一元及多元线性回归分析实验报告

姓名：刘若水

班级：信息管理与信息系统-22

学号：2022312123

linear regression

2024-06-11

# 一、实验目的

通过实验本节将学习如何使用R来实现一元线性回归和多元线性回归的的分析，通过学习将掌握一元及多元线性回归的一般分析过程，能够选择合适的方析，并解释分析结果.同时把核心代码、图、表、解释结果放在报告之中.

# 二、数据来源

问题1的数据来源openintro数据包里面的starbucks；

问题2的数据来源为Nels2.csv.

# 三、实验内容

* 1. **一元线性回归**

openintro包中的starbucks数据集包含了77周星巴克的食品营养信息.对于此问题，你将探索这些物品中卡路里（calories）和碳水化合物（carb）之间的关系.

* + 1. 使用描述统计相关函数（例如，describe()）,对数据集进行描述性分析

1.starbucks

library(openintro)

(1)describe starbucks

library(psych)

## Warning: 程辑包'psych'是用R版本4.3.3 来建造的

describe(starbucks)

## vars n mean sd median trimmed mad min max range skew  
## item\* 1 77 39.00 22.37 39 39.00 28.17 1 77 76 0.00  
## calories 2 77 338.83 105.37 350 344.76 103.78 80 500 420 -0.57  
## fat 3 77 13.77 7.10 13 13.60 7.41 0 28 28 0.20  
## carb 4 77 44.87 16.55 45 44.51 20.76 16 80 64 0.07  
## fiber 5 77 2.22 2.11 2 2.03 2.97 0 7 7 0.46  
## protein 6 77 9.48 8.08 7 8.51 5.93 0 34 34 1.09  
## type\* 7 77 2.51 2.03 1 2.17 0.00 1 7 6 1.07  
## kurtosis se  
## item\* -1.25 2.55  
## calories -0.53 12.01  
## fat -0.77 0.81  
## carb -1.05 1.89  
## fiber -0.99 0.24  
## protein 0.52 0.92  
## type\* -0.26 0.23

3.1.2 创建此数据的散点图，并在屏幕上显示卡路里 （Y轴）和碳水化合物克 (X轴)的关系

(2)Draw geom\_point

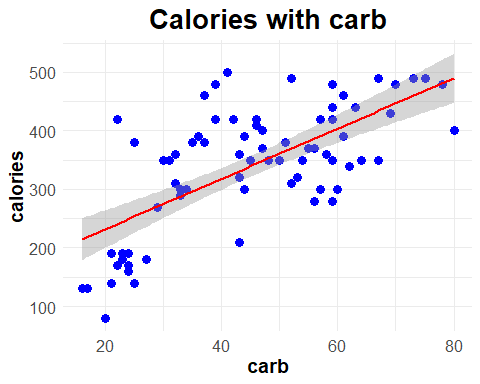
library(ggplot2)

##   
## 载入程辑包：'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

ggplot(starbucks, aes(x = carb, y = calories)) +  
 geom\_point(color = 'blue', shape = 16, size = 3) +  
 geom\_smooth(method = 'lm', color = 'red') +  
 ggtitle("Calories with carb") +  
 xlab("carb") +  
 ylab("calories") +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(hjust = 0.5, size = 20, face = "bold"),  
 axis.title = element\_text(size = 14, face = "bold"),  
 axis.text = element\_text(size = 12)  
 )

## `geom\_smooth()` using formula = 'y ~ x'

 (3)linear

regression for starbucks

· **意义**：通过散点图，可以直观地看到卡路里与碳水化合物之间的线性关系。线性回归线帮助我们了解数据的趋势和模型的拟合效果。

· **输出结果解释**：图中红线表示拟合的线性回归模型，散点显示了每个食品的卡路里和碳水化合物含量。由图中趋势线可以看出两者间存在正相关关系。

对这个数据进行简单的线性回归，以碳水化合物克为解释变量，以卡路里为因变量.使用lm()功能

model <- lm(calories ~ carb, data = starbucks)  
summary(model)

##   
## Call:  
## lm(formula = calories ~ carb, data = starbucks)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -151.962 -70.556 -0.636 54.908 179.444   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 146.0204 25.9186 5.634 2.93e-07 \*\*\*  
## carb 4.2971 0.5424 7.923 1.67e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 78.26 on 75 degrees of freedom  
## Multiple R-squared: 0.4556, Adjusted R-squared: 0.4484   
## F-statistic: 62.77 on 1 and 75 DF, p-value: 1.673e-11

