一、某地29名13岁儿童身高（cm），体重（kg）和肺活量（L）数据见“homework-6.1-data.txt”，求：

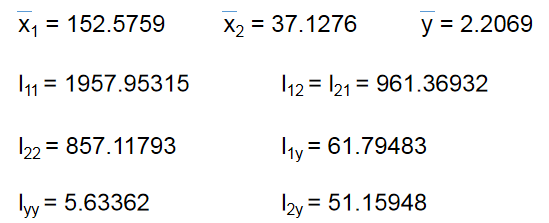
（1）由身高，体重推算肺活量的回归方程；

（2）求出的方程是否有意义；

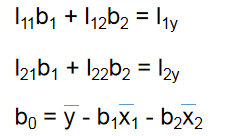
（3）剩余标准差

参考答案：

（1）设要求的回归方程为：y = b0 + b1x1 + b2x2 则由表格数据可计算：



由

可得b1 = 0.005017，b2 = 0.054061，b0 = -0.5657

结论：回归方程为y = -0.5657 + 0.0050x1 + 0.0541x2

1. 检验方程是否有意义，需做F检验

H0：β1=0，β2=0

H1：β1，β2不都等于0



|  |  |  |  |
| --- | --- | --- | --- |
|  | 自由度 | 平方和 | 均方 |
| U | 2 | 3.0758 | 1.5379 |
| Q | 26 | 2.5578 | 0.0984 |
| Lyy | 28 | 5.6336 |  |

计算F=15.6327，查表，F0.01（2.26）=5.53，F>F0.01（2.26），所以P<0.01，因此拒绝H0，认为方程有意义。

1. 剩余标准差：

结论：剩余标准差是0.3137.

R code：

>a<-read.table("./data.txt",header = T,sep = "\t")

>ff<-lm(a[,3]~a[,1]+a[,2])

>summary(ff)

Version:1.0 StartHTML:0000000107 EndHTML:0000001535 StartFragment:0000000127 EndFragment:0000001517

Call:

lm(formula = a[, 3] ~ a[, 1] + a[, 2])

Residuals:

Min 1Q Median 3Q Max

-0.54117 -0.25524 -0.00266 0.22039 0.55425

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.565664 1.240127 -0.456 0.65208

a[, 1] 0.005017 0.010575 0.474 0.63920

a[, 2] 0.054061 0.015984 3.382 0.00228 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3137 on 26 degrees of freedom

Multiple R-squared: 0.546, Adjusted R-squared: 0.511

F-statistic: 15.63 on 2 and 26 DF, p-value: 3.485e-05

所以

（1）回归方程为：y=-0.565664+0.005017x1+0.054061x2

（2）因为F=15.63，p=3.485e-05<0.01，所以方程有意义。

（3）剩余标准差为0.3137.

二、某农场通过试验取得早稻收获量与春季降雨量和春季温度的数据如下：

|  |  |  |
| --- | --- | --- |
| 收获量y(kg/mm2) | 降雨量x1(mm) | 温度x2(℃) |
| 2250 | 25 | 6 |
| 3450 | 33 | 8 |
| 4500 | 45 | 10 |
| 6750 | 105 | 13 |
| 7200 | 110 | 14 |
| 7500 | 115 | 16 |
| 8250 | 120 | 17 |

数据见“homework-6.2-data.csv”

建立早稻收获量对春季降雨量和春季温度的二元线性回归方程，计算各回归系数的置信区间，并对回归模型的线性关系和回归系数进行检验（α=0.05）。

参考答案：

1、多元线性回归的汇总输出：

> a <- read.table('homework-6.2-data.csv',sep = ',',header = TRUE)

> fit<-lm(y~x1+x2,data=a)

> summary(fit)

Call:

lm(formula = y ~ x1 + x2, data = a)

Residuals:

1 2 3 4 5 6

-275.101 90.464 216.483 140.280 150.676 -316.599

7

-6.203

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.591 505.004 -0.001 0.9991

x1 22.387 9.601 2.332 0.0801 .

x2 327.672 98.798 3.317 0.0295 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 261.4 on 4 degrees of freedom

Multiple R-squared: 0.9913, Adjusted R-squared: 0.987

F-statistic: 228.4 on 2 and 4 DF, p-value: 7.532e-05

根据R输出结果，得到的多元线性回归方程为：

y= -0.591+22.387x1+327.672x2

2、各回归系数的置信区间：

> confint(fit,level=0.95)

