目录

[联系方式： 1](#_Toc97923134)

[第一次作业 1](#_Toc97923135)

[1.1题目1 2](#_Toc97923136)

[1.2 题目1解答 2](#_Toc97923137)

[1.3题目2 8](#_Toc97923138)

[1.4 题目2解答 8](#_Toc97923139)

[第二次作业 14](#_Toc97923140)

[2.1 题目1题干 14](#_Toc97923141)

[2.2 题目1解答 14](#_Toc97923142)

[线性回归与多重共线性检验 14](#_Toc97923143)

[岭回归与变量删减 16](#_Toc97923144)

# 联系方式：

如有问题可及时联系

QQ：1011412062

手机：15321132863；

Tips:因为图有点多，就没做标注

# 第一次作业

统计分析与R软件，P389，6.1，6.3

## 1.1题目1

表格

描述已自动生成

## 1.2 题目1解答

先画出散点图，感觉像是线性关系；

图表, 散点图

描述已自动生成

然后开始线性回归，使用代码如下：

depth <- c(5.1, 3.5, 7.1, 6.2, 8.8, 7.8,4.5,5.6,8.0,6.4)

migration <- c(1907,1287,2700,2373,3260,3000,1947,2273,3113,2493)

data.frame(name=depth,val=migration)

#plot(depth,migration,main="21901200405Lzx")

lm.reg <- lm(migration~depth+1)

print(summary(lm.reg))

结果如下图：

文本

描述已自动生成

发现截距项不显著，而migration关于depth的系数很显著，所以去掉截距项

（方程改为y~x-1）

文本

描述已自动生成

发现R^2 更接近于1了，虽然SSE变大了，但是幅度很小，所以采用无截距项的方案。

进行回归诊断，做出残差图，发现不对劲：

图表, 折线图

描述已自动生成

所以再看一下QQ图：

图表, 散点图

描述已自动生成

那么去掉7,8点（去掉5号点数据就太少了），结果 如下：

文本

描述已自动生成

R^2增加至0.9997, 更为显著了； 残差范围也显著的减小了；

图表, 折线图

描述已自动生成

所以最终的拟合方程为：migration=377.561\*depth；

将上述结果用于预测：

使用语句：

new <- data.frame(depth=7)

print("Predict value is:")

print(predict(lm.reg, new, interval="prediction", level=0.95))

结果如下图所示：

文本

描述已自动生成

预测值为[2518.581,2767,267]

至此，题目一解答完成。

## 1.3题目2



## 1.4 题目2解答

先画出散点图如下：

图表, 散点图

描述已自动生成

可以看出有同一个x值可能对应不同的y值，并且随着x的增大，y的变化幅度越来越大。

所以猜测logY与x成线性；同时还注意到在8~10附近有一个明显大小不对的点（24号），所以也将其去掉；做出如下

文本, 信件

描述已自动生成

可以看出截距项依旧不明显，将之去掉；

使用的代码是：

x <- c(1,1,1,1,2,2,2,3,3,3,4,4,4,5,6,6,6,7,7,7,8,8,8,9,11,12,12,12)

y<-c(0.6,1.6,0.5,1.2,2.0,1.3,2.5,2.2,2.4,1.2,3.5,4.1,5.1,5.7,3.4,9.7,8.6,4.0,5.5,10.5,17.5,13.4,4.5,30.4,12.4,13.4,26.2,7.4)

i<-1:28

x<-x[-24]

y<-y[-24]

plot(x,y,main="21901200405Lzx")

lm.reg=lm(formula = log(y)~x-1)

print(summary(lm.reg))

结果如下图，可以看出R^2显著提升；

文本, 信件

描述已自动生成

将回归曲线画在散点图上，如下图：

图表, 散点图

描述已自动生成

分析Summary的内容：T检验：t值较大（显著>1），Pr值显著<1，认为通过t检验；

F检验：p值很小，认为通过f检验；

(3)标准化残差图如下图：

图表, 散点图

描述已自动生成

普通残差图如下图：（下图和上图其实横纵坐标一样，只不过叫法不一样）

图表, 散点图

描述已自动生成

等方差性可以使用：white检验，代码如下

# load lmtest package

library(lmtest)

# perform White's test on the model

bptest(model, ~ formula, data)

但是从图形上直观地判断是否存在异方差是更为快捷的方法，

我们从残差图：Residuals-Fitted values(上图)上数据点的分布就可以看出是否存在异方差；本模型中的残差图数据点分布较为均匀，且在0附近，故认为其等方差性较好。

所以最终方程为：log(y)=0.25477x

(4)

对数据进行开方后，画出散点图如下：

图表, 散点图

描述已自动生成

图像特征不变，还是随着x的增大，y的变化范围不断增大；

所以对sqrt(y)取对数，并进行线性回归，使用代码如下：

plot(x,sqrt(y),main="21901200405Lzx")

lm.reg=lm(formula = log(sqrt(y))~x-1)

print(summary(lm.reg))

plot(lm.reg)

结果如下：

文本, 信件

描述已自动生成

R^2变化不大，其实log(sqrt(y))=1/2logy，所以说次模型应该与(3)中最后的模型一样；至于最终模型，应该是完全一样，为：log(sqrt(y))=0.1274x;其余图像也完全一样；至此，题目2解答完毕；

# 第二次作业

## 2.1 题目1题干

电脑屏幕的照片上有文字

描述已自动生成

对法国进口数据进行回归分析：判断是否存在多重共线性，如果存在，试着给出你的处理方案。

## 2.2 题目1解答

### 线性回归与多重共线性检验

先输入数据，然后看下X1+X2+X3与Y的散点图，如下图

图表, 散点图

描述已自动生成

好像还挺像线性的，但是这曲线怪怪的，估计是存在多重共线性等问题；具体问题还是看下回归结果吧，如下图

手机屏幕截图

描述已自动生成

上图中可以看出，R^2挺不错的，但是这三系数都不显著(P值不显著)

