学生学号	0121613570127	实验课成绩	

武汉理工大学

学 生 实 验 报 告 书

实验课程名称	线性回归模型的实现
开 课 学 院	理学院
指导教师姓名	李丹
学 生 姓 名	胡露文
学生专业班级	统计学 1701 班

2019 -- 2020 学年 第 二 学期

实验教学管理基本规范

实验是培养学生动手能力、分析解决问题能力的重要环节；实验报告是反映实验教学水平与质量的重要依据。为加强实验过程管理，改革实验成绩考核方法，改善实验教学效果，提高学生质量，特制定实验教学管理基本规范。

1、 本规范适用于理工科类专业实验课程，文、经、管、计算机类实验课程可根据具体情况参照执行或暂不执行。

2、 每门实验课程一般会包括许多实验项目，除非常简单的验证演示性实验项目可以不写实验报告外，其他实验项目均应按本格式完成实验报告。

3、 实验报告应由实验预习、实验过程、结果分析三大部分组成。每部分均在实验成绩中占一定比例。各部分成绩的观测点、考核目标、所占比例可参考附表执行。各专业也可以根据具体情况，调整考核内容和评分标准。

4、 学生必须在完成实验预习内容的前提下进行实验。教师要在实验过程中抽查学生预习情况，在学生离开实验室前，检查学生实验操作和记录情况，并在实验报告第二部分教师签字栏签名，以确保实验记录的真实性。

5、 教师应及时评阅学生的实验报告并给出各实验项目成绩，完整保存实验报告。在完成所有实验项目后，教师应按学生姓名将批改好的各实验项目实验报告装订成册，构成该实验课程总报告，按班级交课程承担单位（实验中心或实验室）保管存档。

6、 实验课程成绩按其类型采取百分制或优、良、中、及格和不及格五级评定。

附表：实验考核参考内容及标准

	观测点	考核目标	成绩组成
实验预习	1. 预习报告 2. 提问 3. 对于设计型实验，着重考查设计方案的科学性、可行性和创新性	对实验目的和基本原理的认识程度，对实验方案的设计能力	20%
实验过程	1. 是否按时参加实验 2. 对实验过程的熟悉程度 3. 对基本操作的规范程度 4. 对突发事件的应急处理能力 5. 实验原始记录的完整程度 6. 同学之间的团结协作精神	着重考查学生的实验态度、基本操作技能；严谨的治学态度、团结协作精神	30%
结果分析	1. 所分析结果是否用原始记录数据 2. 计算结果是否正确 3. 实验结果分析是否合理 4. 对于综合实验，各项内容之间是否有分析、比较与判断等	考查学生对实验数据处理和现象分析的能力；对专业知识的综合应用能力；事实求实的精神	50%

实验课程名称：_线性回归模型的实现___

实验项目名称	一元线性回归模型的实现			实验成绩	
实 验 者	胡露文	专业班级	统计学 1701 班	组 别	
同 组 者	若有分组，请注明组成员			实验日期	2020 年 5 月 22 日

第一部分：实验要求

1、 用 R 软件载入所需数据（实验 1 数据 skincancer.txt）；

2、 拟合一元线性回归模型；

3、 一元线性回归模型的诊断，包括决定系数，相关系数检验，t 检验，F 检验，失拟检验；

4、 预测：当地区纬度为 40 时，因变量平均值的置信区间（置信度 95%），因变量新值的预测区间；

5、 实验总结。

第二部分：实验过程记录

过程记录（包括操作的步骤或者代码，输出的结果或者图形）：

一、实验原理

回归分析是研究一个或一组变量（自变量）的变动对另一个变量（因变量）的变动之影响程度。因变量处于被解释的特殊地位，为随机变量，自变量一般是非随机变量，是确定的给出的。回归分析可以进行预测和控制。

一元线性回归是一种研究两个连续的定量变量之间统计关系的统计分析方法。

1. 多元线性回归模型的一般形式

自变量 x 和因变量 y 之间的关系用两部分来描述

第一部分是线性的主体部分：反应了 x 的变化引起的 y 的线性变化。

第二部分是随机部分，这部分不可观测，包含其他所有的随机因素对 y 的影响。

用数学表达式表示如下：

$$y = \beta_0 + \beta_1 x + \varepsilon$$

线性部分中的 β_0 为截距，β_1 称之为斜率，我们可以利用样本集的信息，把截距和斜率估计出来得到的估计值记为 b_0 和 b_1，把估计值带入一元线性回归模型中的线性部分，计算得到的结果记为 $\hat{y}$，即 $\hat{y} = b_0 + b_1 x$。该方程为（估计的）回归方程，他是因变量关于自变量的一次函数，表示的是一条直线，因此称此直线为拟合直线。

2. 回归诊断

在利用 OLS 回归拟合模型时，我们有正态性、独立性、线性、同方差性进行统计假设、若在建立实际问题的回归模型时，出现与上述假设违背的情况，则回归模型需要进行改进。

（1）正态性。当预测变量值固定时，因变量成正态分布，则残差值也应该是一个均值为 0 的正态分布。“正态 Q-Q 图”是在正态分布对应的值下，标准化残差的概率图，若满足正态假设，那么图上的点应该落在 45 度角的直线上，否则就违反了正态性的假设。

（2）独立性。无法从这些图中分辨出，因变量值是否相互独立，只能从收集的数据中来验证。如果是从一个个个体抽样得来的，那么可能必须要调整模型独立性的假设。

（3）线性性。若因变量与自变量线性相关，那么残差值与预测(拟合)值就没有任何系统关联。换句话说，除了白噪声，模型应该包含数据中所有的系统方差。观察“残差图与拟合图”(Residuals vs Fitted) 可以对回归模型进行改进修正。