2.5 % 97.5 %

(Intercept) -1402.707516 1401.52552

x1 -4.268921 49.04184

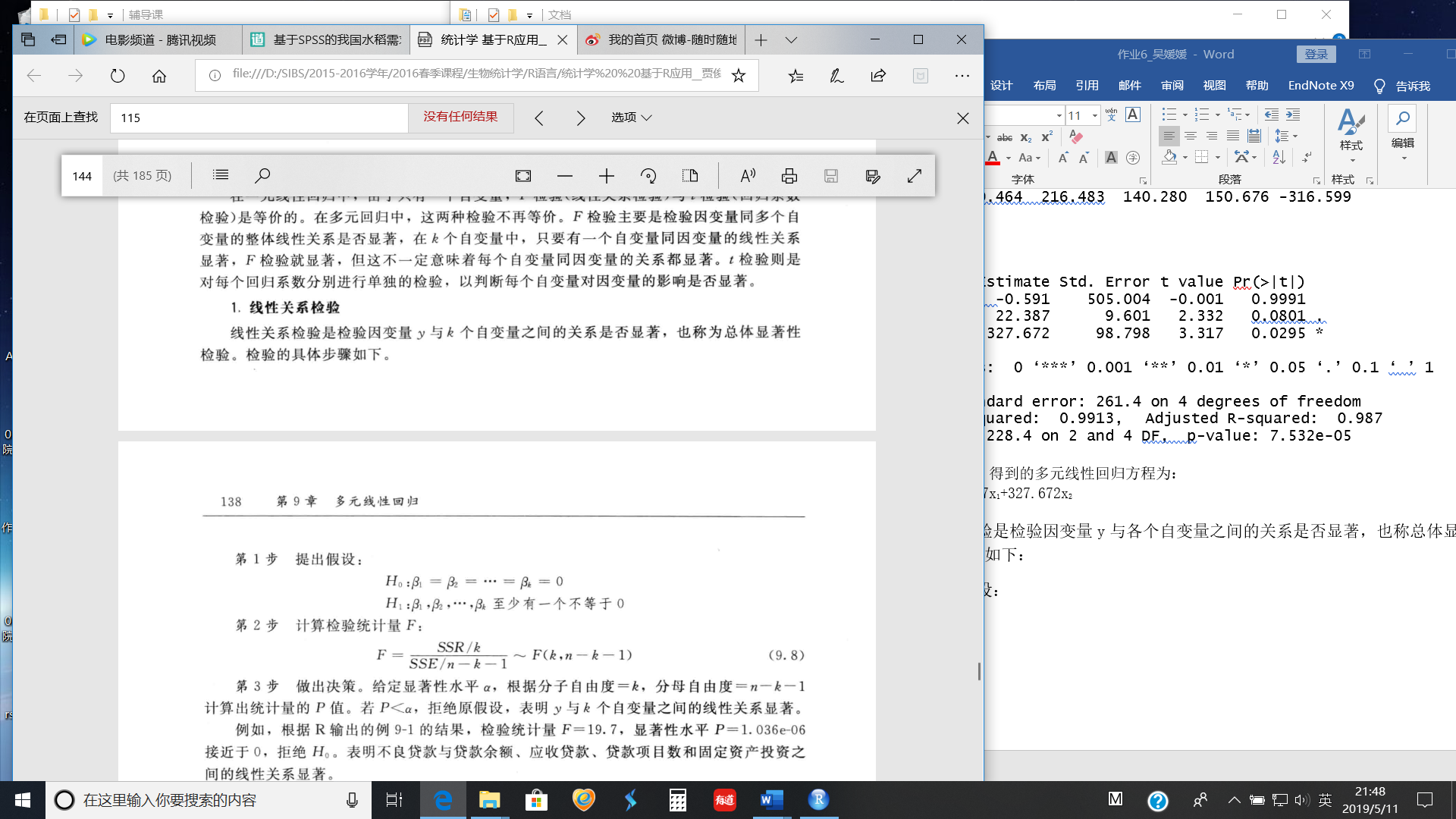
x2 53.364699 601.97873

降雨量x1的置信区间为（-4.268921，49.04184），含义是在温度不变的条件下，降雨量每变动1mm，收获量的平均变动在-4.268921到49.04184 kg/mm2之间。

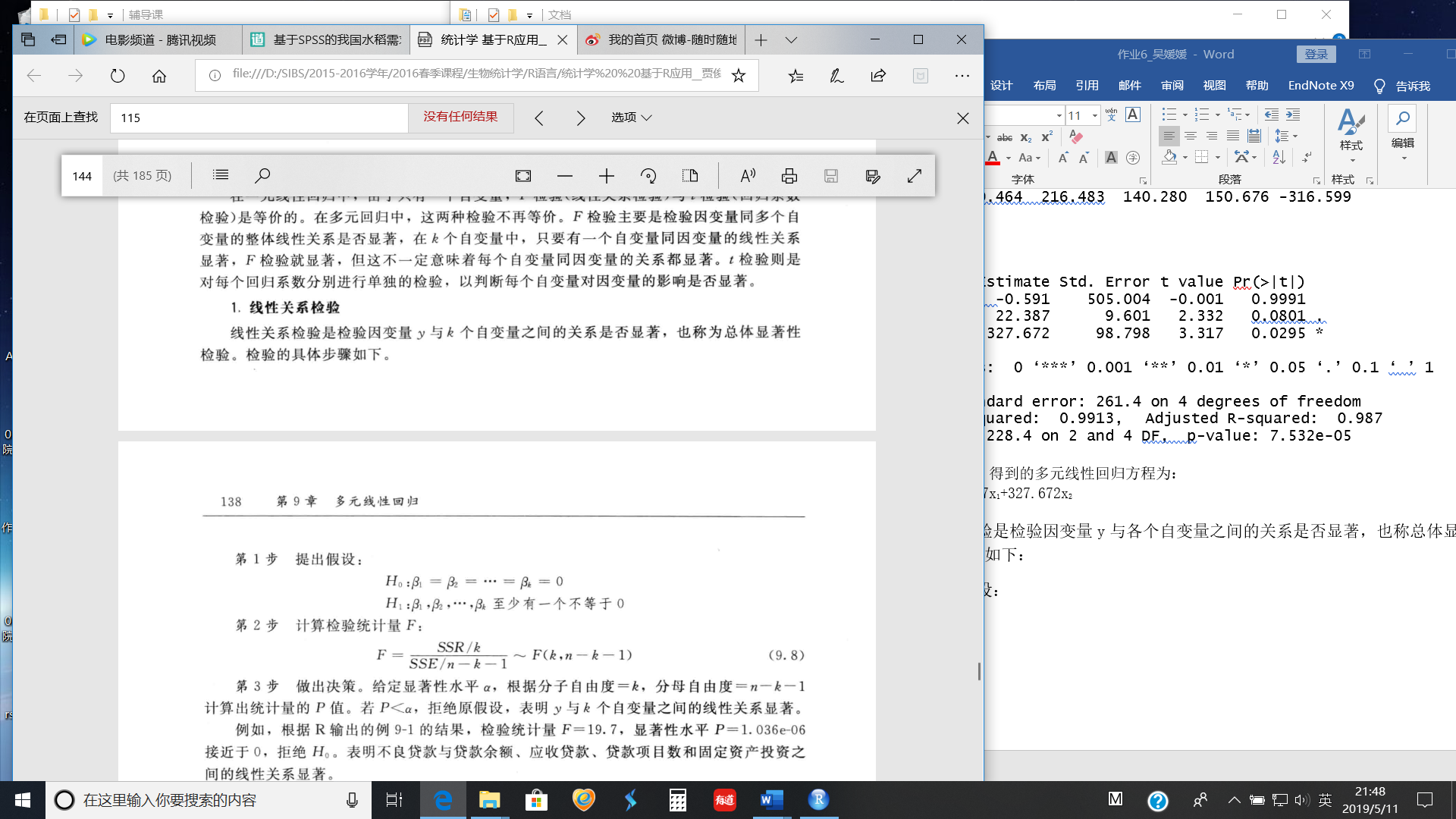
温度x2的置信区间为（53.364699，601.97873），含义是在降雨量不变的条件下，温度每变动1℃，收获量的平均变动在53.364699到601.97873 kg/mm2之间。

