final-education

2154177 何应豪　　2151298 杨滕超　　2151299 苏家铭　　2151294 马威

目录

1 研究问题和假设 **1**
1.1 背景 1
1.2 现有研究 2
1.3 研究问题 2
1.4 研究假设 2

2 数据分析方法选择 **2**

3 数据分析 **2**
3.1 准备工作 2
3.2 数据预处理 3
3.3 高教入学率分析 9
3.4 失业率分析 28

4 分析结果解读 **41**
4.1 原始数据 41
4.2 高教入学率模型 42
4.3 失业率模型 44
4.4 总结 45

1 研究问题和假设

1.1 背景

教育工作在每一个国家和地区都尤为重要，因为它影响着诸多因素，如人才培养等。而全球各个国家和地区的教育情况，一直是统计学家和教育工作者所密切关注的。他们一直致力于通过数据分析，对不同国家和地区的教育动态提供深刻的见解，希望能够给全球教育工作者一个评估、加强和重塑全球教育系统的机会。

1.2 现有研究

已搜集到全球范围内近 200 个国家和地区的教育数据，包括 29 列，内容涵盖失学率、学业完成率、熟练程度、识字率、出生率以及小学和高等教育入学统计等信息。

1.3 研究问题

1. 什么因素影响一个国家或地区的高等教育入学率？
2. 什么因素影响一个国家或地区的失业率？

1.4 研究假设

1. 高教入学率与失学率成反比，与学业完成率成正比
2. 失业率可能与失学率成正比

2 数据分析方法选择

1. 本实验需要进行预测，高等教育入学率、失业率是一个特定数值，且有较多的自变量。适合使用线性回归建立预测模型

3 数据分析

3.1 准备工作

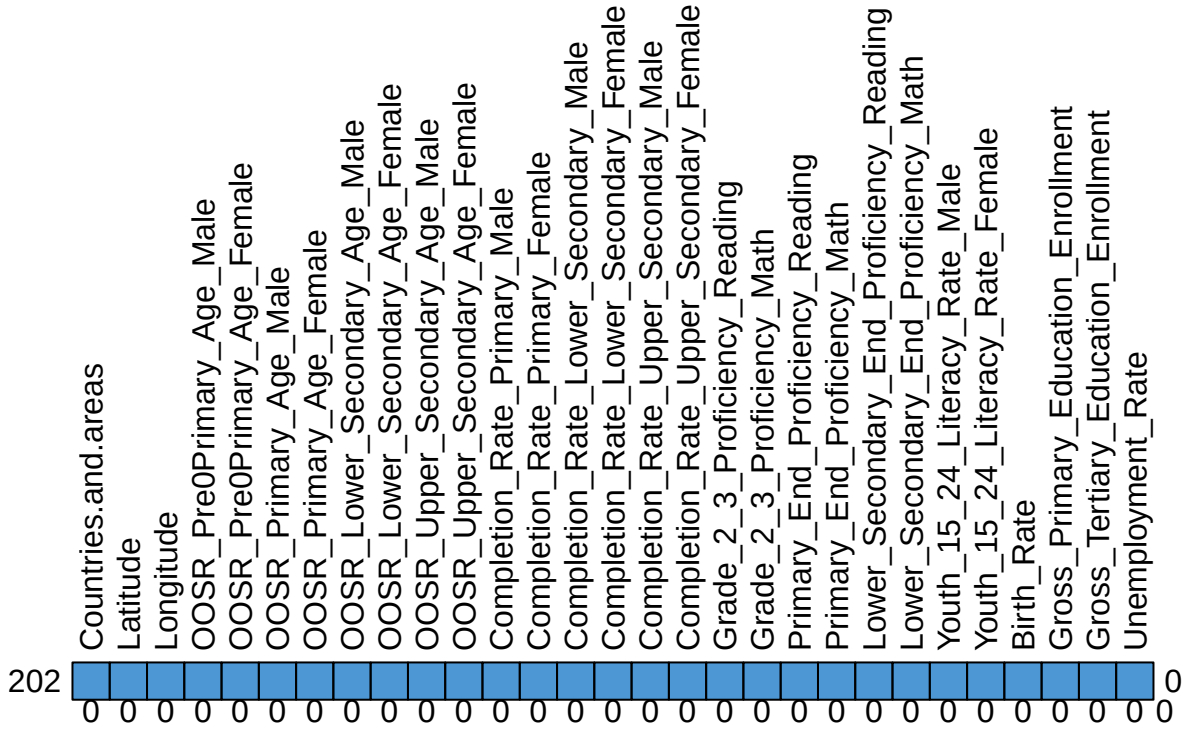
```
# 加载需要的包
library(dplyr)
library(ggplot2)
library(mice)
library(caret)
library(car)

# 清除当前镜像中的数据
rm(list = ls())

# 读取需要数据集
edu <- read.csv("education.csv", stringsAsFactors = F)
```

3.2 数据预处理

3.2.1 处理缺失值（总体）

```
# 缺失值判断：NA
md.pattern(edu, rotate.names = T)
```

```
##  /\     /\
## {  `---'  }
## {  0   0  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```




```
##     Countries.and.areas Latitude Longitude OOSR_Pre0Primary_Age_Male
## 202                   1        1         1                         1
##                       0        0         0                         0
##     OOSR_Pre0Primary_Age_Female OOSR_Primary_Age_Male OOSR_Primary_Age_Female
## 202                           1                     1                       1
##                               0                     0                       0
##     OOSR_Lower_Secondary_Age_Male OOSR_Lower_Secondary_Age_Female
```

```
## 202                              1                                1
##                                  0                                0
##     OOSR_Upper_Secondary_Age_Male OOSR_Upper_Secondary_Age_Female
## 202                              1                                1
##                                  0                                0
##     Completion_Rate_Primary_Male Completion_Rate_Primary_Female
## 202                            1                              1
##                                0                              0
##     Completion_Rate_Lower_Secondary_Male Completion_Rate_Lower_Secondary_Female
## 202                                    1                                      1
##                                        0                                      0
##     Completion_Rate_Upper_Secondary_Male Completion_Rate_Upper_Secondary_Female
## 202                                    1                                      1
##                                        0                                      0
##     Grade_2_3_Proficiency_Reading Grade_2_3_Proficiency_Math
## 202                             1                         1
##                                 0                         0
##     Primary_End_Proficiency_Reading Primary_End_Proficiency_Math
## 202                               1                           1
##                                   0                           0
##     Lower_Secondary_End_Proficiency_Reading
## 202                                      1
##                                          0
##     Lower_Secondary_End_Proficiency_Math Youth_15_24_Literacy_Rate_Male
## 202                                    1                             1
##                                        0                             0
##     Youth_15_24_Literacy_Rate_Female Birth_Rate
## 202                              1          1
##                                  0          0
##     Gross_Primary_Education_Enrollment Gross_Tertiary_Education_Enrollment
## 202                                  1                                   1
##                                      0                                   0
##     Unemployment_Rate
## 202                 1 0
##                     0 0
```

```r
any(is.na(edu))
```

```
## [1] FALSE
```

```r
# 缺失值判断：0
count_zero <- colSums(edu == 0)
print(count_zero)
```

```
##                     Countries.and.areas                                Latitude
##                                       0                                       0
##                               Longitude             OOSR_PreOPrimary_Age_Male
##                                       0                                      52
##            OOSR_PreOPrimary_Age_Female                 OOSR_Primary_Age_Male
##                                      55                                      80
##                OOSR_Primary_Age_Female          OOSR_Lower_Secondary_Age_Male
##                                      85                                      74
##        OOSR_Lower_Secondary_Age_Female          OOSR_Upper_Secondary_Age_Male
##                                      75                                      51
##        OOSR_Upper_Secondary_Age_Female           Completion_Rate_Primary_Male
##                                      51                                      95
##         Completion_Rate_Primary_Female    Completion_Rate_Lower_Secondary_Male
##                                      95                                      95
##  Completion_Rate_Lower_Secondary_Female    Completion_Rate_Upper_Secondary_Male
##                                      95                                      95
##  Completion_Rate_Upper_Secondary_Female          Grade_2_3_Proficiency_Reading
##                                      95                                     131
##             Grade_2_3_Proficiency_Math         Primary_End_Proficiency_Reading
##                                     142                                     158
##           Primary_End_Proficiency_Math Lower_Secondary_End_Proficiency_Reading
##                                     153                                     116
##    Lower_Secondary_End_Proficiency_Math          Youth_15_24_Literacy_Rate_Male
##                                     111                                     123
##       Youth_15_24_Literacy_Rate_Female                              Birth_Rate
##                                     123                                      13
##      Gross_Primary_Education_Enrollment     Gross_Tertiary_Education_Enrollment
##                                      15                                      19
##                     Unemployment_Rate
##                                      26
```

3.2.2 处理缺失值（高教入学率）

```r
# 去掉高教入学率缺失或过高的记录
edu.tertiary <- edu[edu$Gross_Tertiary_Education_Enrollment != 0.0 &
                      edu$Gross_Tertiary_Education_Enrollment <= 100.0, ]

# 影响显著的列
tertiary_remark <- c("Gross_Tertiary_Education_Enrollment")

# 单独对高教入学率和其余列建立线性回归模型
```

```r
tertiary.co <- function(column) {
  model <- lm(
    as.formula(paste("Gross_Tertiary_Education_Enrollment ~ ", column)),
    data = edu[(edu$Gross_Tertiary_Education_Enrollment != 0) & (edu[, column] != 0), ])
  sum <- summary(model)
  p.value <- sum$coefficients[column, "Pr(>|t|)"]
  num <- as.integer(sum$fstatistic["dendf"])
  cat(column, "--", p.value, "--", num)

  if (p.value < 0.05 & num >= 100) {
    tertiary_remark <<- c(tertiary_remark, column)
    cat("\n")
  }
  else {
    cat("(eliminated)", "\n")
  }

}

for (column in names(edu)[-c(1, 28)]) {
  tertiary.co(column)
}
```

```
## Latitude -- 3.602248e-20 -- 181
## Longitude -- 0.8333868 -- 181(eliminated)
## OOSR_Pre0Primary_Age_Male -- 1.170083e-15 -- 134
## OOSR_Pre0Primary_Age_Female -- 2.670608e-14 -- 135
## OOSR_Primary_Age_Male -- 3.133781e-07 -- 110
## OOSR_Primary_Age_Female -- 3.063466e-07 -- 105
## OOSR_Lower_Secondary_Age_Male -- 2.165106e-14 -- 117
## OOSR_Lower_Secondary_Age_Female -- 1.937935e-12 -- 116
## OOSR_Upper_Secondary_Age_Male -- 5.282043e-27 -- 139
## OOSR_Upper_Secondary_Age_Female -- 2.098179e-24 -- 139
## Completion_Rate_Primary_Male -- 8.523871e-15 -- 100
## Completion_Rate_Primary_Female -- 4.859146e-13 -- 100
## Completion_Rate_Lower_Secondary_Male -- 4.188157e-16 -- 100
## Completion_Rate_Lower_Secondary_Female -- 3.740781e-17 -- 100
## Completion_Rate_Upper_Secondary_Male -- 1.142834e-16 -- 100
## Completion_Rate_Upper_Secondary_Female -- 2.45233e-18 -- 100
## Grade_2_3_Proficiency_Reading -- 1.083561e-17 -- 69(eliminated)
## Grade_2_3_Proficiency_Math -- 1.820823e-05 -- 58(eliminated)
```

```
## Primary_End_Proficiency_Reading -- 7.889472e-09 -- 42(eliminated)
## Primary_End_Proficiency_Math -- 4.69121e-11 -- 47(eliminated)
## Lower_Secondary_End_Proficiency_Reading -- 5.074136e-10 -- 83(eliminated)
## Lower_Secondary_End_Proficiency_Math -- 1.572068e-09 -- 88(eliminated)
## Youth_15_24_Literacy_Rate_Male -- 3.040154e-07 -- 74(eliminated)
## Youth_15_24_Literacy_Rate_Female -- 2.24732e-06 -- 74(eliminated)
## Birth_Rate -- 1.386557e-31 -- 181
## Gross_Primary_Education_Enrollment -- 0.9025302 -- 180(eliminated)
## Unemployment_Rate -- 0.6156578 -- 171(eliminated)
```

```
# 对这些列进行均值填补
for (column in tertiary_remark) {
  edu.tertiary[, column][edu.tertiary[, column] == 0] <-
    mean(edu.tertiary[, column][edu.tertiary[, column] != 0])
}

# 剔除单独不显著的列
tertiary_remark <- c(tertiary_remark)
edu.tertiary <- edu.tertiary[, tertiary_remark]
```

3.2.3 处理缺失值（失业率）

```
# 去掉失业率缺失或过高的记录
edu.unemploy <- edu[edu$Unemployment_Rate != 0.0 & edu$Unemployment_Rate <= 100.0, ]

# 影响显著的列
unemploy_remark <- c("Unemployment_Rate")

# 单独对失业率和其余列建立线性回归模型
unemploy.co <- function(column) {
  model <- lm(
    as.formula(paste("Unemployment_Rate ~ ", column)),
    data = edu[(edu$Unemployment_Rate != 0) & (edu[, column] != 0), ])
  sum <- summary(model)
  p.value <- sum$coefficients[column, "Pr(>|t|)"]
  num <- as.integer(sum$fstatistic["dendf"])
  cat(column, "--", p.value, "--", num)

  if (p.value < 0.05 & num >= 100) {
    unemploy_remark <<- c(unemploy_remark, column)
    cat("\n")
```

