学生**学号**	0121714670405	**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	郑广浩
学生专业班级	统计 1701

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	郑广浩	专业班级	统计 1701	组　　别	
同 组 者	陈炜，黄成龙，倪潜峰，于明昊			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

回归分析是指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，是确定给出的。回归分析可以进行预测和控制，是众多用于解决问题的数据分析方法的一种。

1.一元线性回归模型的拟合

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。该模型用数学表达式表示如下：

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

β_0 称为截距，β_1 称为斜率，ε 是影响 y 变化的随机因素，称为随即误差项。

2.一元线性回归模型的诊断

（1）决定系数 r^2

决定系数的数学表达式为：

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

其中，$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$，$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$，分别是一元线性回归模型的回归平方和、残差平方和以及总离差平方和。

决定系数可以解释为：因变量 y 中有的波动中可以被自变量 x 来解释。即当 $r^2 = 1$ 时，表

明所有的样本点完全的落在回归直线上，自变量 x 解释了因变量 y 的所有的波动；当 $r^2=0$ 时，表明估计的回归直线是一条水平直线，自变量 x 解释不了因变量 y 的任何波动。

（2）相关系数检验

皮尔逊相关系数等于决定系数开根号的值：

$$r=\pm\sqrt{r^2}$$

r 的符号取决于估计的斜率的符号。

若 $r=1$， x 和 y 之间完全正相关；

若 $r=-1$， x 和 y 之间完全负相关；

若 $r=0$， x 和 y 之间完全不相关；

对于 r 的其他值，可以理解为 x 和 y 之间是一种不完全的相关关系，越接近于 0，相关性就越弱。

但相关系数仅仅是对样本中的线性关系的强弱进行度量和描述，若换一个不同的样本，极有可能会得到不同的相关系数，并推导出不同的结论。因此，需要对总体相关系数进行假设检验，从而得到总体中两个变量之间的相关性的结论。

以 ρ 来表示总体相关系数，进行假设检验，首先，给出原假设和备择假设：

原假设 $H_0:\rho=0$

备择假设 $H_A:\rho\neq 0$

接着，计算检验统计量的值：

$$T\text{ 检验统计量：}t^*=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

第三步，利用第二步的统计量的值来计算 p 值，p 值是通过自由度为 $n-2$ 的 t 分布来确定的。

最后一步，根据上述的计算得出结论：

若 p 值小于显著性水平 α，拒绝原假设，接受备择假设，即认为：在显著性水平 α 下，有充足的证据来支持总体中两变量之间的线性关系；

若 p 值大于显著性水平 α，不能拒绝原假设，即认为：在显著性水平 α 下，没有充足的证据来证明总体中两变量之间的线性关系。

（3）对 β_1 的 t 检验

首先，给出原假设和备择假设：

原假设 $H_0:\beta_1=0$

备择假设 $H_1:\beta_1\neq 0$

下面计检验统计量的值：

$$t=\frac{b1\sqrt{\sum(x_i-\bar{x})^2}}{\sqrt{MSE}}$$

接着用求得的统计量的值来计算 p 值， p 值是通过自由度为 $n-2$ 的 t 分布来确定的。

最后一步，根据上述的计算得出结论：

若 p 值小于显著性水平 α，拒绝原假设，接受备择假设，即认为：在显著性水平 α 下，有充足的证据来支持总体中两变量之间的线性关系；

若 p 值大于显著性水平 α，不能拒绝原假设，即认为：在显著性水平 α 下，没有充足的证据来证明总体中两变量之间的线性关系。

（4）对 β_1 的 F 检验

均方误差（MSE）的定义如下：

$$MSE=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}=\frac{SSE}{n-2}$$

即，均方误差等于误差平方和除以它的自由度。

回归均方（MSR）的定义如下：

$$MSR=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}=\frac{SSR}{1}$$

由于 $E(MSR)=\sigma^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$ ， $E(MSE)=\sigma^2$ ，所以可以用这些均值来判断 β_1 的系数是否为零：

若 $\beta_1=0$，那么 MSR 与 MSE 之比将会等于 1；

若 $\beta_1\neq 0$，那么 MSR 与 MSE 之比将会大于 1；

因此，可以用统计量 $\frac{MSR}{MSE}$ 进行假设检验，判断 β_1 是否为零：

首先，给出原假设和备择假设：

原假设 $H_0:\beta_1=0$

备择假设 $H_1:\beta_1\neq 0$

下面计检验统计量的值：

$$F^*=\frac{MSR}{MSE}$$

接着用求得的统计量的值来计算 p 值， p 值是通过分布 $F(1,n-2)$ 来确定的。

最后一步，根据上述的计算得出结论：

若 p 值小于显著性水平 α，拒绝原假设，接受备择假设，即认为：在显著性水平 α 下，有充足的证据来支持总体中两变量之间的线性关系；

若 p 值大于显著性水平 α，不能拒绝原假设，即认为：在显著性水平 α 下，没有充足的证据来证明总体中两变量之间的线性关系。

（5）失拟检验

往往我们每个自变量的观测值不止一个，一般数据不会正好的落在估计的回归线上。导致这种现象发生会有两方面原因：1、回归模型并不能完全描述数据的趋势，说明这个模型有点“失拟”；2、数据存在者随机波动。因此我们做如下分解：

