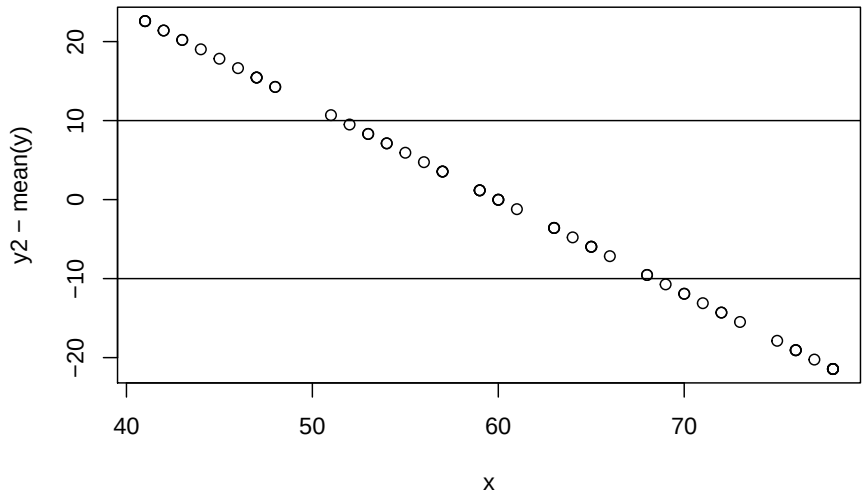
2.1 a

- Plot $Y_i - \hat{Y}_i$ against X_i
- Plot $\hat{Y}_i - \bar{Y}$ against X_i
- SSE and SSR, component of SST
- Imply about R^2

```
y=m_data$mass;y2=reg$fitted.values;x=m_data$age
plot(y=y-y2,x=x);abline(h=10);abline(h=-10)
```




```
plot(y=y2-mean(y),x=x);abline(h=10);abline(h=-10)
```



可以看到 SSE 的点在-10 到 10 的居多，SSR 的点均匀分布。于是 SSR 会比 SSE 大一些，在 SST 的组成成分中占主导地位。

这预示着 R^2 值会大于 0.5

2.2 b

- set up ANOVA table

```
(ans.table=anova(reg))
```

```
## Analysis of Variance Table
##
## Response: mass
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## age        1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.3 c

- F test
- state H_0 and H_1 decision rule and conclusion

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2.3.1 Decision rule:

We reject H_0 when F_0 is large, that is:

$$if F_0 > F(1-\alpha, df_R, df_E) = F_{\alpha,1,n-2}$$

where $F_0 = MSR/MSE$

2.3.2 Conclusion:

```
ans.table$`Pr(>F)`[1]
```

```
## [1] 4.123987e-19
```

可以看到，P-value 非常小，即我们要拒绝原假设，认为 $\beta_1 \neq 0$

2.4 d

- proportion of the total variation in muscle mass "**unexplained**"

其实就是 1-R^2 的值
我们先直接从 ANOVA 中提取出 R^2 值，命令如下：

```
(r2=summary(reg)$r.squared)
```

```
## [1] 0.7500668
```

另一种是通过自己根据 R^2 的定义计算，如下：

```
b1=reg$coefficients[2]
x=m_data$age
y=m_data$mass
r.otherway=b1*sqrt(sum((x-mean(x))^2))/sqrt(sum((y-mean(y))^2))
r.otherway^2
```

```
##       age
## 0.7500668
```

可以看到结果是相同的，都是 0.75。

于是 unexplained 比例就是 $1-0.75=0.25$ 这是一个比较低的数值。算是比较强的解释性。

不过一种说法是作为预测来说 R^2 还不够大。这完全取决于需求。

2.5 e

R^2 已经由 d 给出:

```
(r2=summary(reg)$r.squared)
```

```
## [1] 0.7500668
```

由于此处 $\beta_1 < 0$，于是：

```
(r=-sqrt(r2))
```

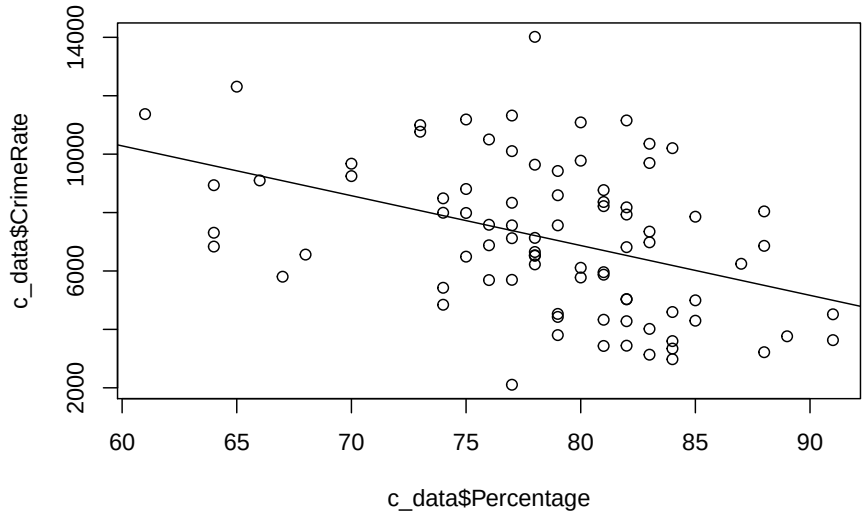
```
## [1] -0.866064
```

3 Problem 3

3.1 Preparations:

```
c_data=read.table("CH01PR28_207305355.txt",header = F)
colnames(c_data)=c("CrimeRate","Percentage")
crime.lm=lm(data=c_data,CrimeRate~Percentage)
```

```
plot(c_data$CrimeRate~c_data$Percentage)
abline(crime.lm)
```




```
x=c_data$Percentage
y=c_data$CrimeRate
```

3.2 a

- set up the ANOVA table

```
(crime.an=anova(crime.lm))
```

```
## Analysis of Variance Table
##
## Response: CrimeRate
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## Percentage  1  93462942 93462942  16.834 9.571e-05 ***
## Residuals  82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 b

- F test
- show the eqivalence
- P-value

3.3.1 t test

由于 t test 是 2.30 的内容，这里仅仅简单的得到结果，不细说其 decision rule

```
m.df=crime.lm$df.residual
T=summary(crime.lm)$coefficients['Percentage','t value']
(t.pvalue=summary(crime.lm)$coefficients['Percentage',4])
```

```
## [1] 9.571396e-05
```

3.3.2 f test

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Decision rule: 如果满足下述条件，我们拒绝原假设

$$F = \frac{MSR}{MSE} > F_{\alpha,1,n-2}$$

```
F=crime.an$`F value`[1]
(f.pvalue=crime.an$`Pr(>F)`[1])
```

```
## [1] 9.571396e-05
```

可以看到 p 值小于 0.01，即 $\beta_1 \neq 0$

3.3.3 相等性

首先可以得到 $T^2 = F$：

```
data.frame(T^2,F)
```

```
##        T.2        F
## 1 16.83376 16.83376
```

其次从 pvalue 来看, 也是相同的

```
data.frame(f.pvalue,t.pvalue)
```

```
##       f.pvalue     t.pvalue
## 1 9.571396e-05 9.571396e-05
```

于是可以得出结论，在 simple linear 的模型下，t test 和 f test 在数值上是等价的。

3.3.4 b 题评论

如果是单边检验（题目的意思应该是双边…但实际是单边更合理），需要修改 p-value，此时仍然相等 (1/2 pvalue)，即：

```
pt(T,df=m.df)#1/2 of the original value
```

```
## [1] 4.785698e-05
```