```
  }
  else {
    cat("(eliminated)", "\n")
  }
}

for (column in names(edu)[-c(1, 29)]) {
  unemploy.co(column)
}
```

```
## Latitude -- 0.5539362 -- 174(eliminated)
## Longitude -- 0.00195735 -- 174
## OOSR_Pre0Primary_Age_Male -- 0.1185538 -- 129(eliminated)
## OOSR_Pre0Primary_Age_Female -- 0.1139885 -- 128(eliminated)
## OOSR_Primary_Age_Male -- 0.4324843 -- 108(eliminated)
## OOSR_Primary_Age_Female -- 0.7787355 -- 103(eliminated)
## OOSR_Lower_Secondary_Age_Male -- 0.3404487 -- 114(eliminated)
## OOSR_Lower_Secondary_Age_Female -- 0.2088867 -- 113(eliminated)
## OOSR_Upper_Secondary_Age_Male -- 0.06961998 -- 133(eliminated)
## OOSR_Upper_Secondary_Age_Female -- 0.04676744 -- 133
## Completion_Rate_Primary_Male -- 0.01214726 -- 101
## Completion_Rate_Primary_Female -- 0.008500464 -- 101
## Completion_Rate_Lower_Secondary_Male -- 0.01525886 -- 101
## Completion_Rate_Lower_Secondary_Female -- 0.004335273 -- 101
## Completion_Rate_Upper_Secondary_Male -- 0.05199858 -- 101(eliminated)
## Completion_Rate_Upper_Secondary_Female -- 0.01594323 -- 101
## Grade_2_3_Proficiency_Reading -- 0.4038369 -- 69(eliminated)
## Grade_2_3_Proficiency_Math -- 0.4385651 -- 58(eliminated)
## Primary_End_Proficiency_Reading -- 0.4732432 -- 42(eliminated)
## Primary_End_Proficiency_Math -- 0.6360228 -- 47(eliminated)
## Lower_Secondary_End_Proficiency_Reading -- 0.2610986 -- 83(eliminated)
## Lower_Secondary_End_Proficiency_Math -- 0.1341811 -- 88(eliminated)
## Youth_15_24_Literacy_Rate_Male -- 0.5469852 -- 73(eliminated)
## Youth_15_24_Literacy_Rate_Female -- 0.4119828 -- 73(eliminated)
## Birth_Rate -- 0.5668487 -- 174(eliminated)
## Gross_Primary_Education_Enrollment -- 0.6488147 -- 173(eliminated)
## Gross_Tertiary_Education_Enrollment -- 0.6156578 -- 171(eliminated)
```

```
# 对这些列进行均值填补
for (column in unemploy_remark) {
  edu.unemploy[, column][edu.unemploy[, column] == 0] <-
    mean(edu.unemploy[, column][edu.unemploy[, column] != 0])
```

```
}

# 剔除单独不显著的列
unemploy_remark <- c(unemploy_remark)
edu.unemploy <- edu.unemploy[, unemploy_remark]

# 对这些列进行均值填补
for (column in tertiary_remark) {
  edu.tertiary[, column][edu.tertiary[, column] == 0] <- mean(edu.tertiary[, column][edu.tertiary[
}

# 剔除单独不显著的列
tertiary_remark <- c(tertiary_remark)
edu.tertiary <- edu.tertiary[, tertiary_remark]
```

3.3 高教入学率分析

3.3.1 基础分析

```
# 统计概要
summary(edu.tertiary)
```

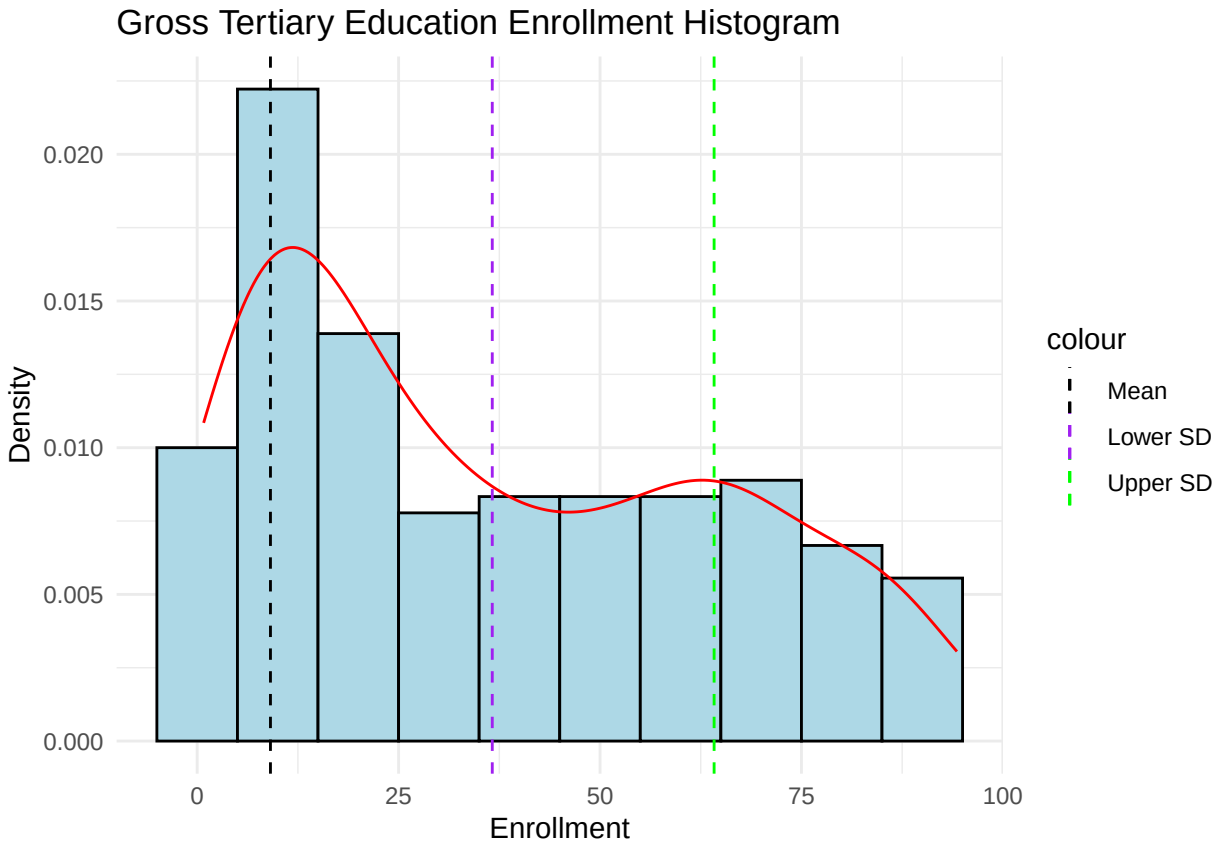
```
##  Gross_Tertiary_Education_Enrollment    Latitude
##  Min.   : 0.80                       Min.   : 0.02356
##  1st Qu.:11.97                       1st Qu.:11.76419
##  Median :28.95                       Median :21.49764
##  Mean   :36.63                       Mean   :25.52722
##  3rd Qu.:61.75                       3rd Qu.:40.19229
##  Max.   :94.30                       Max.   :64.96305
##  OOSR_PreOPrimary_Age_Male OOSR_PreOPrimary_Age_Female OOSR_Primary_Age_Male
##  Min.   : 1.00             Min.   : 1.00               Min.   : 1.000
##  1st Qu.: 8.75             1st Qu.: 7.00               1st Qu.: 3.000
##  Median :27.20             Median :26.88               Median : 8.468
##  Mean   :27.20             Mean   :26.88               Mean   : 8.468
##  3rd Qu.:31.00             3rd Qu.:30.25               3rd Qu.: 8.468
##  Max.   :96.00             Max.   :96.00               Max.   :56.000
##  OOSR_Primary_Age_Female OOSR_Lower_Secondary_Age_Male
##  Min.   : 1.00           Min.   : 1.00
##  1st Qu.: 3.00           1st Qu.: 5.75
##  Median : 9.34           Median :13.77
```

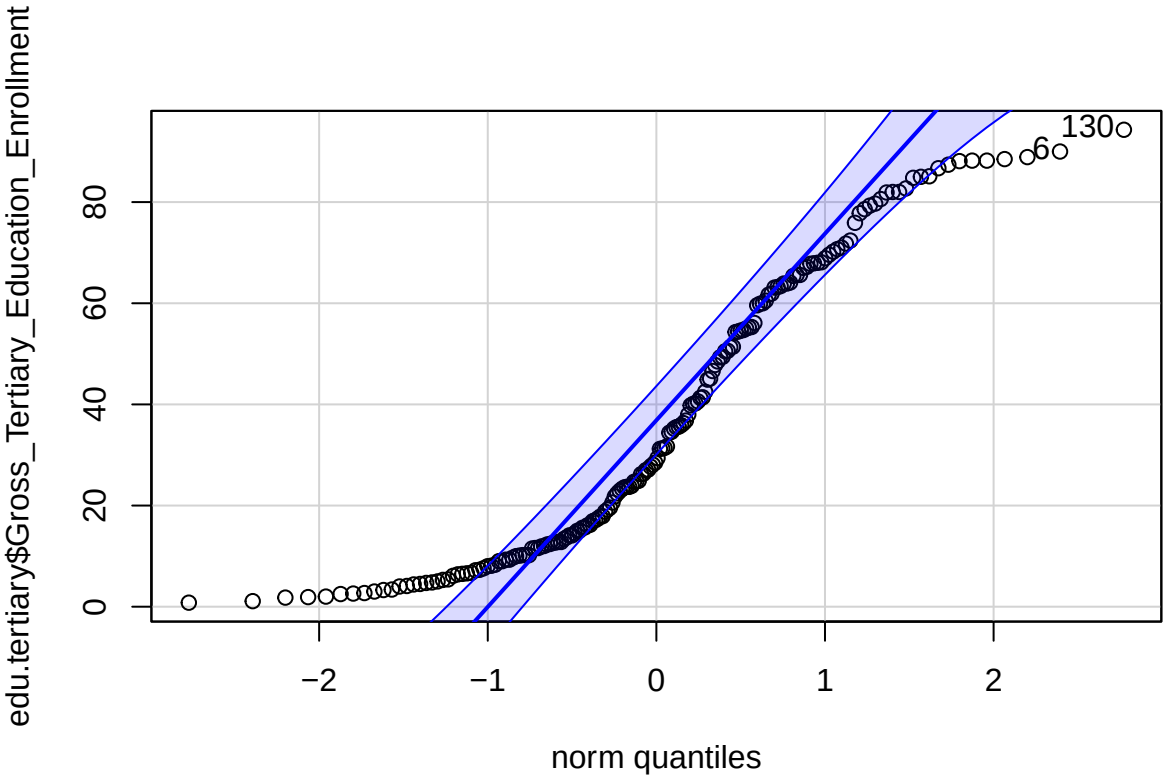
```
##  Mean   : 9.34           Mean   :13.77
##  3rd Qu.: 9.34           3rd Qu.:13.77
##  Max.   :55.00           Max.   :61.00
##  OOSR_Lower_Secondary_Age_Female OOSR_Upper_Secondary_Age_Male
##  Min.   : 1.00                   Min.   : 1.00
##  1st Qu.: 4.00                   1st Qu.:14.75
##  Median :14.04                   Median :26.83
##  Mean   :14.04                   Mean   :26.83
##  3rd Qu.:14.04                   3rd Qu.:33.00
##  Max.   :70.00                   Max.   :84.00
##  OOSR_Upper_Secondary_Age_Female Completion_Rate_Primary_Male
##  Min.   : 1.00                   Min.   : 29.00
##  1st Qu.:11.75                   1st Qu.: 78.87
##  Median :26.66                   Median : 78.87
##  Mean   :26.66                   Mean   : 78.87
##  3rd Qu.:30.25                   3rd Qu.: 91.00
##  Max.   :89.00                   Max.   :100.00
##  Completion_Rate_Primary_Female Completion_Rate_Lower_Secondary_Male
##  Min.   : 24.00                 Min.   : 10.00
##  1st Qu.: 79.60                 1st Qu.: 55.00
##  Median : 79.60                 Median : 61.78
##  Mean   : 79.60                 Mean   : 61.78
##  3rd Qu.: 92.25                 3rd Qu.: 66.50
##  Max.   :100.00                 Max.   :100.00
##  Completion_Rate_Lower_Secondary_Female Completion_Rate_Upper_Secondary_Male
##  Min.   :  4.00                         Min.   :  4.00
##  1st Qu.: 54.50                         1st Qu.: 32.75
##  Median : 62.35                         Median : 43.02
##  Mean   : 62.35                         Mean   : 43.02
##  3rd Qu.: 74.25                         3rd Qu.: 45.00
##  Max.   :100.00                         Max.   :100.00
##  Completion_Rate_Upper_Secondary_Female   Birth_Rate
##  Min.   :  1.00                         Min.   : 6.40
##  1st Qu.: 30.50                         1st Qu.:11.38
##  Median : 43.64                         Median :18.01
##  Mean   : 43.64                         Mean   :20.30
##  3rd Qu.: 43.64                         3rd Qu.:28.82
##  Max.   :100.00                         Max.   :46.08
```

```
# 高等教育总入学率直方图
z <- data.frame(values = edu.tertiary$Gross_Tertiary_Education_Enrollment)
```

```
ggplot() +
  geom_histogram(data = z, aes(x = values, y = after_stat(density)),
                 binwidth = 10, fill = "lightblue", color = "black") +
  geom_density(data = z, aes(x = values, after_stat(density)),
               color = "red") +
  geom_vline(data = z, aes(xintercept = mean(values),
                           color = "Mean"), linetype = "dashed") +
  geom_vline(data = z, aes(xintercept = mean(values) - sqrt(var(values)),
                           color = "Lower SD"), linetype = "dashed") +
  geom_vline(data = z, aes(xintercept = mean(values) + sqrt(var(values)),
                           color = "Upper SD"), linetype = "dashed") +
  labs(title = "Gross Tertiary Education Enrollment Histogram",
       x = "Enrollment", y = "Density") +
  theme_minimal() +
  scale_color_manual(values = c("black", "purple", "green"),
                     labels = c("Mean", "Lower SD", "Upper SD"),
                     guide = guide_legend())

# 高等教育总入学率 QQ 图
qqPlot(edu.tertiary$Gross_Tertiary_Education_Enrollment)
```

```
## [1] 130   6
```


Gross Tertiary Education Enrollment Histogram
Density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
Enrollment
colour
Mean
Lower SD
Upper SD


edu.tertiary$Gross_Tertiary_Education_Enrollment
0
20
40
60
80
−2
−1
0
1
2
norm quantiles
130
6

3.3.2 主要分析：建立预测模型（所有因素下 Akaike 的 backward）

```
enroll_model_all <- lm(Gross_Tertiary_Education_Enrollment ~., data = edu.tertiary)
backward <- step(enroll_model_all, direction = 'backward', scope = formula(enroll_model_all), trac
backward$anova
```

