学生学号	0121906280724	实验课成绩	

武汉理工大学

学 生 实 验 报 告 书

实验课程名称	实用回归分析
开 课 学 院	理学院
指导教师姓名	李丹
学 生 姓 名	张逸敏
学生专业班级	统计 2001

2022 -- 2023 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：_实用回归分析____

实验项目名称	非线性回归模型的实现			实验成绩	
实 验 者	张逸敏	专业班级	统计 2001	组　　别	
同 组 者	刘璇、马钟森、李耀祖、危景熙、焦鼎云			实验日期	2023 年 4 月 14 日

第一部分：实验数据及要求

数据：us_census.txt

1、 构建乘性误差项模型：先对模型进行线性化，然后拟合模型；

2、 构建加性误差项模型：选取参数的初始值，用非线性最小二乘法做拟合；

3、 比较乘性误差项模型与加性误差项模型。

第二部分：实验过程记录（可加页）（包括实验原始数据记录，实验现象记录，实验过程发现的问题等）

一、 构建乘性误差项模型

1.1 乘性误差项构建和线性化

根据生长曲线模型，构建乘性误差项模型如下：

$$y=\frac{\beta_1}{1+e^{\beta_2+\beta_3 x\ +\ \varepsilon}}$$

其中，x是年份，y是人口数，$\beta_1\ >\ 0$, β_2无限制, $\beta_3<0$.

对乘性误差项模型进行线性化如下：

$$\frac{\beta_1}{y}-1=e^{\beta_2+\beta_3 x+\varepsilon}$$

$$\ln(\frac{\beta_1}{y}-1)=\beta_2+\beta_3 x+\varepsilon$$

令$y^*=\ln\left(\frac{\beta_1}{y}-1\right)$，原非线性模型转化为求解线性模型：

$$y^*=\beta_2\ +\ \beta_3 x+\varepsilon$$

考虑β_1的现实意义：$x\to\infty$时，y的值，可以事先人为给定$\beta_1=350$。

1.2 乘性误差项模型求解

读入数据

```
data=read.table("./us_census.txt", header = T)
data
```

year	population
<int>	<dbl>
1790	3.929
1800	5.308
1810	7.240
1820	9.638
1830	12.866
1840	17.069
1850	23.192
1860	31.443
1870	39.818
1880	50.156
1890	62.948
1900	75.995
1910	91.972
1920	105.711
1930	122.775
1940	131.669
1950	150.697
1960	179.323
1970	203.302
1980	226.542
1990	248.710

将自变量年份转化为0～ n - 1

```
data["x"] = (data["year"] - 1790)/10
```

给定$\beta_1=350$，计算$y^*=\ln\left(\frac{\beta_1}{y}-1\right)$

```
beta1 = 350
data["y1"] = log(beta1 / data["population"] - 1)
```

求解一元线性回归模型

```
md1 = lm(y1~x, data=data)
summary(md1)
```

```
Call:
lm(formula = y1 ~ x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16365 -0.12760  0.01944  0.08913  0.18816

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.290095   0.050425   85.08   <2e-16 ***
x           -0.260910   0.004313  -60.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1197 on 19 degrees of freedom
Multiple R-squared:  0.9948,    Adjusted R-squared:  0.9946
F-statistic:  3659 on 1 and 19 DF,  p-value: < 2.2e-16
```

整体方程的 F 检验$p < 2.2\times10^{-16}$，单个系数的 t 检验$p<2\times10^{-16}$，在1%的显著性水平上显著。决定系数$R^2=0.9948$，调整决定系数$adj-R^2=0.9946$，模型拟合效果非常好。

采用 gvlma()函数对线性模型的假设进行综合验证：

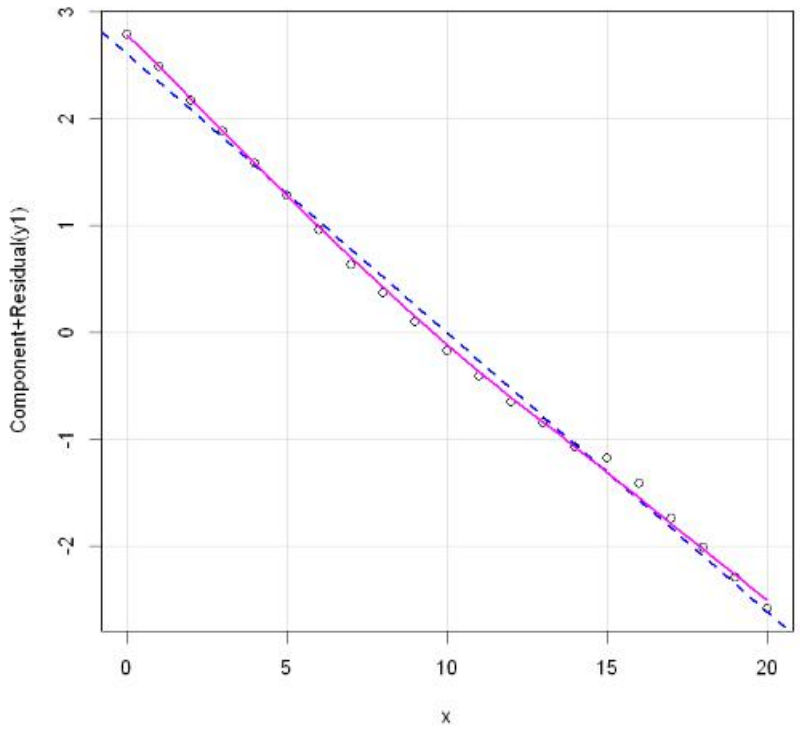
```
library(gvlma)
gvmodel = gvlma(md1)
summary(gvmodel)
```