（4）同方差性。若满足不变方差假设，那么在“位置尺度图”(Scale - Location Graph)中，水平线周围的点应该随机分布。

最后我们可以通过观察“残差与杠杆图”(Residuals vs Leverage)关注的单个观测点的信息，并从图形可以鉴别出离群点、高杠杆值点和强影响点。若一个观测点是离群点，表明拟合回归模型对其预测效果不佳(产生了巨大的或正或负的残差);若一 个观测点有很高的杠杆值，表明它是一个异常的预测变量值的组合，因变量值不参与计算一个观测点的杠杆值;若一个观测点是强影响点(influential observation)， 表明它对模型参数的估计产生的影响过大，非常不成比例。强影响点可以通过 Cook 距离即 Cook'sD 统计量来鉴别。

3. 改进措施

若出现违背回归假设的问题，我们可以采取以下措施:

（1）删除离群点通常可以提高数据集对于正态假设的拟合度，而强影响点会干扰结果，通常也会被删除。删除最大的离群点或者强影响点后，模型需要重新拟合。若离群点或强影响点仍然存在，重复以上过程直至获得比较满意的拟合。

（2）当模型不符合正态性、线性或者同方差性假设时，变量的变换通常可以改善或调整模型效果。当模型违反正态假设时，通常可以对响应变量（因变量）尝试某种变换。当违反了线性假设时，对预测变量（自变量）进行变换常常会比较有用。

二、R 软件实现

1.程序代码

```
rm(list = ls())
#导入数据
#data<-read.table("C:\\Users\\DELL\\Desktop\\skincancer.txt",header=T,sep='\t')
#data<-read.table("C:\\Users\\DELL\\Desktop\\skincancer.txt",header=T)
data<-read.table("C:/Users/DELL/Desktop/skincancer.txt",header=T)
data

#dim(data)
attach(data)
cor=cor(data[,2:5])
plot(data)
#State
#Lat
#Mort
#Ocean
#Long
class(Lat)
class(Mort)
Mort=as.numeric(Mort)
LM=lm(Mort~Lat)
LM
plot(Mort~Lat)
abline(LM)
summary(LM)

r=cor[1,2]
r2=r^2
r2
```

```
t=(r*((49-2)^0.5))/((1-r2)^0.5)
t
p=2*pt(t,47)
p

Full=lm(Mort~as.factor(Lat))
anova(LM,Full)

confint(LM)

MSE=sum((residuals(LM)-mean(residuals(LM)))^2)/49
MSE
var(Mort)
confint(LM)
par(mfrow=c(2,2))
plot(LM)
par(mfrow=c(2,2))
#anova(D)

par(mfrow=c(1， 1))
residplot<-function(LM,nbreaks=49){
   a<-rstudent(LM)
   hist(a,breaks=nbreaks,freq=FALSE,xlab="Studentized Residual",main="Distribution of Errors")
   rug(jitter(a),col="brown")
   curve(dnorm(x,mean=mean(a),sd=sd(a)),add=TRUE,col="blue",lwd=2)
   lines(density(a)$x,density(a)$y,col="red",lwd=2,lty=2)
   legend("topright",legend=c("Normal Curve","Kernel Density
Curve"),lty=1:2,col=c("blue","red"),cex=0.7)
}
residplot(LM)

install.packages("car")
library(car)
outlierTest(LM)

par(mfrow=c(1,1))

hat.plot<-function(LM){
   p<-length(coefficients(LM))
   n<-length(fitted(LM))
   plot(hatvalues(LM),main="Index Plot of Hat Values")
   abline(h=c(2,3)*p/n,col="red",lty=2)
   identify(1:n,hatvalues(LM),names(hatvalues(LM)))
}
hat.plot(LM)

cutoff<-4/(nrow(data)-length(LM$coefficients)-2)
cutoff
par(mfrow=c(1,1))
plot(LM,which=4,cook.levels=cutoff)
abline(h=cutoff,col="red")

avPlots(LM,ask=FALSE,onepage=TRUE,id.method="identify")
influencePlot(LM,id.method="identify",main="InfluencePlot",sub="circle size is proportional to
cook's distance")

new<-data.frame(Lat=40)
```

```
new
lm.conf<-predict(LM,new,interval="confidence",level=0.95)
lm.conf
lm.pred<-predict(LM,new,interval="prediction",level=0.95)
lm.pred
```

2.运行结果几结果分析

（1）一元回归模型

导入原始数据

```
> data<-read.table("C:/Users/DELL/Desktop/skincancer.txt",header=T)
> data
           State  Lat Mort Ocean  Long
1        Alabama 33.0  219     1  87.0
2        Arizona 34.5  160     0 112.0
3       Arkansas 35.0  170     0  92.5
4     California 37.5  182     1 119.5
5       Colorado 39.0  149     0 105.5
6    Connecticut 41.8  159     1  72.8
7       Delaware 39.0  200     1  75.5
8      Wash,D.C. 39.0  177     0  77.0
9        Florida 28.0  197     1  82.0
10       Georgia 33.0  214     1  83.5
```

