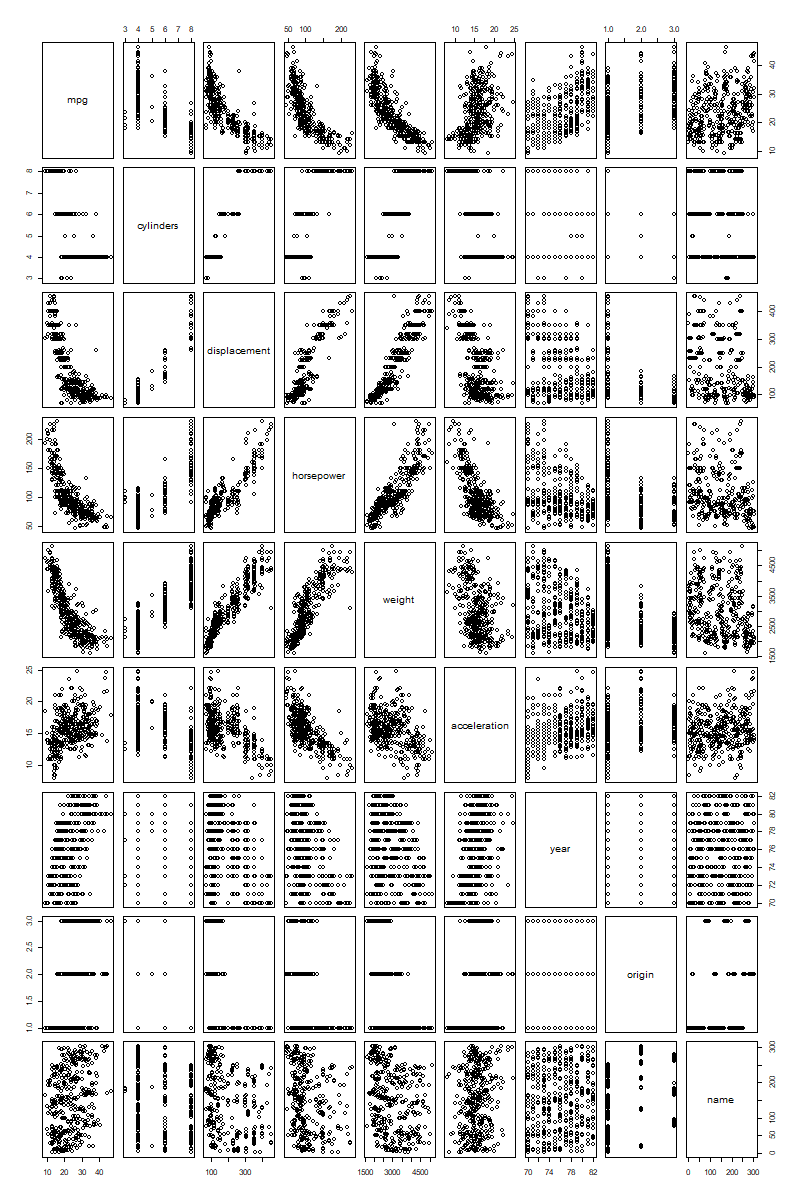
9.

# (a) 加载ISLR包

library(ISLR)

# 绘制所有变量散点图

pairs(Auto)



# (b) 计算相关系数矩阵

## 提取除name列外的变量

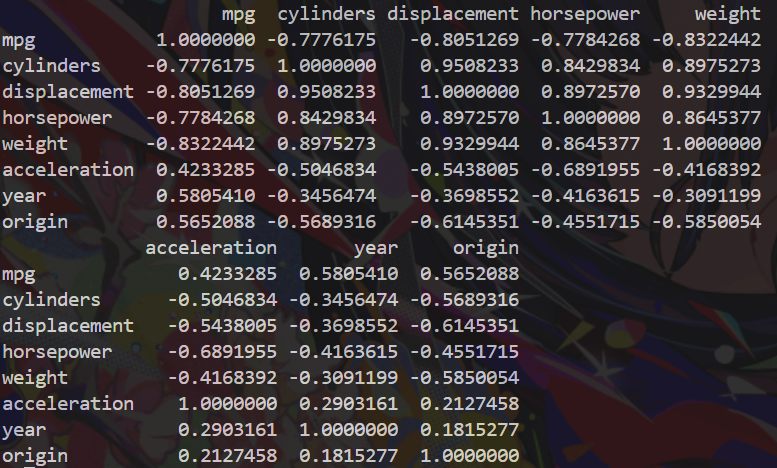
x <- Auto[, -which(names(Auto) == "name")]