```
##                                      Step Df     Deviance Resid. Df Resid. Dev
## 1                                        NA           NA       163   42221.34
## 2          - OOSR_PreOPrimary_Age_Female  1   0.04154230       164   42221.38
## 3      - OOSR_Lower_Secondary_Age_Female  1   0.07928437       165   42221.46
## 4      - OOSR_Upper_Secondary_Age_Female  1   4.40591309       166   42225.86
## 5         - Completion_Rate_Primary_Male  1  12.59180953       167   42238.45
## 6 - Completion_Rate_Lower_Secondary_Male  1   5.64475628       168   42244.10
## 7 - Completion_Rate_Upper_Secondary_Male  1 110.45441082       169   42354.55
## 8                - OOSR_Primary_Age_Male  1 295.89761653       170   42650.45
## 9              - OOSR_Primary_Age_Female  1 238.47945357       171   42888.93
##        AIC
## 1 1016.390
## 2 1014.391
## 3 1012.391
## 4 1010.410
## 5 1008.463
## 6 1006.487
## 7 1004.957
## 8 1004.211
## 9 1003.214
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = Gross_Tertiary_Education_Enrollment ~ Latitude +
##     OOSR_PreOPrimary_Age_Male + OOSR_Lower_Secondary_Age_Male +
##     OOSR_Upper_Secondary_Age_Male + Completion_Rate_Primary_Female +
##     Completion_Rate_Lower_Secondary_Female + Completion_Rate_Upper_Secondary_Female +
##     Birth_Rate, data = edu.tertiary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.545 -10.025  -1.195   8.944  48.574
##
## Coefficients:
```

```
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          52.21651   11.87536   4.397 1.92e-05 ***
## Latitude                              0.45673    0.09086   5.027 1.25e-06 ***
## OOSR_PreOPrimary_Age_Male            -0.18055    0.06956  -2.596   0.0103 *
## OOSR_Lower_Secondary_Age_Male         0.49272    0.19033   2.589   0.0105 *
## OOSR_Upper_Secondary_Age_Male        -0.59084    0.12742  -4.637 7.00e-06 ***
## Completion_Rate_Primary_Female        0.26456    0.19103   1.385   0.1679
## Completion_Rate_Lower_Secondary_Female -0.47819    0.21509  -2.223   0.0275 *
## Completion_Rate_Upper_Secondary_Female  0.32715    0.14461   2.262   0.0249 *
## Birth_Rate                           -0.92552    0.21768  -4.252 3.48e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.84 on 171 degrees of freedom
## Multiple R-squared:  0.6839, Adjusted R-squared:  0.6691
## F-statistic: 46.25 on 8 and 171 DF,  p-value: < 2.2e-16
```

3.3.3 主要分析：建立预测模型（所有因素下 Akaike 的 forward）

```
enroll_model_1 <- lm(Gross_Tertiary_Education_Enrollment ~ 1, data = edu.tertiary)
forward <- step(enroll_model_1, direction = 'forward', scope = formula(enroll_model_all), trace = 
forward$anova
```

```
##                               Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                                   NA        NA       179  135693.52 1194.536
## 2                     + Birth_Rate -1 75136.003       178   60557.52 1051.311
## 3                       + Latitude -1  8741.228       177   51816.29 1025.251
## 4 + OOSR_Upper_Secondary_Age_Male -1  4815.050       176   47001.24 1009.695
## 5     + OOSR_PreOPrimary_Age_Male -1  1024.668       175   45976.57 1007.727
## 6 + OOSR_Lower_Secondary_Age_Male -1  1632.688       174   44343.88 1003.219
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = Gross_Tertiary_Education_Enrollment ~ Birth_Rate +
##     Latitude + OOSR_Upper_Secondary_Age_Male + OOSR_PreOPrimary_Age_Male +
##     OOSR_Lower_Secondary_Age_Male, data = edu.tertiary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.065 -10.204  -0.546   9.156  47.597
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    58.59069    5.01174  11.691  < 2e-16 ***
## Birth_Rate                     -0.99080    0.19350  -5.120 8.02e-07 ***
## Latitude                        0.44735    0.08922   5.014 1.31e-06 ***
## OOSR_Upper_Secondary_Age_Male -0.54811    0.12406  -4.418 1.75e-05 ***
## OOSR_Pre0Primary_Age_Male     -0.17907    0.06938  -2.581   0.0107 *
## OOSR_Lower_Secondary_Age_Male  0.45795    0.18093   2.531   0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.96 on 174 degrees of freedom
## Multiple R-squared:  0.6732, Adjusted R-squared:  0.6638
## F-statistic: 71.69 on 5 and 174 DF,  p-value: < 2.2e-16
```

3.3.4 主要分析：建立预测模型（指定显著因素下 Akaike 的 backward）

```
enroll_model2 <- lm(Gross_Tertiary_Education_Enrollment ~
  Latitude +
  OOSR_Pre0Primary_Age_Male +
  OOSR_Pre0Primary_Age_Female +
  OOSR_Primary_Age_Male +
  # OOSR_Primary_Age_Female +
  # OOSR_Lower_Secondary_Age_Male +
  # OOSR_Lower_Secondary_Age_Female +
  OOSR_Upper_Secondary_Age_Male +
  OOSR_Upper_Secondary_Age_Female +
  Completion_Rate_Primary_Male +
  Completion_Rate_Primary_Female +
  # Completion_Rate_Lower_Secondary_Male +
  # Completion_Rate_Lower_Secondary_Female +
  Completion_Rate_Upper_Secondary_Male +
  Completion_Rate_Upper_Secondary_Female +
  Birth_Rate
, data = edu.tertiary)
summary(enroll_model2)
```

```
##
## Call:
## lm(formula = Gross_Tertiary_Education_Enrollment ~ Latitude +
```

```
##     OOSR_PreOPrimary_Age_Male + OOSR_PreOPrimary_Age_Female +
##     OOSR_Primary_Age_Male + OOSR_Upper_Secondary_Age_Male + OOSR_Upper_Secondary_Age_Female +
##     Completion_Rate_Primary_Male + Completion_Rate_Primary_Female +
##     Completion_Rate_Upper_Secondary_Male + Completion_Rate_Upper_Secondary_Female +
##     Birth_Rate, data = edu.tertiary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.109 -10.408   0.606  10.196  48.209
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          64.21363   12.45676   5.155 7.06e-07 ***
## Latitude                              0.43195    0.09705   4.451 1.55e-05 ***
## OOSR_PreOPrimary_Age_Male            -0.16589    0.19634  -0.845   0.3994
## OOSR_PreOPrimary_Age_Female           0.02797    0.19723   0.142   0.8874
## OOSR_Primary_Age_Male                 0.08998    0.18909   0.476   0.6348
## OOSR_Upper_Secondary_Age_Male        -0.50444    0.29338  -1.719   0.0874 .
## OOSR_Upper_Secondary_Age_Female       0.14861    0.26625   0.558   0.5775
## Completion_Rate_Primary_Male         -0.07416    0.32806  -0.226   0.8214
## Completion_Rate_Primary_Female       -0.05291    0.29564  -0.179   0.8582
## Completion_Rate_Upper_Secondary_Male  0.02803    0.26140   0.107   0.9147
## Completion_Rate_Upper_Secondary_Female 0.08251   0.25178   0.328   0.7435
## Birth_Rate                           -1.02409    0.23112  -4.431 1.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.44 on 168 degrees of freedom
## Multiple R-squared:  0.6652, Adjusted R-squared:  0.6433
## F-statistic: 30.35 on 11 and 168 DF,  p-value: < 2.2e-16
```

```
backward <- step(enroll_model2, direction = 'backward', scope = formula(enroll_model2), trace = 0)
backward$anova
```

```
##                                      Step Df   Deviance Resid. Df Resid. Dev
## 1                                         NA         NA       168   45428.97
## 2   - Completion_Rate_Upper_Secondary_Male  1   3.109029       169   45432.08
## 3            - OOSR_PreOPrimary_Age_Female  1   5.071935       170   45437.15
## 4          - Completion_Rate_Primary_Male  1  11.150232       171   45448.30
## 5                - OOSR_Primary_Age_Male  1  59.968034       172   45508.27
## 6       - OOSR_Upper_Secondary_Age_Female  1  71.172977       173   45579.45
## 7        - Completion_Rate_Primary_Female  1 322.357547       174   45901.80
```

