高级统计方法 第3次作业:

序号： 03 姓名： 潘晨楷 学号： 20212241116 班级：软2107

**概念**

1.问题（略）

在表3.4中，“TV”的零假设是在存在radio ads和newspaper ads的情况下，TV ads对sales没有影响。类似地，、对于“Radio”的零假设是，在TV和newspaper ads存在的情况下，Radio ads对Sales没有影响。(对Newspaper的零假设也是一样)TV和Radio的低p值表明，零假设对TV和Radio来说是错误的。Newspaper的高p值表明，零假设对Newspaper来说是正确的。

1. 问题（略）

由题意可得这样的一个公式：

Y = 50 + 20\*GPA + 0.07\*IQ + 35\*Gender + 0.01\*(GPA \* IQ) - 10 \*(GPA \* Gender)

1. 问题（略）

男性：y = 50 + 20\*GPA + 0.07\*IQ + 0.01\*(GPA \* IQ);

女性：y = 50 + 20\*GPA + 0.07\*IQ + 35 + 0.01\*(GPA\*IQ) - 10\*GPA;

比较可得，iii的结论是正确的。

1. 问题（略）

Y = 50 + 20\*4.0 + 0.07\*110 + 35 + 0.01\*(4.0\*110) - 10\*4.0 = 137.1

1. 问题（略）

错误。必须检验回归系数的p值来确定交互项是否具有统计显著性。

1. 问题（略）
2. 问题（略）

一般来说，三次回归的训练RSS会比线性回归的训练RSS小，因为三次回归会对数据进行贴近训练集的拟合。

1. 问题（略）

题目中明确说明该数据的实际模型是线性拟合，所以用三次拟合会产生过拟合，而线性拟合有更好的泛化能力，所以线性回归的测试RSS小。

1. 问题（略）

同（a）问，一般来说，三次回归的训练RSS会比线性回归的训练RSS小，因为三次回归会对数据进行贴近训练集的拟合。

1. 问题（略）

由已知条件不知道实际情况，无法判断。

、、、、、、

**应用**

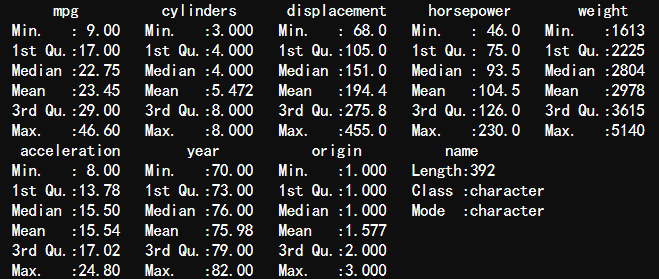
8.问题（略）

（a）问题（略）

> Auto = read.csv("E:\\study\\Auto.csv", header=T, na.strings="?")

> Auto = na.omit(Auto)

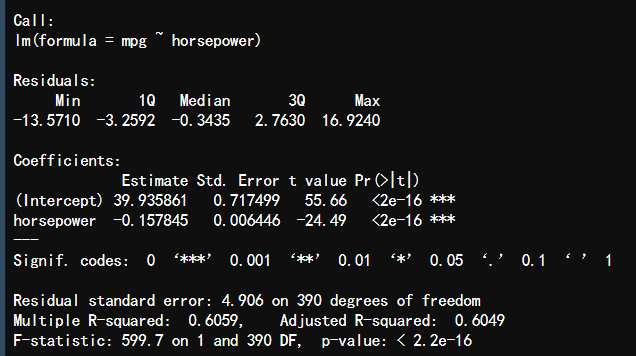
> summary(Auto)



> attach(Auto)

> lm.fit = lm(mpg ~ horsepower)

> summary(lm.fit)



i:horsepower和mpg之间存在关系，这是通过检验所有回归系数等于零的零假设来确定的。由于f统计量远远大于1，并且f统计量的p值接近于零，我们可以拒绝原假设，并说明horsepower和mpg之间存在统计上显著的关系。

ii：mpg的平均值为23.4459。Lm.fit的RSE为4.906，误差百分比为20.9248%。Lm.fit的R2约为0.6059，这意味着60.5948%的MPG方差可以用horsepower来解释。

iii:mpg和horsepower之间的关系是负相关的。汽车的horsepower越大，线性回归表明汽车的mpg越低。

iv：

1. predict(lm.fit,data.frame(horsepower=c(98)),interval="confidence")

