****

课程设计报告书

**回归分析课程设计**

**学院 数学学院**

**专业 数学与应用数学（统计学）**

**学生姓名 邓睿**

**学生学号 202130320342**

**指导教师 王绍臣**

**课程编号 040102511**

**课程学分 2**

**起始日期 2023.12.22**

|  |  |
| --- | --- |
| 教  师  评  语 | 教师签名：  日期： |
| 成  绩  评  定 |  |
| 备  注 |  |

**目录**

[回归分析课程设计 1](#_Toc157711666)

[一、基本操作 1](#_Toc157711667)

[1.1代码 1](#_Toc157711668)

[1.2运行结果及说明 5](#_Toc157711669)

[二、应用实例 10](#_Toc157711670)

[2.1背景介绍 10](#_Toc157711671)

[2.2问题动机 10](#_Toc157711672)

[2.3数据初步统计性描述 10](#_Toc157711673)

[2.4模型及结论和建议 13](#_Toc157711674)

[三、理论背景 21](#_Toc157711675)

[3.1模型建立 21](#_Toc157711676)

[3.2统计分析 23](#_Toc157711677)

[3.3模型选择 26](#_Toc157711678)

[四、课程设计总结 27](#_Toc157711679)

[附录 28](#_Toc157711680)

# 回归分析课程设计

## 一、基本操作

### 1.1代码

|  |
| --- |
| data("faithful")  summary(faithful) #summarizes data  library(ggplot2)  library(car)  library(IMTest)  library(openxlsx)  library(stats)  dev.off()  p <- ggplot(data = faithful, aes(x = waiting, y = eruptions)) + geom\_point()  show(p)  R\_pearson <- cor.test(faithful$eruptions,faithful$waiting,method = 'pearson')  model <- lm(waiting ~ eruptions, data = faithful)  conf\_interval <- confint(model, level = 0.95) # 计算系数95%置信区间  print(conf\_interval)# 打印置信区间  hist(faithful$waiting, breaks = 20, main = "等待时间直方图", xlab = "waiting")# 绘制直方图  boxplot(faithful$waiting, main = "等待时间箱线图", xlab = "waiting", col = "blue")  # 绘制箱线图  mylm <- function(data, formula, newdata = NULL) {  # 拟合线性模型  fit <- lm(formula, data = data)    # 提取系数估计  beta0 <- fit$coefficients[1]  beta1 <- fit$coefficients[2]    # 计算95%置信区间  confint <- confint(fit)    # 如果提供了新数据，则生成预测值  predictions <- if (!is.null(newdata)) {  predict(fit, newdata = newdata)  } else {  NULL  }    # 返回结果  return(list(beta0 = beta0, beta1 = beta1, confint = confint, predictions = predictions))  }  # 使用自定义函数拟合模型  new\_waiting\_times <- data.frame(waiting = seq(50, 80, length.out = 10))  result <- mylm(faithful, eruptions ~ waiting, newdata = new\_waiting\_times)  # 使用R自带函数拟合线性模型并提取参数估计和置信区间  base\_model <- lm(eruptions ~ waiting, data = faithful)  base\_coef <- coef(base\_model)  base\_conf\_interval <- confint(base\_model, level = 0.95)  summary(base\_model)  # 创建散点图  plot(faithful$waiting, faithful$eruptions, main = "散点图", xlab = "等待时间", ylab = "喷发次数")  # 添加拟合直线  abline(a=result$beta0,b=result$beta1, col = "red",lwd=5)  abline(base\_model, col="blue",lwd=2)  # 计算RSS（回归平方和）  rss <- sum((faithful$eruptions - (result$beta0 + result$beta1 \* faithful$waiting))^2)  # 计算TSS（总平方和）  tss <- sum((faithful$eruptions - mean(faithful$eruptions))^2)  # 计算SSE（平方和回）  sse <- sum((result$beta0 + result$beta1 \* faithful$waiting - mean(faithful$eruptions))^2)  mynewlm <- function(x, y, newdata) {  # 1. 计算β0和β1的最小二乘估计  fit <- lm(y ~ x)  beta0 <- coef(fit)[1]  beta1 <- coef(fit)[2]    # 2. 计算给定新观测值的拟合值和预测区间  newdata <- data.frame(x = newdata)  predict\_fit <- predict(fit, newdata = newdata, interval = "confidence")    # 3. 返回结果  return(list(beta0 = beta0, beta1 = beta1, fitted\_value = predict\_fit[, 1],  lower\_ci = predict\_fit[, 2], upper\_ci = predict\_fit[, 3]))  }  # 提取自变量和应变量  x <- faithful$waiting  y <- faithful$eruptions  # 生成新观测值  new\_x <- 85 + rnorm(272)  # 使用mynewlm函数拟合模型并获得结果  my\_result <- mynewlm(x, y, new\_x)  # 使用predict函数拟合模型并获得结果  model <- lm(eruptions ~ waiting, data = faithful)  result\_pred <- predict(model, newdata = data.frame(waiting = new\_x))  # 对比结果  mynewml <- data.frame(my\_result$fitted\_value)  predict <- data.frame(result\_pred)  # 将数据框导出为Excel文件  write.xlsx(mynewml, "data.xlsx", sheetName = "Sheet1")  write.xlsx(predict, "data1.xlsx", sheetName = "Sheet1")  # 定义函数mymlm()  mymlm <- function(formula, data) {  # 拟合线性模型  fit <- lm(formula, data = data)    # 提取系数估计  beta0 <- fit$coefficients[1]  beta1 <- fit$coefficients[2]    # 计算检验统计量  # 使用OLS估计的β1的标准误差计算z统计量  statistic <- abs(beta1 - 0.05) / (fit$coefficients[2] %o% fit$coefficients[2])    # 输出检验结果  cat("Test statistic:", statistic, "\n")  cat("p-value:", 2 \* (1 - pnorm(abs(statistic))), "\n") # 使用正态分布的累积分布函数计算p值  }  # 对faithful数据集进行检验  mymlm(eruptions ~ waiting, data = faithful)  # 加载必要的库  library(MASS)  # 使用lm()函数进行线性回归拟合  fit <- lm(eruptions ~ waiting, data = faithful)  # 计算学生化残差和Cook统计量  studentized\_residuals <- residuals(fit) / sqrt(diag(vcov(fit)))  cook\_stats <- cooks.distance(fit)  studentizedresiduals <- data.frame(residuals(fit) / sqrt(diag(vcov(fit))))  cookstats <- data.frame(cooks.distance(fit))  # 将数据框导出为Excel文件  write.xlsx(studentizedresiduals, "studentized\_residuals.xlsx", sheetName = "Sheet1")  write.xlsx(cookstats, "cook\_stats.xlsx", sheetName = "Sheet1")  # 绘制散点图  plot(cook\_stats ~ studentized\_residuals,  xlab = "Studentized Residuals",  ylab = "Cook's Statistics",  main = "Cook vs. Studentized Residuals Plot",  col = "blue")  # 计算残差  residuals <- residuals(lm(eruptions ~ waiting, data = faithful))  # 绘制残差图  ggplot(data = faithful, aes(x = waiting, y = residuals)) +  geom\_point() +  theme\_minimal() +  labs(x = "Waiting Time", y = "Residuals", title = "Residuals vs. Waiting Time")  # 绘制残差直方图  hist(residuals, data = faithful, main = "Residuals Histogram", xlab = "Residuals") |

### 1.2运行结果及说明

通过使用ggplot绘制faithful数据集两个变量的散点图，如图1所示。

图 1 waiting，eruptions 散点图

使用皮尔逊法得到相关系数为34.08904，并用confint得到其95% 置信区间为[31.20069,35.74810]。

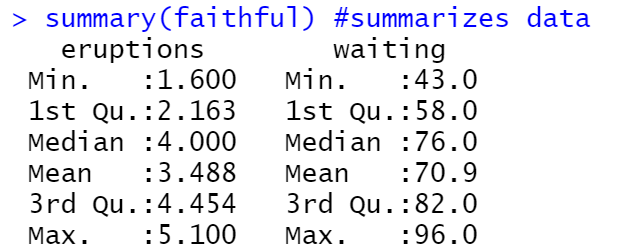
运用summary命令得到faithful数据集的一些描述性统计量，如图2所示。

图 2 描述性统计量

为了更直观得到数据的特性，运用直方图及箱线图将数据进行可视化，如图3所示。

|  |  |
| --- | --- |
|  |  |

图 3 数据可视化

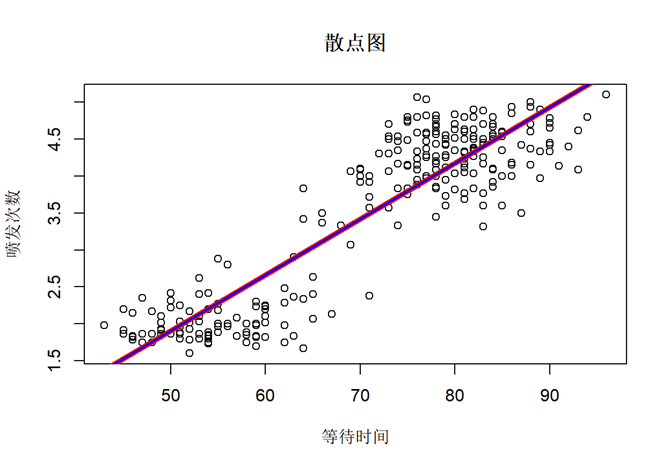
建立线性模型，构建函数mylm（），求解得，的最小二乘估计分别为，，相应的95%置信区间分别为[-2.1893,-1.5587],[0.0713,0.0800]。使用abline指令将自己构建的线性模型以及使用R语言自带的lm函数拟合成的线一同加在散点图上，其中红色的线为自己构建的线性模型，蓝色的线为lm函数拟合成的，如图4所示。

图 4 拟合直线与散点图

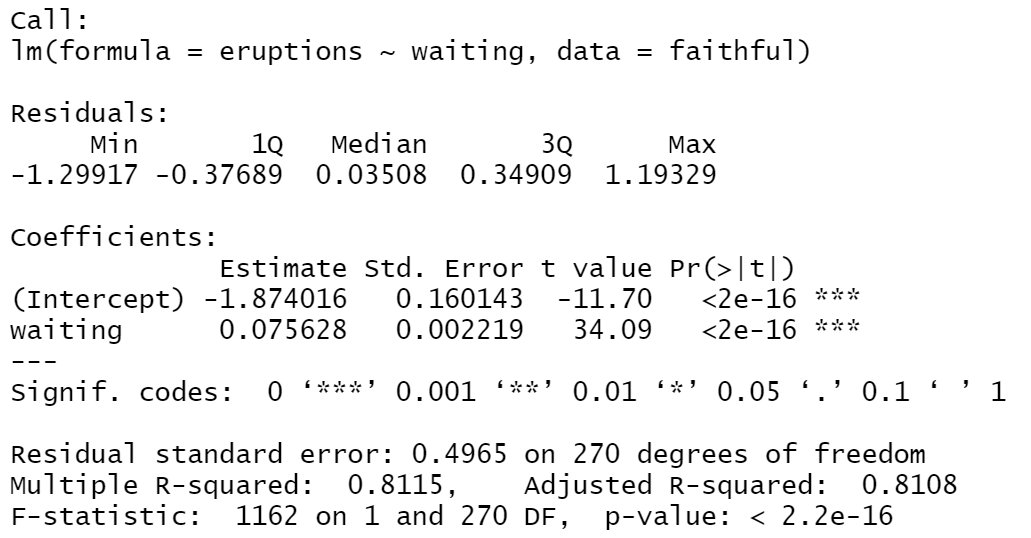
根据R自带函数lm()建立线性模型，使用summary得到以下结果，如图5所示。其中Residuals为该模型残差的最小值、第一四分位数、中位数、第三四分位数和最大值。Coefficients为模型的系数估计值，截距（Intercept）的系数为-1.874016，等待时间（waiting）的系数为0.075628。t value为每个系数的t统计量的值。Pr(>|t|)为t统计量的p值，用于检验系数的显著性，由此可以得到模型的整体显著性非常高（p值 < 2.2e-16），等待时间对喷发次数有显著的正影响（p值 < 2e-16）。Residual standard error为残差的标准误差。Multiple R-squared为多重决定系数，表示模型解释的变异百分比。Adjusted R-squared为调整后的多重决定系数，为0.8101比较接近1，由此可见模型拟合效果较好。