```
## 8 - Completion_Rate_Upper_Secondary_Female  1  74.768917       175   45976.57
##        AIC
## 1 1019.571
## 2 1017.583
## 3 1015.603
## 4 1013.647
## 5 1011.885
## 6 1010.166
## 7 1009.435
## 8 1007.727
```

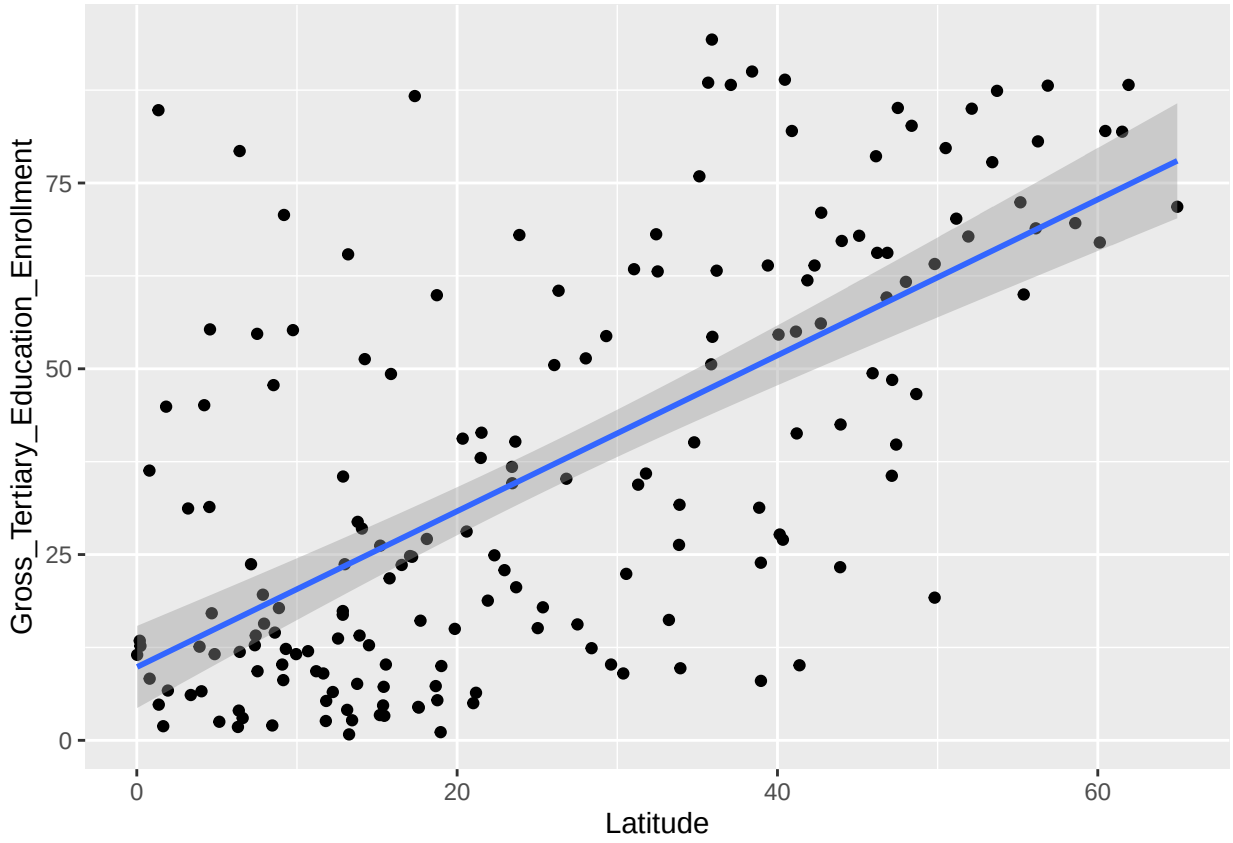
```
summary(backward)
```

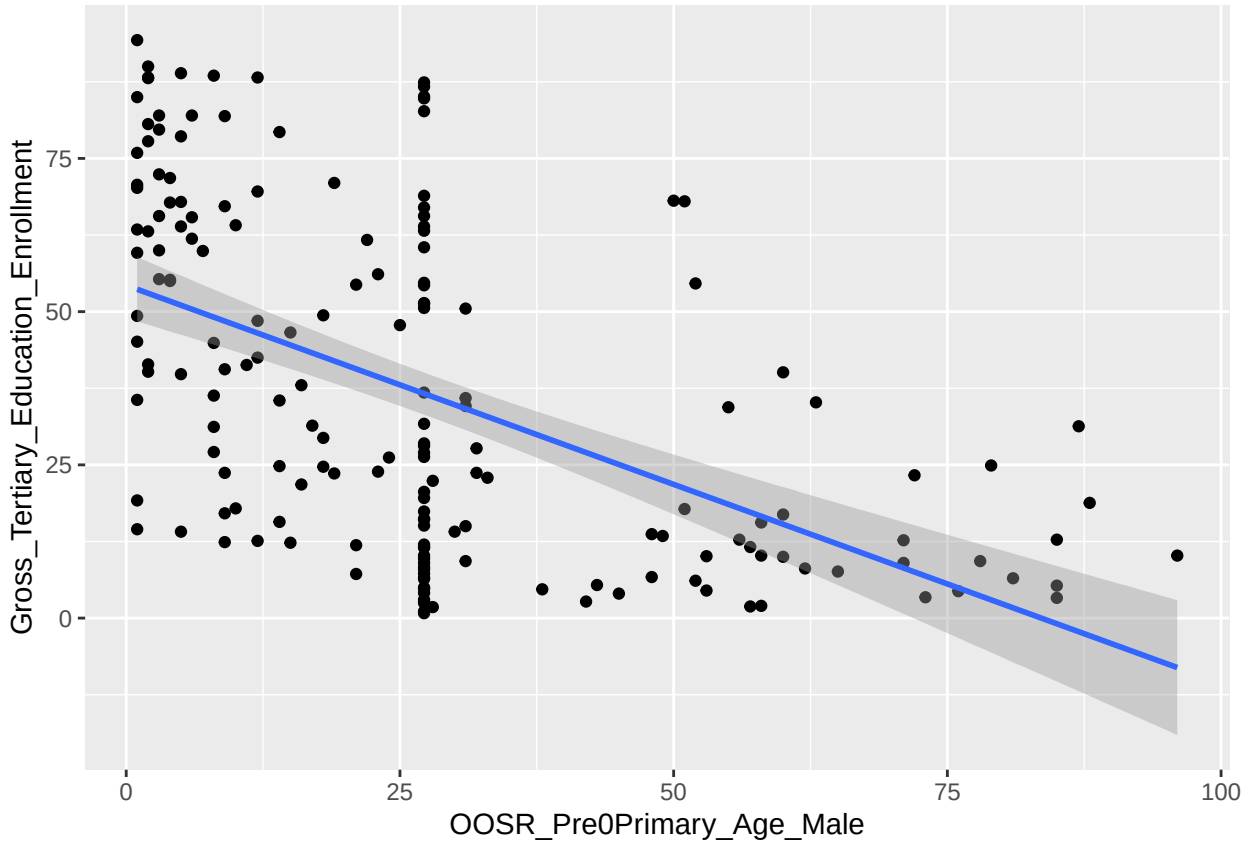
```
##
## Call:
## lm(formula = Gross_Tertiary_Education_Enrollment ~ Latitude +
##     OOSR_PreOPrimary_Age_Male + OOSR_Upper_Secondary_Age_Male +
##     Birth_Rate, data = edu.tertiary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.439 -10.675  -0.554  10.473  48.280
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   57.33641    5.06363  11.323  < 2e-16 ***
## Latitude                       0.46126    0.09042   5.102 8.71e-07 ***
## OOSR_PreOPrimary_Age_Male     -0.13457    0.06814  -1.975 0.049854 *
## OOSR_Upper_Secondary_Age_Male -0.34756    0.09692  -3.586 0.000435 ***
## Birth_Rate                    -0.96061    0.19609  -4.899 2.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.21 on 175 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6534
## F-statistic: 85.37 on 4 and 175 DF,  p-value: < 2.2e-16
```

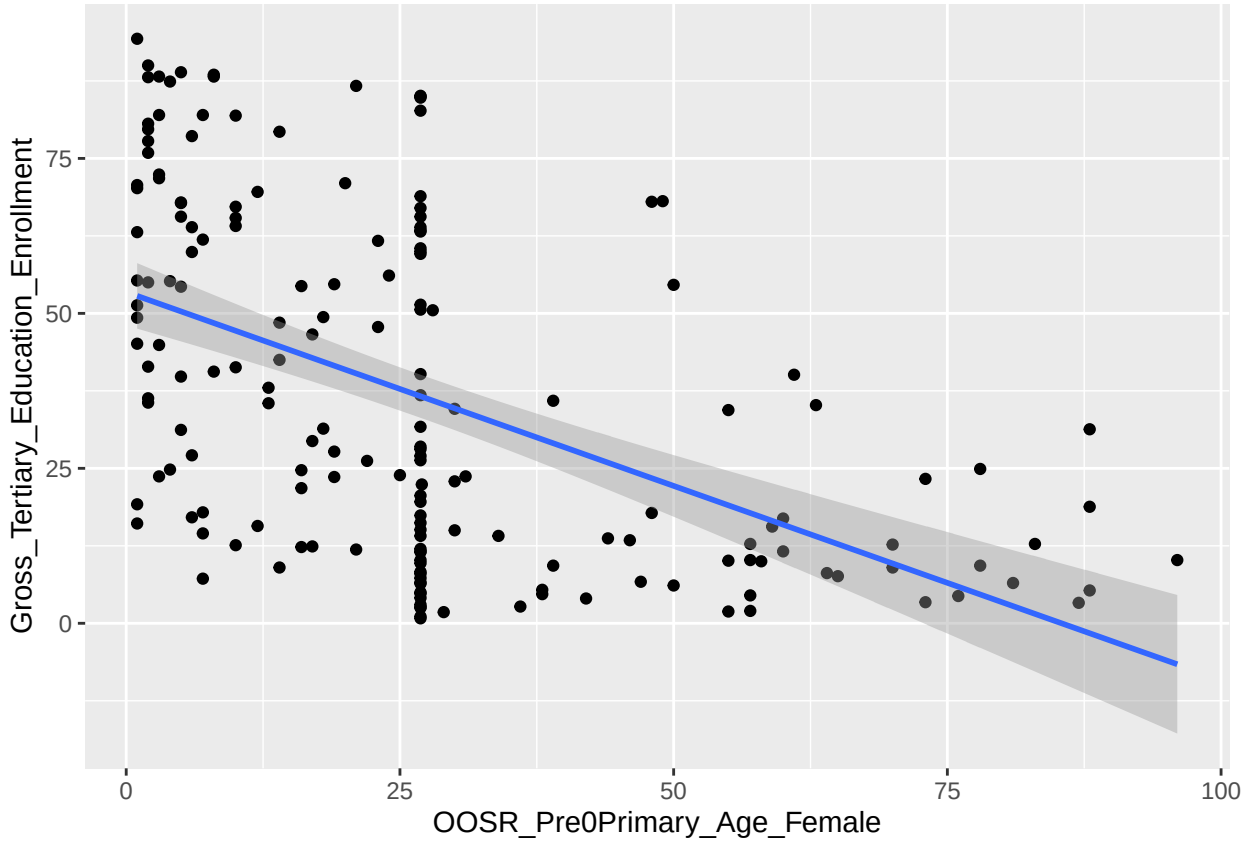
```
enroll_model2 <- backward
```

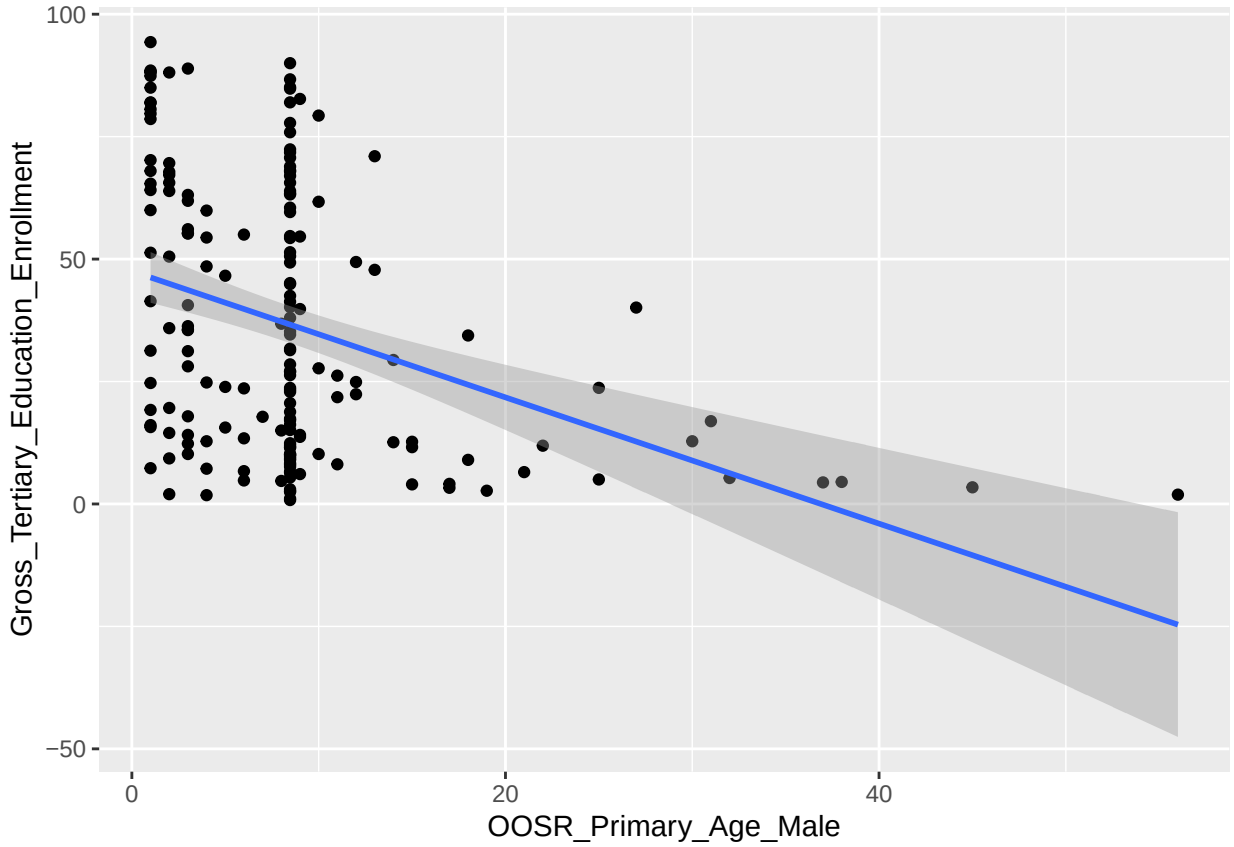
3.3.5 单独影响显著因素散点图

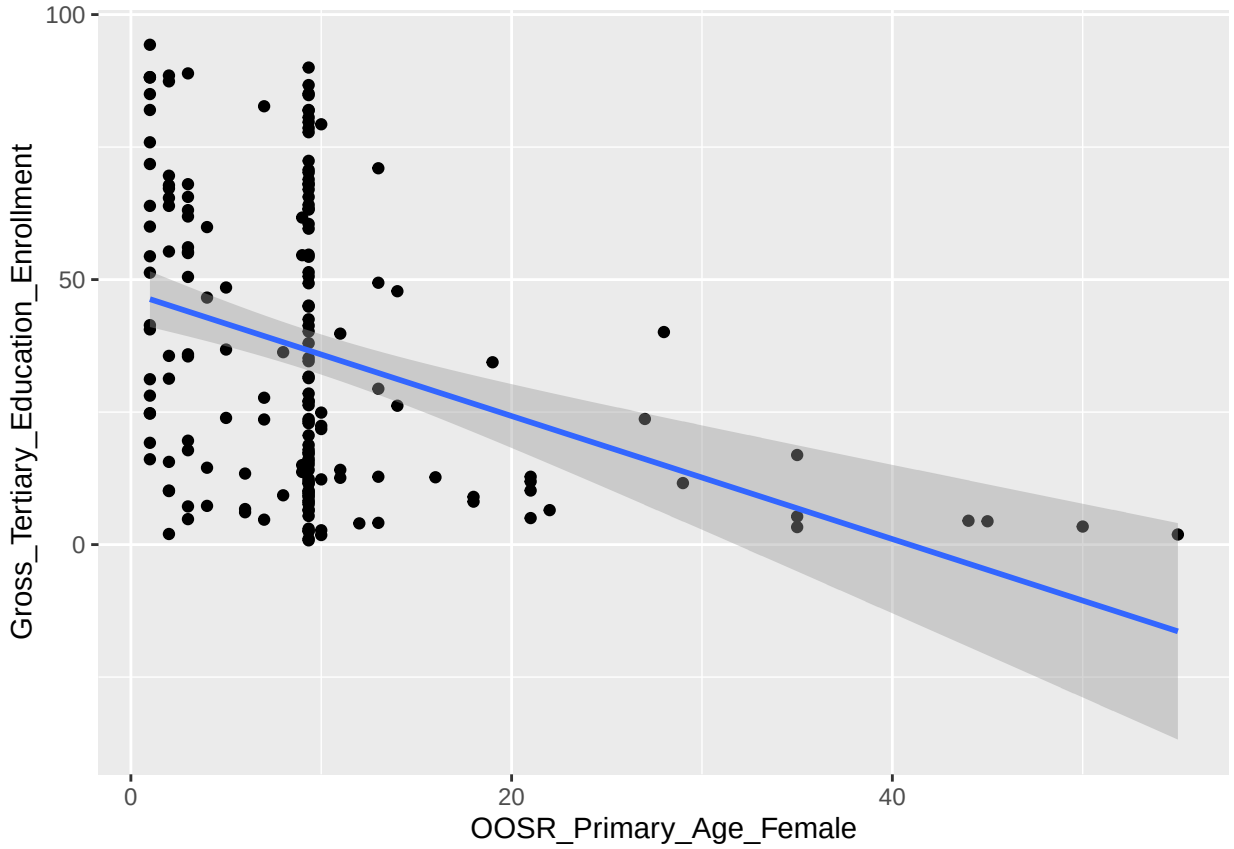
```
visualize <- function(column) {
  plot <- ggplot(data = edu.tertiary, aes(x = !!as.name(column), y = Gross_Tertiary_Education_Enro
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm", se = T)