1695037781624

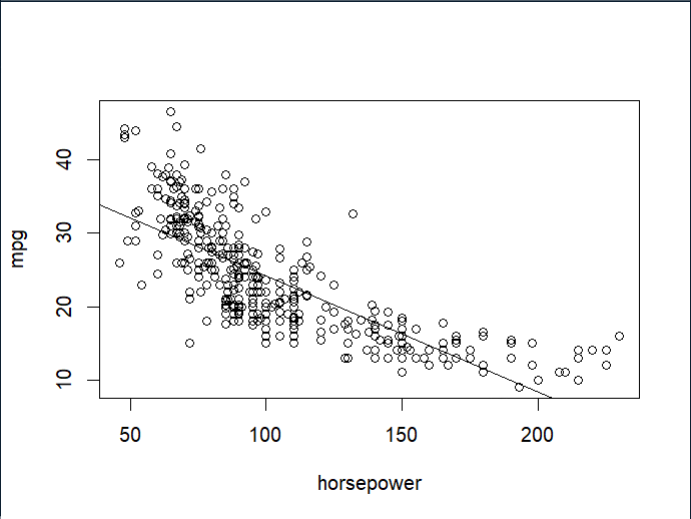
1. predict(lm.fit,data.frame(horsepower=c(98)),interval="prediction")

1695037859131

（b）问题（略）

> plot(horsepower, mpg)

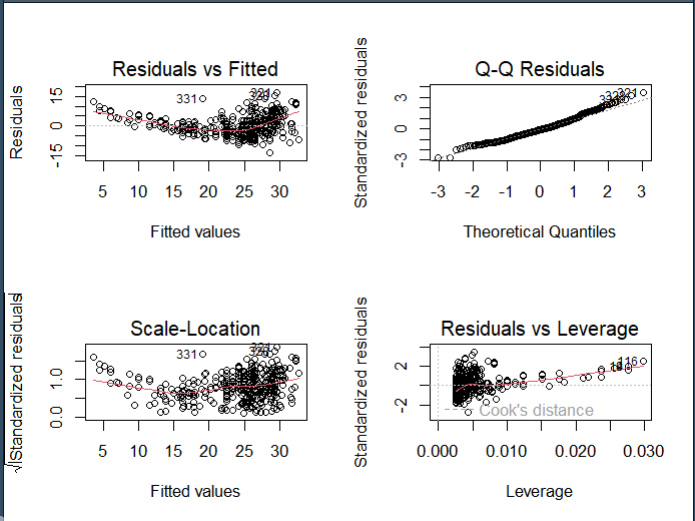
> abline(lm.fit)



（c）问题（略）

> par(mfrow=c(2,2))

> plot(lm.fit)

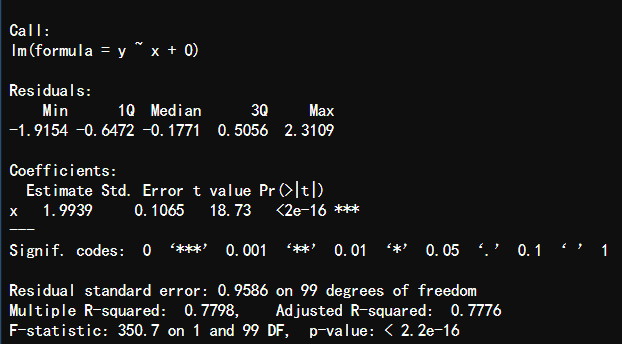


根据残差图，得到了非线性的证据。

1. 问题（略）
2. 问题（略）

> lm.fit = lm(y~x+0)

> summary(lm.fit)