3、线性关系检验是检验因变量y与k个自变量之间的关系是否显著，也称总体显著性检验。具体步骤如下：

第1步 提出假设：



第2步 计算检验统计量F：

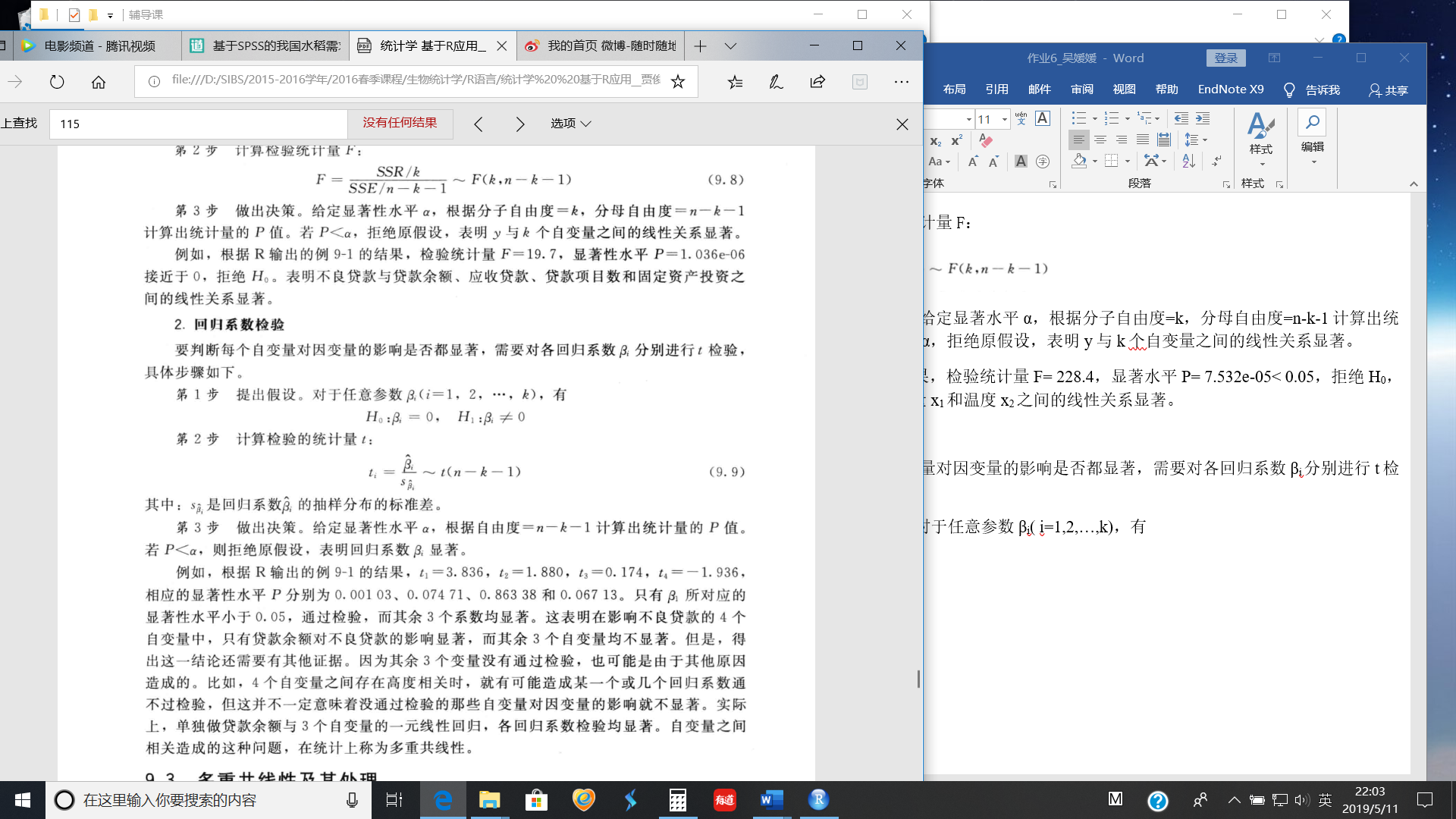


第3步 做出决策。给定显著水平α，根据分子自由度=k，分母自由度=n-k-1计算出统计量的P值。若P< α，拒绝原假设，表明y与k个自变量之间的线性关系显著。

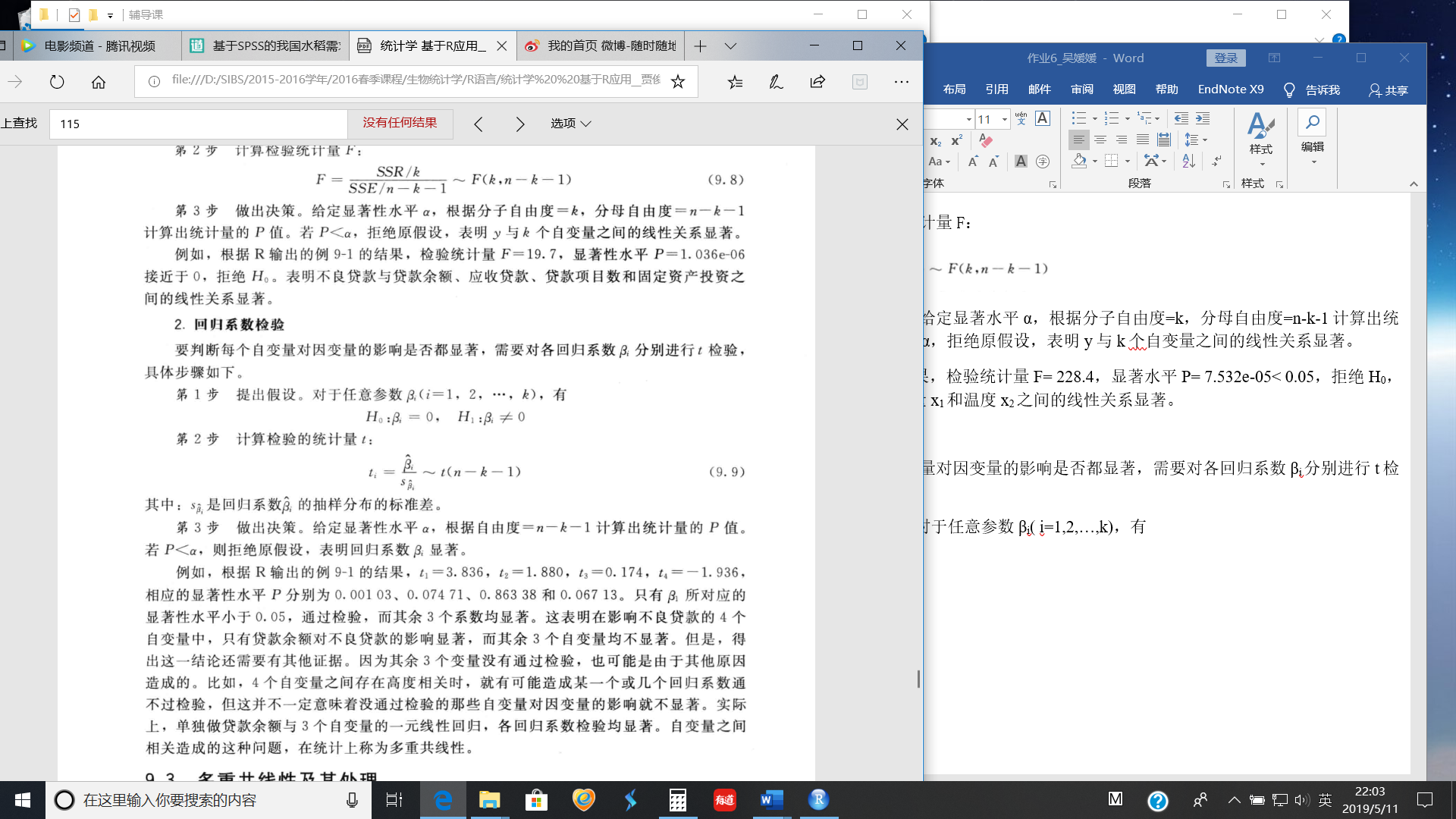
根据以上R输出结果，检验统计量F= 228.4，显著水平P= 7.532e-05< 0.05，拒绝H0，即收获量y与降雨量x1和温度x2之间的线性关系显著。