## 计算相关系数矩阵

cor\_matrix <- cor(x)

## 打印相关系数矩阵

print(cor\_matrix)

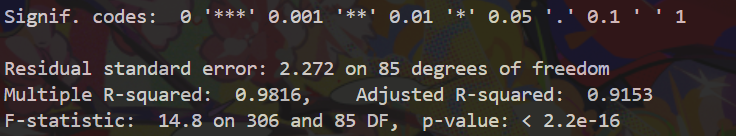


# (c) 多元线性回归建模并分析结果

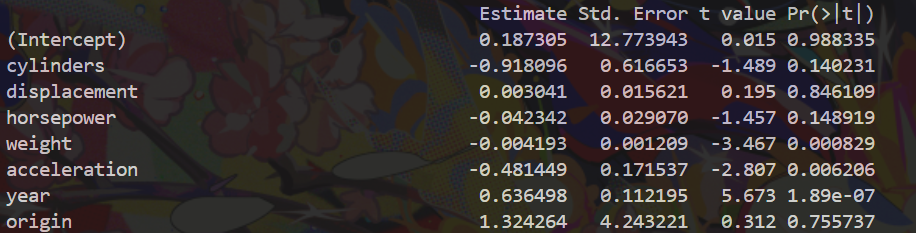
model\_all <- lm(mpg ~ ., data = Auto)

print("多元线性回归建模结果:")

print(summary(model\_all))