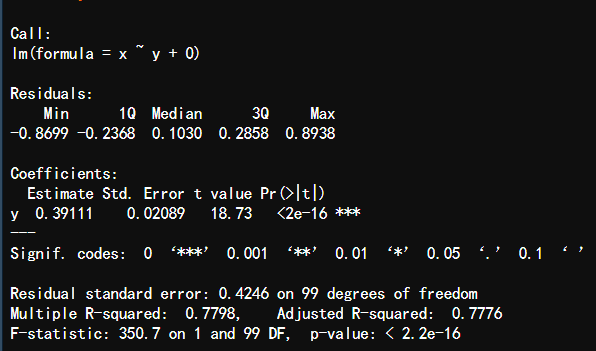


t统计量的p值接近于零，因此零假设被拒绝。

1. 问题（略）

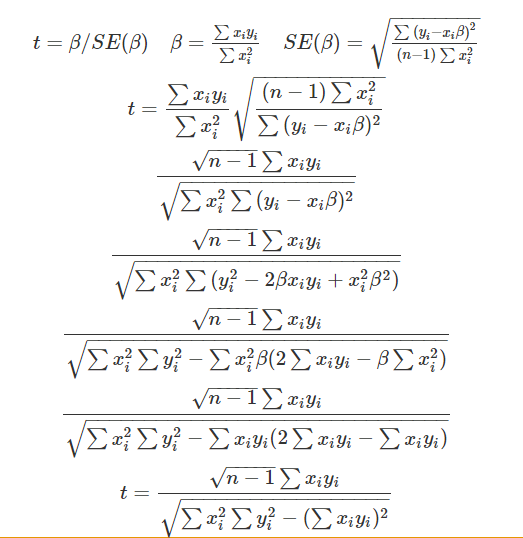
> lm.fit = lm(x~y+0)

> summary(lm.fit)



t统计量的p值接近于零，因此零假设被拒绝。

1. 问题（略）
2. 和(b)中的结果都反映了11（a）中创建的同一行。换句话说，y=2x+ϵ也可以写成x=0.5∗(y-ϵ)。
3. 问题（略）



> (sqrt(length(x)-1) \* sum(x\*y)) / (sqrt(sum(x\*x) \* sum(y\*y) - (sum(x\*y))^2))



1. 问题（略）

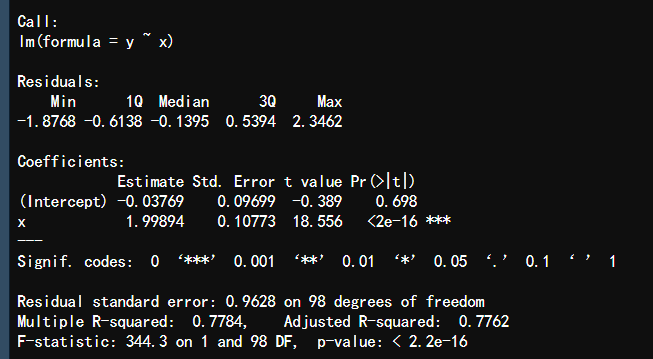
把t(x,y)换成t(y,x)就会得到t(x,y) = t(y,x)

1. 问题（略）

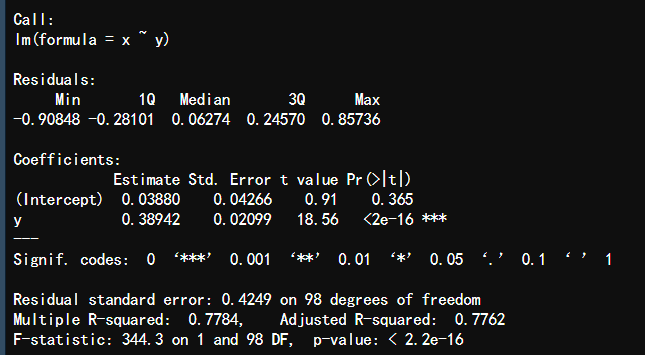
> lm.fit = lm(y~x)

> lm.fit2 = lm(x~y)

> summary(lm.fit)



summary(lm.fit2)



可以看到，t统计量在两个线性回归中是相同的。

13.问题（略）

（a）问题（略）

> set.seed(1)

> x = rnorm(100)

（b）问题（略）

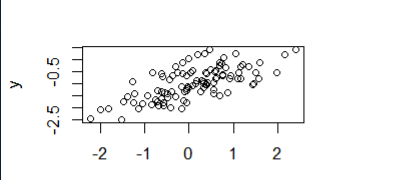
eps = rnorm(100, 0, sqrt(0.25))

（c）问题（略）

向量y的长度为100，β0是-1，β1是0.5.