上图只展示了部分数据。

```
> attach(data)
```

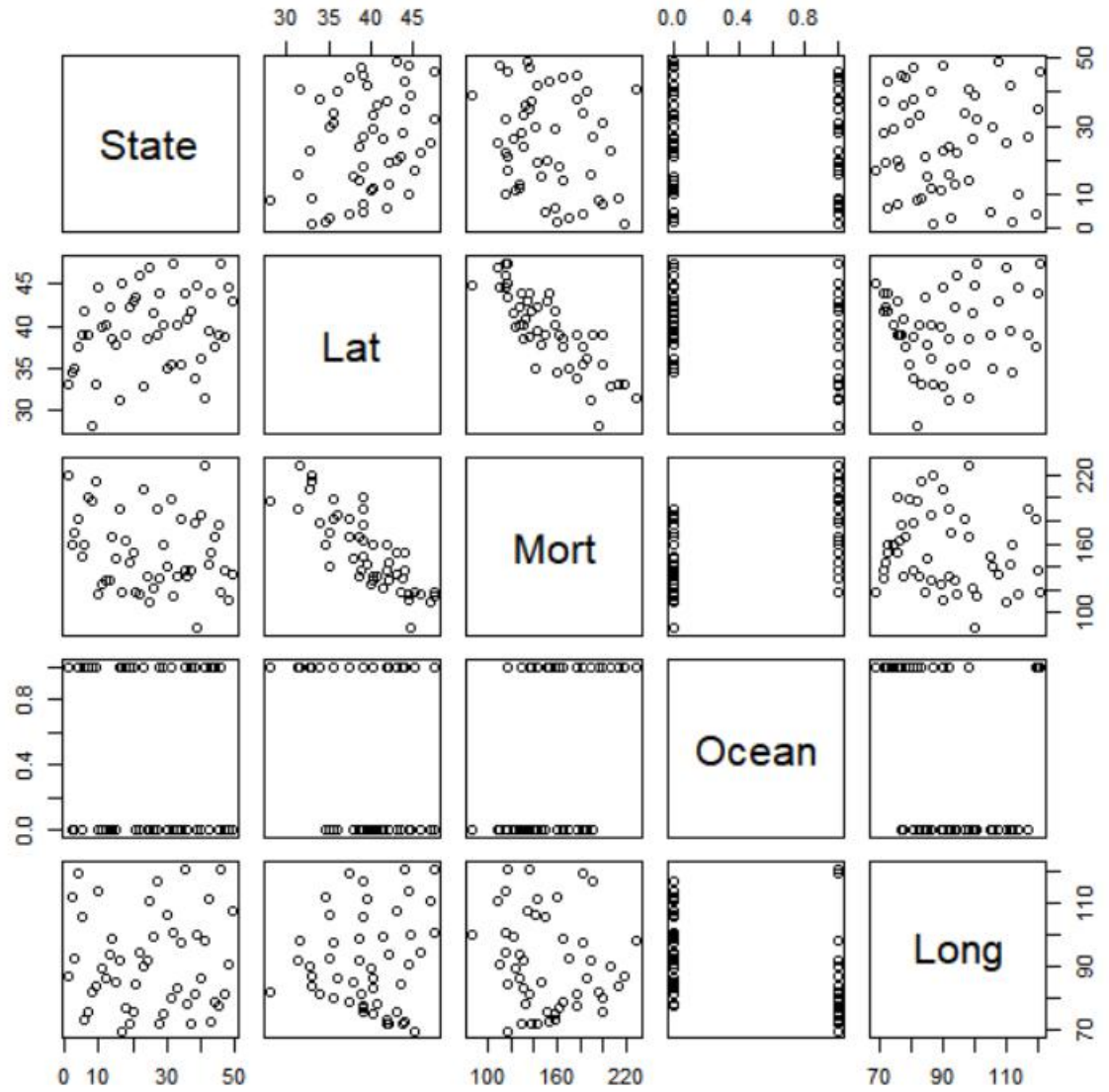
观察变量间的相关性，cor()函数提供了二变量之间的相关系数，我们可以直接调用该函数得到样本相关阵。

```
> cor(data[,2:5])
              Lat        Mort       Ocean        Long
Lat    1.00000000 -0.8245178 -0.2195420  0.09897372
Mort  -0.82451779  1.0000000  0.4733547 -0.14618812
Ocean -0.21954196  0.4733547  1.0000000 -0.38260959
Long   0.09897372 -0.1461881 -0.3826096  1.00000000
```

纬度 Lat 和死亡人数 Mort 之间的相关系数为-0.82451779，说明纬度和死亡人数之间高度线性相关，而其他变量之间的相关系数的绝对值都不超过 0.5，线性相关程度较低。

下面观察两两散点图

```
> plot(data)
```



由两两散点图可以看出，纬度 Lat 和 Mort 之间存在着明显的线性相关关系，而其他变量之间不存在明显的线性相关关系。

于是决定用纬度 Lat 和死亡人数 Mort 进行线性回归，根据实际，以纬度 Lat 为自变量，以死亡人数 Mort 为因变量是合理的。

```
> class(Lat)
[1] "numeric"
> class(Mort)
[1] "integer"
```

发现死亡率不是 numeric 型的变量，因此将其转化为 numeric 型。

```
> Mort=as.numeric(Mort)
```

进行一元线性回归

```
> Mort=as.numeric(Mort)
> LM=lm(Mort~Lat)
> LM
```

```
Call:
lm(formula = Mort ~ Lat)

Coefficients:
(Intercept)          Lat
    389.189       -5.978
```

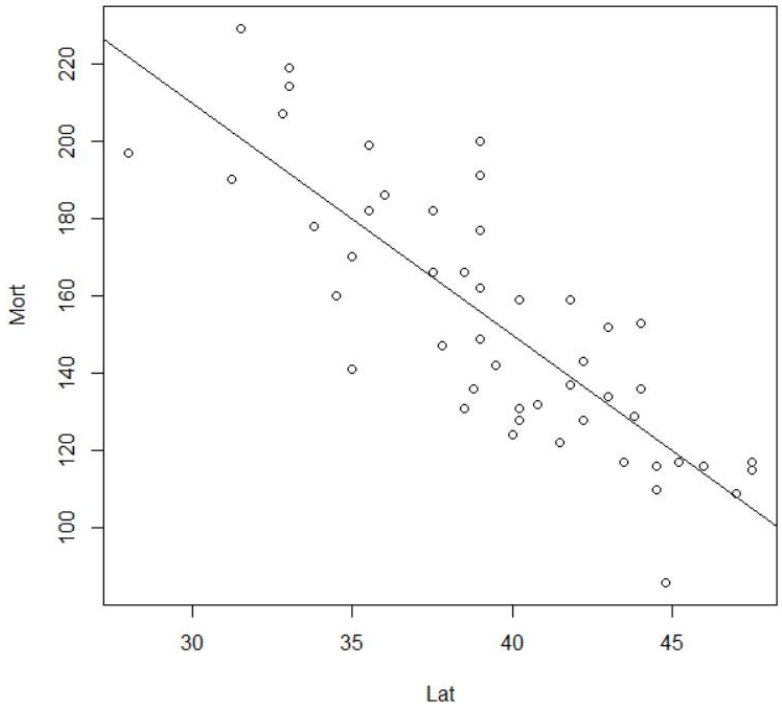
从输出结果可以得到，回归方程中，截距项的估计值为 389.189，斜率项的估计值为-5.978。

