数据预测与建模

刘若水

2024-11-05

library(brms)  
library(dplyr)  
library(ggplot2)  
library(bayesplot)  
library(psych)  
library(car)  
library(lmtest)  
library(MASS)

一、描述性统计

h2\_data <- read.csv("D:\\mycodelife\\workshop\\course\_task\\Salary Data.csv")

检视数据

summary(h2\_data)

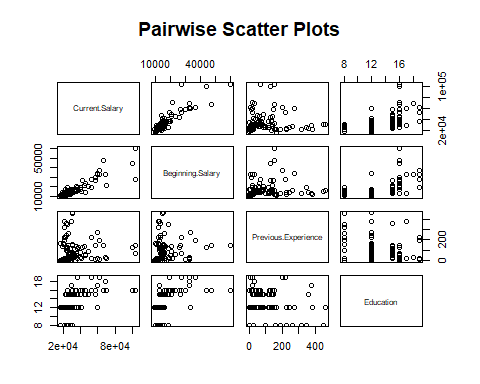
## Current.Salary Beginning.Salary Previous.Experience Education   
## Min. : 16350 Min. : 9750 Min. : 0.00 Min. : 8.00   
## 1st Qu.: 23325 1st Qu.:12000 1st Qu.: 17.75 1st Qu.:12.00   
## Median : 27825 Median :14250 Median : 60.50 Median :12.00   
## Mean : 33633 Mean :16939 Mean : 95.61 Mean :13.22   
## 3rd Qu.: 35775 3rd Qu.:16575 3rd Qu.:139.25 3rd Qu.:15.00   
## Max. :103750 Max. :60000 Max. :460.00 Max. :19.00

describe(h2\_data)

## vars n mean sd median trimmed mad min  
## Current.Salary 1 100 33632.60 17289.55 27825.0 30215.12 8517.54 16350  
## Beginning.Salary 2 100 16938.60 8347.83 14250.0 15107.25 3335.85 9750  
## Previous.Experience 3 100 95.61 105.48 60.5 76.01 78.58 0  
## Education 4 100 13.22 2.72 12.0 13.32 4.45 8  
## max range skew kurtosis se  
## Current.Salary 103750 87400 2.24 5.47 1728.96  
## Beginning.Salary 60000 50250 2.63 8.08 834.78  
## Previous.Experience 460 460 1.55 2.08 10.55  
## Education 19 11 -0.12 -0.30 0.27

二、绘制因变量和自变量的散点图

pairs(h2\_data, main = "Pairwise Scatter Plots")

 该图能反映Current Salary与其他变量的关联

三、给出模型形式

1.构建模型

model <- lm(h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience + h2\_data$Education, data = h2\_data)

检视模型

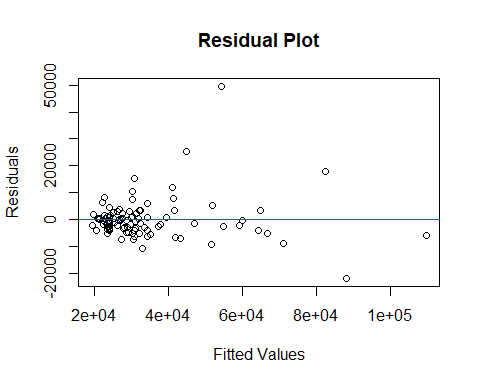
summary(model)

##   
## Call:  
## lm(formula = h2\_data$Current.Salary ~ h2\_data$Beginning.Salary +   
## h2\_data$Previous.Experience + h2\_data$Education, data = h2\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21900 -3583 -577 1418 49548   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4139.2377 4203.3582 -0.985 0.3272   
## h2\_data$Beginning.Salary 1.7302 0.1138 15.203 <2e-16 \*\*\*  
## h2\_data$Previous.Experience -10.9071 7.7710 -1.404 0.1637   
## h2\_data$Education 719.1221 351.7339 2.045 0.0436 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7791 on 96 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.7969   
## F-statistic: 130.5 on 3 and 96 DF, p-value: < 2.2e-16

2.是否满足Gauss-Markov假设

（1）检查线性性是否成立

# 绘制残差图  
plot(model$fitted.values, model$residuals, main = "Residual Plot", xlab = "Fitted Values", ylab = "Residuals")  
abline(h = 0, col = "red")

 如图，图中残差在 0 附近随机分布，则线性性假设可能成立。

（2）检查是否存在多重共线性

vif(model)

## h2\_data$Beginning.Salary h2\_data$Previous.Experience   
## 1.472233 1.095796   
## h2\_data$Education   
## 1.490282

VIF大于1且没有超过5，说明具有相关性且不存在多重共线性

（3）检查同方差性

bptest(model)

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 7.2635, df = 3, p-value = 0.06396

p>0.05，可能满足同方差假设

（4）检查残差的独立性

dwtest(model)

##   
## Durbin-Watson test  
##   
## data: model  
## DW = 1.6712, p-value = 0.04704  
## alternative hypothesis: true autocorrelation is greater than 0