误差平方和SSE=模型失拟+随机误差

判断该模型的好坏，我们得看误差中模型失拟的程度，要是大部分误差都来源于模型的失拟而并非随机误差，此时应该考虑换一个模型。

对于误差的具体分解，我们有如下公式：

残差平方和SSE=失拟平方和SSLF+纯误差平方和SSPE

其中：

$$SSE=\sum_i\sum_j(y_{ij}-\hat{y}_{ij})^2$$

$$SSLF=\sum_i\sum_j(\bar{y}_{ij}-\hat{y}_{ij})^2$$

$$SSPE=\sum_i\sum_j(y_{ij}-\bar{y}_i)^2$$

注意失拟平方和放映的是回归直线与当 $x=x_i$ 时， y_i 均值之间的差距；而纯误差平方和与回归直线没有任何关系。它们都有自由度，其中 SSE 的自由度是 n-2,SSLF 的自由度为 c-2,SSPE 的自由度是 n-c。

基于方差分析的思想，我们构造下面的失拟 F 检验的统计量，首先与前面方差分析类似，用各自的自由度做分母，得出均方误差：

失拟均方： $MSLF=\dfrac{\sum\sum(\bar{y}_{ij}-\hat{y}_{ij})^2}{c-2}=\dfrac{SSLF}{c-2}$

纯误差均方： $MSPE=\dfrac{\sum\sum(y_{ij}-\bar{y}_i)^2}{n-c}=\dfrac{SSPE}{n-c}$

于是我们构造下面的 F 统计量：

$$F^*=\frac{MSLF}{MSPE}$$

下面为失拟 F 检验的步骤：

1) 提出原假设与备选假设

H_0：模型是合理的，不存在失拟；

H_1：模型不合理，存在失拟。

2) 计算 F 统计量的值

$$F^* = \frac{MSLF}{MSPE}$$

本文是借助统计软件 R 来计算的。

3) 利用第二步计算的 F 统计量的值来计算 P 值。此 p 值参考 F(c-2,n-c)分布得到。

4) 得出结论：若 P 值小于显著性水平 α，拒绝原假设，接受备择假设，认为在显著水平 α 时，有充足证据认为此一元模型是失拟的；若 p 值大于显著性水平 α，则不能拒绝原假设，认为在显著性水平 α 时，没有充足证据认为此一元线性模型是失拟的。

（6）因变量均值的置信区间和新值的预测区间估计：

1) 因变量均值的置信区间

置信区间的计算公式如下：

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times (\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2})}$$

其中 $\hat{y}_h$ 是当自变量取 x_h 时，因变量的拟合值，$t_{(\alpha/2,n-2)}$ 是 t 分布的临界值，$\sqrt{MSE \times (\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2})}$ 是拟合的标准差。

2) 因变量新值的预测区间

预测区间的计算公式如下：

$$\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times (1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2})}$$

其中 $\sqrt{MSE \times (1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2})}$ 是预测标准差。

特别注意，上面两个标准差极为相似，但预测标准差比拟合标准差多了项 MSE。

（7）残差分析：

残差：第 i 个样本点观测值 y_i 与估计值 $\hat{y}_i$ 之间的差，记为：

$$e_i = y_i - \hat{y}_i$$

此值是可以观测的。

我们可以用每个点对应的残差来估计未知而又真实存在的误差项。故而残差分析的基本思想就是通过分析能够观测到的残差来判断这些残差是否表现合理，即验证残差是否满足线性、独立性、正态性以及同方差这四个基本假设。下面是分别对四个基本假设的检验方法：

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呈 45 度角的直线上；若不是如此，那就违反了正态性假设。
2) 独立性：你无法从这些图中辨别出因变量值是否相互独立，只能从搜集的数据中来验证。假若你发现数据只从一个个体抽样得来的，那么可能必须要调整模型独立性的假设。
3) 线性：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。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-拟合值点图”可以对回归模型线性进行检验，残差大致分布在 0 上下随机分布，说明 x 与 y 之间存在线性。
4) 同方差性：若满足不变方差假设，那么在残差-拟合值点图中点应该分布在 0 水平线为中心的带型区域内。

二、实验过程

Step 1：选取因变量与自变量

本次实验因变量 y 为州含有的发病数，自变量 x 为州对应的纬度(度)。

Step 2：读取数据

打开 R，在主窗口中直接输入命令，如下所示：

```
setwd("F:/实用回归分析/实验一")
library(readr)
library(ggplot2)
skincancer <- read_table2('实验1数据skincancer.txt')
```

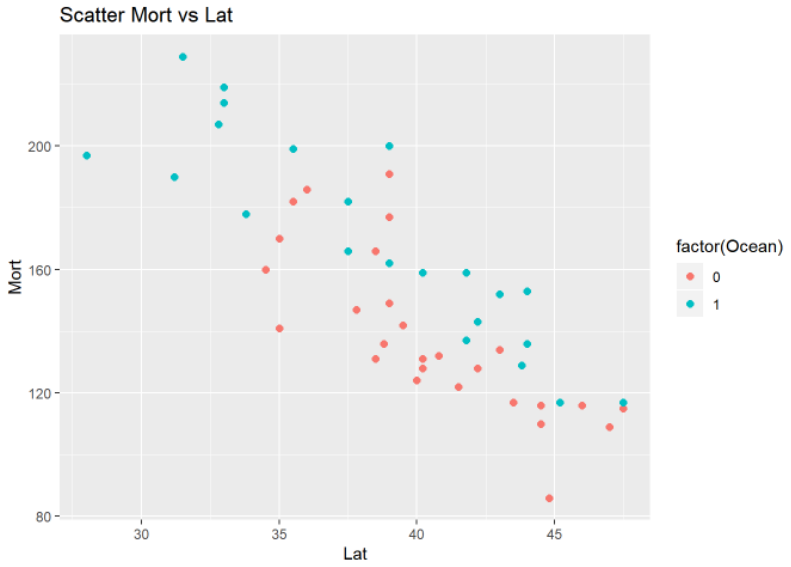
```
## Parsed with column specification:
## cols(
##   State = col_character(),
##   Lat = col_double(),
##   Mort = col_double(),
##   Ocean = col_double(),
##   Long = col_double()
## )
```