即：$\hat{Mort} = 389.189 - 5.978 * Lat$

这意味着，当纬度增加 1 时，死亡人数下降 5.978（每百万人）。

下面将纬度 Lat 和死亡人数 Mort 的散点图和拟合直线图画出来

```
> plot(Mort~Lat)
> abline(LM)
```



从上图直观看出，模型的拟合效果很好。下面我们进一步进行回归诊断。

（2）回归诊断

如果自变量和因变量不存相关关系，那么回归方程就没有任何意义了，如果因变量和自变量是有相关关系的，即因变量会随着自变量的变化而线性变化，这个时候一元线性回归方程才有意义。所以，我们需要用假设检验的方法，来验证相关性的有效性。

1. 相关系数检验

用 t 检验来检验总体相关系数是否存在，按照标准的假设检验过程，首先给出原假设和对立假设：

原假设 $H_0 : \rho = 0$

对立假设 $H_1 : \rho \neq 0$

T 检验统计量 $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ 服从自由度为 n-2 的 t 分布。

```
> r=cor[1,2]
> r2=r^2
> r2
[1] 0.6798296
> t=(r*((49-2)^0.5))/((1-r2)^0.5)
> t
[1] -9.989836
> p=2*pt(t,47)
> p
[1] 3.309456e-13
```

以上结果显示，t 统计量的值为-9.989836，p 值为3.309456×10^{-13}

所以在显著性水平$\alpha=0.01$时，p 值小于 0.01，拒绝原假设，接受备选假设，在显著性水平$\alpha=0.01$下，有充足的证据来支持总体中死亡人数 Mort 和纬度 Lat 之间的线性关系，二者高度相关。

2. T 检验，F 检验、决定系数检验

summary()函数提供了最小值、最大值、四分位数和数值型变量的均值，以及频数统计，因此调用该函数获取相关描述性统计量。

```
> summary(LM)

Call:
lm(formula = Mort ~ Lat)

Residuals:
    Min      1Q  Median      3Q     Max
-38.972 -13.185   0.972  12.006  43.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
Lat          -5.9776     0.5984   -9.99 3.31e-13 ***


---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.12 on 47 degrees of freedom
Multiple R-squared:  0.6798,    Adjusted R-squared:  0.673
F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

首先介绍一下上表提供的信息：

Call，列出了回归模型的公式。

Residuals，列出了残差的最小值点，1/4 分位点，中位数点，3/4 分位点，最大值点。

Coefficients，表示参数估计的计算结果。

Estimate，为参数估计列。Intercept 行表示常数参数 a 的估计值 ，x 行表示自变量 x 的参数 b 的估计值。

Std. Error，为参数的标准差，sd(a), sd(b)

t value，为 t 值，为 T 检验的值

Pr(>|t|) ，表示 P-value 值，用于 T 检验判定，匹配显著性标记

显著性标记，***为非常显著，**为高度显著, **为显著，·为不太显著，没有记号为不显著。

Residual standard error，表示残差的标准差，自由度为 n-2。

Multiple R-squared，为相关系数 R^2 的检验，越接近 1 则越显著。

Adjusted R-squared，为相关系数的修正系数，解决多元回归自变量越多，判定系数 R^2 越大的问题。

F-statistic，表示 F 统计量，自由度为(1,n-2)，p-value:用于 F 检验判定，匹配显著性标记。

根据上表可以得到以下结果

①T 检验法：T 检验是检验模型某个自变量对于因变量的显著性，通常用 P-value 判断显著性，对于一元线性回归来说，原假设和备择假设如下：

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

检验统计量 t 如下，t 服从自由度为 n-2 的 t 分布

$$t = \frac{b_1 \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sqrt{MSE}}$$

对模型的斜率进行 t 检验，得到的 P 值为3.31×10^{-13}非常接近于 0，远远小于 0.01，所以不拒绝原假设，模型的自变量纬度 Lat 对于因变量 Mort 显著。对于截距项做 t 检验，可以得到截距项也非常显著。

②F 检验法：F 检验用于对所有的自变量在整体上看对于因变量的线性显著性，也是所以自变量对因变量的解释能力，也是用 P-value 判断显著性，小于 0.01 时说明整体上自变量与因变量相关关系显著。

检验统计量 $F = \frac{SSR/1}{SSE/(n-2)}$ 服从 $F(1, n-2)$

由上图知，F 统计量的值为 99.8，P 值为3.309×10^{-13}，非常接近于 0，所以整体上自变量纬度 Lat 与因变量死亡人数 Mort 的相关关系显著自变量对因变量的解释能力很强。

事实上，由于本题是一元线性回归，所以以上三种检验是等价的，所以得到的 P 值也都是相等的。

③决定系数：

决定系数为 SSR/SST，它用来判断回归方程的拟合程度，其取值在 0，1 之间，越接近 1 说明拟合程度越好。

由上图知，决定系数 R^2 为 0.6798，说明自变量纬度 Lat 的波动可以解释因变量死亡人数 Mort 67.98%的波动。

调整后的决定系数为 0.673，这个数字还可以接受，所以认为拟合程度还可以。

3. 失拟检验

原假设：模型是合理的，不存在失拟

备择假设：模型不合理，存在失拟

F 统计量的值为 MSLF / MSPE，服从 F(c-2,n-c)。