，好假；

使用下列代码进行VIF检验：

install.packages("car")

library(car)

vif(lm.sol)

结果如下：

文本

描述已自动生成

可以看得出其共线性明显；

再算一下方形矩阵X^T X的条件数

使用如下代码：

#计算条件数

scaleXX = cor(scale(Solution[1:3]) )#求标准化X1,X2,X3的相关阵

kappa(scaleXX,exact=TRUE)

#求相关阵的值和特征向量

eigen(scaleXX)

print(scaleXX)

求出相关矩阵如下：

文本

中度可信度描述已自动生成

可以看出X3和X1的相关性很强。Kappa值为1956，显著>1000

所以，我们可以直接断定，该模型有很强的多重共线性，X1与X3，也就是Dropped和consume有很强的相关性；

### 岭回归与变量删减

确认了多重共线性，首先尝试使用岭回归修改模型

scaleXY = scale(Solution)

x1=scaleXY[,1]

x2=scaleXY[,2]

x3=scaleXY[,3]

y=scaleXY[,4]

library(MASS)

lam<-seq(0,1,by=0.01)

plot(lm.ridge(y~x1+x2+x3,lambda=lam))#画出岭迹图,这里是否有常数项结果一样，因为常数项十分小，可忽略不计

select(lm.ridge(y~x1+x2+x3,lambda=lam))#选出k

先选个k吧，k的三种取值如下图：

文本

描述已自动生成

岭迹图长下图这样

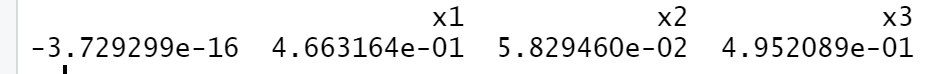
图表

描述已自动生成

所以我们选个k大点的0.41

使用代码进行岭回归：print(lm.ridge(y~x1+x2+x3,lambda=0.41))

打印结果如下图



可以看出x2的数量级最小，将其去掉；

再进行回归，结果如下：

手机屏幕截图

描述已自动生成

还是不显著，尽管岭迹图（下图）很平缓了

图表

描述已自动生成

其是可以看出X1和X3相关性很强，直觉上直接删除一个就行，不过还是走个流程吧，使用step()之类的来删除变量；

使用如下代码进行逐步回归：

##逐步回归

lm.step<-step(lm.sol)#逐步回归

lm.step$anova#可查看逐步回归的SSE和AIC的变化情况

print(summary(lm.step))#显示逐步回归的结果

结果如下：

手机屏幕截图

描述已自动生成

也就是去掉了X1

Summary如下图

文本, 信件

描述已自动生成

X2还是不显著；

不过先看一下VIF吧，结果如下图：

文本

低可信度描述已自动生成

看起来还行。

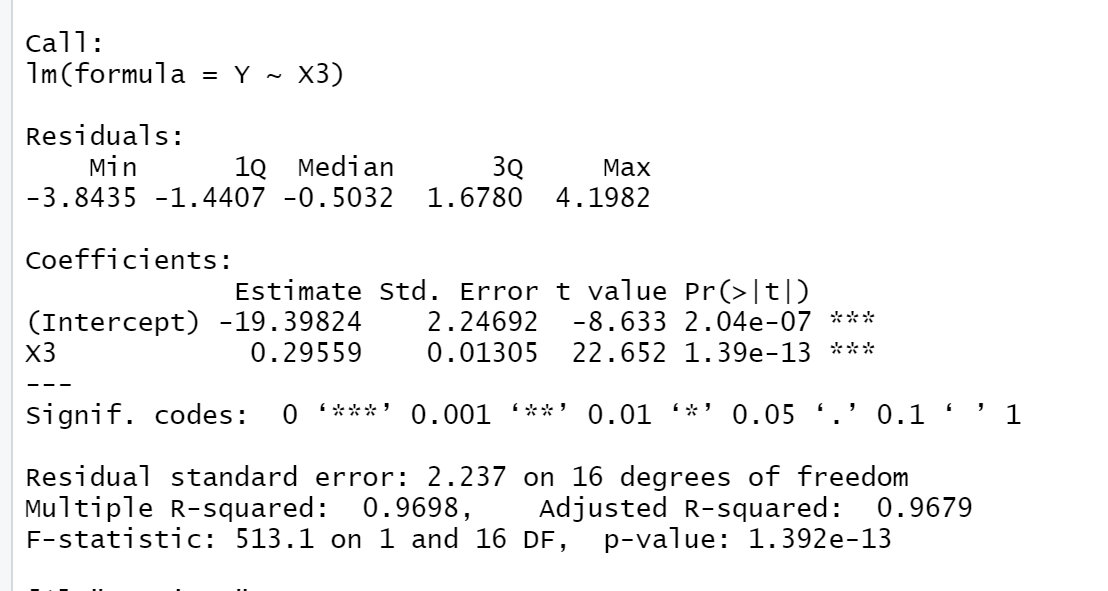
那么就剩X2的问题了，画一下X2与Y的散点图看看，如下图

日历

描述已自动生成

好像确实关系没那么大，起码权重蛮小的。

那么直接删了X2试试：结果如下图；



R^2下降了一点，所以为了看起来比较科学，实际应用中我或许会将最终模型写为：

Y=0.292\*X3+0.417\*X2-20.3；又或者写作：IMPORT=0.292\*CONSUME+0.417\*STOCK-20.3；

至此，作业一，作业二解答完毕；