整体p检验的值为 2.2e-16，远小于0.05，说明至少有一个回归变量与mpg存在显著的线性关系



Pr(>|t|)项中，小于0.05的项有weight, acceleration, year这些变量与mpg的线性关系显著

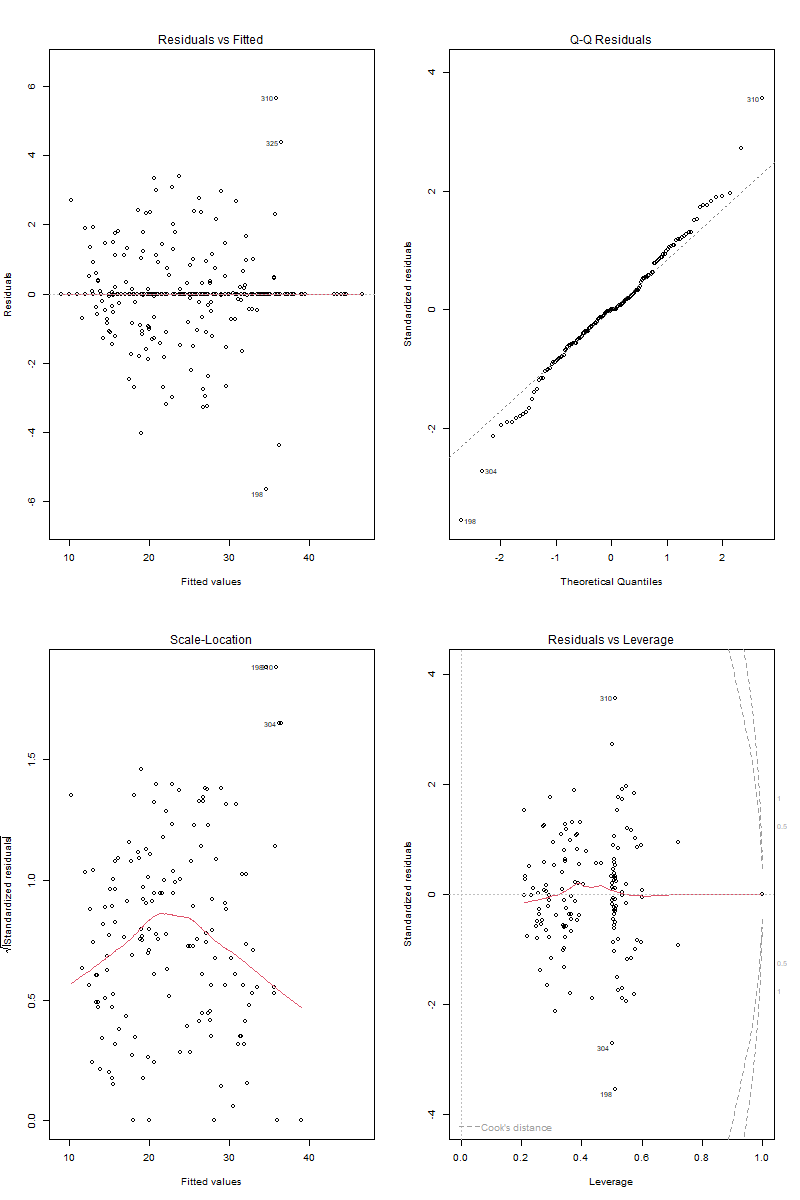
year的Estimate项值为0.63，表示每增加一年车龄，汽车的油耗将上升

# (d) 回归诊断图与异常点分析

## 生成诊断图

par(mfrow = c(2, 2))

plot(model\_all)



残差图存在离群点

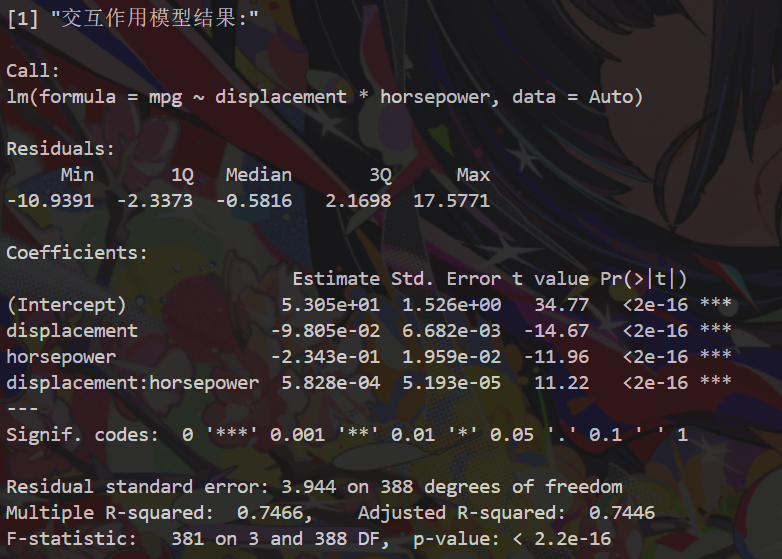
存在点198、304、310有异常高杠杆作用

# (e) 交互作用模型与显著性检验

model\_interact <- lm(mpg ~ displacement \* horsepower, data = Auto)

print("交互作用模型结果:")

print(summary(model\_interact))



p值为2.2e-16，远小于0.05，可认定displacement 和 horsepower的交互作用显著

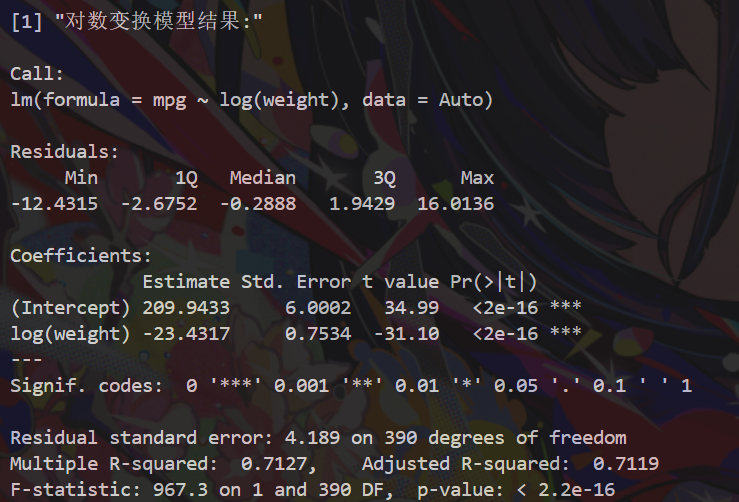
# (f) 预测变量的变换与结果分析

## 1. 对数变换

model\_log <- lm(mpg ~ log(weight), data = Auto)

print("对数变换模型结果:")

print(summary(model\_log))



## 2. 平方根变换

model\_sqrt <- lm(mpg ~ sqrt(weight), data = Auto)

print("平方根变换模型结果:")

print(summary(model\_sqrt))