  print(plot)
}

for (x in tertiary_remark[-1]) {
  visualize(x)
}
```

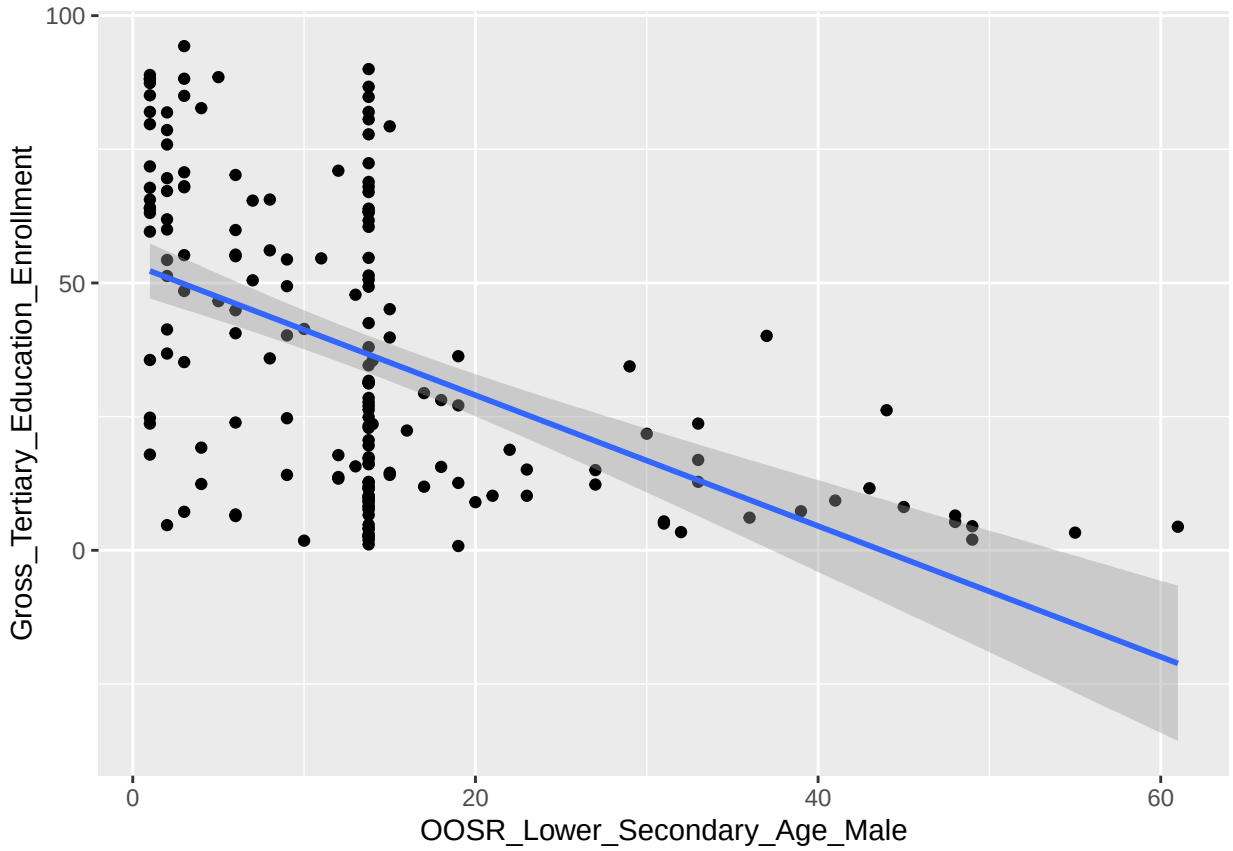


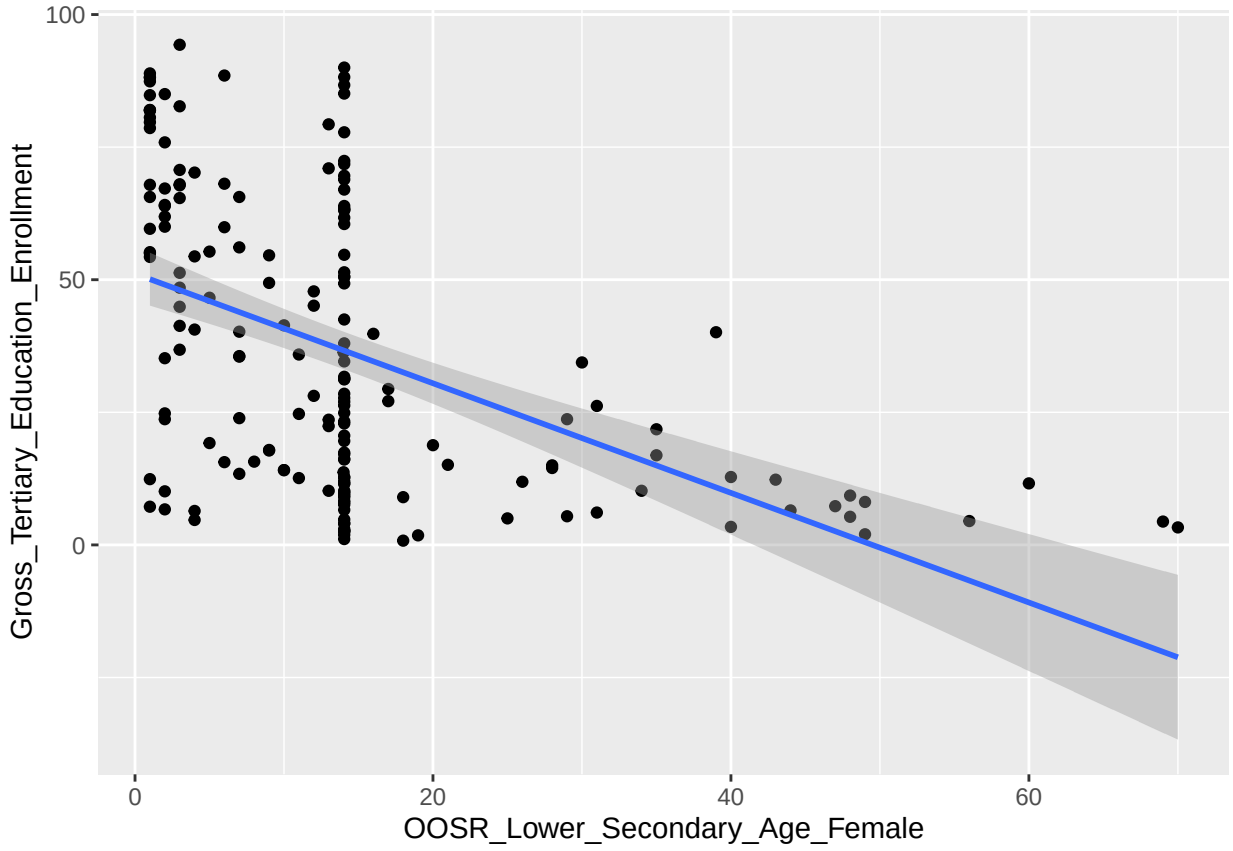

Gross_Tertiary_Education_Enrollment
75
50
25
0
0
25
50
75
100
OOSR_Pre0Primary_Age_Male

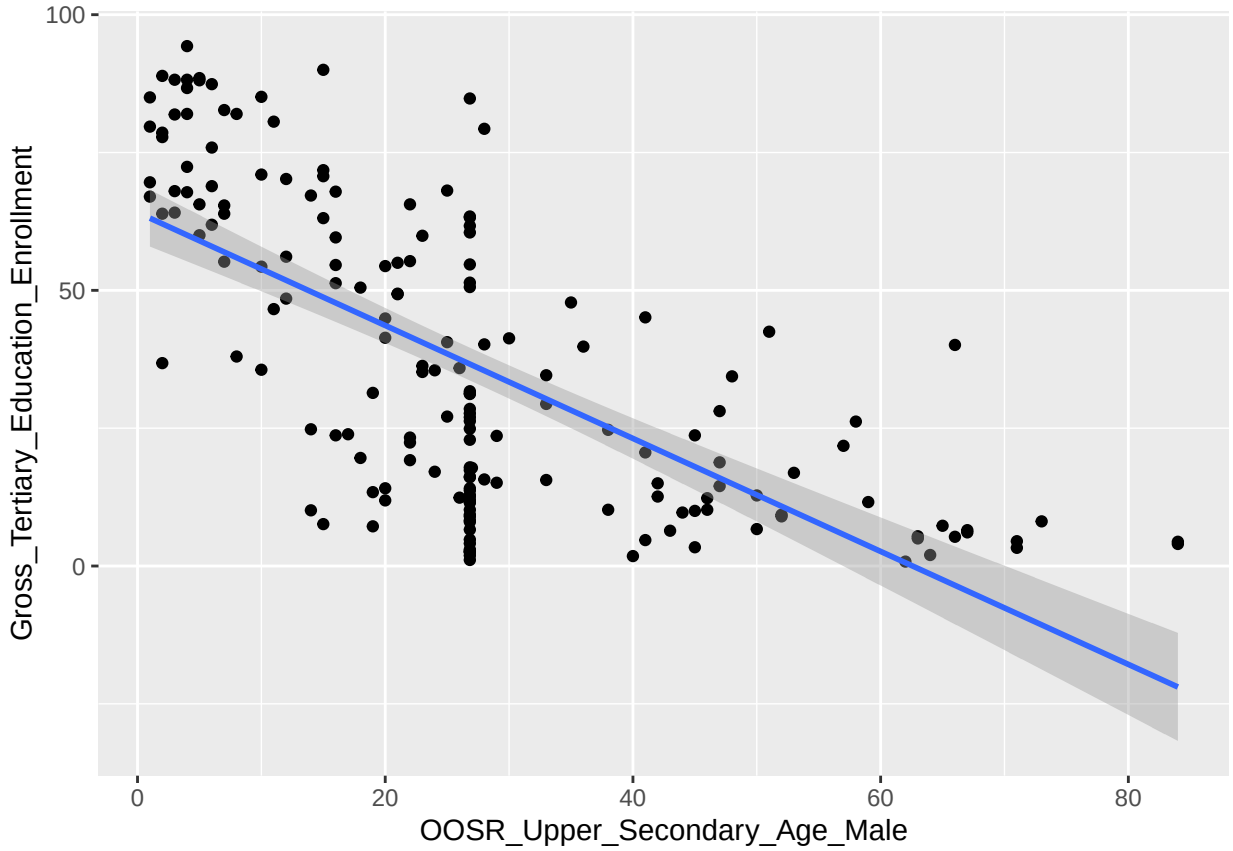

Gross_Tertiary_Education_Enrollment
75
50
25
0
0
25
50
75
100
OOSR_Pre0Primary_Age_Female

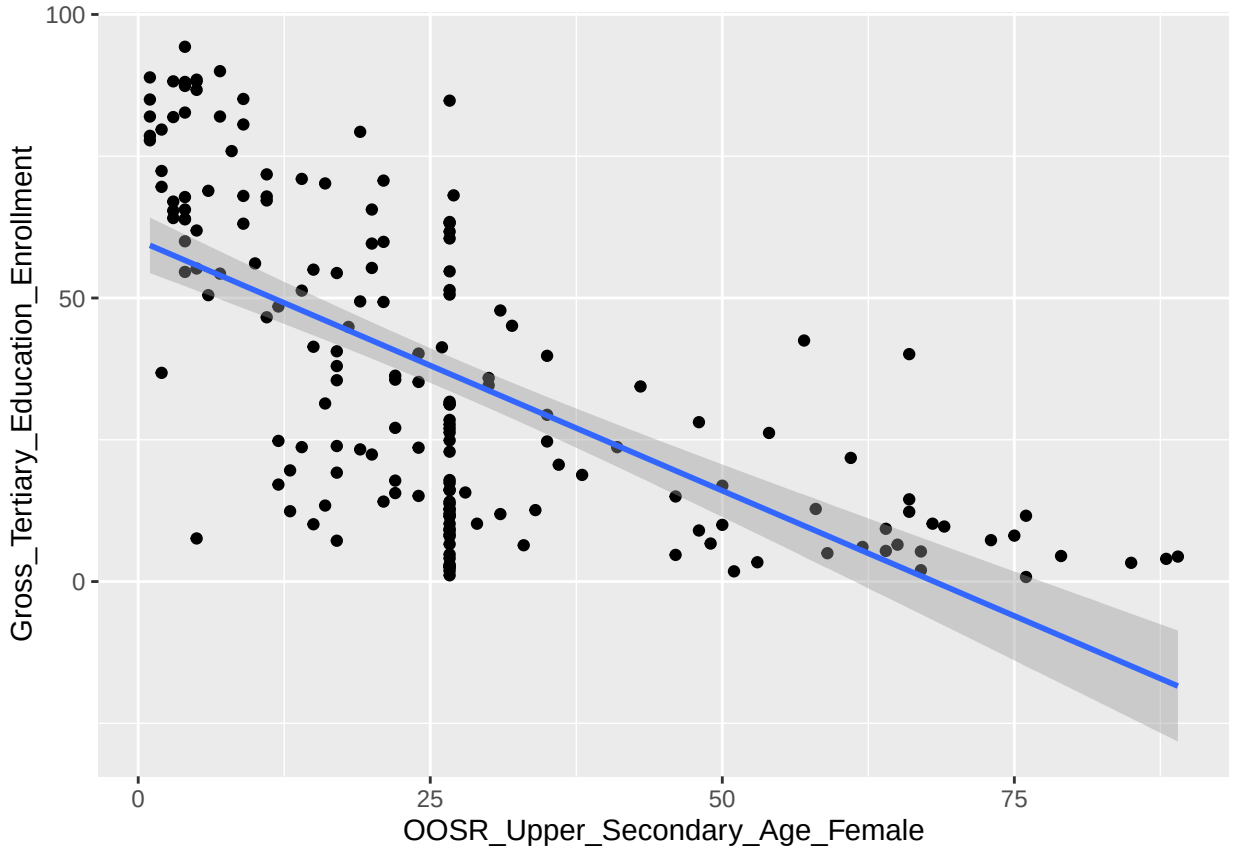

100
50
0
−50
Gross_Tertiary_Education_Enrollment
0
20
40
OOSR_Primary_Age_Male

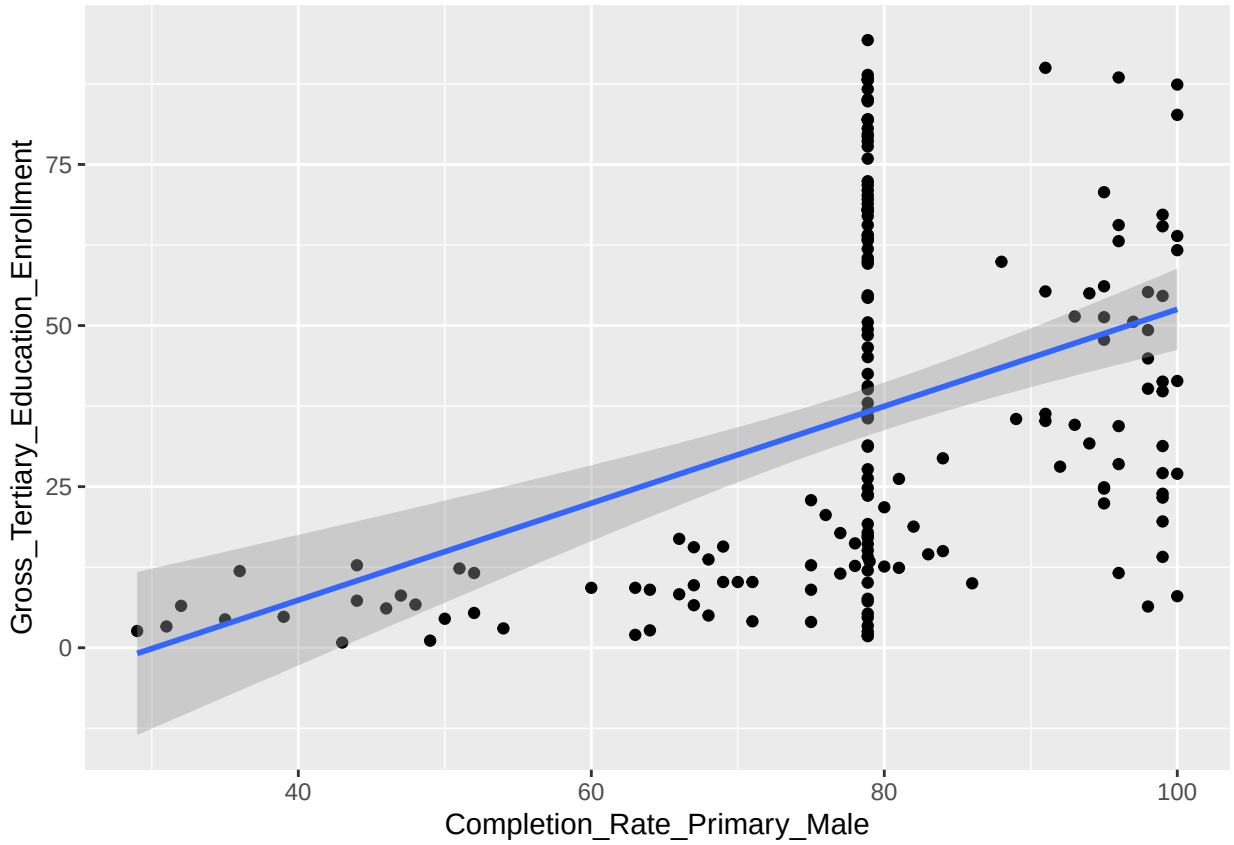

100
50
0
Gross_Tertiary_Education_Enrollment
0
20
40
OOSR_Primary_Age_Female

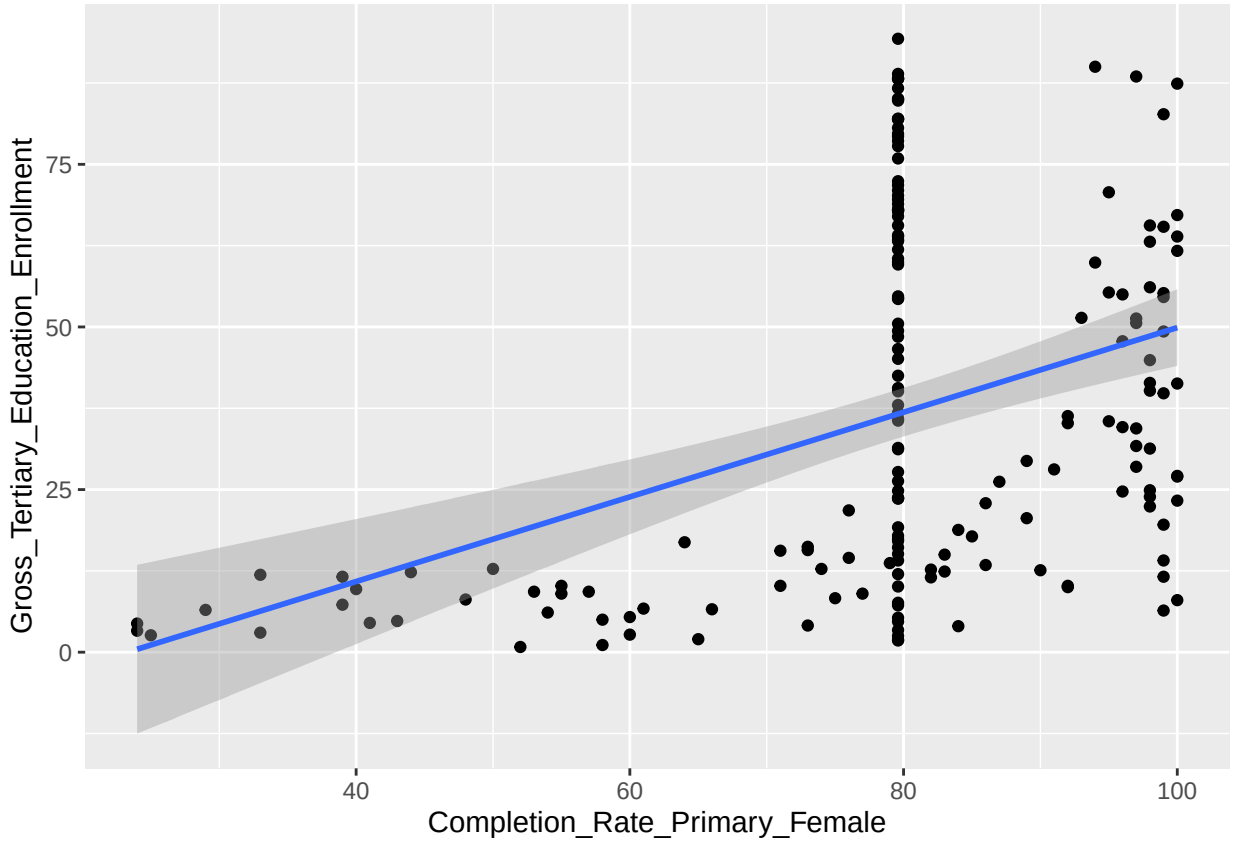