（d）问题（略）

plot(x, y)

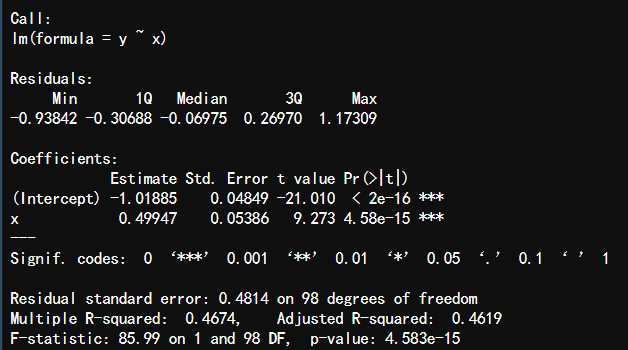


观察到x和y之间的线性关系斜率为正，方差与预期一致。

1. 问题（略）

> lm.fit = lm(y~x)

> summary(lm.fit)



线性回归拟合的模型接近所构建的系数的真实值。该模型的f统计量很大，p值接近于零，因此可以拒绝原假设。

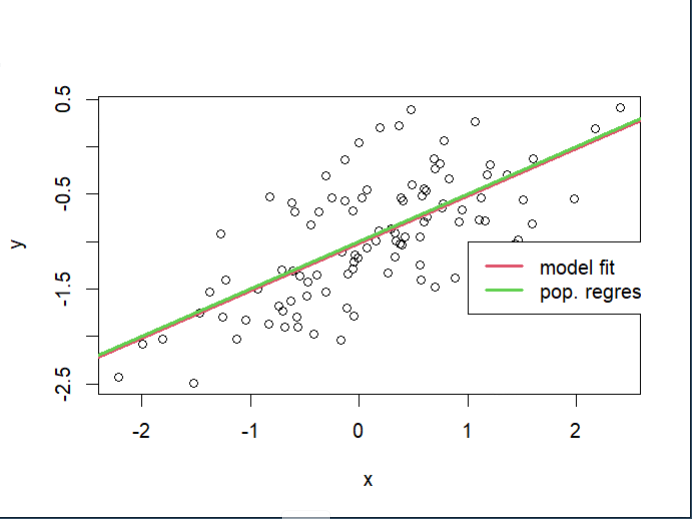
(f)问题（略）

plot(x, y)

abline(lm.fit, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

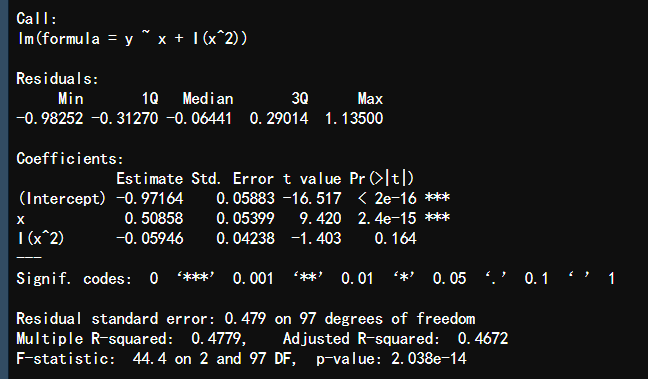
legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



（g）问题（略）

lm.fit\_sq = lm(y~x+I(x^2))

summary(lm.fit\_sq)



由于R2和RSE的轻微增加，模型拟合比训练数据有所增加。但t统计量的p值表明y和x2之间没有关系。

（h）问题（略）

set.seed(1)

eps1 = rnorm(100, 0, 0.125)

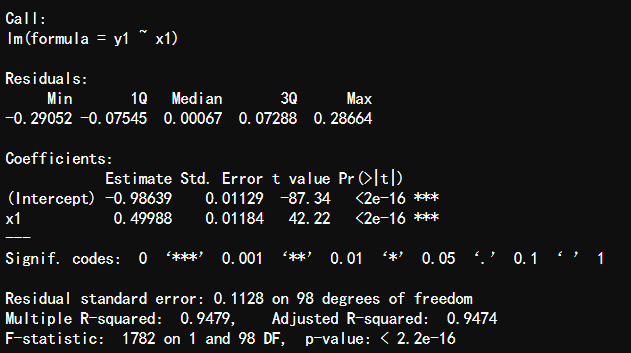
x1 = rnorm(100)

y1 = -1 + 0.5\*x1 + eps1

plot(x1, y1)

lm.fit1 = lm(y1~x1)