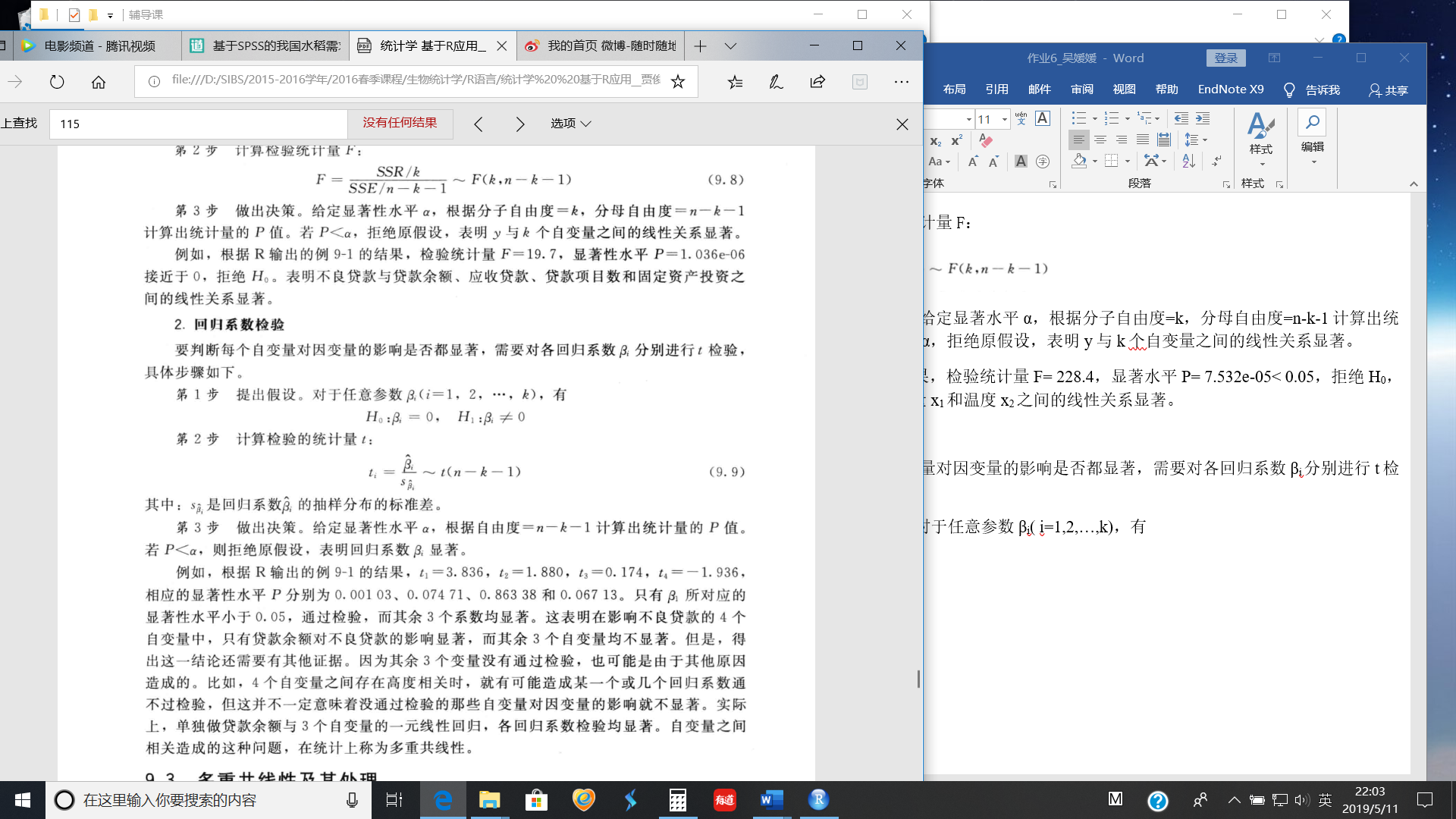
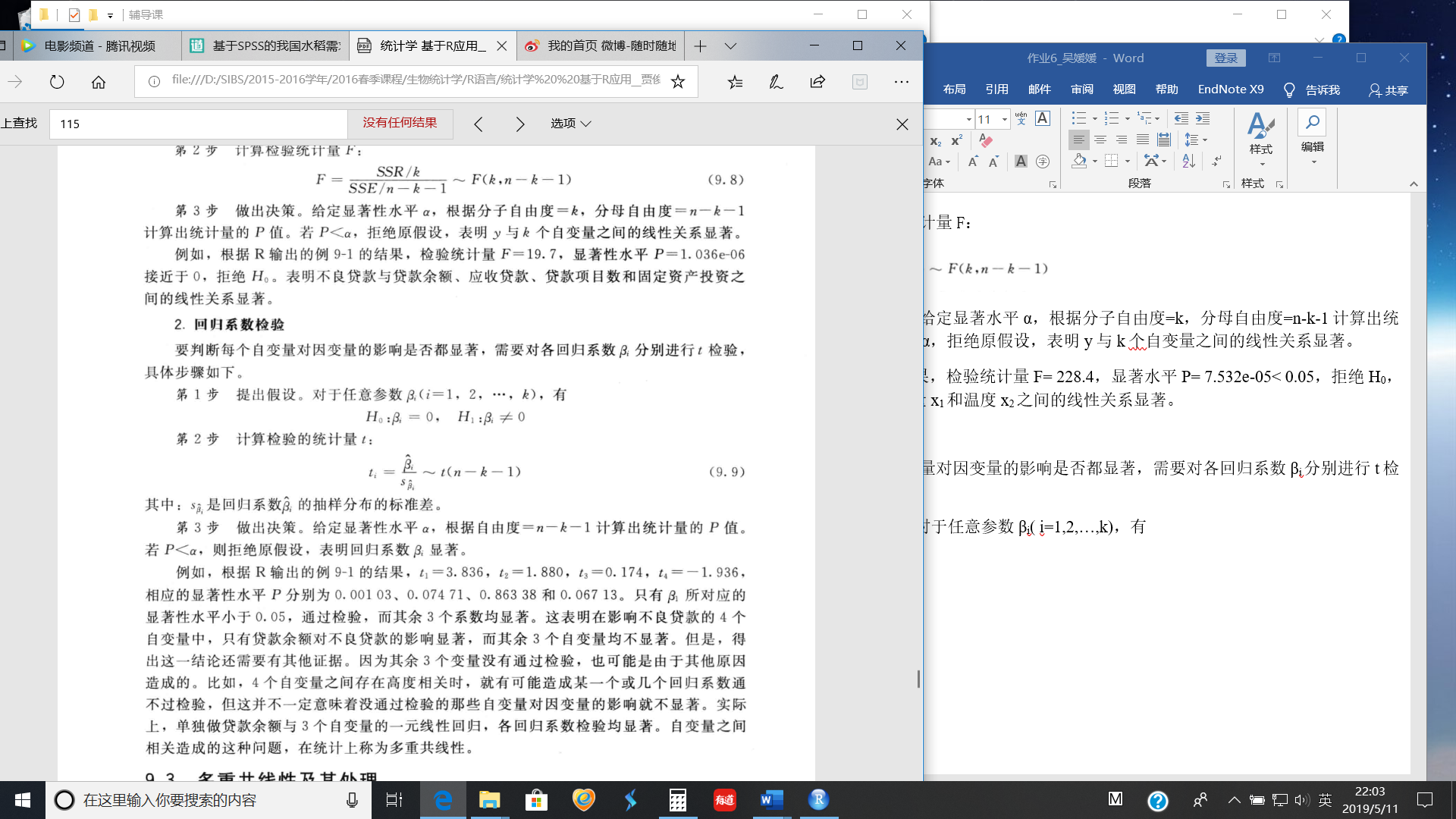
4、要判断每个自变量对因变量的影响是否都显著，需要对各回归系数βi分别进行t检验，具体步骤如下：