100
50
0
Gross_Tertiary_Education_Enrollment
0
20
40
60
OOSR_Lower_Secondary_Age_Male

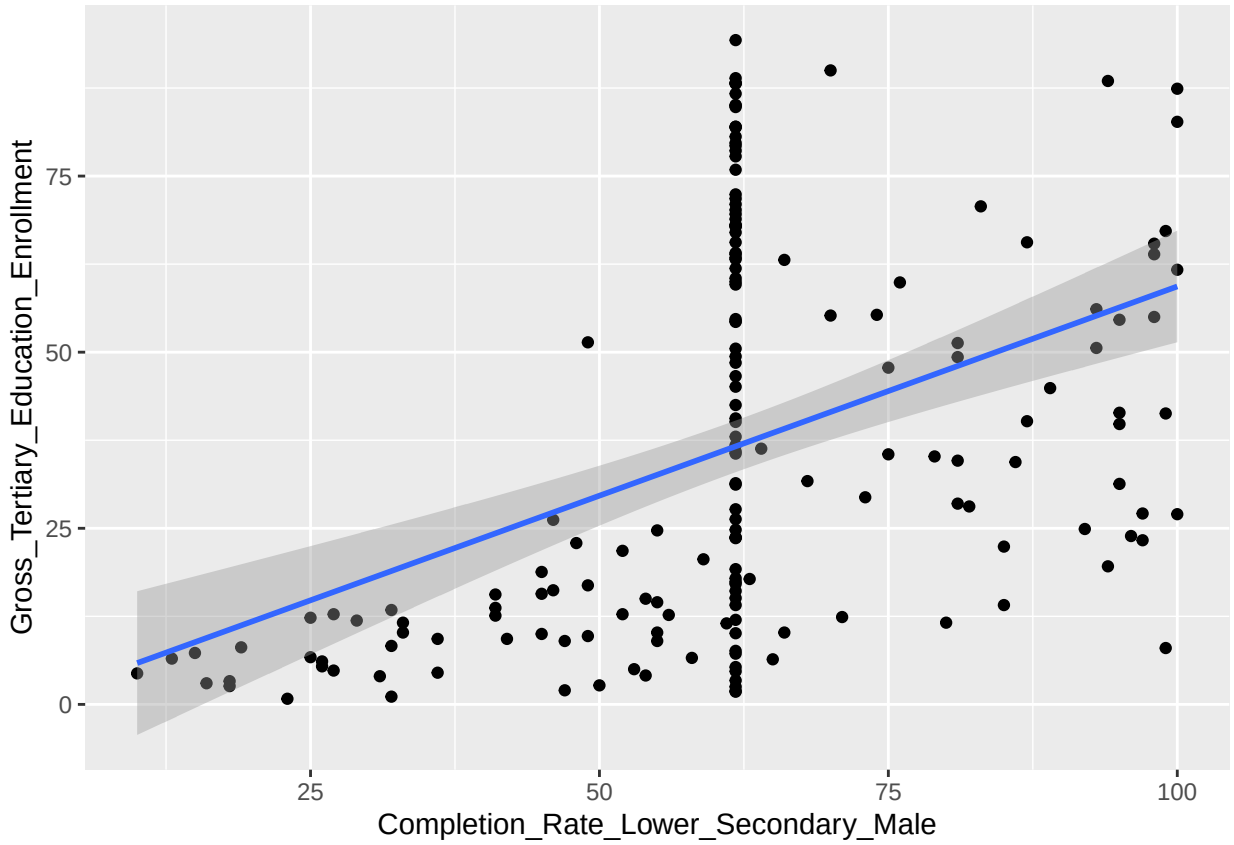

100
50
0
Gross_Tertiary_Education_Enrollment
0
20
40
60
OOSR_Lower_Secondary_Age_Female

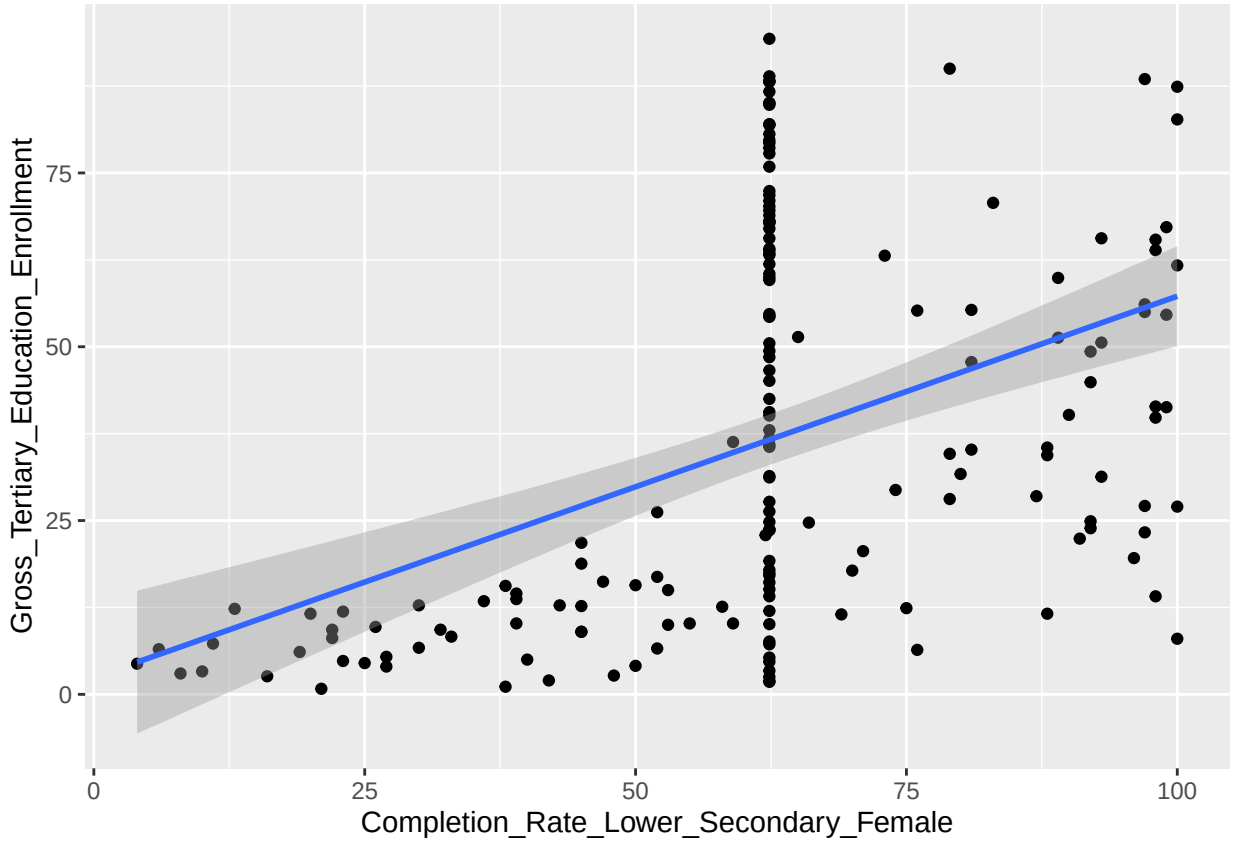

100
50
0
Gross_Tertiary_Education_Enrollment
0
20
40
60
80
OOSR_Upper_Secondary_Age_Male

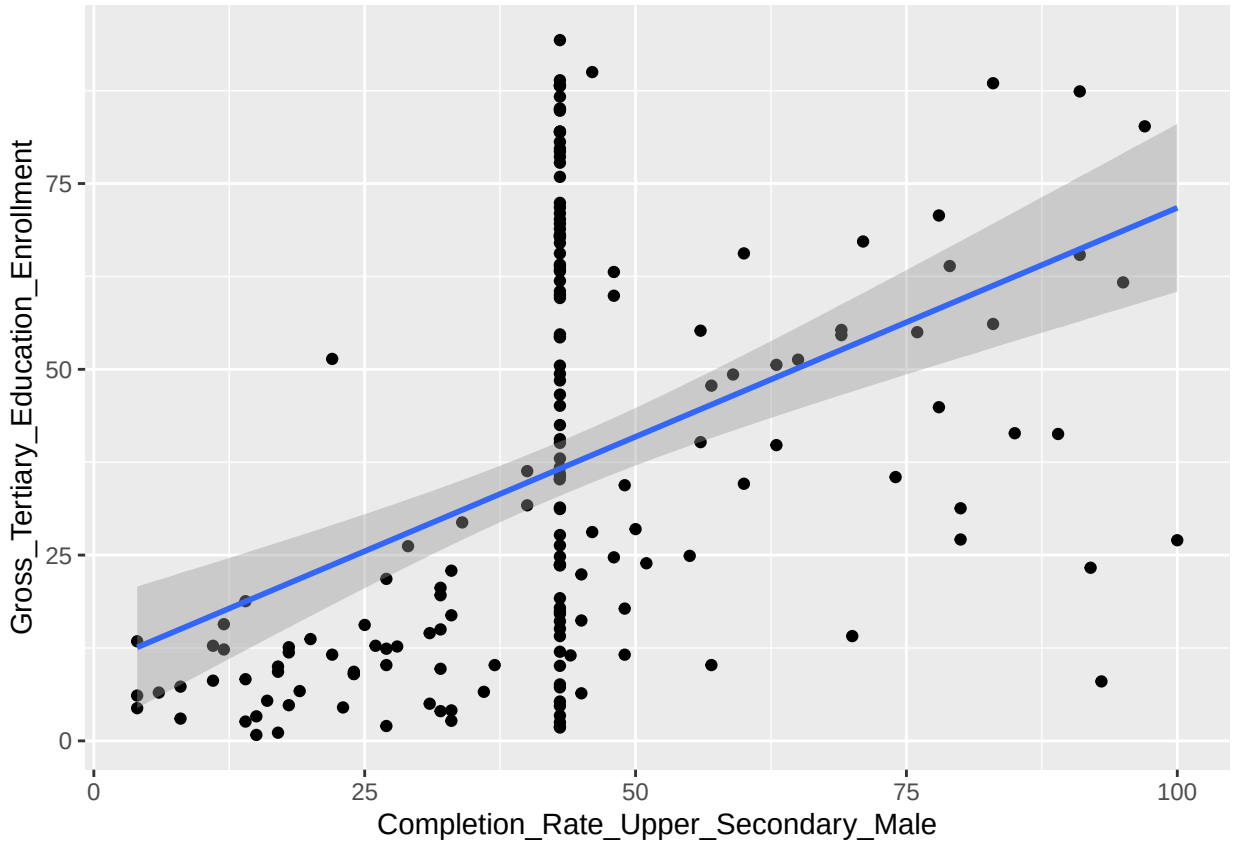

100
50
0
Gross_Tertiary_Education_Enrollment
0
25
50
75
OOSR_Upper_Secondary_Age_Female

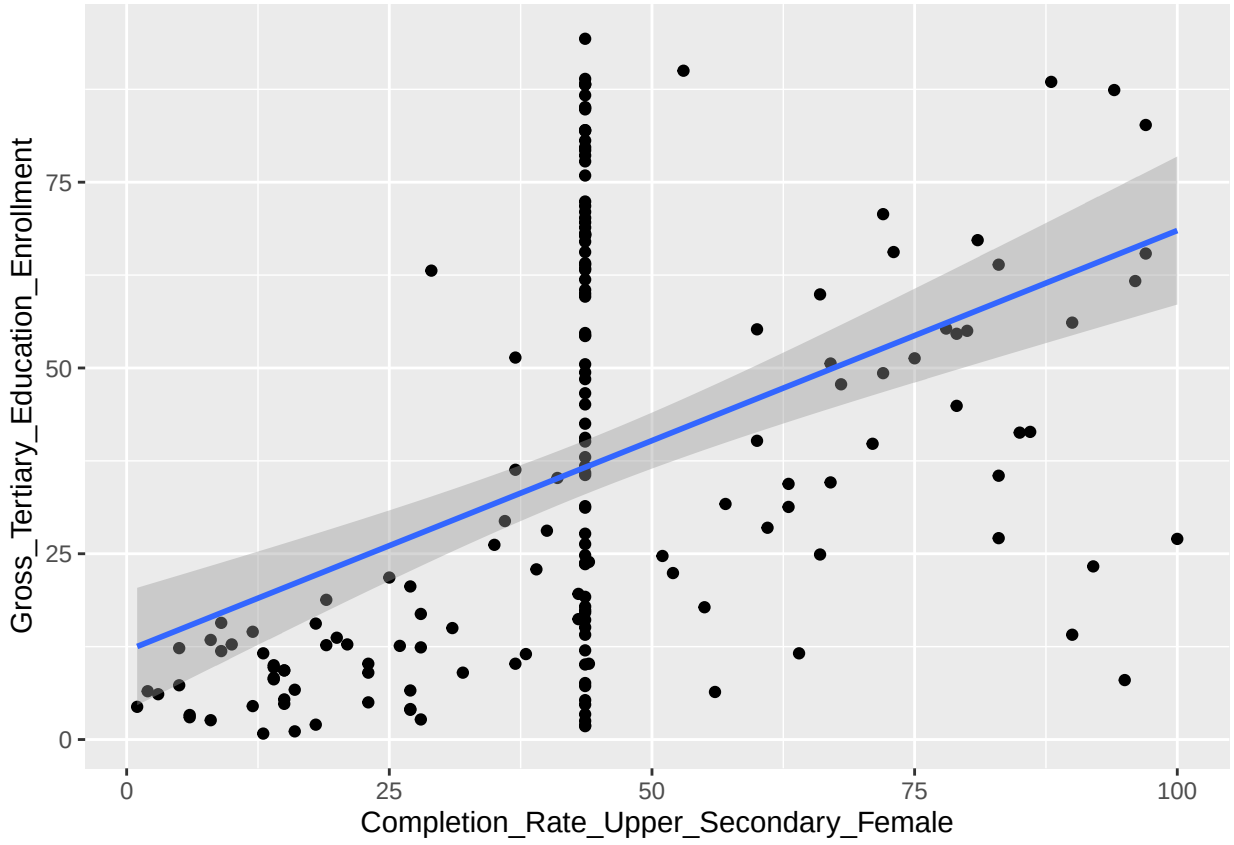

Gross_Tertiary_Education_Enrollment
75
50
25
0
40
60
80
100
Completion_Rate_Primary_Male

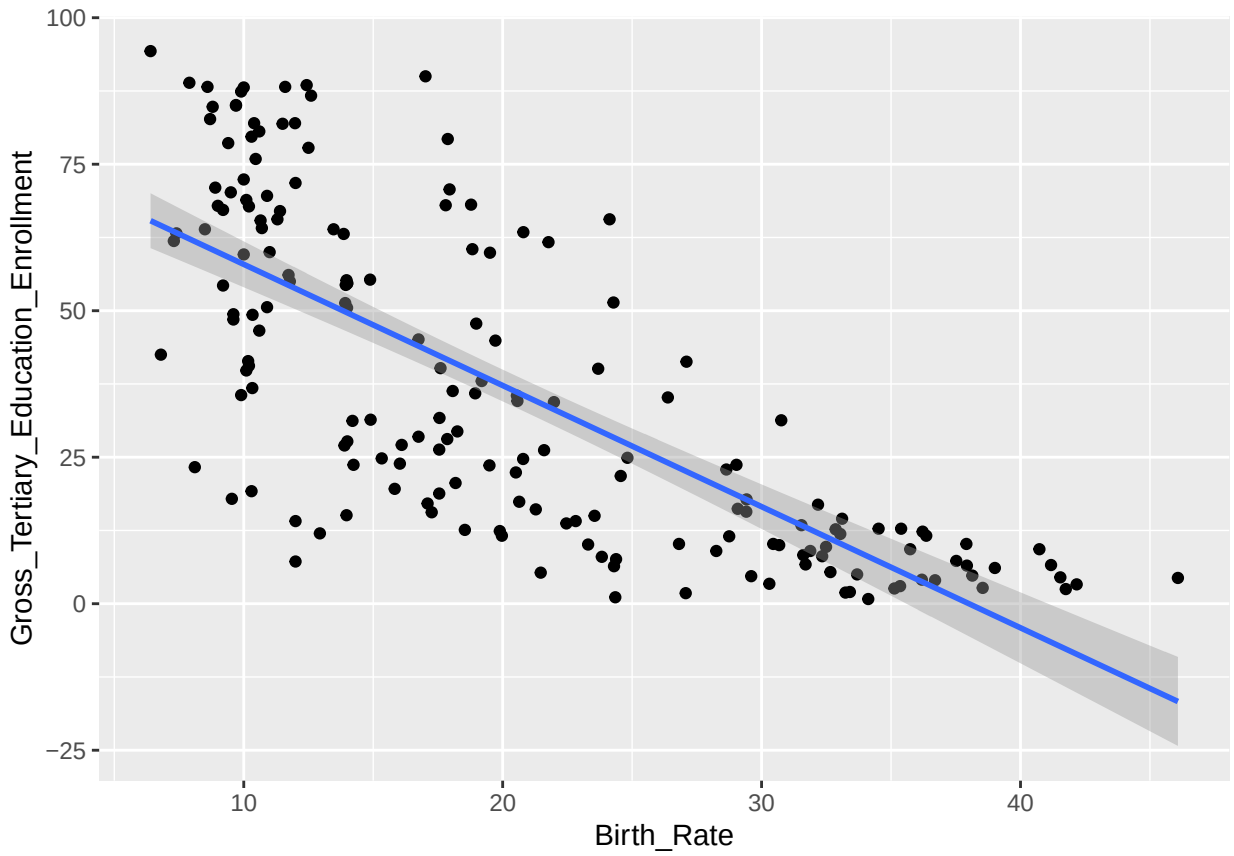