结果如下：

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = mdl)

                      Value   p-value                   Decision
Global Stat        14.888848 0.0049374 Assumptions NOT satisfied!
Skewness            0.005978 0.9383722    Assumptions acceptable.
Kurtosis            1.580344 0.2087112    Assumptions acceptable.
Link Function      12.896839 0.0003291 Assumptions NOT satisfied!
Heteroscedasticity  0.405687 0.5241666    Assumptions acceptable.
```

综合验证显示模型的线性性没有得到满足，下面利用成分残差图检验因变量和自变量是否呈现线性关系。

```
library(car)
crPlots(mdl)
```



成分残差图呈线性，说明因变量和自变量呈线性关系。

综上，给定$\beta_1 = 350$的情况下，模型求解结果为：

$$y = \frac{350}{1 + e^{4.29 - 0.26x}}$$

1.3 采用三和值法确定β_1初值

利用三和法确定β_1的初值，步骤如下：

(1)将 n 个数据分成三组(假设 n=3r)
(2)求各组的y_i值得对数和，即求：

$$S_1=\sum_{i=1}^{r}\ln y_i,\ S_2=\sum_{i=r+1}^{2r}\ln y_i,\ S_3=\sum_{i=2r+1}^{3r}\ln y_i$$

(3)利用下列公式计算β_1的值

$$\beta_1=\exp\left\{\frac{\frac{1}{r}(S_1S_3-S_2{}^2)}{S_1+S_3-2S_2}\right\}$$

代码如下：

```
s1 = 0
s2 = 0
s3 = 0
for (i in 1 : (n / 3))
    s1 = s1 + log(data["population"][i,])
for (i in (n / 3 + 1) : (n / 3 * 2))
    s2 = s2 + log(data["population"][i,])
for (i in (n / 3 * 2 + 1) : n)
    s3 = s3 + log(data["population"][i,])
s1
s2
s3
beta1 = exp(1/(n/3) * (s1 * s3 - s2 * s2) / (s1 + s3 - 2 * s2))
beta1
```

```
15.8185943990524
```

```
28.7028032784076
```

```
36.1490149324191
```

```
750.48745953463
```

利用三和值法求出的$\beta_1=750.49$，在此基础上求解 1.2 中的一元线性回归模型，结果如下：

```
data["y2"] = log(beta1 / data["population"] - 1)
md12 = lm(y2~x, data=data)
summary(md12)
```

```
Call:
lm(formula = y2 ~ x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26005 -0.20056  0.02907  0.15109  0.33141

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.915682   0.085661   57.38   <2e-16 ***
x           -0.226482   0.007327  -30.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2033 on 19 degrees of freedom
Multiple R-squared:  0.9805,    Adjusted R-squared:  0.9795
F-statistic: 955.4 on 1 and 19 DF,  p-value: < 2.2e-16
```

F 检验、t 检验均显著，决定系数$R^2=0.9805$, $adj-R^2=0.9795$，模型拟合效果非常好。

模型求解结果为

$$y=\frac{750.49}{1+e^{4.92-0.23x}}$$

二、构建加性误差项模型

根据生长曲线模型，构建加性误差项模型如下：

$$y=\frac{\beta_1}{1+e^{\beta_2+\beta_3 x}+\varepsilon}$$

其中，x是年份，y是人口数，$\beta_1>0$, β_2无限制, $\beta_3<0$.

2.1 确定β_1,β_2,β_3的初值

采用三和值法估计的初值时，程序抛出“奇异矩阵”的错误，所以采用$\beta_1=750.49$时，乘性误差项模型β_2,β_3的估计值作为非线性最小二乘的初值，即$\beta_1=750.49,\beta_2=4.92,\beta_3=-0.226$。

2.2 非线性最小二乘求解加性误差项模型