图 5 线性模型摘要信息

根据定义求解RSS，TSS，SS回，其中RSS≈66.56，TSS≈353.04，SS回≈286.48，由数值易验证TSS=RSS+ SS回。

用waiting=85+作为新的观测值，其中N(0,1)。将，的最小二乘估计分别为，，新的观测值waiting带入，得到每个waiting值对应的预测值。构建函数mynewml()用于求在自己构建的模型下新的观测值的预测值，并于R自带的predict()预测的预测值进行对比。部分预测值的比较见表1，其余部分见附录。

表 1两种方法预测值部分比较

|  |  |
| --- | --- |
| mynewml()预测值 | predict()预测值 |
| 4.425504812 | 4.425504812 |
| 4.463547279 | 4.463547279 |
| 4.536149509 | 4.536149509 |
| 4.53151439 | 4.53151439 |
| 4.421704784 | 4.421704784 |
| 4.592882866 | 4.592882866 |
| 4.477108943 | 4.477108943 |
| 4.542405916 | 4.542405916 |
| 4.539493237 | 4.539493237 |
| 4.607044192 | 4.607044192 |
| 4.549369547 | 4.549369547 |

为检验=0.05，使用OLS估计的β1的标准误差计算z统计量，求解得统计量的值为4.480734，p值为7.438664e-06，由于p值非常小因此拒绝原假设，即=0.05的概率很小。

运用R自带的函数求学生化残差和cook统计量，部分结果如表2所示，完整结果见附录。

表 2 学生化残差与cook统计量比较

|  |  |
| --- | --- |
| 学生化残差 | cook统计量 |
| -3.125899707 | 0.002560285 |
| -184.7579646 | 0.003256009 |
| -2.431897909 | 0.001199359 |
| -239.7596792 | 0.003048608 |
| -0.133377976 | 7.18583E-06 |
| 269.3115554 | 0.006426575 |
| -0.507317085 | 0.000129858 |
| -430.1743332 | 0.014345489 |
| -0.206123882 | 2.61957E-05 |
| -92.11438865 | 0.000657781 |

通过观察观察waiting和eruptions之间的散点图（图1），我们可以看到这两个变量之间存在明显的线性关系，因此可以初步判断线性关系假设成立。残差图（图6（a））显示了实际观测值与模型预测值之间的差异。从残差图中，我们没有看到明显的模式或集群，因此可以初步判断误差的独立性假设成立。同时我们可以看到残差的波动似乎没有随着waiting的增加而明显变化，因此可以初步判断同方差性假设成立。从残差的直方图（图6（b））中，我们可以看到残差分布似乎接近正态分布。因此，我们可以判断误差的无偏性和正态分布假设成立，故符合Gauss-Markov 假设。残差图和残差直方图如图6所示。

|  |  |
| --- | --- |
| （a）残差图 | （b）残差直方图 |

图 6 残差图和残差直方图

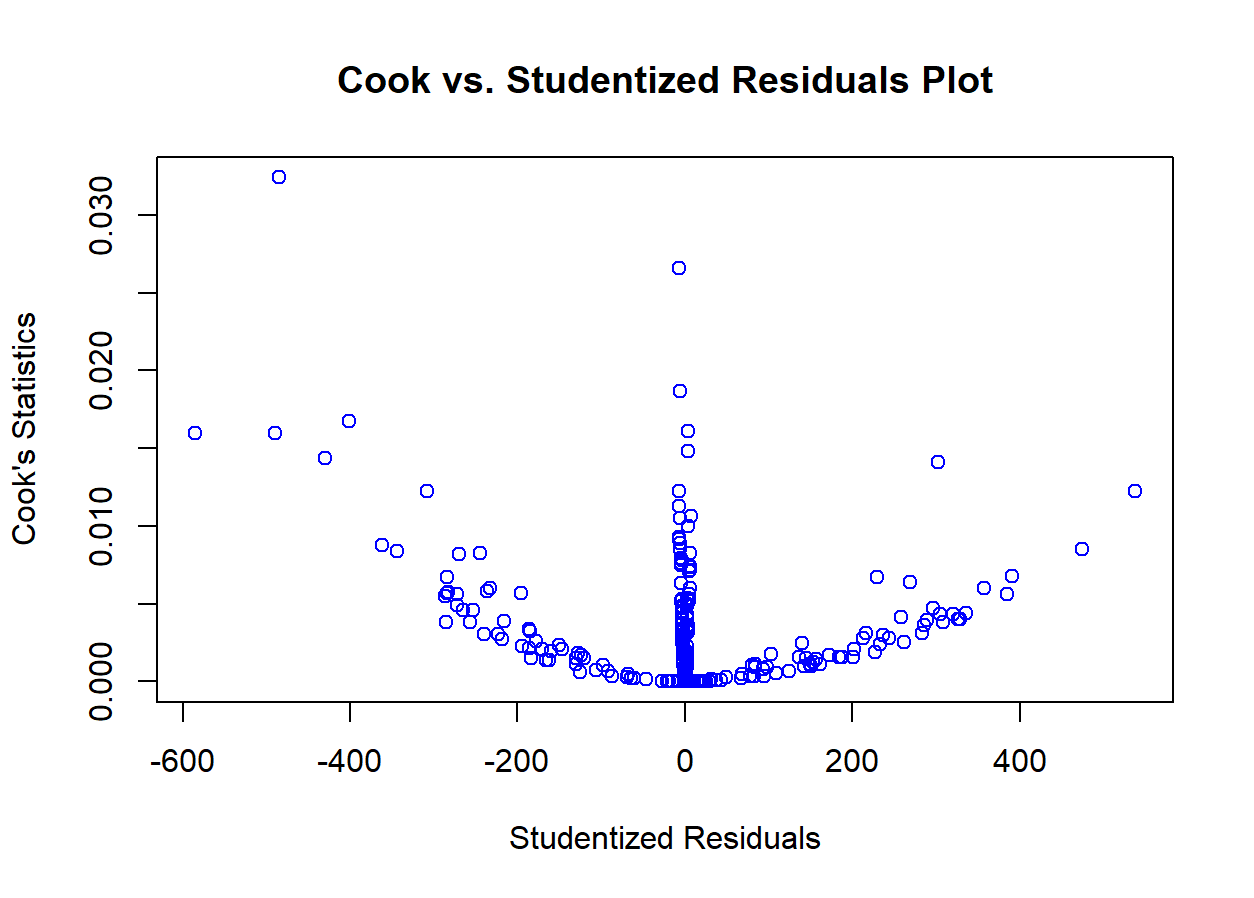
运用R自带的函数计算学生化残差和Cook统计量。若cook统计量过高，则该观察值可能会对模型造成较大影响，可能是异常值。学生化残差是残差与它的标准误的比值。如果学生化残差的观察值远离中心线，则可能是异常值。绘制出每个点的学生化残差与cook统计量的散点图（图7），从图中可以看出，大部分的点都聚集在（0,0）附近，有少部分点偏离原点比较多，对于学生化残差和cook统计量均较大的视为异常值，为了更好的拟合效果，可以在进行拟合的时候去掉偏离比较远的数据。尽管存在一些异常值，但它们不会严重影响回归模型的拟合效果。因此，无需对数据进行变换。

图 7 学生化残差与cook统计量散点图

## 二、应用实例

### 2.1背景介绍

Affair数据集是一个用于研究婚外情的数据库，包含了与婚外情相关的各种信息，如婚姻状况、情感状态、个人特征以及社会经济背景等。该数据集的目的是为了帮助研究者了解婚外情的成因、影响和后果，从而为预防和解决婚外情问题提供科学依据。该数据集是通过调查收集而来的，包含了多个国家和地区的受访者，具有一定的全球性和多样性。数据集中的每个受访者都提供了自己的个人信息、婚姻状况、情感状态、家庭背景、教育经历等方面的信息，以及与婚外情相关的经历和感受。这些信息有助于研究者了解不同因素对婚外情的影响，以及婚外情对个人和家庭的影响。通过研究Affair数据集，可以深入了解婚外情的成因和影响因素，探究婚外情对个人和家庭的影响，为制定有效的预防和干预措施提供科学依据。同时，该数据集还可以用于比较不同国家和地区在婚外情问题上的异同，为国际间的婚姻家庭研究和政策制定提供参考。

### 2.2问题动机

对affair数据集进行广义线性拟合的问题动机在于探究多个因素对婚外情的影响，并描述这些因素与婚外情之间的非线性关系。具体来说，通过广义线性模型，我们可以更好地理解个人特征、婚姻状况、情感状态、家庭背景、教育经历等对婚外情的影响程度和方式。这种深入的了解有助于我们为预防和解决婚外情问题提供更有针对性的建议，并为比较不同国家和地区在婚外情问题上的异同提供科学依据。同时，广义线性模型还可以处理多种类型的因变量，并考虑不同因素之间的交互作用和效应大小，从而更全面地分析婚外情问题。这种分析不仅有助于政策制定者制定更有效的预防和干预措施，还可以为学术研究者提供强大的工具来探索和分析婚外情的多维度影响因素。总的来说，对affair数据集进行广义线性拟合可以为我们深入了解婚外情问题提供科学依据，为预防和解决这一问题提供更科学的指导。

### 2.3数据初步统计性描述

Affairs数据集中的变量涵盖了多个方面，包括个人特征、婚姻状况、情感状态、家庭背景、教育经历等，这些因素都可能对是否发生婚外情产生影响。通过分析这些变量之间的关系，可以深入了解婚外情的成因和影响因素，探究婚外情对个人和家庭的影响，为预防和解决婚外情问题提供科学依据。

表 3 Affairs数据集（部分）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| affairs | gender | age | yearsmarried | children |
| 0 | male | 37 | 10 | no |
| 0 | female | 27 | 4 | no |
| 0 | female | 32 | 15 | yes |
| 0 | male | 57 | 15 | yes |
| 0 | male | 22 | 0.75 | no |
| 0 | female | 32 | 1.5 | no |
| religiousness | education | occupation | rating | religiousness |
| 3 | 18 | 7 | 4 | 3 |
| 4 | 14 | 6 | 4 | 4 |
| 1 | 12 | 1 | 4 | 1 |
| 5 | 18 | 6 | 5 | 5 |
| 2 | 17 | 6 | 3 | 2 |
| 2 | 17 | 5 | 5 | 2 |

查看响应变量的各个取值的占比情况，结果见表4。

表 4 变量affairs分布情况

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 取值 | 0 | 1 | 2 | 3 | 7 | 12 |
| 频率 | 451 | 34 | 17 | 19 | 42 | 38 |

结果显示，有婚外情的有出现过1次，2次等等多次的，他们都属于属于婚外情。对于这个数据集，我们关注是否出轨，即这是一个二值型结果（出轨过/从未出轨）。因此，我们接下来将'affaris'特征转化为二值型因子'ynaffair'，该二值型因子即可以作为Logistic回归的结果变量。观察数据，有婚外情的数量与没有婚外情的数量差异不会太远，则不需要对分组进行调整修改，即运用这些数据建立拟合模型合理。运用summary函数对affairs数据集得到一些统计性描述，见表5。

表 5 affairs数据集统计性描述

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| affairs | gender | age | yearsmarried | children |
| Min. : 0.000 | female:315 | Min. :17.50 | Min. : 0.125 | no :171 |
| 1st Qu.: 0.000 | male:286 | 1st Qu.:27.00 | 1st Qu.: 4.000 | yes:430 |
| Median : 0.000 |  | Median :32.00 | Median : 7.000 |  |
| Mean : 1.456 |  | Mean :32.49 | Mean : 8.178 |  |
| 3rd Qu.: 0.000 |  | 3rd Qu.:37.00 | 3rd Qu.:15.000 |  |
| Max. :12.000 |  | Max. :57.00 | Max. :15.000 |  |
| religiousness | education | occupation | rating | ynaffair |
| Min. :1.000 | Min. : 9.00 | Min. :1.000 | Min. :1.000 | No :451 |
| 1st Qu.:2.000 | 1st Qu.:14.00 | 1st Qu.:3.000 | 1st Qu.:3.000 | Yes:150 |
| Median :3.000 | Median :16.00 | Median :5.000 | Median :4.000 |  |
| Mean :3.116 | Mean :16.17 | Mean :4.195 | Mean :3.932 |  |
| 3rd Qu.:4.000 | 3rd Qu.:18.00 | 3rd Qu.:6.000 | 3rd Qu.:5.000 |  |
| Max. :5.000 | Max. :20.00 | Max. :7.000 | Max. :5.000 |  |