Gross_Tertiary_Education_Enrollment
75
50
25
0
40
60
80
100
Completion_Rate_Primary_Female


Gross_Tertiary_Education_Enrollment
75
50
25
0
25
50
75
100
Completion_Rate_Lower_Secondary_Male


Gross_Tertiary_Education_Enrollment
75
50
25
0
0
25
50
75
100
Completion_Rate_Lower_Secondary_Female


Gross_Tertiary_Education_Enrollment
0
25
50
75
0
25
50
75
100
Completion_Rate_Upper_Secondary_Male


Gross_Tertiary_Education_Enrollment
0
25
50
75
0
25
50
75
100
Completion_Rate_Upper_Secondary_Female



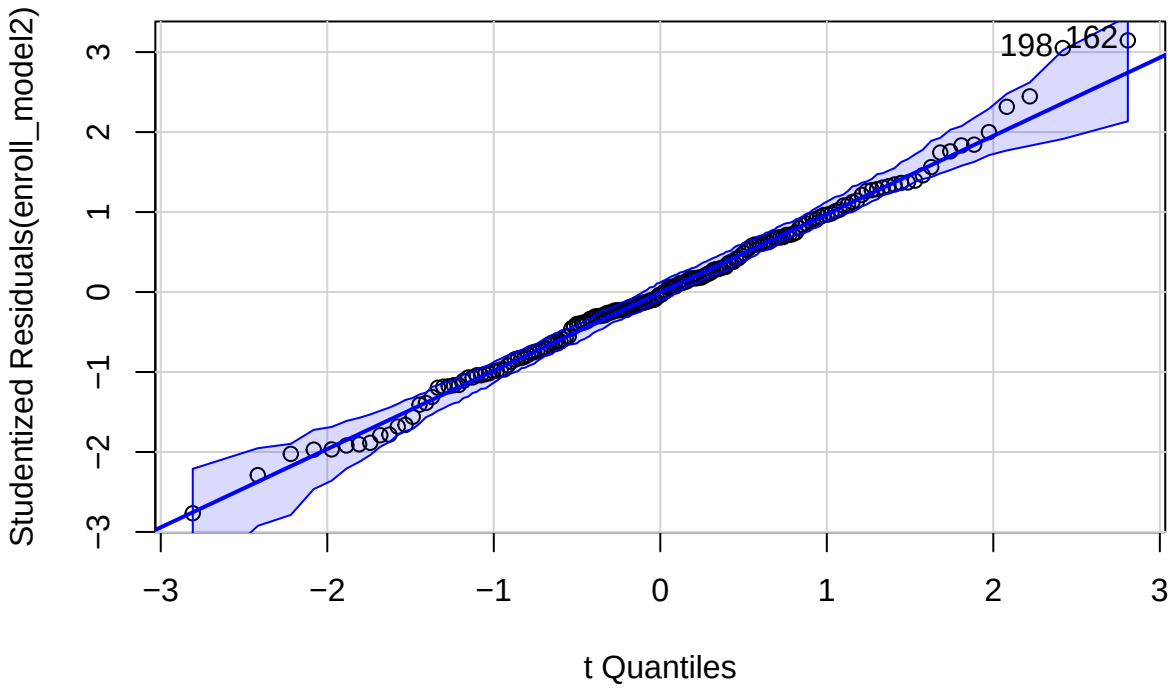
3.3.6 模型评估

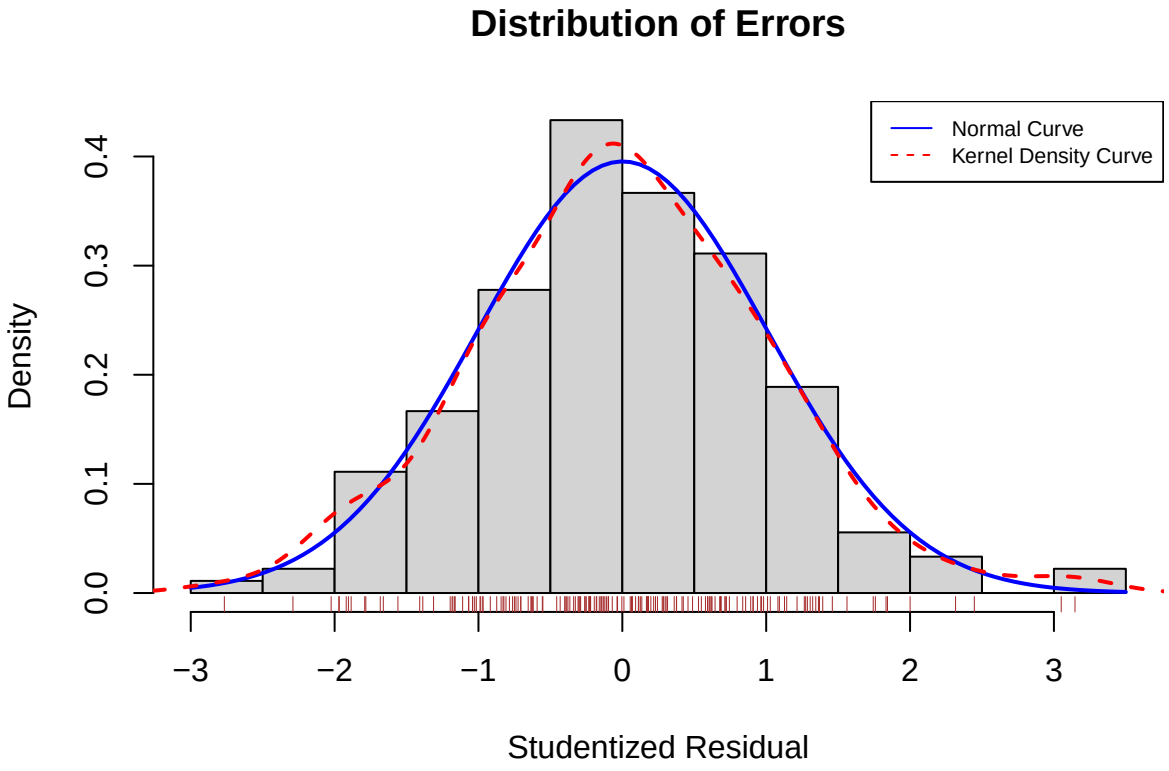
```
# 模型的 QQ 图
qqPlot(enroll_model2, simulate = T)
```