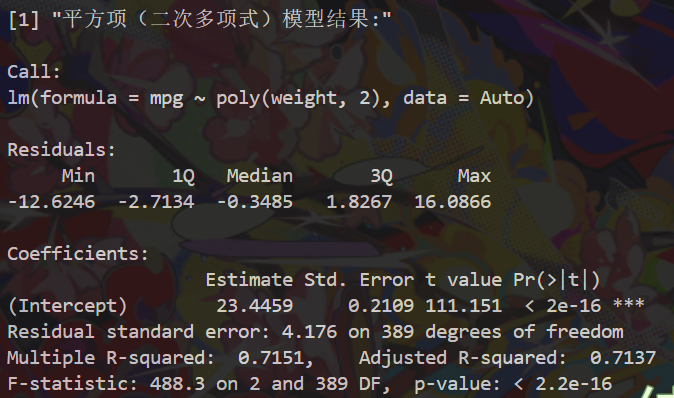


## 3. 平方项（二次多项式）

model\_poly <- lm(mpg ~ poly(weight, 2), data = Auto)

print("平方项（二次多项式）模型结果:")

print(summary(model\_poly))



从这三种变换中，R方最大的是平方项变换，模型的可解释度更高

10.

# (a) 拟合多元回归模型

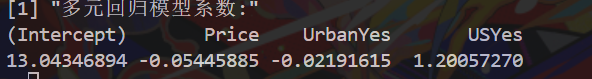
library(ISLR)   # 加载包含Carseats数据集的包

data(Carseats)  # 加载数据集

fit\_a <- lm(Sales ~ Price + Urban + US, data = Carseats)  # 拟合模型

print("多元回归模型系数:")

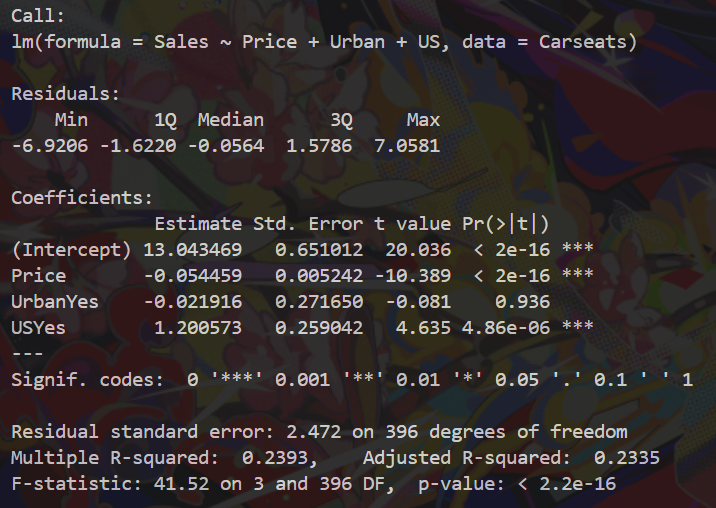
print(coef(fit\_a))



(b)

| 变量 | 系数估计值 | 含义 |
| --- | --- | --- |
| (Intercept) | 13.0435 | 当所有变量为0时的预期销售量（基准情况） |
| Price | -0.0545 | 价格每增加1单位，销售量平均下降约 0.0545 单位 |
| UrbanYes | -0.0219 | 城市店铺相比非城市店铺，销售量平均低 0.0219 单位 |
| USYes | 1.2006 | 在美国销售的产品比不在美国销售的平均多卖 1.2006 单位 |

(c)Sales = 13.04346894 - 0.05445885 × Price - 0.02191615 × UrbanYes + 1.20057270 × USYes + ε

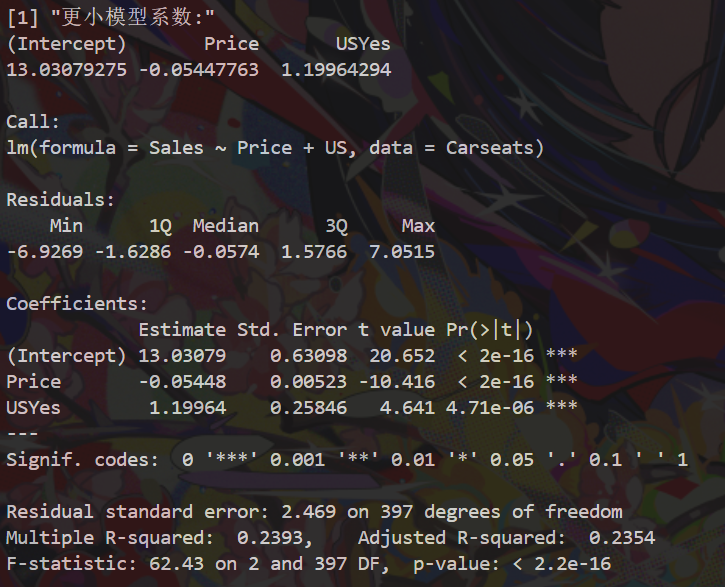
(d)