· **意义**：线性回归模型提供了因变量（卡路里）和自变量（碳水化合物）之间的线性关系。lm() 函数用于拟合线性模型，summary() 提供了模型的详细信息。

· **输出结果解释**：

· 截距（Intercept）：146.02，表示当碳水化合物为0时，预测的卡路里值。

* 斜率（carb）：4.30，表示碳水化合物每增加1克，卡路里预计增加4.30。
* p值：<0.001，表明碳水化合物对卡路里的影响在统计上显著。
* R2R^2R2：0.456，表示45.6%的卡路里变化可以用碳水化合物的变化来解释。

confint(model,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 94.387896 197.652967  
## carb 3.216643 5.377526

analysis of Variance table

anova(model)

## Analysis of Variance Table  
##   
## Response: calories  
## Df Sum Sq Mean Sq F value Pr(>F)   
## carb 1 384453 384453 62.772 1.673e-11 \*\*\*  
## Residuals 75 459342 6125   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

coefficients <- coef(model)  
intercept <- coefficients[1]  
slope <- coefficients[2]  
  
cat("Intercept :", intercept, "\n")

## Intercept : 146.0204

cat("Slope :", slope, "\n")

## Slope : 4.297084

模型的判定系数R²

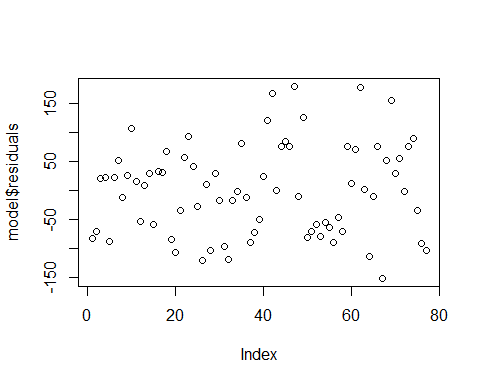
r\_squared <- summary(model)$r.squared  
r\_squared

## [1] 0.4556237

解释：约45.56%的卡路里变化可以用碳水化合物的变化来解释。

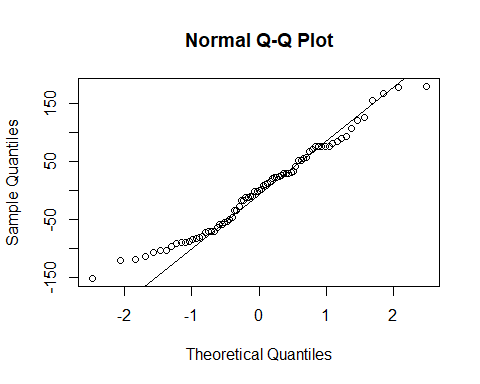
检验一元线性回归的假定 残差图

plot(model$residuals)



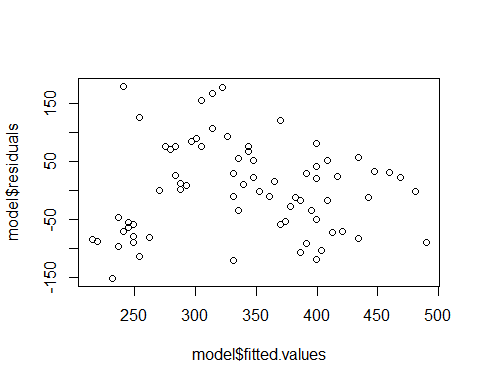
正态性检验

qqnorm(model$residuals)  
qqline(model$residuals)



方差齐性检验

plot(model$fitted.values, model$residuals)