其中，setwd()将 R 语言的工作路径更改至数据所在文件夹；使用 readr 包中的 read_table2()函数读取数据，并将其命名为 skincancer，读取的数据集含有 State,Lat,Mort,Ocean,Long 五个变量。

Step 3：绘制数据散点图

为了直观起见，可画一张“散点图”，以x为横坐标，y为纵坐标，每一数据对(x_i, y_i)为x-y坐标系中的一个点，R 语言代码与散点图如下所示：

```
n <- dim(skincancer)[1]
#绘制散点图
ggplot(data = skincancer,mapping = aes(x = Lat,y = Mort,color = factor(Ocean)))+geom_point(size = 2)+
  ggtitle('Scatter Mort vs Lat')
```



从散点图上可以发现，无论是 Ocean 值为 0 还是 1，n 个点基本在一条直线附近，从而可以认为y与x的关系基本上是线性的，而这些点与直线的偏离是由其他一切不确定因素的影响造成的，为此可以构建一元线性回归理论模型，即：

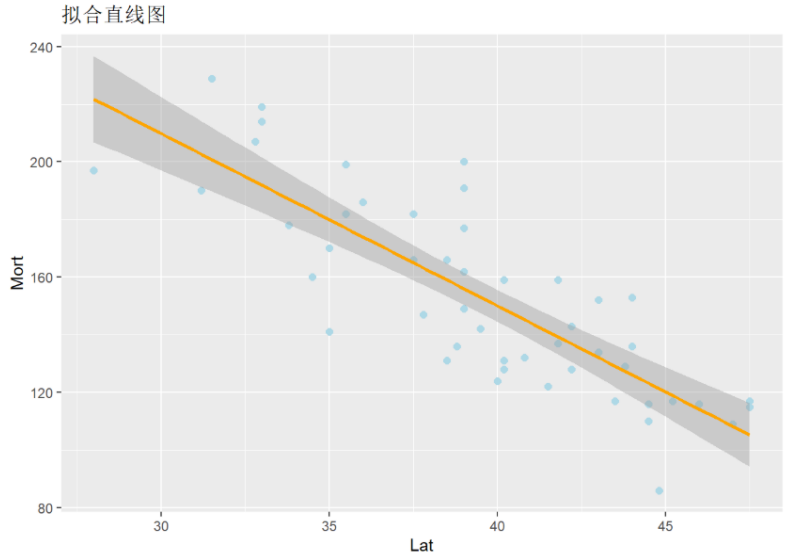
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Step 4：<u>回归参数的估计与显著性检验</u>

本次实验使用最小二乘法估计回归参数，该方法的思想是找到使残差平方和最小的参数作为估计值。在 R 中一元线性模型的拟合可以使用 lm()函数，该函数可以非常方便地求出回归参数$\hat{\beta}_0, \hat{\beta}_1$和作相应的检验。使用 lm 拟合模型，并调用 ggplot 绘制回归直线图，结果如下：

```
linModel <- lm(Mort~Lat,data = skincancer)
ggplot(data = skincancer,aes(x = Lat,y = Mort))+geom_point(size = 2,color = 'lightblue')+geom_smooth(method = 'lm',se = T,color = 'orang
e')+ggtitle('拟合直线图')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
summary(linModel)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = skincancer)
##
## 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
```

在上述操作中，第一行函数 lm()表示作线性模型，第二行 ggplot()表示绘制绘制直线与 95%的置信区间，第三行函数 summary()提取模型的计算结果。

- 在计算结果的第一部分(call)列出了相应的回归模型的公式。
- 在计算结果的第二部分(Residuals)列出的是残差的最小值点、1/4 分位点、中位数、3/4 分位点和最大值点，可以看到残差都集中在[-45,45]区间内。
- 在计算结果的第三部分(Coefficients)中：①Estimate 表示回归方程参数的估计，即 $\hat{\beta}_0, \hat{\beta}_1$。②Std. Error 表示回归参数的标准差。③t value 为t值，需要说明的是，t value 为 Estimate 除以 Std. Error 得到的。④Pr(>|t|)表示p值，即概率$P(t>|T|)$，该值后面还有显著性标记，其中***说明极为显著($p=0$)，**说明高度显著($p<0.001$),*说明显著($p<0.01$)，·说明不太显著($p<0.1$)，没有记号为不显著。由这一部分的结果可以知道，拟合得到的两个参数都是极为显著的拒绝原假设，也即这这两个参数都是显著非 0。
- 在计算结果的第四部分中：①Residual standard error 表示残差的标准差，可以用该值对残差作标准化以找出异常值。②R-squared 的值为 0.6798，说明引入自变量之后，因变量的波动减少了 67.98%，同时说明 x 与 y 之间存在较强的线性相关性，但是也可能存在着能够拟合得更好的非线性关系模型。③F-statistic 表示 F 统计量，其自由度为(1,47)，在此处用以检验$\hat{\beta}_1$是否为 0。

此外，由于我们计算出来的相关系数式基于样本的，只是总体的部分观测值，它只是总体相关系数的一个估计，这个估计值难免存在一定的误差，因此对因变量和自变量的相关系数作显著性检验是非常必要的。相关系数显著性检验的 R 语言代码与结果如下所示：

```
r <-  cor(skincancer$Mort,skincancer$Lat,method = 'pearson')
cor.test(skincancer$Mort,skincancer$Lat,method = 'pearson',alternative = 'two.sided')
```

```
##
##  Pearson's product-moment correlation
##
## data:  skincancer$Mort and skincancer$Lat
## t = -9.9898, df = 47, p-value = 3.309e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8976036 -0.7073128
## sample estimates:
##       cor
## -0.8245178
```