从以上摘要可以看到，男性受访者有286人(占整体受访者的48%)，有子女的受访者有430人(占整体受访者的72%)，受访者的平均年龄为32.5岁。此外，我们发现，451名受访者声称在过去的一年里没有婚外情。这意味着25%的受访者有婚外情，最大的报告数字是12。总而言之，6%的受访者每个月有1次出轨。

对各个变量进行求Pearson相关系数，由于只关注ynaffair与其他变量之间的关系，故只关心ynaffair变量与其他变量的相关系数，结果如表6所示。

表 6 affairs变量与其他变量的相关系数

|  |  |  |  |
| --- | --- | --- | --- |
| gender | age | Yearsmarried | children |
| 0.0437638 | 0.105396 | 0.16711917 | 0.1325887 |
| religiousness | education | occupation | rating |
| -0.13957892 | 0.0337891 | 0.038417120 | -0.2573273 |

由相关系数可见gender、education、occupation对是否有婚外情的影响比较不显著，断言在模型中若包含这三个变量，拟合效果可能会没那么好，因此在构建模型时可以考虑将其去掉。

### 2.4模型及结论和建议

广义线性模型（Generalized Linear Models，GLM）是一类统计模型，它是在线性模型的基础上进行了扩展，以适应非线性的数据分布情况。广义线性模型能够拟合各种类型的数据，包括离散数据和连续数据，以及计数数据和比例数据等。在广义线性模型中，因变量（响应变量）和自变量（解释变量）之间的关系被假定为线性关系，但这种关系的解释形式可以进行调整，以适应不同的数据类型和分布情况。通过选择适当的链接函数（link function），广义线性模型能够模拟各种不同的数据分布特征。

这个模型被设计用来探索因变量ynaffair与多个自变量之间的关系，这些自变量包括：性别（gender）、年龄（age）、婚姻持续年数（yearsmarried）、是否有孩子（children）、宗教信仰程度（religiousness）、教育程度（education）、职业（occupation）和婚姻满意度评分（rating）。运用R语言自带的函数glm()来构建一个广义线性模型（Generalized Linear Model，简称GLM）。该模型假设这些自变量和因变量之间的关系是线性的，即可以用一系列的线性组合来预测或解释ynaffair的变化。我们可以使用广义线性模型来分析婚姻不忠行为的数据。因变量ynaffair表示是否发生婚姻不忠行为，是一个二分类变量（0表示没有发生，1表示发生了）。自变量包括性别、年龄、婚姻持续年数、是否有孩子、宗教信仰程度、教育程度、职业和婚姻满意度评分等。

在拟合广义线性模型时，我们假设因变量与自变量之间的关系是线性的，并使用逻辑回归的形式来拟合数据。具体来说，我们使用二项分布的分布族，并指定连接函数为逻辑连接函数（logit）。这样，我们可以通过拟合模型来估计自变量对因变量的影响，以及预测给定自变量值时因变量的概率分布。

在处理Affairs数据集时，首先需要加载数据框并查看数据的基本情况。然

后，对因变量进行编码，将二分类变量转换为因子型变量。接下来，使用glm()函数拟合广义线性模型，指定因变量和自变量，并选择二项分布的分布族。最后，通过模型拟合的结果分析自变量对因变量的影响，以及模型的预测精度和泛化能力。首先对所有因变量建立模型，得到结果如图8。

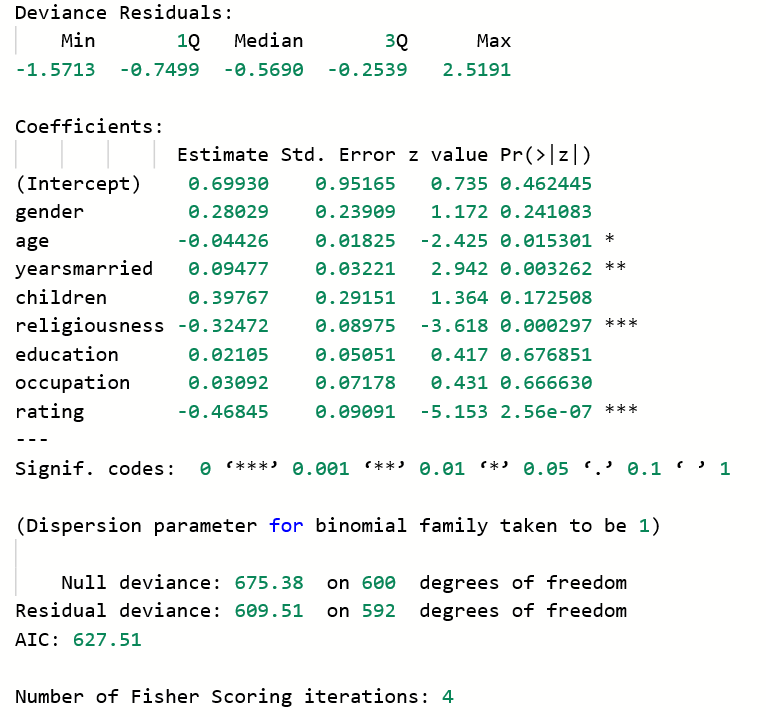
回归系数是统计学中的重要概念，它用于描述当其他预测变量保持不变时，某一预测变量的变化对响应变量影响的程度。在这个问题中，当年龄变化一单位时，出现婚外情的可能性减少了0.04426。这个数值就是回归系数，它表示了年龄与出现婚外情之间的关联性。

图 8 运行结果

从回归系数的分析中，我们可以看到与出现婚外情有正相关关系的因素包括性别、是否有孩子、结婚年份、学历和职业。这意味着这些因素的出现增加了出现婚外情的可能性。而年龄、信仰和婚姻满意评分度与出现婚外情有负相关关系，即这些因素的出现减少了出现婚外情的可能性。通过观察回归系数的p值，我们可以判断这些预测变量对方程的贡献是否显著。在这次分析中，性别、是否有孩子、学历和职业的p值都较大，说明这些因素对出现婚外情的预测贡献并不显著，即它们无法显著地改变方程的结果。为了更好地理解方程，我们原假设所有预测变量都对方程有贡献，并且它们的贡献是相同的。然而，由于某些预测变量的贡献并不显著，我们无法拒绝这些参数为0的假设。为了提高方程的精度和解释性，我们决定去除这些不显著的因子，并重新拟合方程。运行结果被整理成图9并呈现出来。

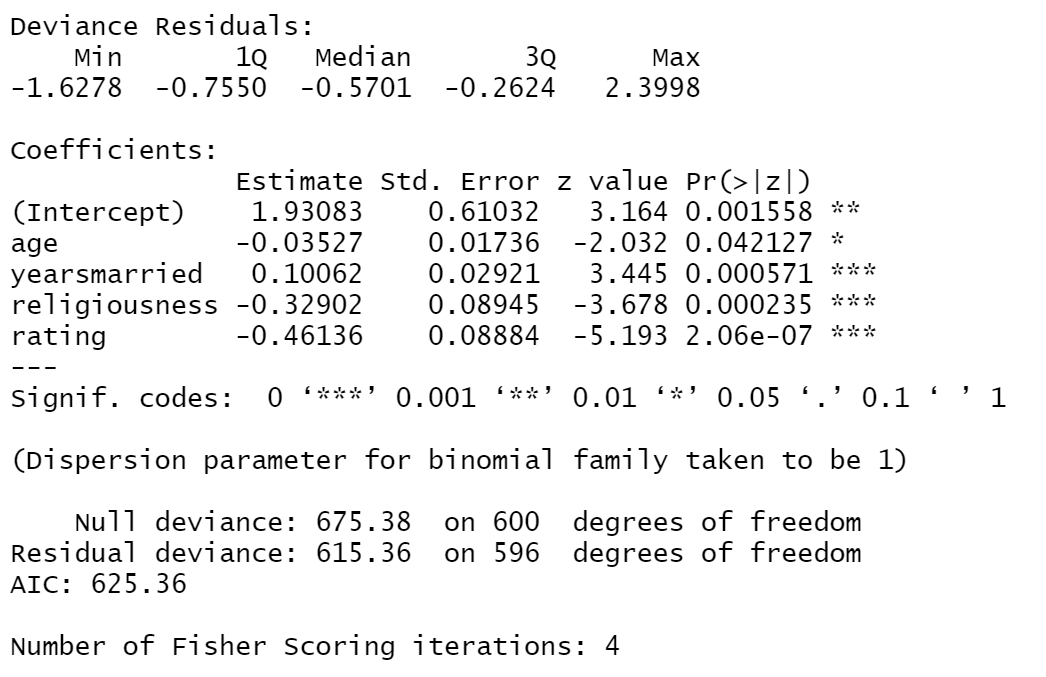
基于第二个模型的回归系数，我们可以深入探讨各个因素对发生婚外情几率的影响。首先，我们发现结婚年限的回归系数为正，这意味着随着结婚时间的增加，发生婚外情的几率也在逐步上升。这一发现可能与夫妻间随着时间推移可能出现的情感冷淡、沟通不畅或生活琐事等问题有关，这些问题可能增加出轨的风险。另一方面，年龄、宗教信仰和幸福感的自我评价的回归系数为负，表明这些因素与发生婚外情的几率呈负相关。年龄的增长可能伴随着成熟和家庭责任感，减少了寻求婚外刺激的可能性。有强烈宗教信仰的人可能更重视婚姻的神圣性和忠诚度，从而降低了出轨的倾向。而婚姻幸福感的提高，可能意味着夫妻间的情感联系更为紧密，降低了出轨的风险。通过观察每个回归系数的正负号，我们可以更直观地理解各个因素对婚外情几率的影响方向。总的来说，我们可以得出结论：结婚时间越长，发生婚外情的几率越大；而婚姻越幸福，出轨的可能性就越小。

图 9 修改后模型的运行结果

逻辑回归实际上是在建模响应变量为Y=1的优势之对数，即log(odds)。这里的“优势”指的是某一事件发生的相对可能性。因此，当我们说log(odds)，实际上是在说事件发生的相对几率的自然对数。回归系数在逻辑回归中的意义是：它表示当自变量单位增加时，log(odds)的预期变化量。这里的关键是“保持所有其他预测变量不变”。这意味着我们是在单一地考察每个自变量对因变量的独立影响，而不是在多个自变量共同作用的情况下观察其影响。指数化结果如表7所示。

表 7 指数化结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | Age | Yearsmarried | Religiousness | Rating |
| 6.8952321 | 0.9653437 | 1.1058594 | 0.7196258 | 0.6304248 |

对于结婚年限这一变量，每增加一年的结婚时间，发生婚外情的几率会增加1.106倍。这意味着，如果其他条件（如年龄、宗教信仰和幸福感评分）保持不变，夫妻间的结婚年限会对婚外情的发生概率产生显著影响。随着结婚时间的增长，夫妻间的情感可能逐渐平淡，日常生活的琐事和矛盾也可能增加，这些都可能成为导致婚外情的因素。

对于年龄这一变量，年龄的增长与婚外情几率的关系呈现出相反的趋势。每增加一年的年龄，发生婚外情的几率会增加0.965倍。这个回归系数实际上表明，随着年龄的增长，发生婚外情的几率是在下降的，即每增加一岁，出轨的几率会下降-3.5%。这可能是因为随着年龄的增长，人们通常会更加成熟、稳定，并且家庭责任感更强，从而减少了寻求婚外刺激的可能性。