**输出结果解释**：残差图和Q-Q图显示残差分布较为随机且接近正态分布，拟合值与残差图显示残差方差较为一致，符合线性回归模型假设。

## 多元线性回归

NELS2.csv数据集中包含对某地高中生学习情况的长期调研，试图研究哪些因素会影响高中生的学习成绩，该数据集包含下列变量：

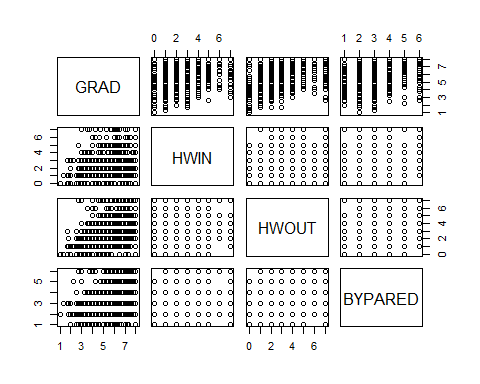
* GRAD：每位高中生英语、数学、自然科学、社会科学课程的平均分
* HWIN：每位高中生在学校花在写作业的时间
* 0=无；
* 1=不足1小时；
* 2=2-3小时；
* 3=4-6小时；
* 4=7-9小时；
* 5=10-12小时；
* 6=13-15小时；
* 7=超过15小时
* HWOUT：每位高中生在家里花在作业上的时间
  + 0=无；
  + 1=不足1小时；
  + 2=2-3小时；
  + 3=4-6小时；
  + 4=7-9小时；
  + 5=10-12小时；
  + 6=13-15小时；
  + 7=超过15小时
* BYPARED：每个高中生的父亲或母亲的最高学历
  + 1=高中未毕业；
  + 2=高中毕业；
  + 3=大专；
  + 4=本科；
  + 5=硕士；
  + 6=博士；

nels2 <- read.csv("D:\\mycodelife\\workshop\\course\_task\\NELS2.csv")  
library(psych)  
describe(nels2)

## vars n mean sd median trimmed mad min max range  
## ID 1 909 46230.19 26347.15 45851.00 46959.15 34087.94 1249 91991 90742  
## GRAD 2 909 5.69 1.47 5.75 5.76 1.48 1 8 7  
## HWIN 3 909 2.07 1.52 2.00 1.87 1.48 0 7 7  
## HWOUT 4 909 2.54 1.73 2.00 2.36 1.48 0 7 7  
## BYPARED 5 909 3.21 1.27 3.00 3.16 1.48 1 6 5  
## skew kurtosis se  
## ID -0.21 -1.45 873.88  
## GRAD -0.44 -0.33 0.05  
## HWIN 1.30 1.74 0.05  
## HWOUT 0.89 0.25 0.06  
## BYPARED 0.40 -0.14 0.04

1. 创建散点图 创建学生成绩（Y轴）与三个自变量（X轴）的散点图

pairs(~GRAD + HWIN + HWOUT + BYPARED, data = nels2)

 3. 相关性分析

cor(nels2[, c("GRAD", "HWIN", "HWOUT", "BYPARED")])

## GRAD HWIN HWOUT BYPARED  
## GRAD 1.00000000 0.09639305 0.3231861 0.30429242  
## HWIN 0.09639305 1.00000000 0.2750130 0.05898967  
## HWOUT 0.32318606 0.27501303 1.0000000 0.27097207  
## BYPARED 0.30429242 0.05898967 0.2709721 1.00000000

1. 线性模型拟合

model\_multi <- lm(GRAD ~ HWIN + HWOUT + BYPARED, data = nels2)  
  
summary(model\_multi)

##   
## Call:  
## lm(formula = GRAD ~ HWIN + HWOUT + BYPARED, data = nels2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.7823 -0.9372 0.1180 0.9986 3.1818   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.24239 0.13538 31.337 < 2e-16 \*\*\*  
## HWIN 0.01160 0.03057 0.379 0.704   
## HWOUT 0.21792 0.02801 7.780 1.97e-14 \*\*\*  
## BYPARED 0.27052 0.03668 7.375 3.72e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.35 on 905 degrees of freedom  
## Multiple R-squared: 0.1553, Adjusted R-squared: 0.1525   
## F-statistic: 55.45 on 3 and 905 DF, p-value: < 2.2e-16

**斜率解释**：

* HWIN: 学校写作业时间对成绩影响不显著。
* HWOUT: 家里写作业时间每增加一单位，成绩预计增加0.218。
* BYPARED: 家长学历每增加一个等级，成绩预计增加0.271。

· 截距（Intercept）：4.24，表示当所有自变量为0时，预测的成绩。

· HWIN：0.0116，表示在学校作业时间对成绩影响不显著。

· HWOUT：0.218，表示家里作业时间每增加一个单位，成绩预计增加0.218。

· BYPARED：0.271，表示家长学历每增加一个等级，成绩预计增加0.271。

· p值：显示HWOUT和BYPARED的影响显著，HWIN的影响不显著。

· R2R^2R2：0.155，表示15.5%的成绩变化可以用这些变量来解释。

1. 模型的判定系数R²及调整后的R²

r\_squared\_multi <- summary(model\_multi)$r.squared  
r\_squared\_multi

## [1] 0.1552722

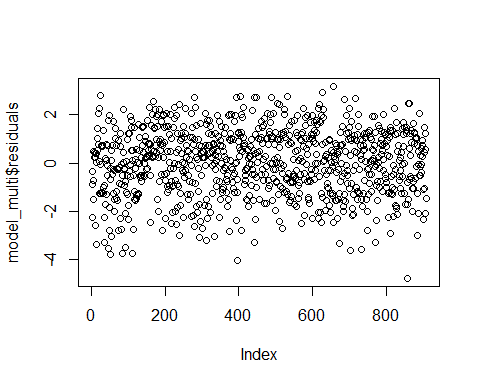
adj\_r\_squared\_multi <- summary(model\_multi)$adj.r.squared  
adj\_r\_squared\_multi