```
> Full=lm(Mort~as.factor(Lat))
> anova(LM,Full)
Analysis of Variance Table

Model 1: Mort ~ Lat
Model 2: Mort ~ as.factor(Lat)
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     47 17173.1
2     17  4310.5 30     12863 1.691  0.128
```

由上图知，F 的值为 1.691，P 值为 0.128，大于 0.01，所以 P 值大于显著性水平，不能拒绝原假设，认为在显著性水平为 0.01 时，没有充足的证据认为此一元线性模型是失拟的。

通过 confint()函数得到模型参数的置信区间如下：

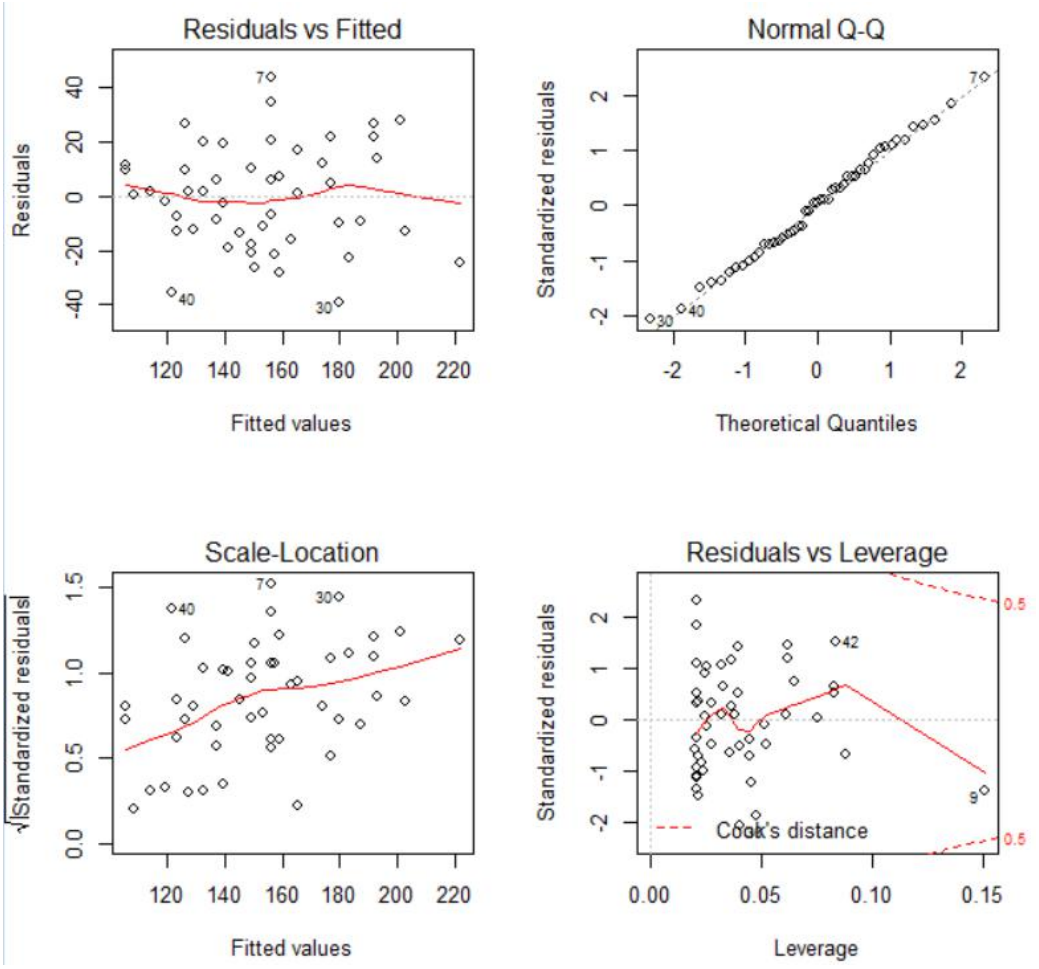
```
> confint(LM)
                2.5 %     97.5 %
(Intercept) 341.285151 437.093552
Lat          -7.181404  -4.773867
```

由上表知，截距项和斜率项的 95%置信区间，都不包含 0，截距项和斜率项显著。

4. 模型适用性检验

利用 lm()函数返回的对象 LM 使用 plot()函数，生成评价模型拟合情况的四幅图形，同时利用 par(mfrow=c(2,2))将，plot()函数绘制的四幅图形组合在一个大的 2*2 的图中。

```
> par(mfrow=c(2,2))
> plot(LM)
```



①对残差和拟合值作图，横坐标是拟合值，纵坐标是残差。图中的红色线没有明显的形状特征。

残差值在残差等于 0 的直线上下随机分布，这一点表明线性性的假设是满足的。

残差值均落入以残差为 0 的水平线为中心的带型区域内，这就表明误差项的方差都是相等的。

图中存在几个被标注出来的离群点，但是这些离群点并没有明显跳出以上残差点的分布模式，所以这些点是否需要删除要其他图来决定。

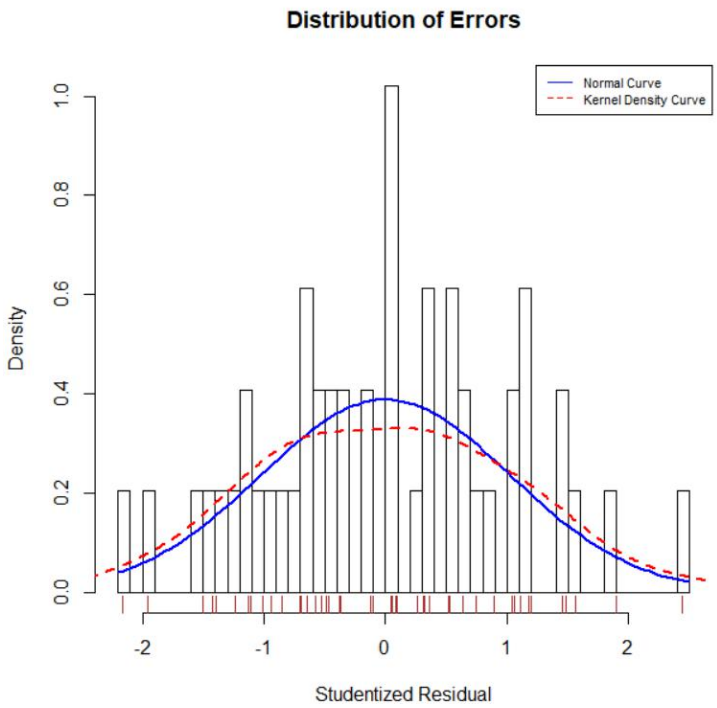
红色线呈现出一条平稳的曲线并没有明显的形状特征，综上说明残差数据表现非常好。

②残差 QQ 图，用来描述残差是否符合正态分布。图中的数据点按对角直线排列，趋于一条直线，并被对角直接穿过，直观上符合正态分布。

事实上，我们还可以通过以下方法来检验残差的正态性：

学生化残差是残差除以它的标准差后得到的数据，用以直观地判断误差项服从正态分布这一假设是否成立，若假定成立，学生化残差的分布也应服从正态分布。