```
nonlin = nls(population~A/(1+exp(B+C*x)),start=list(A=750.49,B=4.92, C=-0.226), data=data)
summary(nonlin)
```

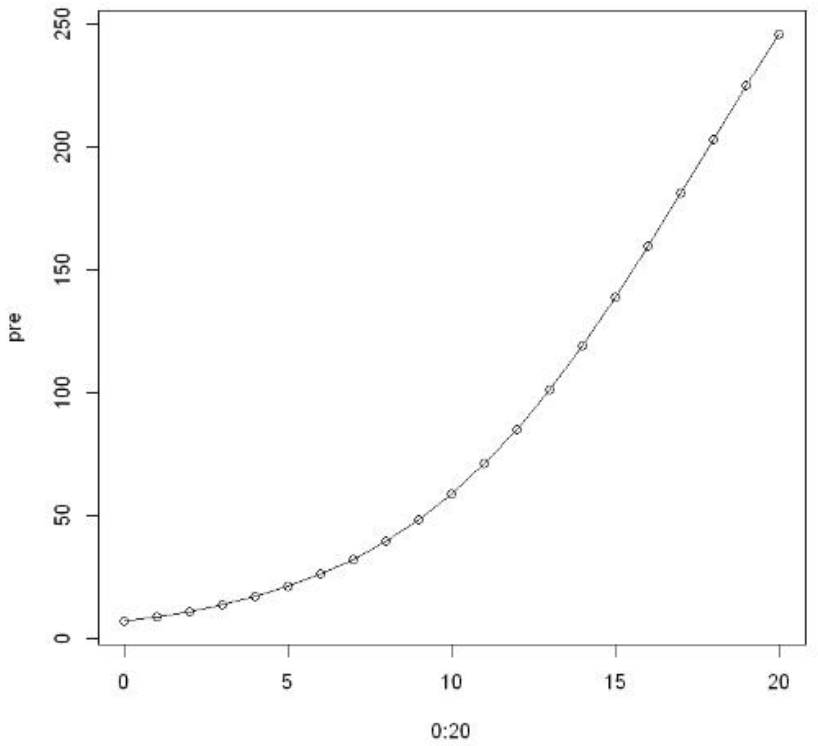
```
Formula: population ~ A/(1 + exp(B + C * x))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
A 389.16621   30.81214   12.63  2.2e-10 ***
B   3.99034    0.07032   56.74  < 2e-16 ***
C  -0.22662    0.01086  -20.87  4.6e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.45 on 18 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 7.833e-06
```

模型 t 检验显著，求解结果为

$$y=\frac{389.167}{1+e^{3.99-0.227x}}$$

模型图像如下：

```
pre=predict(nonlin,interval="prediction")
plot(0:20, pre)
lines(0:20, pre)
```



三、对比乘性误差项模型和加性误差项模型

乘性误差项模型和加性误差项模型所得的结果有一定差异，其中乘性误差项模型认为$\frac{\beta_1}{y_t}-1$是异方差的，而$\ln\left(\frac{\beta_1}{y_t}-1\right)$是同方差的。加性误差项认为$\frac{\beta_1}{y_t}-1$是同方差的。从统计性质看两者的差异，前者淡化了近期数据的作用，强化了早期数据的作用，对早期数据拟合的效果较好，而后者对近期数据拟合的效果较好。

影响模型拟合效果的统计性质主要是异方差、自相关、共线性三个方面。异方差可以选择乘性误差项模型和加性误差项模型解决，必要时还可以使用加权最小二乘。时间序列数据通常都存在自相关，使用自回归方法可以改进模型的拟合效果。

教师签字_____

第三部分　结果与讨论（可加页）

(1) 常见的生长曲线模型有皮尔模型、林德诺模型、龚帕兹模型等，本文采用的是皮尔生长曲线模型。

(2) 皮尔生长曲线模型中的β_1是时间充分大时人口的极限值，可以人为给定一个比样本数据都大的一个初始值。

(3) 初始值也可以使用三和值法进行估计。在正态误差假定下，非线性回归的最小二乘估计与最大似然估计是相同的，而最大似然估计具有良好的大样本性质，如渐近无偏性、渐近正态性、一致性等。因而非线性最小二乘估计比三和值更精确，可以把三和值的参数估计值作为求解非线性最小二乘估计值的初值。

(4) 对于可转化为线性模型的曲线回归问题，通常的处理方法都是先转化为线性模型，然后用普通最小二乘法求出参数的估计值，最后经过适当的变换得到所求的回归曲线。通过对因变量做变换使曲线线性化的方法，当然会对估计参数的性质产生影响，比如不具有无偏性等。

(5) 对于不能线性化的模型，参数估计就要采用非线性最小二乘法，一般采用迭代法求解。

(6) 关于确定模型曲线类型，一种方法是根据样本散点图的形状大致确定曲线类型。此外，还可以根据专业知识来确定曲线类型，如商品销量和广告费用的关系，一般用 S 形曲线来描述；在农业生产中，粮食的产量与种植密度往往服从抛物线关系。