自变量与因变量的积差 Pearson 相关系数为-0.8245，其 95%置信区间为[-0.8976,-0.7073]，

不包含 0 值，说明两者之间高度相关。同时相关系数 t 检验的统计量取值为-9.9898，p 值为 0，因此没有理由接受原假设成立，说明相关系数显著不为 0。

从计算结果可以看出回归方程通过了回归参数的检验、回归方程和相关系数的检验，因此拟合的一元回归方程为：

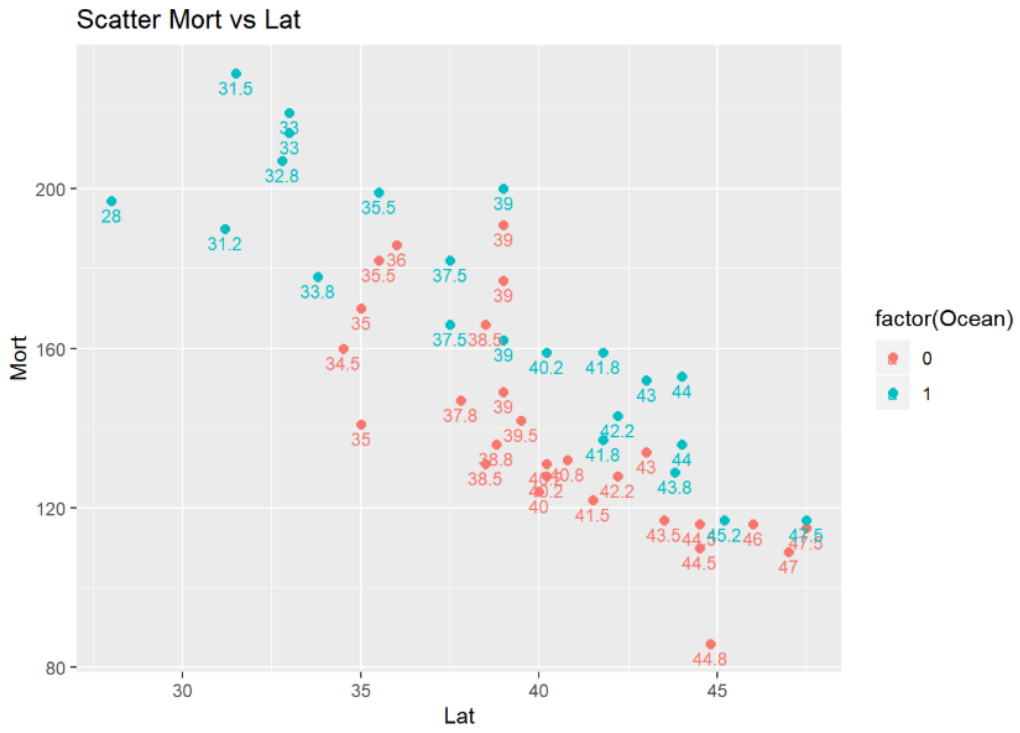
$$y = 389.1894 - 5.9776x + \varepsilon$$

Step 5：<u>失拟检验</u>

失拟检验是一种用来判断回归模型是否可以接受的检验。倘若失拟误差显著大于纯误差，那么久放弃模型；如果前者不显著大于纯误差，那么就可以接受该模型。

观察下图，图中点的标签为点的横坐标值，不难发现，不少点的 x 坐标值重复，也即对于同一个 x，存在着多组观察值，符合失拟检验的要求，也有必要对模型做失拟检验。

```
#绘制带标签的散点图
ggplot(data = skincancer,mapping = aes(x = Lat,y = Mort,color = factor(Ocean)))+geom_point(size = 2)+
  ggtitle('Scatter Mort vs Lat')+geom_text(aes(label=skincancer$Lat),size=3,hjust=0.5,vjust=1.5)
```



将 x 转为因子以保证相同的 x 为一个类别，接着对所有观测到的类别作一元回归，然后调用 R 语言中的 anova()函数，得到结果如下图所示。

```
# 失拟检验
linModel2 <- lm(Mort~factor(Lat),data = skincancer)
anova(linModel,linModel2)
```

```
## Analysis of Variance Table
##
## Model 1: Mort ~ Lat
## Model 2: Mort ~ 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
```

可以将未将 *x* 转为因子时得到的模型(Model 1)记为简模型，转为因子后得到的模型(Model 2)记为全模型。图中，Res.Df 为自由度，RSS 为对应模型的残差平方和，Df 为有序平方和的自由度，Sum of Sq 为有序平方和与其自由度的比值，F 为得到的 *F* 统计量的值，Pr(>F)为在原假设成立的情况下拒绝原假设的最小显著性水平，即 *p* 值，由于 p=0.128>0.05，没有理由拒绝原假设，**模型通过失拟检验，因此可以认为 *E*(*y*)是 *x* 的线性函数**。