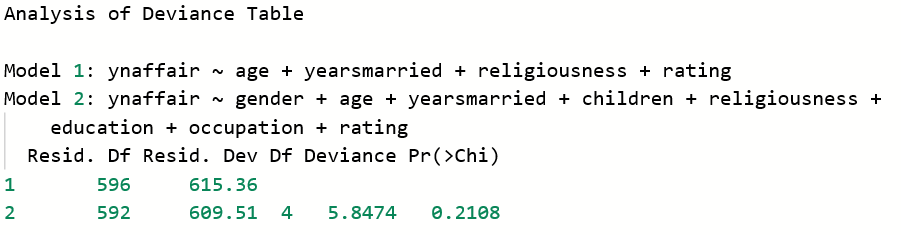
当我们比较两个回归模型时，特别是当因变量是分类变量时，我们可以使用卡方检验（Chi-square test）来评估模型拟合的好坏。对于广义线性回归模型，卡方检验尤其适用，因为它允许我们比较观察到的频数与期望频数之间的差异。卡方检验的基本思想是通过计算实际观测频数与期望频数之间的差异程度来检验模型拟合的优劣。如果观察到的频数与期望频数之间的差异很大，那么我们可能会拒绝该模型，认为它不能很好地拟合数据。对以上两个模型运用卡方检验来比较两个模型的好坏，结果见图10。

图 10 卡方检验结果

卡方值不显著（p=0.21），表明四个预测变量的新模型与九个完整预测变量的模型拟合程度一样好，故为了使回归模型中所有回归系数都具有显著贡献度可以将不显著贡献的变量去掉。

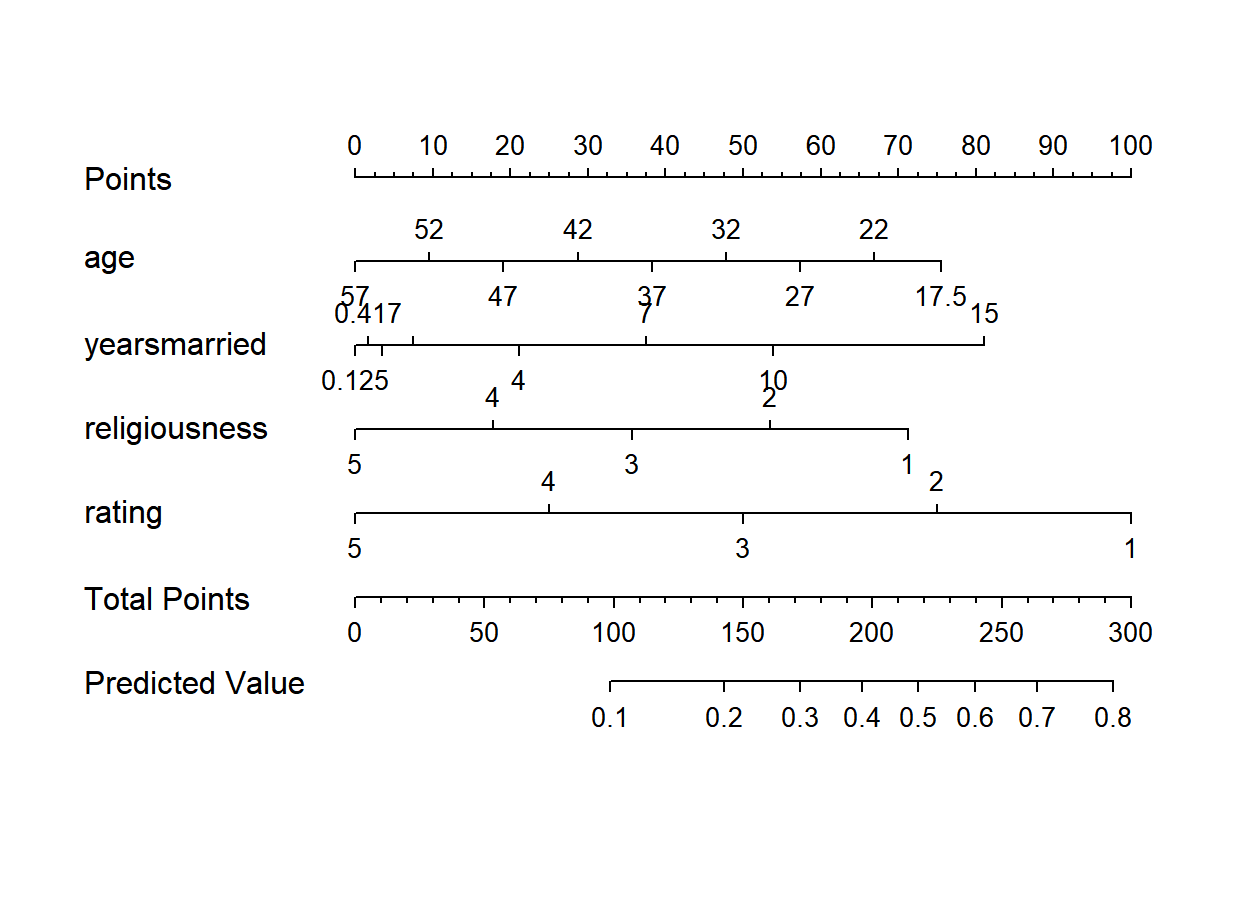
运用R自带的函数，nomogram绘制各个变量不同取值的评分值以及出现婚外情的概率。Nomogram上的刻度表示输入变量的值与输出变量之间的关系（图11）。假设有个人，22岁结婚，结婚2年，无信仰，婚姻满意度较坏。然后我们根据每个变量的值来计算潘金莲每个特征的得分：20岁结婚（67.5）+结婚2年（12.5）+无信仰（70）+婚姻满意度较坏（100）=250分，总分250分的出轨发生率大于70%。通过旋转或移动这些刻度，可以找到与特定输入值对应的预测概率，在没有计算或解读复杂数学公式的情况下快速获得预测结果。

图 11 输入变量的值与输出变量之间的关系

在预测模型的评估中，校准曲线是一个重要的可视化工具。它实际上是一个散点图，展示了模型预测的概率与实际发生概率之间的关系。通过观察校准曲线，我们可以直观地评估模型的预测准确度。校准曲线为我们提供了 Hosmer-Lemeshow 拟合优度检验的结果。这个检验帮助我们判断模型的预测概率与实际发生率之间的匹配程度。简单来说，如果预测概率与实际发生率越接近，即越靠近理想状态 Y = X，并且 Hosmer-Lemeshow 拟合优度检验的 p 值大于 0.05，那么我们可以认为模型的校准效果较好。当校准曲线与 Y = X 线几乎重合时，这表明模型的预测与实际结果非常一致，模型校准得非常好。因此，除了关注校准曲线本身，我们还应结合 Hosmer-Lemeshow 拟合优度检验的结果来全面评估模型的性能。运用R语言自带的函数calibrate（），来校准已经拟合的逻辑回归模型fit.lmreduced。method='boot'表示使用Bootstrap方法进行校准。Bootstrap是一种重抽样技术，通过它可以从数据中产生多个样本，然后计算每个样本的统计量。在这个情境下，Bootstrap用于估计模型的校准曲线。B=100表示进行100次Bootstrap抽样。增加抽样次数可以提高校准估计的精度。运行结果为：Mean absolute error=0.013，平均绝对误差是衡量预测值与实际值之间差异的指标。数值越小，表示预测值与实际值越接近。Mean squared error=0.00025，均方误差是预测值与实际值差的平方的平均值。这也是一个衡量预测准确性的指标，数值越小表示预测越准确。0.9 Quantile of absolute error=0.025这表示在所有

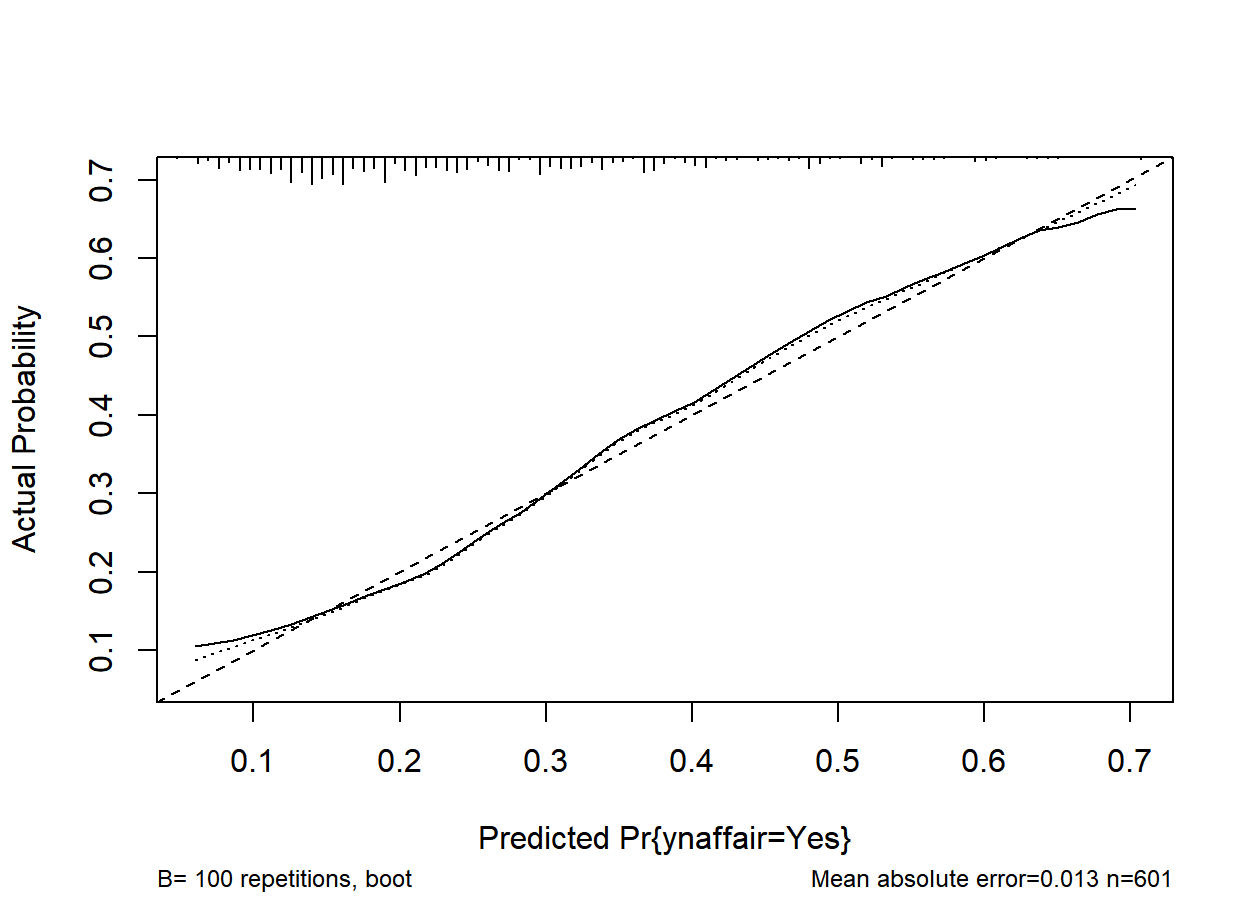
数据点中，有90%的点的误差绝对值小于0.025。

图 12 校准曲线

C-Index是一个广泛应用于评价回归模型判别能力的通用指标。它的取值范围在0.5到1.0之间，通过这个指标，我们可以清晰地了解模型的预测效果。具体来说，C-Index的值越接近0.5，意味着模型的预测结果与实际结果存在较大的不一致性，模型的判别能力较差，预测效果不理想。而当C-Index的值越接近1.0时，说明模型的预测结果与实际结果高度一致，模型的判别能力极佳。为了方便理解和应用，我们可以将C-Index的值划分为不同的精度范围：0.50-0.70被视为低精度，0.71-0.80为中等精度，0.80以上为高精度，而0.9以上则被认为是极高精度。用fit.lmreduced可以查看模型的C-Index值，运行结果如图13。C-Statistics：读取该模型中的Rank Discrim，模型参数中的参数C，即C-Statistics = 0.704，