```
> residplot<-function(LM,nbreaks=49){
+   a<-rstudent(LM)
+   hist(a,breaks=nbreaks,freq=FALSE,xlab="Studentized Residual",main="Distribution of Errors")
+   rug(jitter(a),col="brown")
+   curve(dnorm(x,mean=mean(a),sd=sd(a)),add=TRUE,col="blue",lwd=2)
+   lines(density(a)$x,density(a)$y,col="red",lwd=2,lty=2)
+   legend("topright",legend=c("Normal Curve","Kernel Density Curve"),lty=1:2,col=c("blue","red"),cex=
+ }
> residplot(LM)
```



两条曲线贴近程度较高，因此认为误差较好的服从正态分布。

③对标准化残差平方根和拟合值作图（左下图），横坐标是拟合值，纵坐标是标准化后的残差平方根。与残差和拟合值对比图的判断方法类似，数据随机分布，红色线稍微向上倾斜，但并不明显，标准化后的残差基本上随机分布在水平线 0 附近的带型区域内，残差表现较好。

④对标准化残差和杠杆值作图，虚线表示的 Cook 距离等高线，通常用 Cook 距离度量的回归影响点。图中出现了红色等高线，说明数据中国可能有影响回归结果的异常点，对此我们进行分析。

5.异常值检验

①对于异常观测值，因为异常值在一定程度上与其他观测点不同，可能对结果产生较大的负面影响，因此我们还需要对异常值进行分析，包括离群点、高杠杆值点和强影响点。

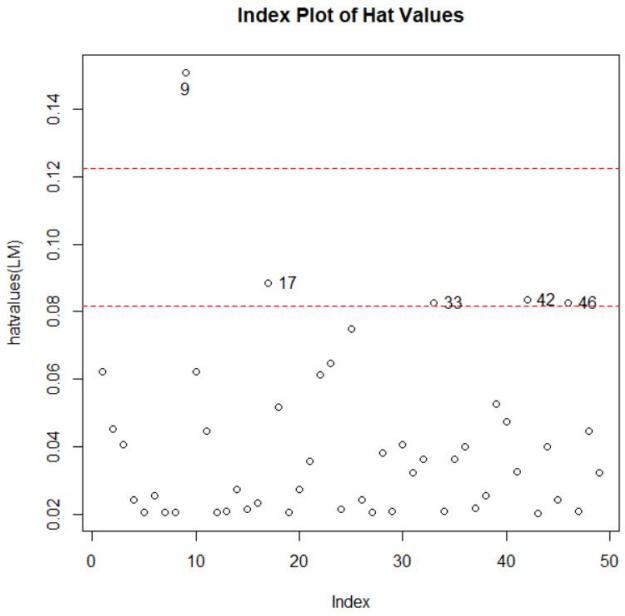
对于离群点：

car 包也提供了一种离群点的统计检验方法。outlierTest()函数可以求得最大标准化残差绝对值 Bonferroni 调整后的 p 值。（该函数只是根据单个最大（或正或负）残差值的显著性来判断是否有离群点。）

```
> library(car)
> outlierTest(LM)
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
  rstudent unadjusted p-value Bonferroni p
6 2.752595          0.0087682       0.3858
```

由于 P 值较大，为 0.3858，大于 0.01，也大于 0.1，所以认为不显著，即没有离群点。

②高杠杆值观测点，即是与其他预测变量有关的离群点。换句话说，它们是由许多异常的预测变量值组合起来的，与响应变量值没有关系。高杠杆值的观测点可通过帽子统计量（hat statistic）判断。对于一个给定的数据集，帽子均值为 p/n，其中 p 是模型估计的参数数目（包含截距项），n 是样本量。一般来说，若观测点的帽子值大于帽子均值的 2 或 3 倍，即可以认定为高杠杆值点。




```
> hat.plot<-function(LM){
+   p<-length(coefficients(LM))
+   n<-length(fitted(LM))
+   plot(hatvalues(LM),main="Index Plot of Hat Values")
+   abline(h=c(2,3)*p/n,col="red",lty=2)
+   identify(1:n,hatvalues(LM),names(hatvalues(LM)))
+ }
> hat.plot(LM)
```

可以看到第 9 个点是高杠杆值点。这个点是弗罗里达州，这个点在原始数据散点图上没有异常。

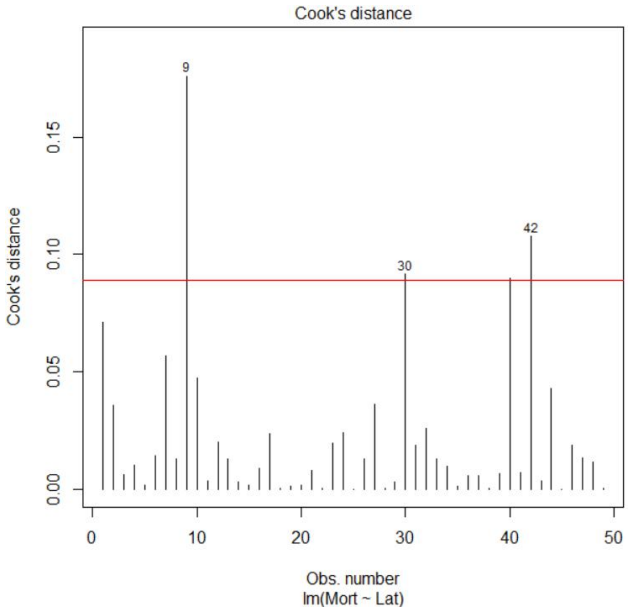
③强影响点，即对模型参数估计值影响有些比例失衡的点。例如，若移除模型的一个观测点时模型会发生巨大的改变，那么你就需要检测一下数据中是否存在强影响点了。有两种方法

可以检测强影响点：Cook 距离，或称 D 统计量，以及变量添加图（added variable plot）。一般来说，Cook's D 值大于 4/(n-k-1)，则表明它是强影响点，其中 n 为样本量大小，k 是预测变量数目。可通过如下代码绘制 Cook's D 图形