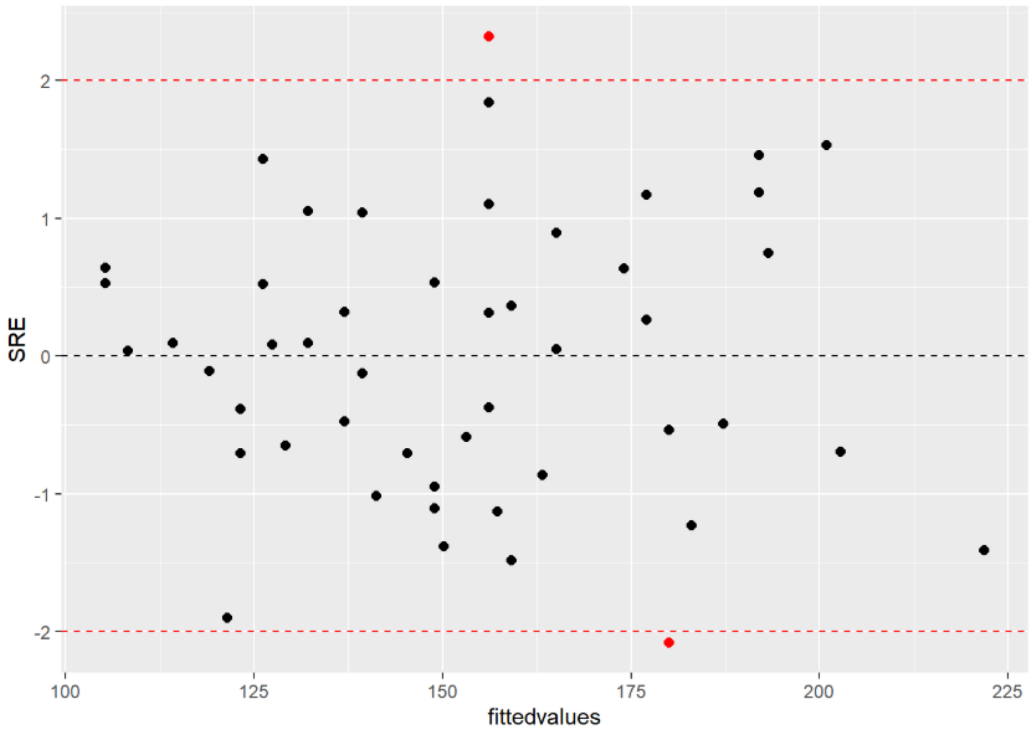
***Step 6*：残差分析**

一个线性回归方程通过了 t 检验或 F 检验，只是表面变量 x 与 y 之间的线性关系式显著的，或者说线性回归方程是有效的，但不能保证数据拟合得很好，也不能排除由于意外原因而导致的数据不完全可靠，比如异常值出现、周期性因素的干扰等。因此，在利用回归方程做分析和预测之前，应该用残差图帮助我们诊断回归效果与样本数据的质量，检查模型是否满足基本假定。

在本次实验的残差分析部分中，主要做了四小块的内容——残差图、QQ 图、直方图以及误差的独立性检验。

```
RSAdata <- data.frame(fittedvalues = predict(linModel), SRE = rstandard(linModel))
RSAdata <- mutate(RSAdata, new = if_else(abs(SRE) > 2, 1, 0))

#残差图
ggplot(data = RSAdata, aes(x = fittedvalues, y = SRE, color = new))+geom_point(size = 2)+
  geom_hline(yintercept = c(-2, 0, 2), linetype = 'dashed', color = c('red', 'black', 'red'))+
  scale_color_gradient(low = 'black', high = 'red', guide = F)
```

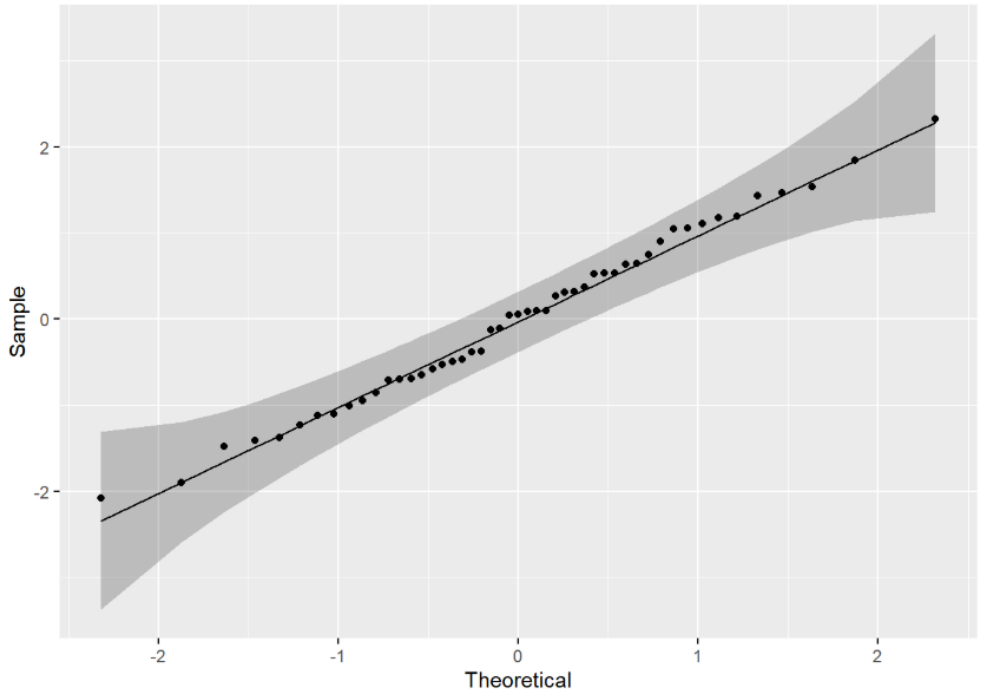


代码中的前两行实现的是将模型的残差转化为学生化残差，这么做的目的是：普通残差的方差不等，直接用它们做判断和比较会带来一定的麻烦，而学生化残差解决了普通残差受单位影响以及方差不等的问题。