Price和 USYes的p值小于0.05可以拒绝零假设

# (e) 拟合更小模型

fit\_e <- lm(sales ~ Price + US, data = Carseats)  # 仅保留显著变量

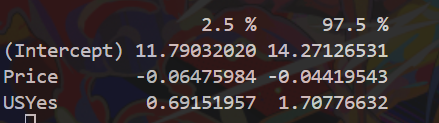
(f)



更小模型的R方更大，拟合度比完整模型更好

# (g) 计算95%的置信区间

print(confint(fit\_e))



# (h) 离群点和高杠杆点检测

rstudent\_fit\_e <- rstudent(fit\_e)

outliers <- abs(rstudent\_fit\_e) > 3

print(sum(outliers))  # 统计离群点数量

hatvals <- hatvalues(fit\_e)

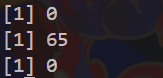
high\_leverage <- hatvals > (2 \* 2)/400  # 2k/n = 0.01

print(sum(high\_leverage))  # 统计高杠杆点数量

cookd <- cooks.distance(fit\_e)

influential <- cookd > 1

print(sum(influential))  # 统计强影响点数量



无离群点，但存在高杠杆点

14.

(a)

线性模型的函数形式为：

y=β0+β1x1+β2x2+ε

其中，β0,β1,β2是回归系数，ε是随机误差项（由 rnorm(100)生成，均值为0，方差默认1）。

根据代码 y = 2 + 2\*x1 + 0.3\*x2 + rnorm(100)，真实回归系数为：

β0=2，β1=2，β2=0.3。

(b)

# (b)

# 计算相关系数

cor\_x1x2 <- cor(x1, x2)

cat("x1与x2的相关系数：", cor\_x1x2, "\n")