第1步 提出假设。对于任意参数βi( i=1,2,…,k)，有



第2步 计算检验的统计量t：



其中：是回归系数的抽样分布的标准差。

第3步 做出决策。给定显著性水平α，根据自由度= n-k-1计算出统计量的P值。若P< α，则拒绝原假设，表明回归系数βi显著。

根据R输出结果，降雨量x1和温度x2的回归系数相应的显著水平分别为0.0801和0.0295，只有温度对应的显著性水平小于0.05通过检验，这表明影响收获量的自变量中，只有温度对收获量的影响显著，而降雨量对收获量的影响不显著。

三、某葡萄酒爱好者想探索葡萄酒的品质与哪些因素相关。他有一个数据集包含了（1 -固定酸度，2 -挥发性酸度，3 -柠檬酸，4 -残余糖，5 -氯化物，6 -自由二氧化硫量，7 -二氧化硫总量，8 -密度，9 - pH值，10 -硫酸盐，11 -酒精浓度，和12 -品质(0 - 10分)。数据见“homework-6.3-winequality-red.csv”.

1. 查看数据集的前五行和数据集的总结
2. 通过直方图展示固定酸度的分布和展示挥发性酸度与品质的散点图
3. 计算这些变量与品质的相关性
4. 通过方差分析不同品质的葡萄酒的酒精浓度是否有差异
5. 通过多元线性回归建立一个品质预测模型，并说明哪些变量与品质显著相关。

Answer

> wine\_data <- read.csv("../homework-6.3-winequality-red.csv", sep=",",header=T)

> #1 answer

> head(wine\_data,5)

fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide

1 7.4 0.70 0.00 1.9 0.076 11

2 7.8 0.88 0.00 2.6 0.098 25

3 7.8 0.76 0.04 2.3 0.092 15

4 11.2 0.28 0.56 1.9 0.075 17

5 7.4 0.70 0.00 1.9 0.076 11

total.sulfur.dioxide density pH sulphates alcohol quality

1 34 0.9978 3.51 0.56 9.4 5

2 67 0.9968 3.20 0.68 9.8 5

3 54 0.9970 3.26 0.65 9.8 5

4 60 0.9980 3.16 0.58 9.8 6

5 34 0.9978 3.51 0.56 9.4 5

> summary(wine\_data)

fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide

Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900 Min. :0.01200 Min. : 1.00

1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900 1st Qu.:0.07000 1st Qu.: 7.00

Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200 Median :0.07900 Median :14.00

Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539 Mean :0.08747 Mean :15.87

3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600 3rd Qu.:0.09000 3rd Qu.:21.00

Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500 Max. :0.61100 Max. :72.00

total.sulfur.dioxide density pH sulphates alcohol quality

Min. : 6.00 Min. :0.9901 Min. :2.740 Min. :0.3300 Min. : 8.40 Min. :3.000

1st Qu.: 22.00 1st Qu.:0.9956 1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50 1st Qu.:5.000

Median : 38.00 Median :0.9968 Median :3.310 Median :0.6200 Median :10.20 Median :6.000

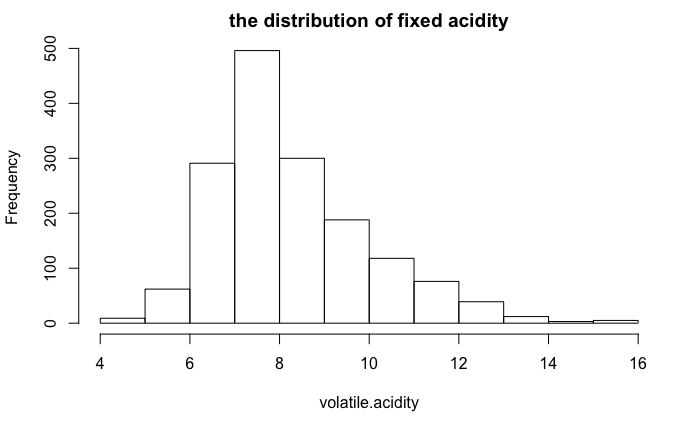
Mean : 46.47 Mean :0.9967 Mean :3.311 Mean :0.6581 Mean :10.42 Mean :5.636