调用 ggplot()等函数绘制残差-拟合值散点图，一般认为超过±2 的学生化残差为异常值，即图中在红色虚线以外的红点，因此总共有两个异常值。

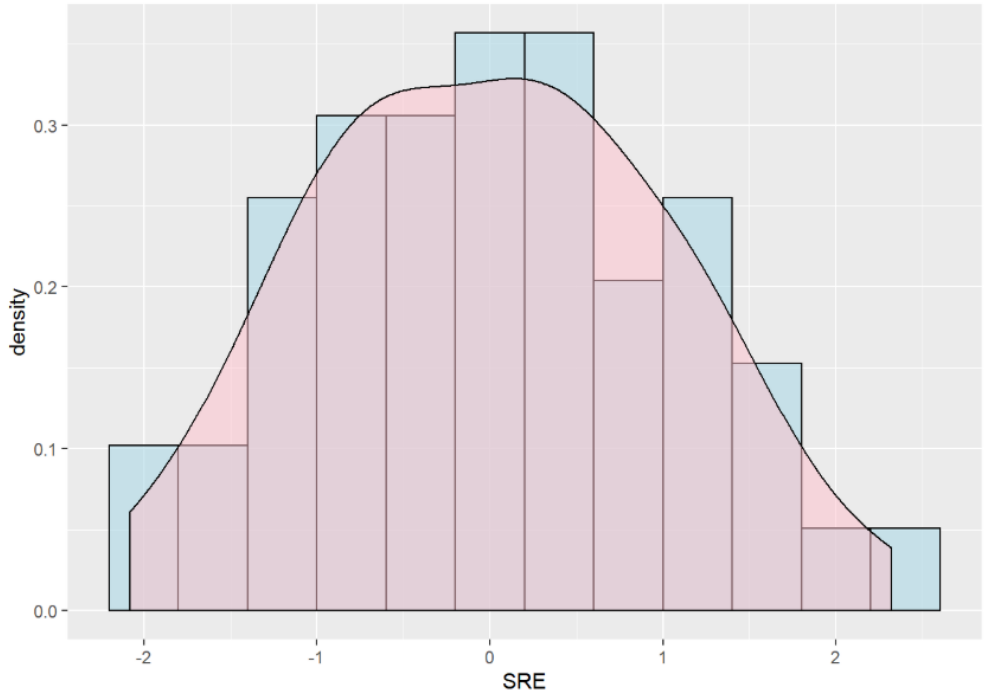
总的来看，几乎所有的点都在 0 附近随机波动，因此可以认为残差满足**同方差性与线性性**假设。接着，通过绘制直方图与 QQ 图检验模型是否满足正态性假设。

```
#QQ图
ggqqplot(RSAdata$SRE,ggtheme = ggplot2::theme_gray())
```



观察 QQ 图，横轴为理论分位数值，纵轴为样本分位数值，图中包含两端散点在内的所有散点几乎都分布在黑色直线，即 $y=x$ 附近，说明样本分位数与理论上正态分布的分位数基本一致，因此从 QQ 图角度来看残差满足**正态性**假设。接着分析一下直方图。

```
#直方图
ggplot(data = RSAdata,aes(x = SRE,y = ..density..))+geom_histogram(bins = 12,fill = 'lightblue',color = 'black',alpha = 0.6)+
  geom_density(fill = 'pink',alpha = 0.5)+geom_line(stat = 'density')
```



横轴为学生化残差值，纵轴为学生化残差对应的密度值，不难发现黑色曲线均值大致为 0，最大值位于(3,3.5)区间内，与标准正态分布的概率密度曲线一致，因此从直方图角度来看也可以认为残差满足**正态性**假设。

接着，检验误差的独立性。判断因变量(或残差)是否相互独立，最好的方法是依据收集数据方式的先验知识。在 R 语言中的 car 包提供了一个可做 Durbin-Watson 检验的函数，能够检测序列的相关性

```
durbinWatsonTest(linModel)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.08680286      2.130728   0.638
##  Alternative hypothesis: rho != 0
```

Autocorrelation 值为-0.0868，说明自相关性很弱；D-W 统计量为 2.32，对应的 p 值为 0.282>0.05，说明在检验水平为 α=0.05 的条件下接受原假设，说明残差无相关性，误差项之间相互独立，满足残差**独立性**假设。

***Step 7*：寻找异常值**

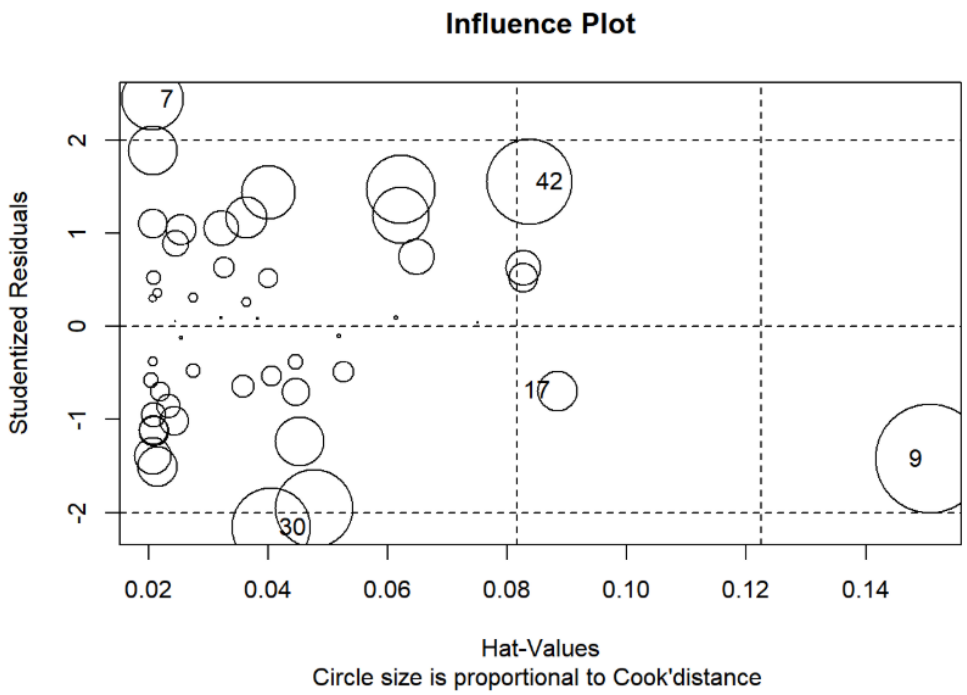
在 R 原因中，可以利用 car 包中的 influencePlot()函数将离群点、杠杆值和强影响点的信息整合到一幅图中。