# 绘制散点图（辅助观察线性趋势）

plot(x1, x2,

    main = "Scatter Plot of x1 vs x2",

    xlab = "x1",

    ylab = "x2",

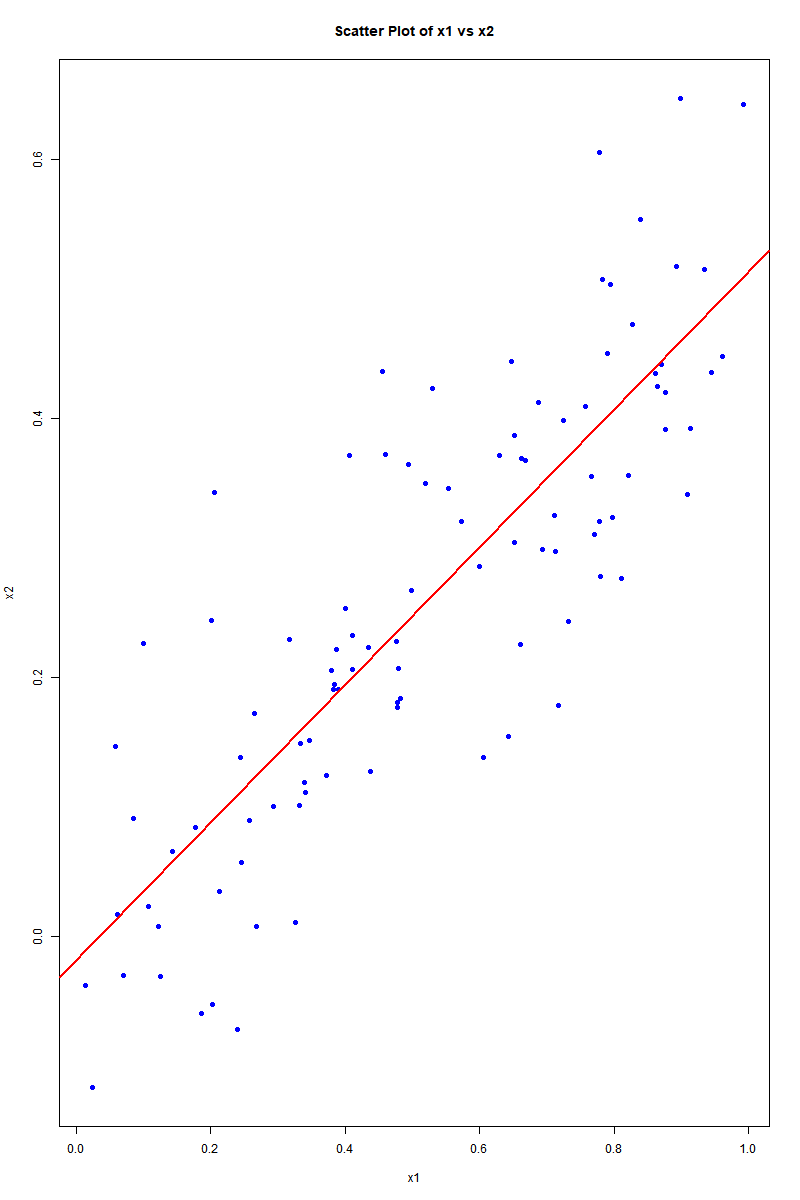
    pch = 16,

    col = "blue"

)

# 添加回归线（辅助观察线性趋势）

abline(lm(x2 ~ x1), col = "red", lwd = 2)



# (c)

# 拟合模型：y ~ x1 + x2

fit\_c <- lm(y ~ x1 + x2)

# 输出回归结果摘要（含系数、显著性检验等）

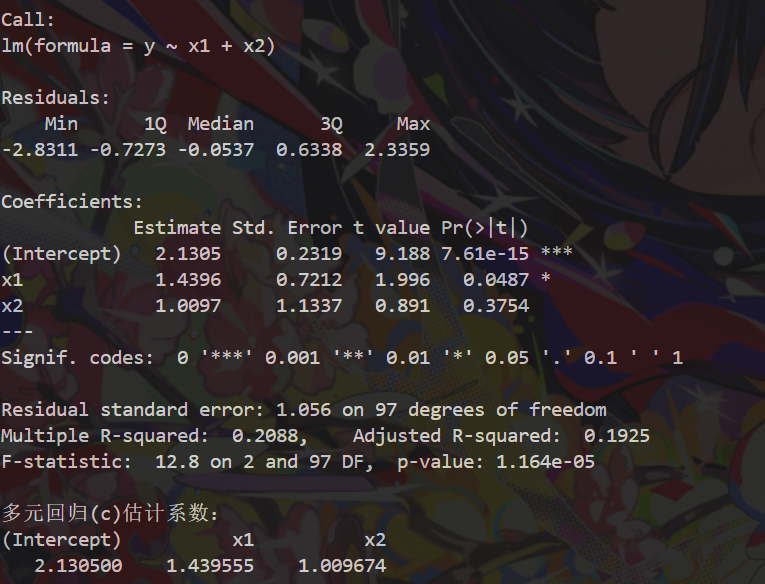
print(summary(fit\_c))

# 提取并打印系数（也可通过coef(fit\_c)直接获取）

coef\_c <- coef(fit\_c)

cat("多元回归(c)估计系数：\n")

print(coef\_c)



p值为 1.164e-05 小于0.05 可以拒绝零假设

β0 = 2.13， β1=1.43， β2=1.00

与真实值接近，存在误差

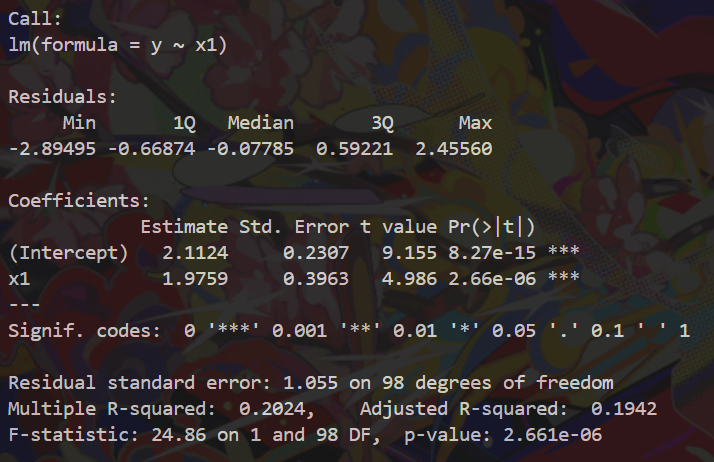
# (d)

# 拟合模型：y ~ x1

fit\_d <- lm(y ~ x1)