表示该模型的预测结果与实际结果存在一定的不一致性。根据C-Index的精度范围划分，0.71-0.80被视为中等精度，因此该模型的判别能力属于中等水平。

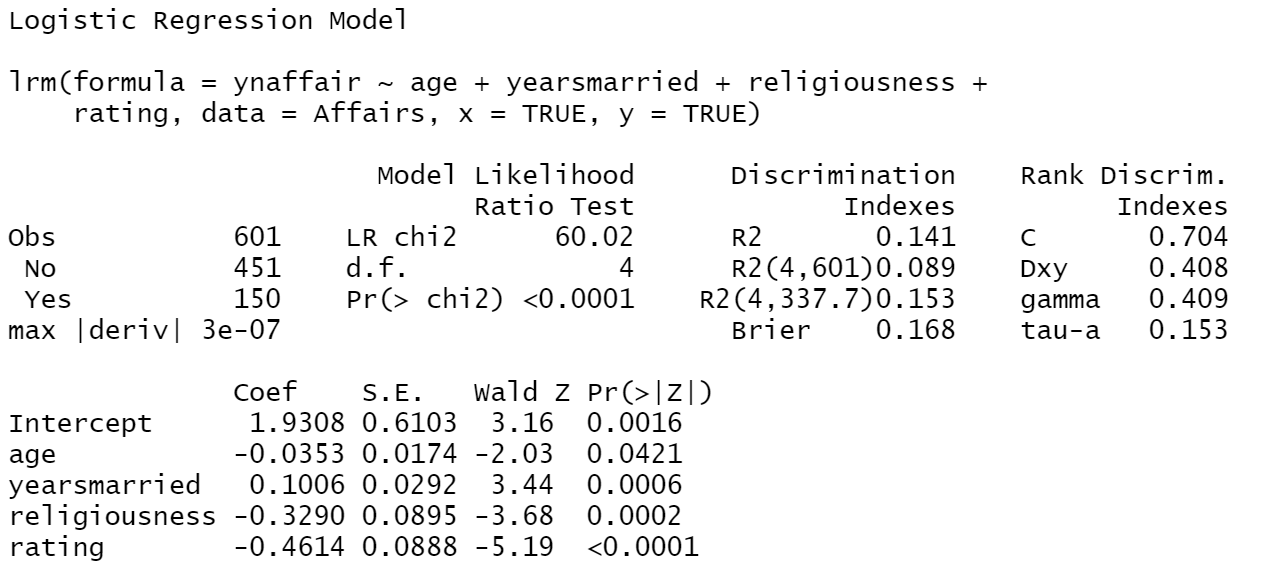
为识别出模型可能存在的偏差和不足，对模型进行交叉验证。交叉验证通过对数据进行重复使用，以不同的训练策略进行数据分割，将训练数据集分为训练集、验证集和测试集，然后基于不同的训练策略对模型进行训练和验证，最后进行评估以得到综合性的准确性评估结果。其中准确率（Accuracy）和Kappa系数都是评估分类模型性能的重要指标。准确率表示分类器正确预测的样本数与总样本数之比，是分类器性能的直接反映。然而，准确率只考虑了正确预测的正例和负例，而没有考虑到预测错误的情况，因此可能会受到样本不平衡的影响。Kappa系数则是一种更全面的评价指标，它同时考虑了模型预测的正例和负例的准确性。Kappa系数的值介于-1和1之间，其中1表示完美分类，0表示分类准确度与随机猜测相同，负值表示分类准确度低于随机猜测。相对于准确率，Kappa系数更能反映分类器在处理不平衡数据集时的性能。

图 13 fit.lmreduced运行结果

对模型进行交叉验证，首先创建了一个交叉验证控制参数control，指定了使用10折交叉验证方法。这意味着将数据集分成10份，每次使用其中的9份作为训练集，剩下的1份作为验证集，进行10次重复的训练和验证过程。接着，使用train函数训练了一个广义线性模型fit.reduced。该模型基于年龄、结婚年数、宗教信仰程度和婚姻评分这四个预测变量，预测婚外情的发生与否。使用逻辑回归作为分析方法，并指定了二项分布族。在训练模型的过程中，通过将之前定义的交叉验证控制参数control传递给trControl参数，实现了对模型的交叉验证。在每次迭代中，模型会使用不同的训练集和验证集进行训练和验证，然后对验证集上的性能进行评估。

交叉验证得到模型准确率为0.7504372，表示在多次交叉验证中，该模型的平均准确率达到了75.04%。这说明模型具有一定的预测能力。准确率达到75%以上通常被认为是相对较好的性能，特别是在实际应用中，这样的准确率可能已经能够满足一些需求的预测要求。模型性能有待进一步提高。虽然75%的准确率具有一定的应用价值，但仍然存在提升的空间。可以考虑通过增加特征、调整模型参数、采用集成学习等技术来进一步提高模型的准确率。

过度离势（Overdispersion）是指在模型中观测到的响应变量的方差大于期望的方差。过度离势可能导致奇异的标准误检验和不精确的显著性检验。过度离势可以通过比较模型的残差偏差与残差自由度来检测，如果比值比1大很多，说明存在过度离势。

使用 pchisq() 函数来比较 fit.od 的离散参数与基于 fit.aff 的残差自由度得到的期望离散参数。代码首先从 fit.od 的摘要中提取离散参数，然后乘以 fit.aff 的残差自由度。这给出了基于 fit.aff 的残差自由度的期望离散参数。然后，使用这个期望的离散参数，计算一个 p 值，并与设定的显著性水平（通常为0.05）进行比较。如果 p 值小于显著性水平，则认为存在过度离势。运行结果得过度离势检验的结果为0.340122，这个值大于显著性水平（通常为0.05），因此我们可以认为数据中不存在过度离势。这意味着模型拟合可能不会受到离散性的影响，因此模型选择可能是合理的。

模型优点：

1、适用性强：该模型适用于多种类型的因变量，包括二分类、多分类等，因此具有较广的应用范围。

2、预测效果好：交叉验证准确率达到75.04%，说明模型在多数情况下能够较好地预测结果。

3、易于理解和解释：该模型采用了线性回归的形式，使得我们可以直观地了解自变量与因变量之间的关系，并方便地进行结果解释。

4、相对可靠的评估指标：交叉验证是一种评估模型性能的常用方法，通过将数据集分成多个部分，并多次使用不同的部分进行训练和验证，可以获得更为可靠和稳定的准确率估计。

模型缺点：

1、没有进行特征选择或变量筛选的步骤，可能存在一些不相关或冗余的变量，影响模型的性能和解释能力。

2、没有对模型进行超参数调优，可能存在一些潜在的改进空间。

3、未提供其他评估指标：仅凭交叉验证准确率一个指标难以全面评价模型的性能，需要结合其他评估指标如精度、召回率、F1分数等来进行综合评价。

改进建议：

1、在模型评估部分，可以使用更多的指标来评估模型的性能，如准确率、召回率、F1分数等。

2、可以尝试使用特征选择方法（如LASSO回归、逐步回归等）来选择最相关的变量，以提高模型的预测能力和解释性。

3、可以尝试使用网格搜索等方法来调优模型的超参数，以找到最佳的模型配置。

4、可以探索其他模型算法的效果，比如随机森林、支持向量机等，以比较不同算法的性能。

## 三、理论背景

### 3.1模型建立

在建立广义线性模型时，首先需要明确响应变量和预测变量的选择。响应变量是模型试图解释的目标变量，而预测变量则是对响应变量有影响的因素。在挑选预测变量时，需要考虑这些变量与响应变量之间的关联程度，即它们的相关性。相关性强的预测变量对响应变量的影响力较大，因此在构建模型时应当优先考虑将这些变量纳入。这样可以增强模型的解释力和预测能力。数据的可用性也是选择预测变量时必须考虑的重要因素。如果某些预测变量的数据缺失或存在异常值，这可能会对模型的稳定性和准确性造成影响。因此，在选择预测变量时，我们不仅要考虑它们与响应变量的相关性，还要确保这些数据在实际操作中是易于获取且可靠的。

在建立广义线性模型时，选择合适的连接函数和数据分布类型是至关重要的。连接函数是用于将线性预测值与响应变量的期望值关联起来的函数，它决定了模型如何将预测变量转换为响应变量的预测值。常见的连接函数包括对数几率函数、逆函数和平方根函数等。对数几率函数是最常用的连接函数之一，适用于二分类或多分类的响应变量。它的形式为：,其中μ是期望的响应概率。对数几率函数的特点是在预测变量增加时，响应概率以指数方式增加。逆函数适用于泊松分布的响应变量，它的形式为：。泊松分布是一种离散概率分布，适用于计数数据的分析。在泊松模型中，预测变量与响应变量的期望值之间呈线性关系，因此选择逆函数作为连接函数能够将线性预测值转换为期望值。除了对数几率函数和逆函数，平方根函数也是常用的连接函数之一。它的形式为：。平方根函数适用于响应变量为连续变量的情况，并且当预测变量增加时，响应变量的预测值以二次方的速度增加。

除了连接函数的选择外，数据分布类型的选择也是广义线性模型中需要考虑的重要因素。数据分布类型决定了误差项的分布，进而影响模型的稳定性和拟合效果。常见的数据分布类型包括泊松分布、二项分布和对数正态分布等。泊松分布适用于计数数据的分析，其概率质量函数为，其中λ是泊松分布的参数。二项分布适用于二元分类问题的分析，其概率质量函数为，其中n是试验次数，p是成功概率。对数正态分布适用于响应变量是对数变换后呈正态分布的情况，其概率密度函数为，其中μ是均值，σ是标准差。

常用以下一些方法来确定合适的连接函数和数据分布类型：

**连接函数**的选择：

1、理解问题：首先，需要深入理解问题的背景和目标。连接函数的选择应与问题的性质和预期的模型输出相匹配。

2、查看数据的形状：观察响应变量的分布，看是否呈现特定的形状（如正态分布）。这可以帮助你选择合适的连接函数。

3、检验线性关系：使用相关系数、散点图等工具来检查预测变量与响应变量之间是否存在线性关系。如果存在线性关系，线性回归可能是合适的选择。

4、考虑模型的适用性：不同的连接函数适用于不同类型的问题。例如，逻辑回归适用于二元分类问题，而线性回归适用于连续变量的预测。

5、交叉验证：交叉验证是一种用于评估模型泛化能力的统计学方法，它通过将数据集分成多个部分，将模型在其中的一部分数据上进行训练，并在另一部分数据上进行测试，以评估模型的预测精度和稳定性。通过比较不同模型组合的交叉验证结果，可以确定最佳的模型组合。比较不同模型组合的交叉验证结果，可根据均方误差（MSE）、准确率等选择拟合效果和预测精度最佳的模型组合。

6、参数估计和模型诊断：使用参数估计来检验模型的假设，并使用模型诊断工具（如残差图）来检查模型是否存在异常。

7、考虑其他因素：如果数据存在异常值或缺失值，可能会影响连接函数的选择。在这种情况下，需要处理数据或在选择连接函数时考虑这些因素。

**数据分布类型**的选择：

1、理解数据：查看数据的直方图或其他可视化工具，了解数据的分布情况。

2、理论依据：根据所研究领域的理论和先验知识，选择适合的数据分布类型。例如，如果数据来自正态分布的过程，那么可以选择正态分布作为数据分布类型。

3、假设检验：使用假设检验（如卡方检验）来检验数据是否符合特定的分布类型。这将帮助你选择合适的数据分布类型。

4、模型诊断：通过模型诊断工具（如Q-Q图）来检查数据是否符合所选的数据分布类型。如果不符合，则需要选择不同的数据分布类型。

5、参数估计：参数估计的结果可以帮助你验证所选的数据分布类型的合理性。如果参数估计的结果不满足预期，可能需要选择不同的数据分布类型。

6、考虑其他因素：如果数据存在异常值或缺失值，这可能会影响数据分布类型的选择。在这种情况下，你可能需要处理数据或在选择数据分布类型时考虑这些因素。

### 3.2统计分析

**一、参数估计**：用于估计未知的参数向量，以拟合模型并反映预测变量与响应变量之间的关系。常见的参数估计方法包括极大似然估计、矩估计、最小二乘估计和贝叶斯估计等。

**极大似然估计**：通过寻找使观测到的样本数据出现的概率最大的参数值来估计总体参数。具体步骤如下：

1、确定似然函数：根据样本的联合概率分布，写出样本的联合概率似然函数。

2、将似然函数转换为对数似然函数：为了便于计算，通常将似然函数转换为对数似然函数。

3、求对数似然函数的最大值：通过求导并解似然方程，求得对数似然函数的最大值，即为参数的极大似然估计值。

优点：最合理的参数估计量应该使得从模型中抽取该n组样本的观测值的概率最大，也就是概率分布函数或者似然函数最大。

缺点：当总体类型已知时，没有充分利用分布提供的信息（有一定随意性）。

**矩估计**是基于样本矩与总体矩之间的关系来进行参数估计的。它不需要对总体分布做出具体的假设，适用范围较广。具体步骤如下：

1、确定样本矩：根据样本数据计算样本矩，如样本的均值、方差等。

2、确定总体矩：根据已知的总体分布或假设，确定总体矩，如总体的均值、方差等。

3、利用样本矩和总体矩的关系：通过样本矩和总体矩之间的关系，建立方程并求解未知参数。

优点：原理简单，操作方便，且不需要事先知道总体的分布。在样本数量充分大时，矩估计的精度较高。

缺点：当总体类型已知时，矩估计没有充分利用分布提供的信息，因此不一定是理想的估计。在样本数量较少时，矩估计的结果可能不准确。矩估计可能不唯一，如指数分布的参数的矩估计不唯一。在某些分布下，如柯西分布，矩估计可能不存在。