```
> cutoff<-4/(nrow(data)-length(LM$coefficients)-2)
> cutoff
[1] 0.08888889
```

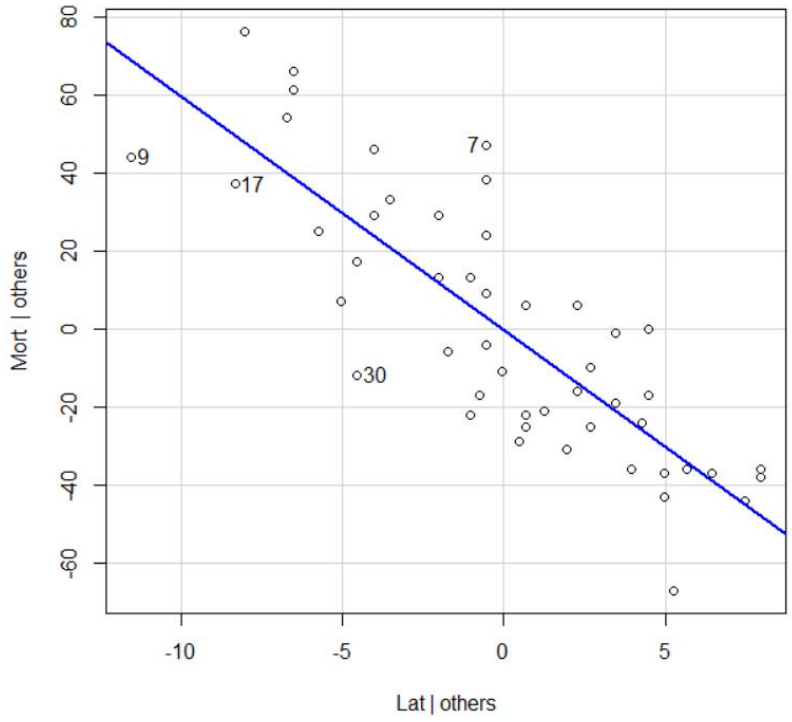
即 cook's D 的值为 0.08888889

```
> par(mfrow=c(1,1))
> plot(LM,which=4,cook.levels=cutoff)
> abline(h=cutoff,col="red")
```



上图可以看到第 9 个点是强影响点，此外，30、40、42 也是强影响点。

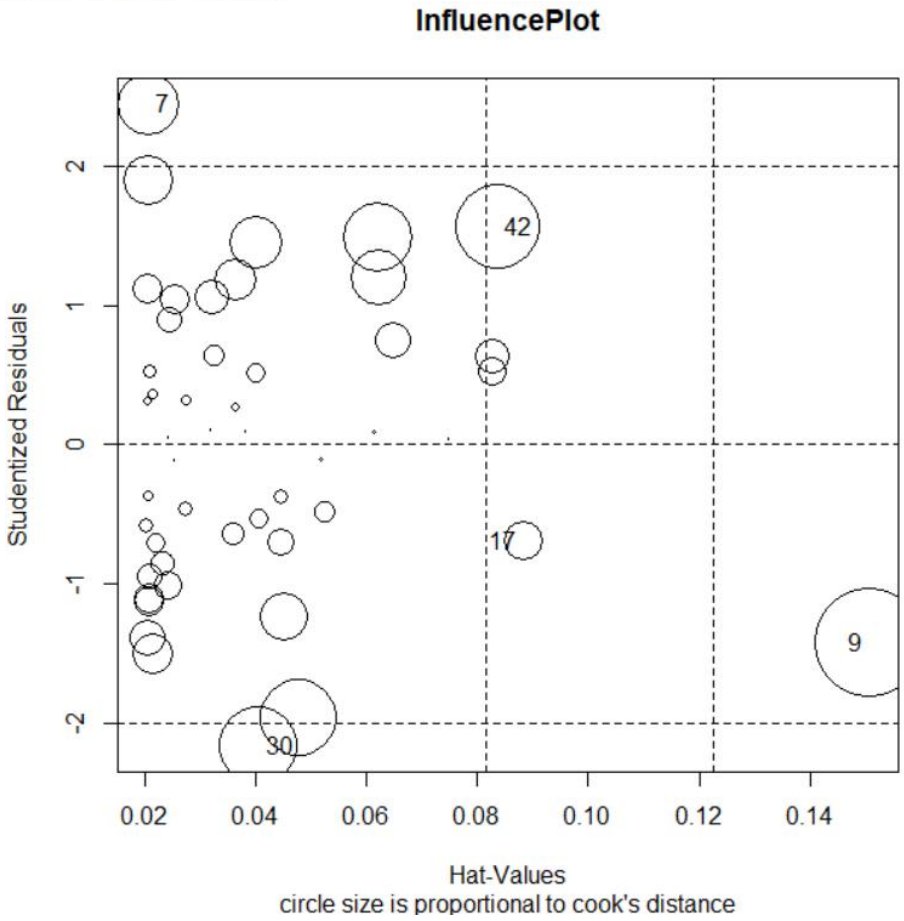
```
avPlots(LM,ask=FALSE,onepage=TRUE,id.method="identify")
```



由上图可以直观看出点对模型的影响。

还可以将离群点、杠杆值和强影响点的信息整合到一幅图形中.