# 输出回归结果摘要

print(summary(fit\_d))



能够拒绝零假设

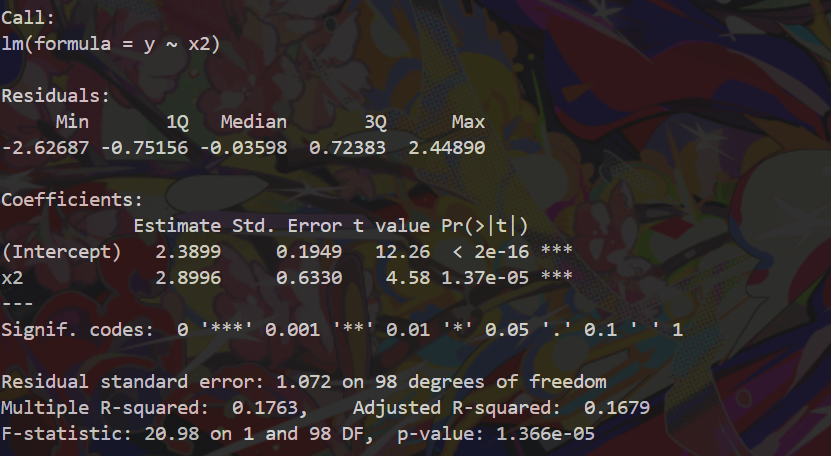
# (e)

# 拟合模型：y ~ x2

fit\_e <- lm(y ~ x2)

# 输出回归结果摘要

print(summary(fit\_e))



能够拒绝零假设

(f)

​(c)(d)(e)结果不矛盾,是多重共线性的典型表现：x1与 x2高度相关，导致多元回归中各自系数的“解释权”被分散，估计值与单变量回归差异大；单变量回归中 x1,x2均显著，多元回归中系数估计虽变但仍有显著性（因共线性未完全消除解释力），本质是共线性引发系数估计不稳定。

(g)

# (g)

# --------------- 分析新观测的“离群点”与“高杠杆点”属性 ---------------

## 1. 离群点：学生化残差（rstudent）绝对值大（通常>2或3）

rstudent\_c\_new <- rstudent(fit\_c\_new)  # 多元回归的学生化残差

rstudent\_d\_new <- rstudent(fit\_d\_new)  # 单变量x1的学生化残差

rstudent\_e\_new <- rstudent(fit\_e\_new)  # 单变量x2的学生化残差

## 2. 高杠杆点：杠杆值（hatvalues）远大于阈值（2p/n，p=2, n=101 → 2 \* 2/101≈0.0396）

hatvalues\_c\_new <- hatvalues(fit\_c\_new)  # 多元回归的杠杆值

# 输出新观测（第101个数据点）的诊断结果

cat("\n===== 新观测对模型的影响分析 =====\n")

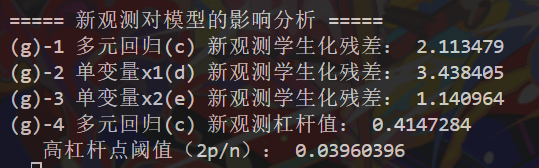
cat("(g)-1 多元回归(c) 新观测学生化残差：", rstudent\_c\_new[101], "\n")

cat("(g)-2 单变量x1(d) 新观测学生化残差：", rstudent\_d\_new[101], "\n")

cat("(g)-3 单变量x2(e) 新观测学生化残差：", rstudent\_e\_new[101], "\n")

cat("(g)-4 多元回归(c) 新观测杠杆值：", hatvalues\_c\_new[101], "\n")

cat("   高杠杆点阈值（2p/n）：", 2 \* 2/101, "\n")



离群点：残差（y−y^）绝对值大。新观测的 y=6远高于原数据均值（ 3.135623），且模型预测值 y^远小于6，残差极大，属于离群点​​。

高杠杆点：杠杆值 hii​反映观测对模型拟合的影响权重。新观测的 (x1​,x2​)=(0.1,0.8)远离原数据分布（原 x1​∈[0,1], x2​∈[0,0.7]），在自变量空间中远离中心，杠杆值 hii​远大于阈值（2p/n≈0.08，p=2，n=101），属于高杠杆点​​。

对模型的影响：高杠杆+离群点会显著扭曲系数估计（尤其共线性下，此类点对回归平面的方向和截距影响极大），导致(c)(d)(e)的系数估计、拟合优度等指标剧烈变化。