3.4 c

```
(r2=summary(crime.lm)$r.squared)
```

```
## [1] 0.170324
```

这是一个比较小的值，说明解释性不强

3.5 d

考虑到 $\beta_1 < 0$，于是：

```
(r=-sqrt(r2))
```

```
## [1] -0.4127033
```

4 Problem 4

4.1 a

- State theh full and reduced models

4.1.1 Full model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

4.1.2 Redeced model

$$Y_i = \beta_0 + \epsilon_i$$

4.2 b

obtain

- $SSE(F)$
- $SSE(R)$
- df_F
- df_R
- test statistic $F*$
- decision rule

```
crime.lm2=lm(c_data$CrimeRate~1)
df.F=crime.lm$df.residual
df.R=crime.lm2$df.residual
SSE.F=anova(crime.lm)$'Sum Sq'[2]
SSE.R=anova(crime.lm2)$'Sum Sq'
F.new=(SSE.R-SSE.F)/(df.R-df.F)/(SSE.F/df.F)
```

```
SSE.F
```

```
## [1] 455273165
```

```
SSE.R
```

```
## [1] 548736108
```

```
df.F
```

```
## [1] 82
```

```
df.R
```

```
## [1] 83
```

```
F.new
```

```
## [1] 16.83376
```

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

i.e. H_0 means reduced models can replace full models, while H_1 means it can't.

We reject H_0 if

$$F.new > F(1-\alpha, df_{ER} - df_{EF}, df_{EF}) = F_{\alpha,1,n-2}$$

And we can get the new p-value as:

```
(f2.pvalue=1-pf(F.new,df.R - df.F, df.F))
```

```
## [1] 9.571396e-05
```

4.3 c

可以看到，t-test，f-test 和这里的 general linear test 的 p 值是相同的。即在这个模型中，他们是等价的：

```
data.frame(t.pvalue,f.pvalue,f2.pvalue)
```

```
##       t.pvalue     f.pvalue    f2.pvalue
## 1 9.571396e-05 9.571396e-05 9.571396e-05
```

5 Problem 5

$$E(MSR) = E(b_1^2 S_{xx}) = S_{xx} E(b_1^2)$$

where

$$E(b_1^2) = var(b_1) + E(b_1)^2$$

$$var(b_1) = \frac{\sigma^2}{S_{xx}}, E(b_1) = \beta_1$$

so we can conclude that:

$$E(MSR) = S_{xx}(\frac{\sigma^2}{S_{xx}} + \beta_1^2)$$

$$= \sigma^2 + \beta_1^2 S_{xx}$$

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

6 Problem 6

6.1 a

$$E(MSR) = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

$$E(MSE) = \sigma^2$$

于是：

```r
x=c(1,4,10,11,14)
(E.MSR=0.36+9*sum((x-mean(x))^2))
```

```
## [1] 1026.36
```

```r
(E.MSE=0.36)
```

```
## [1] 0.36
```

6.2 b

- 为了测试是否负相关，用 X=c(6,7,8,9,10) 好吗？
- 如果是为了预测 X=8 的值呢？

如果为了测试负相关与否，需要用到 t-test 或者 f-test

以 t-test 为例，统计量 $T = \frac{b1-0}{s(b1)}$. 其中 $s(b1) = \frac{x}{\sqrt{s_{xx}}}$, S_{xx} 越大，精度越高，效果越好。

所以应该尽可能分散 X 的取值，用 6，7，8，9，10 不好

如果为了预测 X=8 的值，我们关注的是 $\sigma^2(\hat{\mu}_h) = \sigma^2[\frac{1}{n} + \frac{(X_h-\bar{X})^2}{\sum(X_i-\bar{X})^2}]$

由于原先的样本 $\bar{X} = 8$, 新样本也是 $\bar{X} = 8$ 所以两种样本下，$\sigma^2(\hat{\mu}_h)$ 都为：

$$\sigma^2(\hat{\mu}_h) = \sigma^2[\frac{1}{n}]$$

即如果为了预测 X=8 的值，两种选择无好坏之分。

7 后记

这是我第一篇CTeX 文章，第一次用 pandoc 把 r mardown 转为中文的 $CTeX$

CTeX 好像排版不是很好看…以后回去尝试一些新的 ^_^ 加油！