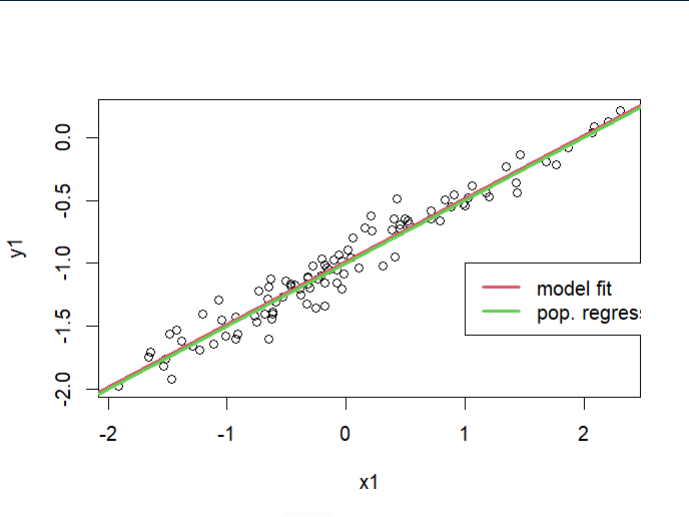
summary(lm.fit1)



abline(lm.fit1, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



和预测的结果相似，在R2和RSE中观察到的误差大大减小。

1. 问题（略）

set.seed(1)

eps2 = rnorm(100, 0, 0.5)

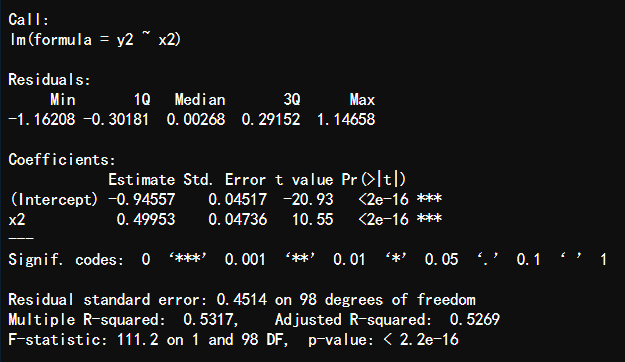
x2 = rnorm(100)

y2 = -1 + 0.5\*x2 + eps2

plot(x2, y2)

lm.fit2 = lm(y2~x2)

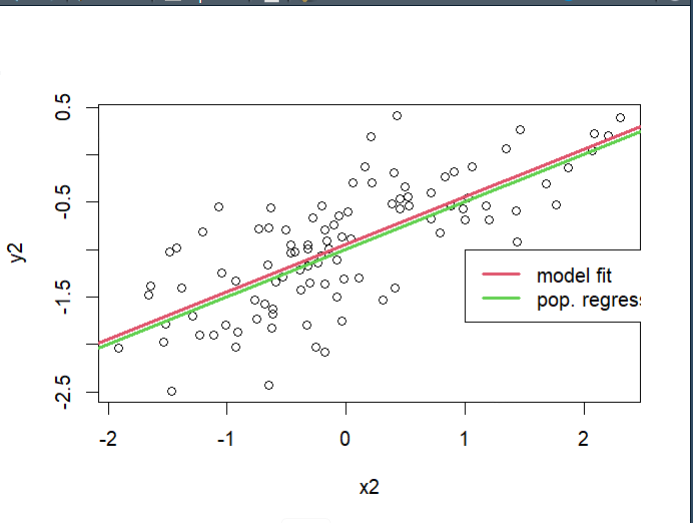
summary(lm.fit2)



abline(lm.fit2, lwd=3, col=2)

abline(-1, 0.5, lwd=3, col=3)

legend(-1, legend = c("model fit", "pop. regression"), col=2:3, lwd=3)



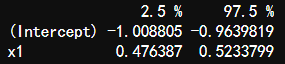
和预测的结果相似，在R2和RSE中观察到的误差大大减小。

（j）问题（略）

confint(lm.fit)



confint(lm.fit1)



confint(lm.fit2)



可以发现所有的区间似乎都以0.5为中心，第二次拟合的区间比第一次拟合的区间

窄，最后一次拟合的区间比第一次拟合的区间宽。