**最小二乘估计**是另一种常用的参数估计方法，其目标是最小化预测值与观测值之间的平方误差。具体步骤如下：

1、构造合适的包含待估参数的统计量U，且统计量的分布已知。

2、根据给定的置信度，按照某种规则（如P（U>u）=α），确定合适的u值。

3、解出统计量U的分布函数F（u），计算F（u）的逆函数。

4、求出待估参数的估计值。

优点：最合理的参数估计量应该使得模型能最好地拟合样本数据，也就是估计值与观测值之差的平方和最小。

缺点：目标最小化估计值与观测值之差的平方和。

**贝叶斯估计**是基于贝叶斯定理的参数估计方法，它结合了样本信息和先验信息来估计未知参数。具体步骤如下：

1、确定先验分布：根据先验信息确定参数的先验分布。

2、确定似然函数：根据样本数据和模型，确定参数的似然函数。

3、确定后验分布函数：利用贝叶斯定理，将先验分布和似然函数结合起来，求出参数的后验分布函数。

4、根据后验分布函数求解参数的贝叶斯估计值。

优点：结合先验信息进行更准确的参数估计。

缺点：需要确定合适的先验分布和先验信息。

**二、诊断分析**：用于评估模型的拟合效果和诊断模型可能存在的问题。这些工具和方法包括残差分析、偏残差图和诊断统计量等。

**残差分析：**通过比较实际观测值与模型预测值之间的差异来评估模型的拟合效果。计算每个观测值的残差（实际观测值与模型预测值之差），然后分析这些残差的分布和特征。观察残差的直方图、QQ图等，检查是否有异常值、离群点或不符合预期的分布形态。如果残差显示出某种模式或结构（如周期性、趋势等），这可能意味着模型未能充分捕捉数据中的某些重要特征，或者存在其他未考虑的影响因素。

**偏残差图**：观察残差与预测变量之间的关系，判断误差项是否满足独立性和恒定方差等假设。通过图形展示残差与一个或多个预测变量之间的关系。检查图中是否有弯曲的线条、趋势或其他结构，这些可能意味着误差项之间存在相关性或方差不恒定。如果偏残差图显示了某种模式或结构，这可能意味着模型的独立性或恒定方差假设不成立，可能需要进一步的模型修正或选择其他模型。

**诊断统计量：**通过一系列统计量检验模型的假设条件是否满足，如正态性、同方差性和无自相关性等。计算并评估诊断统计量的值。例如，Jarque-Bera统计量用于检验残差的正态性，Durbin Watson统计量用于检验残差的自相关性。如果统计量的值不在预期范围内，则表明相应的假设可能不成立。如果诊断统计量的值表明模型的某个假设不成立，这可能意味着模型需要进行修正，例如通过引入新的变量、改变模型的类型或进行其他调整。

**异常值和离群点检测：**识别数据中的异常值和离群点，这些值可能对模型的拟合效果产生负面影响。使用诸如IQR（四分位距）方法、Z分数等方法来识别异常值和离群点。将异常值和离群点与模型预测进行比较，了解它们对模型拟合的影响。若数据中存在异常值或离群点，且这些值对模型拟合有显著影响，可能需要进一步调查这些值的来源，并在建模过程中适当处理它们。

**模型修正和优化：**当诊断分析结果显示模型的拟合效果不佳时，可以考虑引入新的预测变量。选择新变量时，需要基于业务逻辑、先验知识或数据探索来识别可能的候选变量。然后，将这些变量纳入模型中，并重新进行诊断分析，以评估其对模型拟合的改进效果。调整模型参数。例如，在逻辑回归模型中，正则化参数（如L1、L2正则化）或模型复杂度参数（如决策树深度）等。通过调整这些参数，可以优化模型的复杂度，防止过拟合或欠拟合，进一步提高模型的拟合效果。若诊断分析显示当前模型假设不满足或拟合效果不佳，可能需要考虑更换模型类型。不同的模型类型对数据的拟合能力有所不同，选择适合特定数据的模型类型可以提高拟合效果。在选择新的模型类型时，需要考虑其假设、适用场景以及是否能够处理当前数据集的特性。此外，重新进行诊断分析也是必要的，以确保新模型满足要求。诊断分析、模型修正和重新评估应该是一个迭代的过程。在每次迭代中，分析诊断结果，提出修正建议，然后重新评估模型的拟合效果和假设满足情况。这个过程可能需要反复进行，直到模型的拟合效果和假设满足要求。通过迭代改进，可以逐步优化模型，提高其预测性能和准确性。对模型的修正与优化比较耗费时间和计算成本，故在处理实际问题时需要平衡好模型的准确率与成本。

**三、推断分析：**对新数据点的预测以及参数置信区间的估计。这一过程有助于理解模型的不确定性、评估预测结果的可信度以及模型改进和优化。

**新数据点的预测：**推断分析的首要任务是对新数据点进行预测。在广义线性模型的背景下，这一过程通常涉及使用预测方程和连接函数。预测方程基于模型的参数和自变量，用于计算出新数据点的预测值。这些预测值反映了模型对未知数据的预测能力，对于决策制定和预测未来的趋势具有重要意义。通过比较实际观测值与预测值，可以对模型的预测性能进行评估。

**参数置信区间的估计：**推断分析的另一重要方面是参数置信区间的估计。通过利用估计的参数向量和相关的函数，可以构建参数的置信区间。这些区间提供了参数估计的不确定性范围，有助于我们了解参数估计的精度和可靠性。置信区间的一个重要应用是假设检验。通过将参数的置信区间与预期值或假设值进行比较，可以判断特定的参数假设是否成立。例如，在医学研究中，如果某个药物的疗效参数没有落入预期的置信区间内，这可能意味着该药物的疗效与预期不符，需要进一步的研究或重新评估。

**结果解读与决策应用：**通过理解模型的不确定性、评估预测结果的可信度，决策者可以更加准确地评估各种选择和可能的结果。综合应用推断分析的方法和技术，可以帮助我们更好地理解数据、评估模型的性能以及指导决策过程。通过不断地迭代和优化模型，可以提高数据分析的质量和可靠性，为实际应用提供更准确的预测和决策支持。

### 3.3模型选择

在应用广义线性模型时，选择合适的模型对于提高预测精度和解释能力至关重要。模型选择是一个涉及多个因素的复杂过程，需要综合考虑数据的分布特征、连接函数的选择、模型的复杂度以及模型的解释性。

首先，数据的分布特征是选择模型的重要依据。不同的数据分布类型（如正态分布、泊松分布等）适用于不同的模型。例如，泊松回归适用于分析计数数据，特别是当响应变量是离散且呈泊松分布时。而正态回归则适用于分析连续数据，当响应变量呈正态分布时。因此，了解数据的分布特征是选择合适模型的关键步骤。

其次，连接函数的选择也直接影响模型的拟合效果和预测精度。常见的连接函数包括逻辑函数、对数函数、指数函数等。选择合适的连接函数需要考虑其与数据分布类型的匹配程度以及对非线性关系的适应性。例如，当数据呈正态分布时，使用对数连接函数可能更为合适；而当数据呈泊松分布时，使用线性连接函数可能更为合适。

此外，模型的复杂度和解释性也需要权衡考虑。过于复杂的模型可能过拟合数据，导致泛化能力下降；而过于简单的模型则可能无法充分捕捉数据的复杂性和非线性关系。因此，在选择模型时需要找到一个平衡点，既要保证模型的拟合效果和预测精度，又要确保模型的解释性和实用性。

常用的模型选择方法包括交叉验证、网格搜索和贝叶斯模型平均等。

**交叉验证：**将训练数据集分成K份，每次取其中的一份作为验证集，其余K-1份作为训练集。使用训练集训练模型，然后使用验证集评估模型的性能。重复此过程K次，每次使用不同的验证集，取K次评估结果的平均值作为最终性能指标。根据性能指标选择最优的模型或参数组合。

优点：数据利用率较高，能够评估模型在未见数据上的表现。可以选择不同类型的交叉验证（如k-fold、留出交叉验证等）以满足不同需求。

缺点：计算成本较高，特别是当数据集很大或模型复杂时。对于不平衡数据集，可能会存在过拟合或欠拟合的风险。

**网格搜索：**定义参数空间，即所有可能参数值的集合。初始化一个网格，该网格包含参数空间中的一些离散点。对每个参数组合，使用交叉验证或其他评估方法计算模型的性能指标。根据性能指标选择最优的参数组合。可选地，可以进一步细化网格并重复上述过程以找到更精确的最优参数组合。

优点：可以系统地搜索参数空间，找到最优的模型配置。可以为每个参数设置合理的默认值，减少搜索空间。

缺点：当参数空间很大时，计算成本很高。可能陷入局部最优解，无法找到全局最优解。

**贝叶斯模型平均：**每个模型定义一个先验概率。使用数据和贝叶斯推断方法更新每个模型的先验概率。对每个模型进行预测，并计算加权平均预测结果，权重为其先验概率和后验概率。可选地，可以进一步调整先验概率和后验概率以改进预测性能。

优点：可以综合考虑多个模型的不确定性，给出更稳健的预测。可以处理模型的不确定性，提供更全面的预测结果。

缺点：需要为每个模型设定先验概率，这可能需要额外的信息和经验。对于大规模数据集，计算成本较高。

## 四、课程设计总结

本次课程设计题目首先是对R语言的一个基本操作的实现，接着是建立一个广义线性模型并进行相应的评估及改进。在学习的过程中，加深了我对于课堂学习的印象、加强了我R语言以及一些统计模型的掌握以及熟练程度。设计过程中，由于本题建立模型相对容易实现，因此在建模的过程并没有遇到困难。在优化以及评价模型方面遇到了困难，在给出模型优化建议方面做得不够，这说明我还缺少对实际问题建立模型并使之应用与实际中的经验、能力还待提高。

本次课程设计让我对于所学知识和现实生活中问题的联系产生了很大的兴趣，这加深了我对于课本上知识的理解。也在此感谢王老师关于课本上知识以及提供的资料，帮助我顺利地完成了本次课程设计。