3rd Qu.: 62.00 3rd Qu.:0.9978 3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10 3rd Qu.:6.000

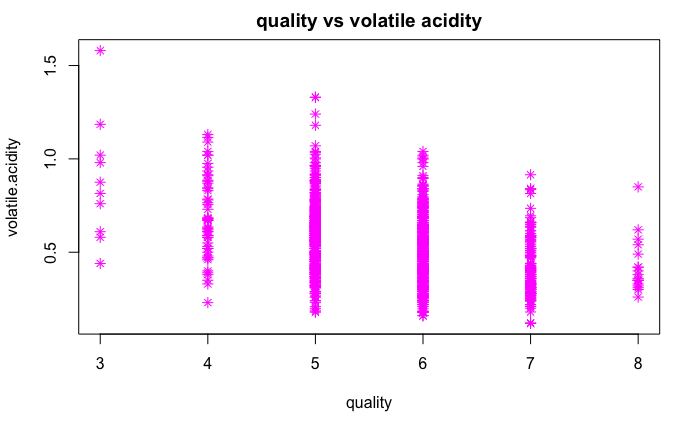
Max. :289.00 Max. :1.0037 Max. :4.010 Max. :2.0000 Max. :14.90 Max. :8.000

> #2 answer

> hist(wine\_data$fixed.acidity,main="the distribution of fixed acidity",xlab="volatile.acidity")



> plot(wine\_data$quality,wine\_data$volatile.acidity,main="quality vs volatile acidity",xlab="quality",ylab="volatile.acidity",col=6,pch=8)



> #3 answer

> apply(wine\_data,2,function(x)cor(x,wine\_data$quality))

fixed.acidity volatile.acidity citric.acid residual.sugar chlorides

0.12405165 -0.39055778 0.22637251 0.01373164 -0.12890656

free.sulfur.dioxide total.sulfur.dioxide density pH sulphates

-0.05065606 -0.18510029 -0.17491923 -0.05773139 0.25139708

alcohol quality

0.47616632 1.00000000

> #4 answer

> shapiro.test(wine\_data$alcohol)

> shapiro.test(wine\_data$quality)

> bartlett.test(wine\_data$alcohol~wine\_data$quality)

> kruskal.test(wine\_data$alcohol~wine\_data$quality)

Kruskal-Wallis rank sum test

data: wine\_data$alcohol by wine\_data$quality

Kruskal-Wallis chi-squared = 412.38, df = 5,

p-value < 2.2e-16

> #5 answer

> linear\_quality = lm(quality ~ fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide+density+pH+sulphates+alcohol, data=wine\_data)

> summary(linear\_quality)

Call:

lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +

residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +

density + pH + sulphates + alcohol, data = wine\_data)

Residuals:

Min 1Q Median 3Q Max

-2.68911 -0.36652 -0.04699 0.45202 2.02498

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.197e+01 2.119e+01 1.036 0.3002

fixed.acidity 2.499e-02 2.595e-02 0.963 0.3357

volatile.acidity -1.084e+00 1.211e-01 -8.948 < 2e-16 \*\*\*

citric.acid -1.826e-01 1.472e-01 -1.240 0.2150

residual.sugar 1.633e-02 1.500e-02 1.089 0.2765

chlorides -1.874e+00 4.193e-01 -4.470 8.37e-06 \*\*\*

free.sulfur.dioxide 4.361e-03 2.171e-03 2.009 0.0447 \*

total.sulfur.dioxide -3.265e-03 7.287e-04 -4.480 8.00e-06 \*\*\*

density -1.788e+01 2.163e+01 -0.827 0.4086

pH -4.137e-01 1.916e-01 -2.159 0.0310 \*

sulphates 9.163e-01 1.143e-01 8.014 2.13e-15 \*\*\*

alcohol 2.762e-01 2.648e-02 10.429 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.648 on 1587 degrees of freedom

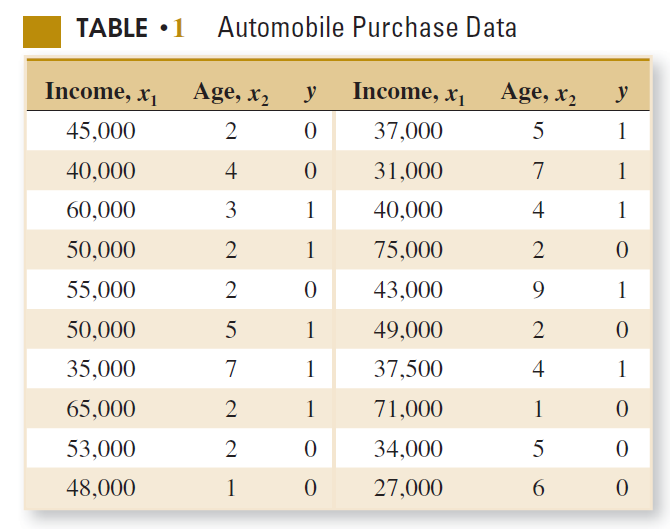
Multiple R-squared: 0.3606, Adjusted R-squared: 0.3561

F-statistic: 81.35 on 11 and 1587 DF, p-value: < 2.2e-16

与红酒品质显著相关的变量有：2 -挥发性酸度，5 -氯化物，6 -自由二氧化硫量，7 -二氧化硫总量，9 - pH值，10 -硫酸盐，11 -酒精浓度。

四、A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and

their total family income. A follow-up survey was conducted six months later to determine if they had actually purchased a new vehicle during that time period (*y* = 1 indicates *yes* and *y* = 0 indicates *no*). The data from this study are shown in the Table1.(“homework-6.4-data.txt”)



(a) Fit a logistic regression model to the data.(40’)