```
> influencePlot(LM,id.method="identify",main="InfluencePlot",sub="circle size is proportional to cook's distan
      StudRes         Hat      CookD
7   2.4423988 0.02068619 0.05698310
9  -1.4240545 0.15074004 0.17612265
17 -0.6913077 0.08844730 0.02344590
30 -2.1612026 0.04054064 0.09153034
42  1.5592161 0.08363628 0.10766696
```



点 9 是高杠杆值点也是强影响点，但是在散点图和拟合直线的图中可以看出，该点虽然稍偏离拟合直线，并且该点对拟合直线的影响较大，但是该点并不反常，该点和其他点的分布模式相同，因此该点保留，同理保留所有的点。

事实上，在尝试删除所有高杠杆值和强影响点之后，再次分析会发现新的异常点，继续删除，直至没有任何高杠杆值和强影响点的情况下，删除的点数量过多，皮肤癌的死亡人数本来就受到各种因素的影响，不仅仅是纬度一个因素，所以在本例中，我选择不对这些点进行删除。

（3）模型确定及模型解释

从输出结果可以得到，回归方程中，截距项的估计值为 389.189，斜率项的估计值为-5.978。

即：$\hat{Mort} = 389.189 - 5.978 * Lat$

这意味着，当纬度增加 1 时，皮肤癌的死亡人数下降 5.978，这个结论是合理的，因为皮肤癌受到紫外线的影响非常大，而在其他条件不变的情况下，纬度增加，紫外线会减弱。

决定系数 R^2 为 0.6798，说明自变量纬度 Lat 的波动可以解释因变量死亡人数 Mort 67.98%的波动。死亡人数的波动无法由纬度解释的部分，可能由生活习惯等因素来解释。

（4）模型预测

我们调用 fitted()函数对模型进行预测，得到了皮肤癌死亡人数的一系列预测值如下

当地区纬度为 40 时，因变量平均值的置信区间（置信度 95%），因变量新值的预测区间；

```
> new<-data.frame(Lat=40)
> new
  Lat
1  40
> lm.conf<-predict(LM,new,interval="confidence",level=0.95)
> lm.conf
       fit      lwr      upr
1 150.0839 144.5617 155.6061
> lm.pred<-predict(LM,new,interval="prediction",level=0.95)
> lm.pred
       fit     lwr      upr
1 150.0839 111.235 188.9329
```

由上图可以看出，当地区纬度为 40 时，皮肤癌死亡人数平均值的置信区间（置信度 95%）[144.5617,155.6061]，单位是每百万人，皮肤癌死亡人数新值的预测区间（置信度 95%）为[111.235,188.9329]，单位是每百万人。

值得注意的是：

同等置信度水平下，因变量平均值的置信区间要比因变量新值的预测区间要小。

当给定自变量的值为平均值，并且增加样本容量 n，置信区间的标准差可以趋近于 0，但预测区间的标准差多了一项 MSE，所以预测区间的标准差不能趋近于 0。

因变量平均值的置信区间和因变量新值的预测区间都在自变量的均值附近都是最小的。

教师签字_____

第四部分　实验总结

实验收获、未解问题和深入探讨：

1. 通过本次上机，我学会了 R 语言的一些基本操作，比如数据的导入还有如何利用 R 软件进行一元线性回归，包括假设检验。也让我认识到了想要学会一门编程语言，必须要熟知代码的意义和其统计学原理，理论是应用的最有力支撑。

2. 每个模型都是基于一定的假设建立的，因此当我们使用一个模型的时候，必须要对其假设进行检验，符合假设的时候才可以采用该模型，否则就要采用变量变换或者变量删除或者剔除异常点等操作来使模型通过假设检验，满足假设条件。

3. 在对模型进行假设检验的时候，有些检验会受到主观因素的影响，比如置信度是人为确定的，以及决定系数多大是可以接受的，所以根据不同的目的和实际，可以对不同的假设检验提出不同的要求，比如所有的假设检验以及区间估计对独立性要求很高，所有的假设检验以及区间估计对等方差的要求较高，截距项和斜率项的假设检验和置信区间对正态性的要求可以适当放宽，预测区间对正态性的要求较高。

需要注意的是，违背基本假设的程度直接影响模型得到结论的偏差程度，对违背基本假设的处理往往取决于你打算如何使用回归模型，例如，如果仅仅考虑的是检验自变量和因变量之间的相关性，也就是斜率是否为 0 ，那么就可以放宽误差服从正态分布的基本假设的要求。如果需要用得到的模型来预测因变量的新值，那么当误差项不服从正态分布时，由模型得到的结果可能是不准确的。所以，我们关注需要重点关注的前提假设，并且掌握相应的补救方法。

4. 不是所有的异常点都需要删除，在本案例中，高杠杆值点也是强影响点，但是在散点图和拟合直线的图中可以看出，该点虽然稍偏离拟合直线，并且该点对拟合直线的影响较大，但是该点并不反常，该点和其他点的分布模式相同，因此该点保留，同理保留所有的点。

事实上，在尝试删除所有高杠杆值和强影响点之后，再次分析会发现新的异常点，继续删除，直至没有任何高杠杆值和强影响点的情况下，删除的点数量过多，皮肤癌的死亡人数本来就受到各种因素的影响，不仅仅是纬度一个因素，所以在本例中，我选择不对这些点进行删除。

5. 对于模型的解释和分析始终都要结合实际情况，比如分析自变量对因变量的影响以及决定异常点是否要删除。

6. 在本例中，所进行的检验基本都通过了。但是以后我可能会遇到没有通过假设检验的情况，比如出现异方差应该怎么处理，我会在下一章继续学习。