## 附录

|  |
| --- |
| **附录1：两种方法预测值比较** |
| 介绍：第一部分中用两种方式预测的值的比较 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | mynewml()  预测值 | predict()  预测值 | mynewml()  预测值 | predict()  预测值 | mynewml()  预测值 | predict()  预测值 | | 4.425504812 | 4.425504812 | 4.681774473 | 4.681774473 | 4.576052598 | 4.576052598 | | 4.463547279 | 4.463547279 | 4.565445438 | 4.565445438 | 4.599036044 | 4.599036044 | | 4.536149509 | 4.536149509 | 4.400845663 | 4.400845663 | 4.530358282 | 4.530358282 | | 4.53151439 | 4.53151439 | 4.518650335 | 4.518650335 | 4.501219488 | 4.501219488 | | 4.421704784 | 4.421704784 | 4.43090665 | 4.43090665 | 4.45093173 | 4.45093173 | | 4.592882866 | 4.592882866 | 4.389615048 | 4.389615048 | 4.499241393 | 4.499241393 | | 4.477108943 | 4.477108943 | 4.558837418 | 4.558837418 | 4.545168909 | 4.545168909 | | 4.542405916 | 4.542405916 | 4.60335596 | 4.60335596 | 4.506320259 | 4.506320259 | | 4.539493237 | 4.539493237 | 4.496774436 | 4.496774436 | 4.599335733 | 4.599335733 | | 4.607044192 | 4.607044192 | 4.453861971 | 4.453861971 | 4.582152953 | 4.582152953 | | 4.549369547 | 4.549369547 | 4.508829239 | 4.508829239 | 4.432694003 | 4.432694003 | | 4.544199127 | 4.544199127 | 4.437096545 | 4.437096545 | 4.578358746 | 4.578358746 | | 4.679504244 | 4.679504244 | 4.607526173 | 4.607526173 | 4.595327601 | 4.595327601 | | 4.582485236 | 4.582485236 | 4.597808073 | 4.597808073 | 4.534595966 | 4.534595966 | | 4.601516675 | 4.601516675 | 4.433710966 | 4.433710966 | 4.552355769 | 4.552355769 | | 4.590275198 | 4.590275198 | 4.507162542 | 4.507162542 | 4.521670293 | 4.521670293 | | 4.597591143 | 4.597591143 | 4.633558391 | 4.633558391 | 4.532171352 | 4.532171352 | | 4.655378598 | 4.655378598 | 4.541649614 | 4.541649614 | 4.478199211 | 4.478199211 | | 4.550040919 | 4.550040919 | 4.555079257 | 4.555079257 | 4.586517168 | 4.586517168 | | 4.572725379 | 4.572725379 | 4.585914611 | 4.585914611 | 4.461729224 | 4.461729224 | | 4.703318977 | 4.703318977 | 4.601765773 | 4.601765773 | 4.509703959 | 4.509703959 | | 4.51427388 | 4.51427388 | 4.645585283 | 4.645585283 | 4.485263017 | 4.485263017 | | 4.651995949 | 4.651995949 | 4.612787123 | 4.612787123 | 4.48849387 | 4.48849387 | | 4.508447511 | 4.508447511 | 4.450177019 | 4.450177019 | 4.526486483 | 4.526486483 | | 4.683693944 | 4.683693944 | 4.526949985 | 4.526949985 | 4.517660937 | 4.517660937 | | 4.503972667 | 4.503972667 | 4.577221804 | 4.577221804 | 4.501422909 | 4.501422909 | | 4.492770285 | 4.492770285 | 4.546110102 | 4.546110102 | 4.535192904 | 4.535192904 | | 4.632740593 | 4.632740593 | 4.389553791 | 4.389553791 | 4.434497036 | 4.434497036 | | 4.627335381 | 4.627335381 | 4.497024929 | 4.497024929 | 4.562757193 | 4.562757193 | | 4.596084348 | 4.596084348 | 4.631088011 | 4.631088011 | 4.602144818 | 4.602144818 | | 4.575020277 | 4.575020277 | 4.566312372 | 4.566312372 | 4.457321468 | 4.457321468 | | 4.57265095 | 4.57265095 | 4.442992865 | 4.442992865 | 4.447489689 | 4.447489689 | | 4.709756925 | 4.709756925 | 4.570689645 | 4.570689645 | 4.723160951 | 4.723160951 | | 4.446019945 | 4.446019945 | 4.542356869 | 4.542356869 | 4.572473031 | 4.572473031 | | 4.479764998 | 4.479764998 | 4.605170519 | 4.605170519 | 4.558382461 | 4.558382461 | | 4.457620847 | 4.457620847 | 4.713674294 | 4.713674294 | 4.539346878 | 4.539346878 | | 4.627019556 | 4.627019556 | 4.439715173 | 4.439715173 | 4.591617529 | 4.591617529 | | 4.535782194 | 4.535782194 | 4.516162672 | 4.516162672 | 4.557176055 | 4.557176055 | | 4.544527317 | 4.544527317 | 4.543865233 | 4.543865233 | 4.513073389 | 4.513073389 | | 4.689227425 | 4.689227425 | 4.392802789 | 4.392802789 | 4.521652312 | 4.521652312 | | 4.510569007 | 4.510569007 | 4.551998485 | 4.551998485 | 4.528306918 | 4.528306918 | | 4.707339394 | 4.707339394 | 4.509499432 | 4.509499432 | 4.530183649 | 4.530183649 | | 4.447766473 | 4.447766473 | 4.723433639 | 4.723433639 | 4.497510146 | 4.497510146 | | 4.654024793 | 4.654024793 | 4.333679036 | 4.333679036 | 4.635591684 | 4.635591684 | | 4.578796568 | 4.578796568 | 4.666910703 | 4.666910703 | 4.561455167 | 4.561455167 | | 4.52497053 | 4.52497053 | 4.630594427 | 4.630594427 | 4.495894132 | 4.495894132 | | 4.541288289 | 4.541288289 | 4.609953061 | 4.609953061 | 4.599074518 | 4.599074518 | | 4.655623496 | 4.655623496 | 4.565462679 | 4.565462679 | 4.580846651 | 4.580846651 | | 4.683214931 | 4.683214931 | 4.500603222 | 4.500603222 | 4.543954728 | 4.543954728 | | 4.427420512 | 4.427420512 | 4.637976224 | 4.637976224 | 4.584677462 | 4.584677462 | | 4.813089954 | 4.813089954 | 4.487372357 | 4.487372357 | 4.485812006 | 4.485812006 | | 4.749324102 | 4.749324102 | 4.606903 | 4.606903 | 4.539321709 | 4.539321709 | | 4.514630577 | 4.514630577 | 4.584810365 | 4.584810365 | 4.47879557 | 4.47879557 | | 4.549490518 | 4.549490518 | 4.635753276 | 4.635753276 | 4.464830436 | 4.464830436 | | 4.504176865 | 4.504176865 | 4.509244007 | 4.509244007 | 4.52929659 | 4.52929659 | | 4.547326108 | 4.547326108 | 4.510420687 | 4.510420687 | 4.56706696 | 4.56706696 | | 4.610313641 | 4.610313641 | 4.57720502 | 4.57720502 | 4.485310855 | 4.485310855 | | 4.562176707 | 4.562176707 | 4.577477931 | 4.577477931 | 4.541120302 | 4.541120302 | | 4.568880685 | 4.568880685 | 4.658272266 | 4.658272266 | 4.415926706 | 4.415926706 | | 4.666198774 | 4.666198774 | 4.591069352 | 4.591069352 | 4.520573768 | 4.520573768 | | 4.387685539 | 4.387685539 | 4.565313544 | 4.565313544 | 4.642894753 | 4.642894753 | | 4.51719276 | 4.51719276 | 4.618085199 | 4.618085199 | 4.467280368 | 4.467280368 | | 4.565281571 | 4.565281571 | 4.513247646 | 4.513247646 | 4.464140589 | 4.464140589 | | 4.495135295 | 4.495135295 | 4.637018811 | 4.637018811 | 4.537731919 | 4.537731919 | | 4.634598289 | 4.634598289 | 4.477026866 | 4.477026866 | 4.59479564 | 4.59479564 | | 4.520666912 | 4.520666912 | 4.579932981 | 4.579932981 | 4.647843989 | 4.647843989 | | 4.521901575 | 4.521901575 | 4.425274347 | 4.425274347 | 4.755056555 | 4.755056555 | | 4.644282112 | 4.644282112 | 4.572998041 | 4.572998041 | 4.614929502 | 4.614929502 | | 4.617455538 | 4.617455538 | 4.435777755 | 4.435777755 | 4.605140242 | 4.605140242 | | 4.542992143 | 4.542992143 | 4.536852858 | 4.536852858 | 4.707727329 | 4.707727329 | | 4.52957199 | 4.52957199 | 4.42144684 | 4.42144684 | 4.531601302 | 4.531601302 | | 4.44483782 | 4.44483782 | 4.505986454 | 4.505986454 | 4.538224412 | 4.538224412 | | 4.501625387 | 4.501625387 | 4.495659679 | 4.495659679 | 4.573513891 | 4.573513891 | | 4.750878074 | 4.750878074 | 4.596316058 | 4.596316058 | 4.446961245 | 4.446961245 | | 4.551529969 | 4.551529969 | 4.510334887 | 4.510334887 | 4.415549282 | 4.415549282 | | 4.623445107 | 4.623445107 | 4.496232319 | 4.496232319 | 4.691003362 | 4.691003362 | | 4.540404229 | 4.540404229 | 4.646740314 | 4.646740314 | 4.510873841 | 4.510873841 | | 4.60047935 | 4.60047935 | 4.680595522 | 4.680595522 | 4.519731 | 4.519731 | | 4.550371969 | 4.550371969 | 4.637057267 | 4.637057267 | 4.479264383 | 4.479264383 | | 4.657510208 | 4.657510208 | 4.441579979 | 4.441579979 | 4.565255504 | 4.565255504 | | 4.516270828 | 4.516270828 | 4.650910242 | 4.650910242 | 4.599066334 | 4.599066334 | | 4.425106876 | 4.425106876 | 4.461447814 | 4.461447814 | 4.537070191 | 4.537070191 | | 4.622315058 | 4.622315058 | 4.548918081 | 4.548918081 | 4.555471675 | 4.555471675 | | 4.536380253 | 4.536380253 | 4.609905156 | 4.609905156 | 4.626311483 | 4.626311483 | | 4.665034102 | 4.665034102 | 4.538183749 | 4.538183749 | 4.553545222 | 4.553545222 | | 4.563766063 | 4.563766063 | 4.54298569 | 4.54298569 | 4.511906335 | 4.511906335 | | 4.664303026 | 4.664303026 | 4.563842005 | 4.563842005 | 4.549775773 | 4.549775773 | | 4.580846536 | 4.580846536 | 4.586717614 | 4.586717614 | 4.490656176 | 4.490656176 | | 4.527963509 | 4.527963509 | 4.585469197 | 4.585469197 | 4.562065176 | 4.562065176 | | 4.609235123 | 4.609235123 | 4.42379843 | 4.42379843 | 4.486451305 | 4.486451305 | | 4.519628329 | 4.519628329 | 4.541075674 | 4.541075674 |  |  | |