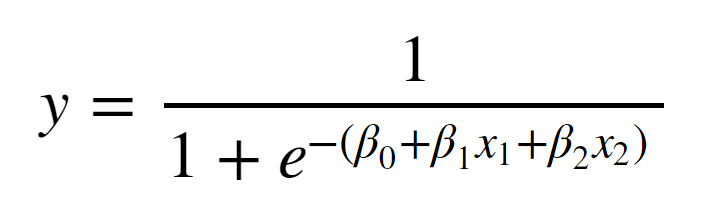
(b) Interpret the model coefficients β1 and β2 and write the logistic regression model formula.(20’)

(c) What is the estimated probability that a family with an income of $45,000 and a car that is five years old will purchase a new vehicle in the next six months?(40’)

**Solution**

**(a)**

Suppose that the model has the form



> data <-read.table("homework-6.4-data.txt",sep=",",header = TRUE) 10’

> fit <-glm(y ~ Income + Age,data=data,family = binomial()) 20’

> summary(fit) 10’

Call:

glm(formula = y ~ Income + Age, family = binomial(), data = data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5635 -0.8045 -0.1397 0.9535 1.7915

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -7.047e+00 4.674e+00 -1.508 0.132

Income 7.382e-05 6.371e-05 1.159 0.247

Age 9.879e-01 5.274e-01 1.873 0.061 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27.726 on 19 degrees of freedom

Residual deviance: 21.082 on 17 degrees of freedom

AIC: 27.082

Number of Fisher Scoring iterations: 5

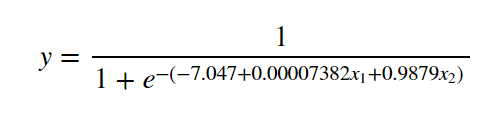
**(b)**

> coef(fit) 10’

(Intercept) Income Age

-7.047061e+00 7.381679e-05 9.878861e-01

β0=−7.047,β1=0.00007382,β2=0.9879

 10’

**(c)**

> predictdata <-data.frame(Income=c(45000),Age=c(5)) 20’

> predictdata

Income Age

1 45000 5

> predictdata$prob <-predict(fit,newdata = predictdata,type="response") 20’

> predictdata

Income Age prob

1 45000 5 0.7710279

So,the estimated probablitity is 0.771

五、数据文件“homework-6.5-Drivers.csv”为对45名司机的调查结果，其中四个变量的含义为：

1）x1：表示视力状况，它是一个分类变量，1表示好，0表示有问题；

2）x2：年龄，数值型；

3）x3：驾车教育，它也是一个分类变量，1表示参加过驾车教育，0表示没有；

4）y：一个分类型输出变量，表示去年是否出过事故，1表示出过事故，0表示没有；

问题：

1. 请在R语言中调用logistic回归函数，计算视力状况、年龄、驾车教育与是否发生事故的logistic回归模型，并以“odds=……”的形式写出回归公式。（10分）
2. 指出（1）得到的模型中哪些因素对是否发生事故有显著性影响。如果存在对是否发生事故没有显著性影响的因素，请去除这些因素后重新计算logistic回归模型，并以“p=……”的形式写出回归公式。（20分）
3. A是一名参加过驾车教育，但视力有问题的50岁老司机；B是一名没有参加过驾车教育，但视力良好的20岁新手。现在A、B都想在某保险公司投保，但按公司规定，被保险人必须满足“明年出事故的概率不高于40%”的条件才能予以承保。请预测A、B两者明年出事故的概率，并告诉保险公司谁可以投保。（20分）

参考答案：

（1）① R语言代码：

data<-read.csv("homework-6.5-Drivers.csv ",header = T,sep = ",")

fit.full<-glm(y~x1+x2+x3, data=data, family=binomial)

summary(fit.full) （**代码正确：5分**）【代码运行结果请见附录】

② 回归公式为：（**公式正确：5分**）

（2）① 题目（1）得到的模型中x1（或者回答“视力状况”）对是否发生事故有显著性影响。（**结论正确：5分**）

② R语言代码：

fit.reduced<-glm(y~x1, data=data, family=binomial)

summary(fit.reduced) （**代码正确：5分**）【代码运行结果请见附录】

③ 回归公式为：（**公式正确：10分**）

（3）① R语言代码：

testdata<- data.frame(x1=c(0,1))

testdata$prob<-predict(fit.reduced,testdata ,type='response')

testdata （**代码正确：10分**）【代码运行结果请见附录】

② 所以A、B两者明年出事故的概率分别为0.65和0.32（**数值正确：5分**）。

③ 因只有B明年出事故的概率不高于40%，故只有B可以投保。（**结论正确：5分**）

注：用（1）的模型计算不得分，但若能够正确调用predict()的可得5分。

附录：

第（1）题代码运行结果----------------------------------------------------------------------------------------

Call:

glm(formula = y ~ x1 + x2 + x3, family = binomial, data = data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5636 -0.9131 -0.7892 0.9637 1.6000

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.597610 0.894831 0.668 0.5042

x1 -1.496084 0.704861 -2.123 0.0338 \*

x2 -0.001595 0.016758 -0.095 0.9242

x3 0.315865 0.701093 0.451 0.6523

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.183 on 44 degrees of freedom

Residual deviance: 57.026 on 41 degrees of freedom

AIC: 65.026

Number of Fisher Scoring iterations: 4

----------------------------------------------------------------------------------------------------------------------

第（2）题代码运行结果----------------------------------------------------------------------------------------

Call:

glm(formula = y ~ x1, family = binomial, data = data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4490 -0.8782 -0.8782 0.9282 1.5096

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.6190 0.4688 1.320 0.1867

x1 -1.3728 0.6353 -2.161 0.0307 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.183 on 44 degrees of freedom

Residual deviance: 57.241 on 43 degrees of freedom

AIC: 61.241

Number of Fisher Scoring iterations: 4

----------------------------------------------------------------------------------------------------------------------

第（2）题代码运行结果----------------------------------------------------------------------------------------

x1 prob

1 0 0.65

2 1 0.32

----------------------------------------------------------------------------------------------------------------------

六、Many digitized image of a fine needle aspirate (FNA) of a breast mass are collected and computed to predict the diagnosis of breast cancer(“homework-6.6-data.csv”).

Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the threelargest values) of these features were computed for each image,resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.

In total, there are 357 benign and 212 malignant samples.

You may need to use proper regression algorithm to train your data, and make predictions.