残差偏离2，可能存在残差相关性

（5）残差的正态性

shapiro.test(model$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model$residuals  
## W = 0.73658, p-value = 4.198e-12

p<0.05，说明残差不满足正态性

综上所述，不满足Gauss-Markov假设，因此OLS估计是有偏的，需要采取其他估计

四、采用Robust回归

前文假定误差协方差矩阵满足，但是经过检验可能为，因此需要采用其他估计方法。

1.M估计

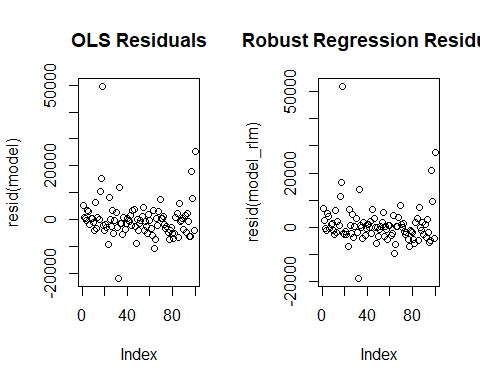
M估计通过调整回归系数估计的损失函数，使模型对离群点不敏感。常见的损失函数包括Huber损失（对小残差使用平方惩罚，对大残差使用线性惩罚）和Tukey损失。

model\_rlm <- rlm(h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience + h2\_data$Education, psi = psi.huber) # 使用Huber损失  
summary(model\_rlm)

##   
## Call: rlm(formula = h2\_data$Current.Salary ~ h2\_data$Beginning.Salary +   
## h2\_data$Previous.Experience + h2\_data$Education, psi = psi.huber)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -18716.17 -2480.96 -31.55 2146.05 51884.32   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) -1694.7814 2287.9118 -0.7408  
## h2\_data$Beginning.Salary 1.6732 0.0619 27.0098  
## h2\_data$Previous.Experience -7.2503 4.2298 -1.7141  
## h2\_data$Education 502.3916 191.4508 2.6241  
##   
## Residual standard error: 3576 on 96 degrees of freedom

显著性不明显。比较其与OLS的区别

# 绘制残差图比较  
par(mfrow = c(1, 2))  
plot(resid(model), main = "OLS Residuals")  
plot(resid(model\_rlm), main = "Robust Regression Residuals")



可见，没有明显的区别。

2.加权最小二乘回归

weights <- 1 / abs(h2\_data$Current.Salary - predict(lm(h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience + h2\_data$Education))) # 权重示例  
  
# 使用lm()函数指定权重  
model\_wls <- lm(h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience + h2\_data$Education, weights = weights)  
summary(model\_wls)

##   
## Call:  
## lm(formula = h2\_data$Current.Salary ~ h2\_data$Beginning.Salary +   
## h2\_data$Previous.Experience + h2\_data$Education, weights = weights)  
##   
## Weighted Residuals:  
## Min 1Q Median 3Q Max   
## -135.13 -53.42 -18.60 40.72 227.03   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.331e+03 9.089e+02 -3.665 0.000406 \*\*\*  
## h2\_data$Beginning.Salary 1.680e+00 5.048e-02 33.281 < 2e-16 \*\*\*  
## h2\_data$Previous.Experience -9.797e+00 2.143e+00 -4.572 1.44e-05 \*\*\*  
## h2\_data$Education 6.883e+02 8.193e+01 8.401 3.99e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 66.92 on 96 degrees of freedom  
## Multiple R-squared: 0.9479, Adjusted R-squared: 0.9463   
## F-statistic: 582.7 on 3 and 96 DF, p-value: < 2.2e-16

效果非常好，所有系数均显著。说明采用加权最小二乘回归效果较好。

五、采用逐步回归

1.前向选择

null\_model <- lm(h2\_data$Current.Salary ~ 1, data = h2\_data)  
stepwise\_model\_forward <- step(model,   
 scope = list(lower = null\_model, upper = model),   
 direction = "forward")

## Start: AIC=1796.06  
## h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience +   
## h2\_data$Education

summary(stepwise\_model\_forward)

##   
## Call:  
## lm(formula = h2\_data$Current.Salary ~ h2\_data$Beginning.Salary +   
## h2\_data$Previous.Experience + h2\_data$Education, data = h2\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21900 -3583 -577 1418 49548   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4139.2377 4203.3582 -0.985 0.3272   
## h2\_data$Beginning.Salary 1.7302 0.1138 15.203 <2e-16 \*\*\*  
## h2\_data$Previous.Experience -10.9071 7.7710 -1.404 0.1637   
## h2\_data$Education 719.1221 351.7339 2.045 0.0436 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7791 on 96 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.7969   
## F-statistic: 130.5 on 3 and 96 DF, p-value: < 2.2e-16

2.后向选择

stepwise\_model\_backward <- step(model, direction = "backward")

## Start: AIC=1796.06  
## h2\_data$Current.Salary ~ h2\_data$Beginning.Salary + h2\_data$Previous.Experience +   
## h2\_data$Education  
##   
## Df Sum of Sq RSS AIC  
## <none> 5.8270e+09 1796.1  
## - h2\_data$Previous.Experience 1 1.1958e+08 5.9466e+09 1796.1  
## - h2\_data$Education 1 2.5372e+08 6.0807e+09 1798.3  
## - h2\_data$Beginning.Salary 1 1.4029e+10 1.9856e+10 1916.7

summary(stepwise\_model\_backward)

##   
## Call:  
## lm(formula = h2\_data$Current.Salary ~ h2\_data$Beginning.Salary +   
## h2\_data$Previous.Experience + h2\_data$Education, data = h2\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21900 -3583 -577 1418 49548   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4139.2377 4203.3582 -0.985 0.3272   
## h2\_data$Beginning.Salary 1.7302 0.1138 15.203 <2e-16 \*\*\*  
## h2\_data$Previous.Experience -10.9071 7.7710 -1.404 0.1637   
## h2\_data$Education 719.1221 351.7339 2.045 0.0436 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7791 on 96 degrees of freedom  
## Multiple R-squared: 0.8031, Adjusted R-squared: 0.7969   
## F-statistic: 130.5 on 3 and 96 DF, p-value: < 2.2e-16

由于变量较少，且本身不具有显著的多重共线性，因此采用逐步回归效果较差，和直接使用OLS没有显著区别。