|  |
| --- |
| **附录2：**学生化残差与cook统计量比较 |
| 介绍：第一部分中学生化残差与cook统计量比较 |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 学生化残差 | cook统计量 | 学生化残差 | cook统计量 | 学生化残差 | cook统计量 | | -3.125899707 | 0.002560285 | -270.22228 | 0.008168817 | -0.956041816 | 0.000318011 | | -184.7579646 | 0.003256009 | -0.252157977 | 4.20293E-05 | -286.7215725 | 0.005489426 | | -2.431897909 | 0.001199359 | 357.0075491 | 0.006015481 | 0.312161921 | 6.00805E-05 | | -239.7596792 | 0.003048608 | -6.603740021 | 0.011274456 | 183.9208575 | 0.001596532 | | -0.133377976 | 7.18583E-06 | 328.5057998 | 0.004016438 | -2.471109537 | 0.002288868 | | 269.3115554 | 0.006426575 | 1.175624305 | 0.000518053 | 102.8289918 | 0.001750108 | | -0.507317085 | 0.000129858 | -21.67194575 | 1.89651E-05 | 0.086774197 | 2.61981E-06 | | -430.1743332 | 0.014345489 | -0.724409685 | 0.00032355 | -46.21105947 | 0.000189218 | | -0.206123882 | 2.61957E-05 | 258.0633569 | 0.004130941 | 3.422885593 | 0.003522512 | | -92.11438865 | 0.000657781 | -2.072623593 | 0.00118706 | -272.1504466 | 0.005654028 | | -2.353474654 | 0.002752839 | -186.7188161 | 0.003376006 | 5.784144131 | 0.007364283 | | -253.1986234 | 0.004611869 | 1.675040622 | 0.001985908 | -170.711997 | 0.00209644 | | 1.092996358 | 0.000293789 | 43.67567996 | 0.000127375 | 0.104056766 | 2.51158E-06 | | 31.32798471 | 0.000152282 | -1.260419854 | 0.000477637 | -8.495580588 | 4.16458E-06 | | 1.853941452 | 0.001195861 | 84.06533605 | 0.001096523 | -7.528354093 | 0.02656856 | | 48.84411461 | 0.00026301 | 1.381689745 | 0.000715579 | 187.8098756 | 0.001570218 | | -6.649774117 | 0.012219278 | -124.242577 | 0.001701719 | 1.667198297 | 0.001713752 | | 144.8106179 | 0.001508534 | 1.373847419 | 0.000821502 | 289.3955602 | 0.00395275 | | -2.863918131 | 0.004711408 | -256.4062066 | 0.00379352 | -3.51972816 | 0.003944316 | | 67.34519119 | 0.00022807 | 5.838020551 | 0.007170878 | 10.1527309 | 6.39384E-06 | | -1.142784971 | 0.000805197 | -129.8298675 | 0.001108485 | -5.464651126 | 0.018679508 | | 31.32798471 | 0.000152282 | 0.26931266 | 3.93539E-05 | -120.4689031 | 0.001488939 | | -3.590309089 | 0.003170018 | 142.6198722 | 0.001022859 | 3.590759263 | 0.003170813 | | -124.9976294 | 0.000589042 | -5.545239357 | 0.010536672 | 80.29166212 | 0.001067025 | | 5.061390806 | 0.00519515 | 171.8030565 | 0.001703127 | 2.608064078 | 0.001577764 | | -361.9962535 | 0.008750126 | 2.557825291 | 0.00432463 | -283.3986451 | 0.005777631 | | -1.988975788 | 0.001826464 | 20.57225952 | 3.28088E-05 | 0.63222461 | 0.000282912 | | 94.33762976 | 0.000375958 | -4.814643708 | 0.007943121 | 43.67567996 | 0.000127375 | | -1.092546184 | 0.000293547 | -130.0924466 | 0.001522613 | -6.947329686 | 0.009298604 | | 149.8318177 | 0.001128922 | 3.014392347 | 0.004857458 | 236.092118 | 0.002998105 | | 4.078695622 | 0.003282336 | 325.7489632 | 0.00400045 | 0.22009373 | 3.42865E-05 | | 233.3352815 | 0.002423728 | 1.877468429 | 0.000817619 | 15.74002144 | 1.00039E-05 | | 1.558425598 | 0.000529314 | -177.6613633 | 0.002596398 | 2.815149376 | 0.001924783 | | -64.55585927 | 0.000224158 | -1.131757812 | 0.000646274 | 161.9496187 | 0.001107974 | | 0.690305722 | 9.66365E-05 | -218.5434928 | 0.002755885 | 1.659355971 | 0.001471935 | | -18.7678743 | 3.88309E-05 | 2.421320923 | 0.005386689 | -196.0797806 | 0.005662141 | | 0.692345438 | 0.000362835 | 77.76471981 | 0.000375113 | -1.782910349 | 0.001467611 | | 296.0414149 | 0.004713992 | -0.115653609 | 6.17548E-06 | 124.5376515 | 0.000655196 | | -4.714733191 | 0.00761688 | -127.3356101 | 0.001813914 | -0.252157977 | 4.20293E-05 | | -67.38631327 | 0.000507996 | 2.10910056 | 0.004087067 | -27.25923629 | 4.60918E-05 | | 1.085154032 | 0.000330027 | -106.4229354 | 0.00075627 | -2.871760457 | 0.004098814 | | -283.7021568 | 0.005701683 | 2.740363754 | 0.003218727 | 308.7253067 | 0.003848603 | | 0.551183579 | 0.000113875 | -236.1332401 | 0.005806384 | -0.155885095 | 5.97595E-06 | | -343.6514537 | 0.008365929 | 1.424539006 | 0.001750108 | 7.395894362 | 2.75066E-06 | | 5.533642514 | 0.006041748 | 25.02736846 | 3.88532E-05 | 0.362400708 | 3.2298E-05 | | -489.5575392 | 0.016003452 | -0.624499169 | 0.000240457 | 109.0969231 | 0.000561747 | | 5.412822797 | 0.007115844 | 136.581835 | 0.001558243 | 3.103842762 | 0.001864077 | | -15.44494696 | 2.44737E-05 | -0.632341494 | 0.000213753 | 202.569169 | 0.00206349 | | 1.907817873 | 0.001176394 | -165.6907972 | 0.001380551 | 4.140414368 | 0.003317051 | | -265.0538453 | 0.004620071 | -0.117693325 | 4.16458E-06 | 93.35268307 | 0.000831252 | | 6.256395838 | 0.008235491 | -194.1189291 | 0.002302538 | -2.791172225 | 0.003390823 | | -97.58633498 | 0.001065356 | 1.283377147 | 0.000532338 | 139.5595239 | 0.002470834 | | -2.353474654 | 0.002752839 | 391.0965889 | 0.006809123 | -3.012922316 | 0.005291456 | | 296.0414149 | 0.004713992 | 2.868005939 | 0.001810558 | -147.3459974 | 0.002070885 | | -2.97791538 | 0.004407441 | -272.7165374 | 0.004891064 | -2.247319731 | 0.002510103 | | 216.311625 | 0.003124395 | 2.852321288 | 0.002280149 | 150.3979084 | 0.001006947 | | 1.382709602 | 0.000368336 | 83.49924526 | 0.000947089 | -0.940357165 | 0.000231699 | | -585.5976599 | 0.015984352 | -1.78756784 | 0.002792209 | -285.4004292 | 0.003796694 | | 3.856945531 | 0.003450584 | -150.6689248 | 0.002329027 | 2.197531117 | 0.001016041 | | 29.3671332 | 4.97632E-05 | 6.766839315 | 0.010621291 | 301.7759403 | 0.014130355 | | -2.216970286 | 0.001684157 | 22.83662272 | 2.3216E-05 | 1.892133222 | 0.001558243 | | 9.586640105 | 6.6113E-06 | -4.007664518 | 0.003683592 | 4.261928682 | 9.01251E-07 | | -0.038250212 | 1.10747E-06 | 156.9284189 | 0.001420982 | 0.178842387 | 1.29196E-05 | | 212.9886976 | 0.002813905 | 0.446048513 | 3.83306E-05 | -222.8832575 | 0.003081422 | | -5.286895415 | 0.00889928 | 261.4599017 | 0.002538588 | -2.209127961 | 0.001941172 | | -308.200338 | 0.012222216 | 1.549563414 | 0.000721916 | 17.81542298 | 1.96874E-05 | | 0.886930918 | 0.000193454 | -485.1760477 | 0.032424407 | -6.619424672 | 0.009142797 | | 304.2701977 | 0.004369527 | -2.087288386 | 0.002329027 | 282.8650497 | 0.003114114 | | -6.087052137 | 0.008497728 | -401.1064931 | 0.016753705 | -0.061777189 | 1.89678E-06 | | 474.715397 | 0.008533448 | 4.188488179 | 0.016118766 | 21.13835032 | 2.98365E-05 | | -1.838826485 | 0.001092849 | -216.352747 | 0.003909977 | -0.498454902 | 4.90221E-05 | | -177.6613633 | 0.002596398 | -3.199665471 | 0.003778935 | 384.5660785 | 0.005600153 | | 2.494066829 | 0.001629875 | -86.52709811 | 0.000353362 | -3.94174108 | 0.007839452 | | 227.3708618 | 0.001911477 | 2.388931764 | 0.001243795 | -161.9171233 | 0.001410163 | | -5.194827224 | 0.007457169 | 319.7109261 | 0.004318021 | 2.631591055 | 0.001334193 | | 537.872277 | 0.012221601 | -3.269226543 | 0.002763158 | 21.13835032 | 2.98365E-05 | | -4.038013962 | 0.005191455 | 98.60377704 | 0.000941487 | -2.255162056 | 0.002179833 | | 244.3209009 | 0.002817324 | -0.784530513 | 0.000353549 | 82.21982875 | 0.000339945 | | 0.058022671 | 7.4105E-07 | -244.4773672 | 0.008232859 | 4.633575276 | 0.005279958 | | -361.9962535 | 0.008750126 | 0.532314093 | 0.00020056 | 24.46127766 | 4.3044E-05 | | 2.091376194 | 0.000920249 | -159.4637984 | 0.001941172 | -4.13632656 | 0.006315241 | | 2.490038829 | 3.84599E-07 | 3.956856047 | 0.003631667 | -69.01096821 | 0.000318011 | | 4.246569291 | 0.003489321 | 28.99000406 | 3.24993E-05 | 3.777955217 | 0.014829166 | | -184.2654913 | 0.001494495 | -0.529824204 | 8.43978E-05 | -186.456237 | 0.002124344 | | 2.623748729 | 0.001358266 | 36.57907868 | 7.7206E-05 | 5.944175475 | 0.00743403 | | 68.40375533 | 0.000453093 | 5.327577074 | 0.005600153 | -60.78218534 | 0.000213176 | | 0.476397957 | 4.99565E-05 | 229.7088425 | 0.006693914 | 3.404016108 | 0.009993083 | | 153.6054916 | 0.001269104 | -3.461647047 | 0.004840324 | -232.3595662 | 0.006040014 | | 2.565667616 | 0.004982694 | 200.3784233 | 0.001562708 | 1.324628489 | 0.001513228 | | -283.9647359 | 0.006735629 | -2.513505998 | 0.002916835 | 335.6024011 | 0.004383551 | | -2.895287433 | 0.002668928 | 285.6218863 | 0.003631667 |  |  | |

|  |
| --- |
| **附录3：**Task 2 代码 |
| 介绍：第二部分求解的代码 |
| data(Affairs, package="AER") # 调用AER包数据集Affairs。  library(openxlsx)  library(ggplot2)  library(rms)  library(caret)  h <- head(Affairs) # 显示数据前6行。  write.xlsx(h, "head.xlsx", sheetName = "Sheet1")    s <- summary(Affairs) # 数据描述。  write.xlsx(s, "summary.xlsx", sheetName = "Sheet1")  # 查看数据响应数据的各个取值的占比情况  table(Affairs$affairs)  # 将非数值型变量转换为数值型  Affairs$gender <- as.numeric(Affairs$gender)  Affairs$children <- as.numeric(Affairs$children)  # 计算Spearman相关系数矩阵  correlation\_matrix <- cor(Affairs, method = "spearman")  # 打印相关系数矩阵  print(correlation\_matrix)  table(Affairs$affairs) # 数据集Affairs中affairs列频数统计。  Affairs$ynaffair[Affairs$affairs > 0] <- 1 # 添加新列ynaffairs，原affairs列中值大于0，赋值1。  Affairs$ynaffair[Affairs$affairs == 0] <- 0 # 添加新列ynaffairs，原affairs列中值等于0，赋值0。  Affairs$ynaffair <- factor(Affairs$ynaffair,  levels=c(0,1),  labels=c("No","Yes")) # 新列转为因子。  table(Affairs$ynaffair) # 统计因子频数。  fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children +  religiousness + education + occupation +rating,  data=Affairs,family=binomial()) # Logistic回归，响应变量为ynaffair，预测变量为gender，age，yearsmarried，children，religiousness，education，occupation，rating。  summary(fit.full) # 返回回归结果。  fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness +  rating, data=Affairs, family=binomial()) # 去除不显著预测变量，重新拟合模型。  summary(fit.reduced) # 返回结果。  anova(fit.reduced, fit.full, test="Chisq") # 两个拟合模型的对比。  coef(fit.reduced) # 查看fit.reduced模型的回归系数。  exp(coef(fit.reduced)) # 指数化结果。  exp(confint(fit.reduced)) # 在优势比尺度上得到系数95%的置信区间。  # 定义变量的范围限制  ddist <- datadist(Affairs)  options(datadist = "ddist")  # 创建逻辑回归模型  fit.lmreduced <- lrm(ynaffair ~ age + yearsmarried + religiousness + rating,  data = Affairs,  x = TRUE,  y = TRUE)  # 创建nomogram  nom <- nomogram(fit.lmreduced, fun = plogis, lp = FALSE)  # 绘制nomogram  plot(nom)  call<-calibrate(fit.lmreduced,method='boot',B=100)  plot(call)  fit.lmreduced  deviance(fit.reduced)/df.residual(fit.reduced)  set.seed(666)  # 创建一个控制交叉验证的参数  control <- trainControl(method = "cv", # 使用交叉验证  number = 10) # 设置交叉验证折数  # 训练模型并进行交叉验证  fit.reduced <- train(ynaffair ~ age + yearsmarried + religiousness + rating,  data = Affairs,  method = "glm",  family = binomial(),  trControl = control)  # 输出交叉验证结果  print(fit.reduced)  fit.aff <- glm(ynaffair ~ age + yearsmarried + religiousness +  rating, family = binomial(), data = Affairs) # 模型fit.aff构建。  fit.od <- glm(ynaffair ~ age + yearsmarried + religiousness +  rating, family = quasibinomial(), data = Affairs) # 模型fit.od构建。  pchisq(summary(fit.od)$dispersion \* fit.aff$df.residual,  fit.aff$df.residual, lower = F) # 过度离势检验。 |