Instructions:

1. Use all mean features(such as: radius\_mean,texture\_mean…) to construct a logistic regression model
2. Then try to reduce the number of features from your last model, construct another regression model, and you will need to write down the equation of your logistic regression model(Tips: Logit P = α+β1X1+β2X2+..+βpXp)
3. Use proper test to test the difference between two models
4. You may split the data properly, use part of them to train your regression model and use another part to make predictions. Lastly, you may try to calculate the accuracy of your model.(Tips: To split the data, you can use the first 398 rows as training data, use the last 171 rows as prediction data.The predict function return a value between 0 and 1, 0.~0.5 belong to the first class, and 0.5~1 belong to second class in binary classification problems)

Answer:

1)

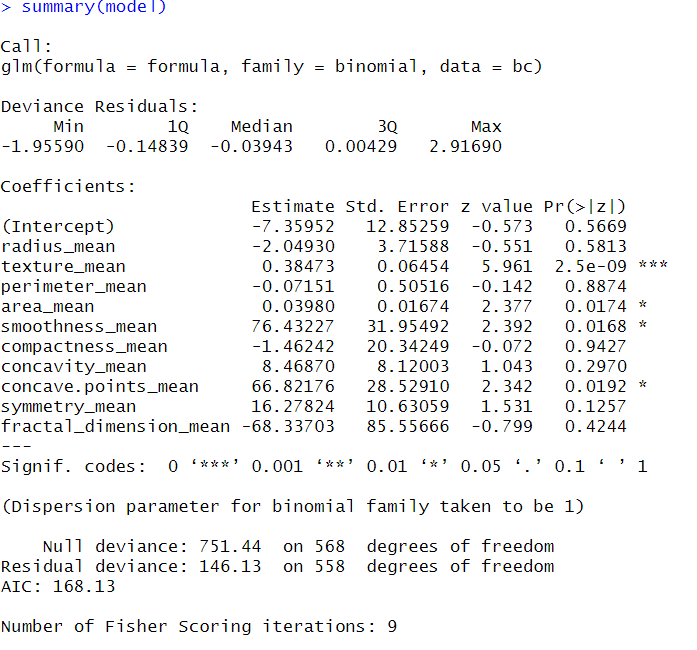
R script

bc<-read.csv("data.csv")

formula<-paste("diagnosis",paste(colnames(bc)[3:12],collapse="+"),sep="~")

model<-glm(formula,bc,family = binomial)

summary(model)



2)

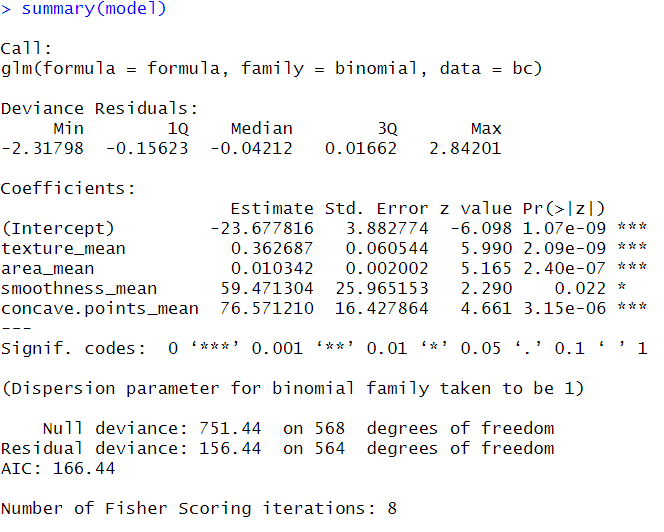
R script

reduced\_feature<-c('texture\_mean','area\_mean','smoothness\_mean','concave.points\_mean')

formula<-paste("diagnosis",paste(reduced\_feature,collapse="+"),sep="~")

reduced.model<-glm(formula,bc,family = binomial)

summary(model)

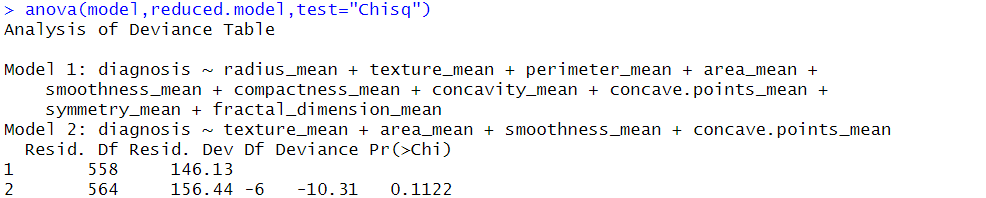


Log(diagnosis)=-23.6778+0.3626\*texture\_mean+0.0103\*area\_mean+59.4713\*smoothness\_mean+76.5712\*concave.points\_mean

3)

R script

anova(model,reduced.model,test="Chisq")



4)

R script

trainData<-bc[1:398,]

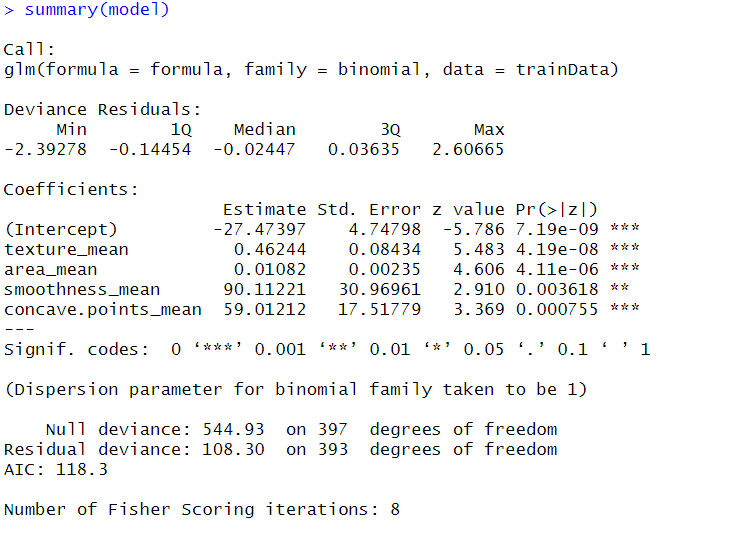
testData<-bc[399:569,]

reduced\_feature<-c('texture\_mean','area\_mean','smoothness\_mean','concave.points\_mean')

formula<-paste("diagnosis",paste(reduced\_feature,collapse="+"),sep="~")

model<-glm(formula,trainData,family = binomial)

summary(model)



pred<-predict(model,testData,type="response")

pred\_num <- ifelse(pred > 0.5, 1, 0)

y\_pred <- factor(pred\_num, levels=c(0, 1))

y\_act <- factor(ifelse(testData$diagnosis=="B",0,1))

mean(y\_pred == y\_act)

1557655242(1)

The accuracy is 0.9064

Notice: the forth question can use caret package to split data and can take account of data imbalance.