离群点指那些模型预测效果不佳的观测点，它们通常有很大的、或正或负的残差，正的残差说明模型低估了响应值，负的残差说明高估了响应值。

高杠杆观测点，即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，与相应变量值没有关系。

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变化，那么就需要检测一下数据中是否存在强影响点了。

```
library(carData)
influencePlot(linModel,id=T,main='Influence Plot',
              sub="Circle size is proportional to Cook'distance ")
```



```
##       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
```

- 纵坐标超过+2 或者小于-2 的州可以被认为是离群点，对应的州为 Connecticut(7)与 NewJersey(30)，它们的学生化残差分别为 2.4424 与-2.1612；
- 水平轴超过$2(p+1)/n$或$3(p+1)/n$的州有高杠杆值，其中 p 为自变量的个数，此处为1，n为样本个数，此处为49，对应的州为Wash,D.C.(9)、Kentucky(17)与 Tennessee(42)，它们的帽子统计量取值依次为 0.1507、0.0884 与 0.0836；
- 圆圈大小与影响成比例，圆圈很大的点可能是对模型参数的估计造成的不成比例影响的强影响点，由于所有点对应的 Cook'D 统计量都小于 0.5，因此认为没有强影响点。

综上所述，在构建一元回归模型的时候，因为自变量或因变量异常而产生的异常点有 5 个，它们的序号、学生化残差、帽子统计量以及 Cook'D 统计量可以在上面的图中看到，在此就不再依次列出。

Step 8：分析模型泛化能力

我们在使用最小二乘法计算模型参数时，是通过使得预测误差平方和最小和对响应变量的结实度(R 平方)最大实现的。由于等式只是最优化已给出的数据，所以在新数据集上表现得并不一定好，因此有必要评价回归方程的泛化能力。

在本次实验中，我们使用的是 k 重交叉验证的方法。在 k 重交叉验证中，样本会被分为 k 个子样本，轮流将 k-1 个子样本组合作为训练集，另外 1 个子样本作为保留集。这样会获得 k 个预测方程，记录 k 个保留样本的预测表现结果，然后求其平均值。

在 R 语言中，可以使用 bootsrap 包中的 crossval()函数实现 k 重交叉验证，我们基于该函数写了一个默认为 10 重交叉检验的函数 shrinkage()，该函数创建了一个包含预测变量和预测值的矩阵，可以获得初始 R 平方以及交叉验证的 R 平方。

```
# 模型泛化能力分析
shrinkage <- function(fit, k= 10){
  require(bootstrap)

  theta.fit <- function(x,y){lsfit(x,y)}
  theta.predict <- function(fit,x){cbind(1,x)%*%fit$coef}

  x <- fit$model[,2:ncol(fit$model)]
  y <- fit$model[,1]

  results <- crossval(x,y,theta.fit, theta.predict,ngroup=k)
  r2 <- cor(y,fit$fitted.values)^2
  r2cv <- cor(y,results$cv.fit)^2
  cat("Original R-square = ",r2,'\n')
  cat(k, "Fold Cross-Validated R-square = ",r2cv,"\n")
  cat("Change = ", r2-r2cv,"\n")
}

shrinkage(linModel)
```

```
## Loading required package: bootstrap
```

```
## Original R-square =  0.6798296
## 10 Fold Cross-Validated R-square =  0.6644529
## Change =  0.01537671
```

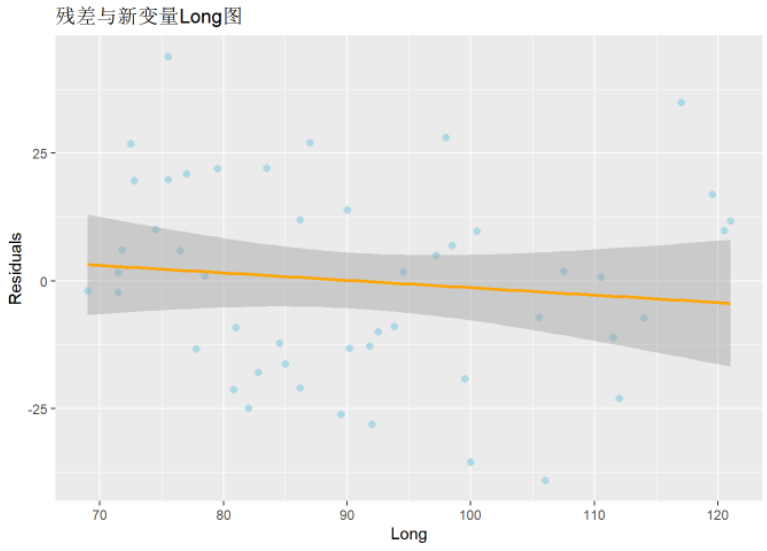
可以看到，基于初始样本的 R 平方为 0.6798，对新数据更好的方差解释率估计是交叉检验后的 R 平方值(0.6645)，两者的差别非常小，因此可以认为**基于初始样本的意愿回归模型具有比较好的泛化能力**。