## [1] 0.152472

调整后的R2R^2R2比普通R2R^2R2更适合多变量的模型。

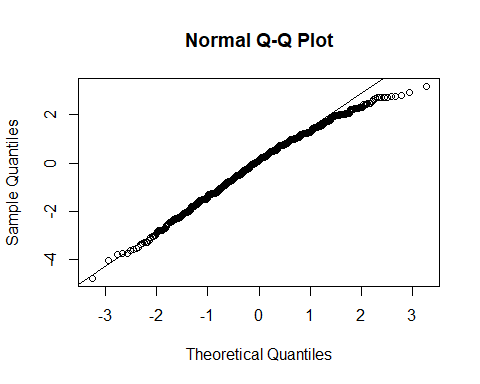
1. 检验多元回归的假定 残差图

plot(model\_multi$residuals)



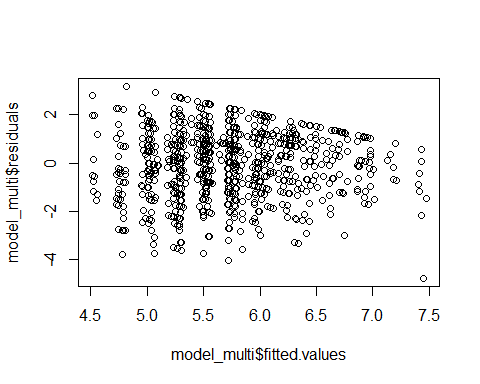
正态性检验

qqnorm(model\_multi$residuals)  
qqline(model\_multi$residuals)



方差齐性检验

plot(model\_multi$fitted.values, model\_multi$residuals)



7. 检验多重共线性

library(car)

vif(model\_multi)

## HWIN HWOUT BYPARED   
## 1.082125 1.163814 1.079548

**输出结果解释**：残差图和Q-Q图显示残差分布接近正态，拟合值与残差图显示方差齐性较好，VIF均小于5，表明无显著多重共线性。

1. 结果整理成规范的表格

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

stargazer(model\_multi, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## GRAD   
## -----------------------------------------------  
## HWIN 0.012   
## (0.031)   
##   
## HWOUT 0.218\*\*\*   
## (0.028)   
##   
## BYPARED 0.271\*\*\*   
## (0.037)   
##   
## Constant 4.242\*\*\*   
## (0.135)   
##   
## -----------------------------------------------  
## Observations 909   
## R2 0.155   
## Adjusted R2 0.152   
## Residual Std. Error 1.350 (df = 905)   
## F Statistic 55.450\*\*\* (df = 3; 905)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

1. 计算置信区间和预测区间

confint(model\_multi)

## 2.5 % 97.5 %  
## (Intercept) 3.97669829 4.50808009  
## HWIN -0.04839155 0.07159145  
## HWOUT 0.16294680 0.27288361  
## BYPARED 0.19852996 0.34251970

# 预测区间  
predict(model\_multi, interval = "prediction")

## Warning in predict.lm(model\_multi, interval = "prediction"): 用当前数据得到的预测结果对\_未来\_响应有用

## fit lwr upr  
## 1 5.271879 2.617847 7.925911  
## 2 5.219269 2.564411 7.874128  
## 3 5.001354 2.346484 7.656224  
## 4 5.707709 3.053353 8.362065  
## 5 5.489794 2.836169 8.143419  
## 6 5.001354 2.346484 7.656224  
## 7 5.542404 2.887257 8.197550  
## 8 5.760319 3.105975 8.414663  
## 9 5.271879 2.617847 7.925911  
## 10 5.271879 2.617847 7.925911

### 实验结论

通过实验，我们学习了一元和多元线性回归的基本分析方法，并使用R语言进行了实际操作。结果表明，碳水化合物对卡路里的影响显著，而家庭作业时间和家长学历对学生成绩的影响显著，这些结论从统计学角度揭示了变量之间的关系。