Step 9： **模型改进措施展望**

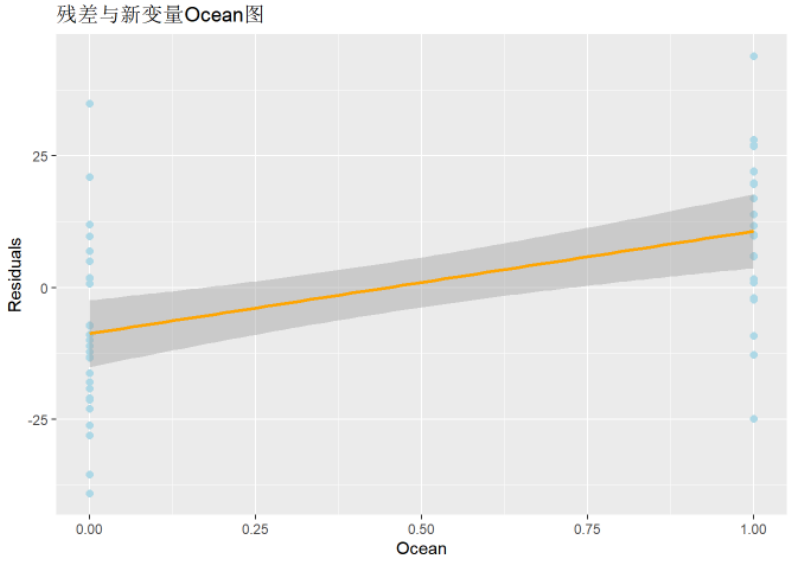
在本次实验中，针对构建好的一元回归模型，我们有以下几个改进方向：

- 删除观测点。在前文中，我们发现了 5 个异常值点。通常来说，删除离群点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。
- 添加新变量。在前文中，我们只选用了 Lat 这一个变量作为自变量，得到的决定系数并未达到 0.7，说明因变量的方差可能并未解释完全，因此可以考虑引入新变量。我们绘制了单独引入 Long 变量或 Ocean 变量得到的残差变量图，结果如下所示：

```
# 引入新变量
ggplot(mapping = aes(x = skincancer$Long, y = linModel$residuals)) + geom_point(size = 2,color = 'lightblue')+
  ggtitle('残差与新变量Long图') + geom_smooth(method = 'lm',color = 'orange')+xlab('Long')+ylab('Residuals')
```




```
ggplot(mapping = aes(x = skincancer$Ocean, y = linModel$residuals)) + geom_point(size = 2,color = 'lightblue')+
  ggtitle('残差与新变量Ocean图') + geom_smooth(method = 'lm',color='orange')+xlab('Ocean')+ylab('Residuals')
```



观察以上的第一幅图，如果引入 Long 变量的话，它对残差的解释可以说比较少，拟合出来的直线几乎为水平直线，说明 Long 变量可以不考虑引入。

观察以上的第二幅图，如果引入 Ocean 变量的话，从总体水平来看，拟合直线有明显的倾斜，说明 Ocean 取 0 与 1 时残差有区别，即 Ocean 能够解释部分残差的方差信息，因此在之后的分析中可以考虑引入该变量。

Step 10：预测新值及预测区间

当训练好模型之后，我们常常需要将模型应用到一些尚未解决的问题上，如如果我们已知某一地的维度，该地的 Mort 值为多少。这个在 R 中实现起来比较容易，直接使用 predict()函数即可。同时可以通过设置 interval 参数为'prediction'或'confidence'选择预测新值或其区间。

```
# 新值预测
newdata <- data.frame(Lat=40)
y1 <- predict(linModel,newdata,interval = 'prediction',level = 0.95)
y2 <- predict(linModel,newdata,interval = 'confidence',level = 0.95)
y1
```

```
##        fit     lwr      upr
## 1 150.0839 111.235 188.9329
```

```
y2
```

```
##        fit      lwr      upr
## 1 150.0839 144.5617 155.6061
```

如上图所示，在纬度为 40 时，得到预测的 Mort 值为 150.0839，它的 95%的置信区间为[144.5617,155.6061]，95%的预测区间为[111.235,188.9329]。

教师签字_____

第四部分　实验总结

通过本次上机实验，首先最大的收获是学会了使用 R 软件，又进一步了解了如何使用 R 软件进行一元线性回归，并对拟合结果进行相关系数检验、t 检验、F 检验、失拟检验与残差分析，并对新值进行预测。在本次实验的过程中，我学习到了很多有关数据处理与数据检验的知识，同时对线性回归有了更加理解与认识，以下是我对这些收获的总结：

1、并不是所有的问题都适合进行线性回归，在进行线性回归之前我们要先进行预判，我们可以通过画散点图的方式观察其是否大致符合线性关系，我们讨论的问题要有意义，回归方程的选择要符合实际需要。

2、拟合都是在一定范围内进行的，即在我们处理的数据的范围内，不能把我们得到的回归方程任意扩大范围。比如我们得到的纬度与死亡人数的关系，当纬度大大超过其数据范围，其结果可能就不太准确，因此我们在进行数据预测时，应当尽量将自变量控制在我们所处理的数据范围之类，以提高预测的精确度。

3、有时候通过散点图判断线性关系并不准确，有些数据看起来大致是一条直线，但并不适合进行线性回归，因此在得到线性回归模型之后，我们需要对拟合的结果进行多重检验，以判断模型的效果。

4、一些异常数据的存在会很大程度上影响模型的拟合度，我们可以通过画残差图和 QQ 图的方式找到它们，如在本次实验残差图中标准化残差值大于 2 和小于-2 的两个点（QQ 图中落在置信区间带外的两个点）便是异常数据。