```
## 162 198
## 146 176
```

```
# 绘制预测残差分布直方图
residplot <- function(fit, nbreaks=10) {
 z <- rstudent(fit)
 hist(z, breaks=nbreaks, freq=FALSE,
 xlab="Studentized Residual",
 main="Distribution of Errors")
 rug(jitter(z), col="brown")
 curve(dnorm(x, mean=mean(z), sd=sd(z)),
 add=TRUE, col="blue", lwd=2)
 lines(density(z)$x, density(z)$y,
 col="red", lwd=2, lty=2)
 legend("topright",
```

```
 legend = c( "Normal Curve", "Kernel Density Curve"),
 lty=1:2, col=c("blue","red"), cex=.7)
}

residplot(enroll_model2)
```





3.4 失业率分析

3.4.1 基础分析

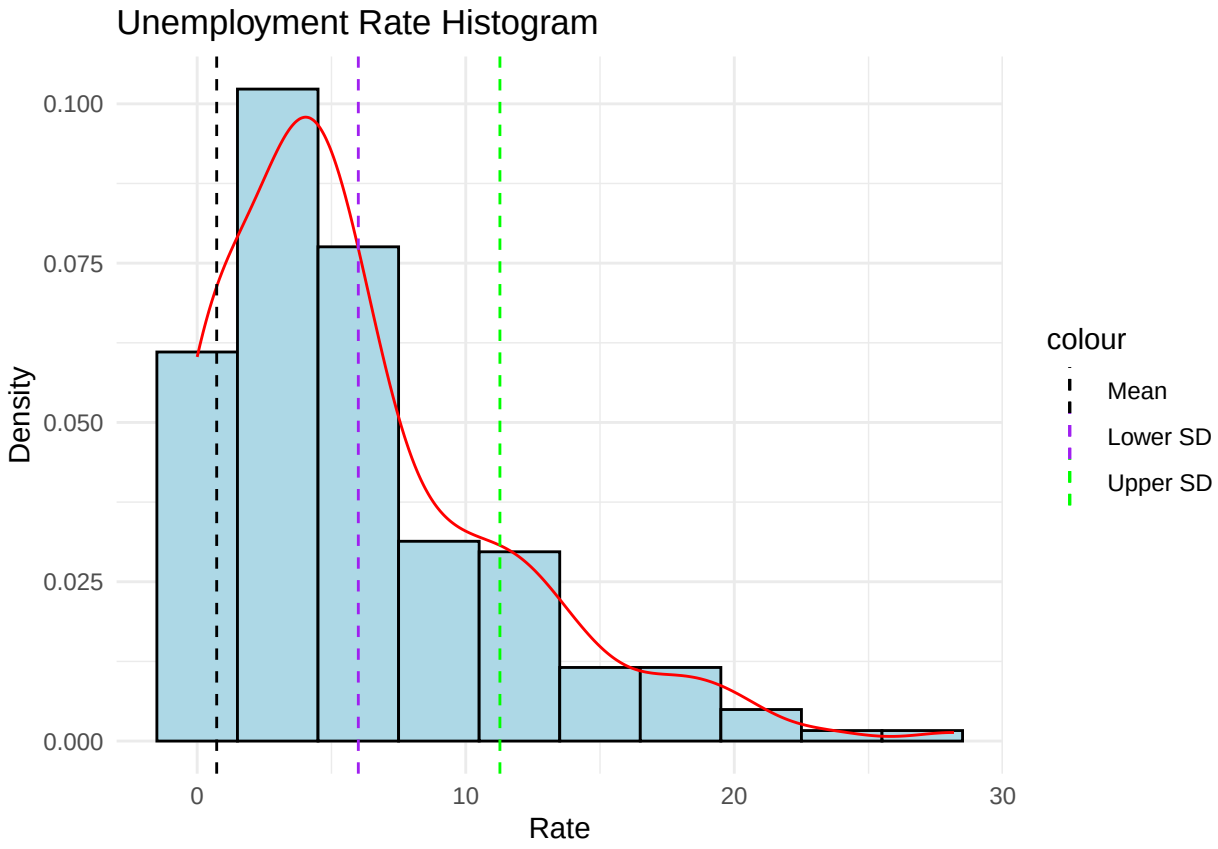
```
# 统计概要
summary(edu.unemploy)
```

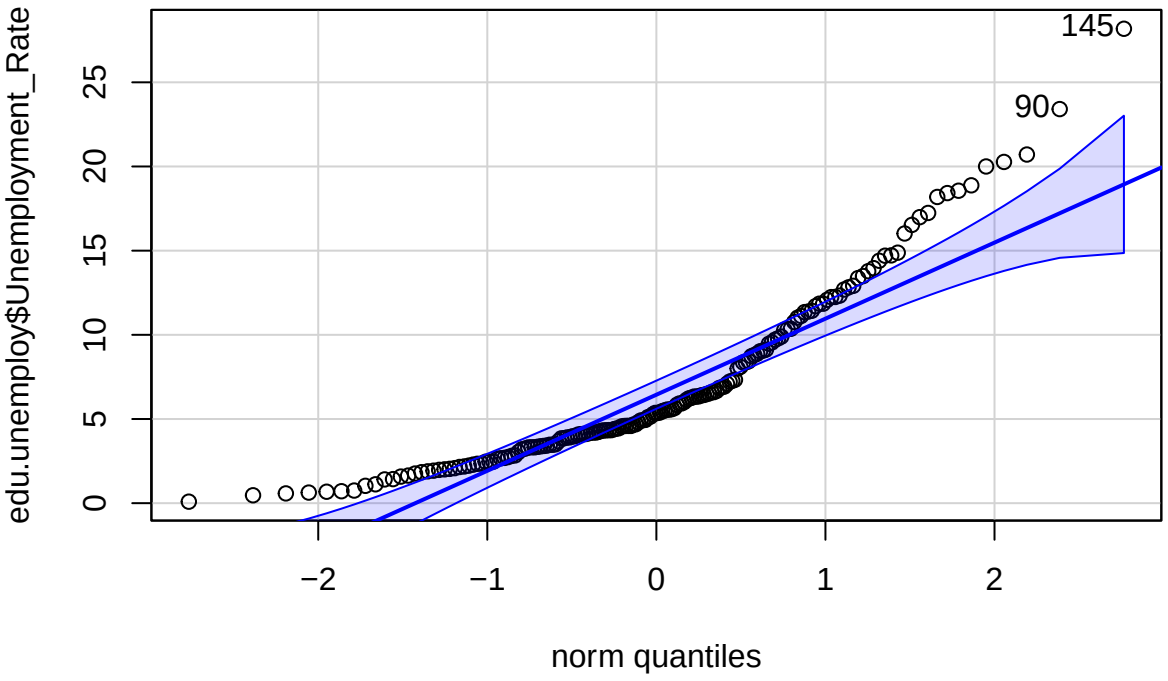
```
##  Unemployment_Rate   Longitude        OOSR_Upper_Secondary_Age_Female
##  Min.   : 0.090    Min.   :  0.8248   Min.   : 1.00
##  1st Qu.: 3.395    1st Qu.: 18.3363   1st Qu.:11.00
##  Median : 5.360    Median : 39.3896   Median :26.81
##  Mean   : 6.886    Mean   : 51.0643   Mean   :26.81
##  3rd Qu.: 9.490    3rd Qu.: 75.5857   3rd Qu.:30.25
##  Max.   :28.180    Max.   :178.0650   Max.   :89.00
##  Completion_Rate_Primary_Male Completion_Rate_Primary_Female
##  Min.   : 29.00               Min.   : 18.00
##  1st Qu.: 78.00               1st Qu.: 79.00
##  Median : 78.41               Median : 79.00
##  Mean   : 78.41               Mean   : 79.00
##  3rd Qu.: 91.00               3rd Qu.: 93.25
```

```
##  Max.   :100.00                Max.   :100.00
##  Completion_Rate_Lower_Secondary_Male Completion_Rate_Lower_Secondary_Female
##  Min.   : 10.0                        Min.   :  4.00
##  1st Qu.: 54.0                        1st Qu.: 52.75
##  Median : 61.4                        Median : 61.84
##  Mean   : 61.4                        Mean   : 61.84
##  3rd Qu.: 68.5                        3rd Qu.: 75.25
##  Max.   :100.0                        Max.   :100.00
##  Completion_Rate_Upper_Secondary_Female
##  Min.   :  1.00
##  1st Qu.: 28.00
##  Median : 43.25
##  Mean   : 43.25
##  3rd Qu.: 43.44
##  Max.   :100.00
```

```r
# 失业率直方图
z <- data.frame(values = edu$Unemployment_Rate)

ggplot() +
  geom_histogram(data = z, aes(x = values, y = after_stat(density)),
                 binwidth = 3, fill = "lightblue", color = "black") +
  geom_density(data = z, aes(x = values, y = after_stat(density)),
               color = "red") +
  geom_vline(data = z, aes(xintercept = mean(values),
                           color = "Mean"), linetype = "dashed") +
  geom_vline(data = z, aes(xintercept = mean(values) - sqrt(var(values)),
                           color = "Lower SD"), linetype = "dashed") +
  geom_vline(data = z, aes(xintercept = mean(values) + sqrt(var(values)),
                           color = "Upper SD"), linetype = "dashed") +
  labs(title = "Unemployment Rate Histogram",
       x = "Rate", y = "Density") +
  theme_minimal() +
  scale_color_manual(values = c("black", "purple", "green"),
                     labels = c("Mean", "Lower SD", "Upper SD"),
                     guide = guide_legend())

# 失业率 QQ 图
qqPlot(edu.unemploy$Unemployment_Rate)
```

```
## [1] 145  90
```


Unemployment Rate Histogram
Density
0.100
0.075
0.050
0.025
0.000
0
10
20
30
Rate
colour
Mean
Lower SD
Upper SD


edu.unemploy$Unemployment_Rate
0
5
10
15
20
25
145
90
−2
−1
0
1
2
norm quantiles

3.4.2 主要分析：建立预测模型（所有因素下 Akaike 的 backward）

```
unemploy_model_all <- lm(Unemployment_Rate ~., data = edu.unemploy)
backward <- step(unemploy_model_all, direction = 'backward', scope = formula(unemploy_model_all),
backward$anova
```

```
##                                       Step Df Deviance Resid. Df Resid. Dev
## 1                                         NA       NA       168   3708.442
## 2 - Completion_Rate_Upper_Secondary_Female  1 3.641297       169   3712.084
## 3         - Completion_Rate_Primary_Female  1 8.018438       170   3720.102
## 4           - Completion_Rate_Primary_Male  1 2.027557       171   3722.130
## 5        - OOSR_Upper_Secondary_Age_Female  1 9.731522       172   3731.861
##        AIC
## 1 552.4274
## 2 550.6002
## 3 548.9799
## 4 547.0758
## 5 545.5354
```

```
summary(backward)
```

```
##
## Call:
## lm(formula = Unemployment_Rate ~ Longitude + Completion_Rate_Lower_Secondary_Male +
##     Completion_Rate_Lower_Secondary_Female, data = edu.unemploy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.206 -3.378 -1.212  2.280 17.729
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          5.390786   1.166498   4.621 7.45e-06
## Longitude                           -0.041195   0.008843  -4.659 6.34e-06
## Completion_Rate_Lower_Secondary_Male -0.132673   0.068172  -1.946  0.05326
## Completion_Rate_Lower_Secondary_Female  0.189913   0.061700   3.078  0.00243
##
## (Intercept)                         ***
## Longitude                           ***
## Completion_Rate_Lower_Secondary_Male   .
## Completion_Rate_Lower_Secondary_Female **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.658 on 172 degrees of freedom
## Multiple R-squared:  0.1734, Adjusted R-squared:  0.159
## F-statistic: 12.03 on 3 and 172 DF,  p-value: 3.466e-07
```

3.4.3 主要分析：建立预测模型（所有因素下 Akaike 的 forward）

```
unemploy_model_1 <- lm(Unemployment_Rate ~ 1, data = edu.unemploy)
forward <- step(unemploy_model_1, direction = 'forward', scope = formula(unemploy_model_all), trac
forward$anova
```

```
##                                       Step Df  Deviance Resid. Df Resid. Dev
## 1                                         NA        NA       175   4514.727
## 2 + Completion_Rate_Lower_Secondary_Female -1 261.08155       174   4253.645
## 3                               + Longitude -1 439.60573       173   3814.040
## 4   + Completion_Rate_Lower_Secondary_Male -1  82.17855       172   3731.861
##        AIC
## 1 573.0524
## 2 564.5684
## 3 547.3690
## 4 545.5354
```

```
summary(forward)
```

```
##
## Call:
## lm(formula = Unemployment_Rate ~ Completion_Rate_Lower_Secondary_Female +
##     Longitude + Completion_Rate_Lower_Secondary_Male, data = edu.unemploy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.206 -3.378 -1.212  2.280 17.729
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          5.390786   1.166498   4.621 7.45e-06
## Completion_Rate_Lower_Secondary_Female  0.189913   0.061700   3.078  0.00243
## Longitude                           -0.041195   0.008843  -4.659 6.34e-06
## Completion_Rate_Lower_Secondary_Male   -0.132673   0.068172  -1.946  0.05326
##
## (Intercept)                          ***
```

```
## Completion_Rate_Lower_Secondary_Female **
## Longitude                               ***
## Completion_Rate_Lower_Secondary_Male    .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.658 on 172 degrees of freedom
## Multiple R-squared:  0.1734, Adjusted R-squared:  0.159
## F-statistic: 12.03 on 3 and 172 DF,  p-value: 3.466e-07
```

3.4.4 主要分析：建立预测模型（指定显著因素下 Akaike 的 backward）

```
unemploy_model2 <- lm(Unemployment_Rate ~
  Longitude +
  OOSR_Upper_Secondary_Age_Female +
  Completion_Rate_Primary_Male +
  Completion_Rate_Primary_Female +
  Completion_Rate_Lower_Secondary_Male +
  Completion_Rate_Lower_Secondary_Female +
  Completion_Rate_Upper_Secondary_Female
, data = edu.unemploy)
summary(unemploy_model2)
```

```
##
## Call:
## lm(formula = Unemployment_Rate ~ Longitude + OOSR_Upper_Secondary_Age_Female +
##     Completion_Rate_Primary_Male + Completion_Rate_Primary_Female +
##     Completion_Rate_Lower_Secondary_Male + Completion_Rate_Lower_Secondary_Female +
##     Completion_Rate_Upper_Secondary_Female, data = edu.unemploy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.685 -3.478 -1.190  2.324 17.249
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          4.147526   2.764467   1.500   0.1354
## Longitude                           -0.042183   0.009052  -4.660 6.41e-06
## OOSR_Upper_Secondary_Age_Female      0.013837   0.021205   0.653   0.5150
## Completion_Rate_Primary_Male         0.075451   0.110594   0.682   0.4960
## Completion_Rate_Primary_Female      -0.065517   0.099326  -0.660   0.5104
```

```
## Completion_Rate_Lower_Secondary_Male   -0.179630   0.094651  -1.898   0.0594
## Completion_Rate_Lower_Secondary_Female  0.251656   0.107512   2.341   0.0204
## Completion_Rate_Upper_Secondary_Female -0.017403   0.042849  -0.406   0.6851
##
## (Intercept)
## Longitude                             ***
## OOSR_Upper_Secondary_Age_Female
## Completion_Rate_Primary_Male
## Completion_Rate_Primary_Female
## Completion_Rate_Lower_Secondary_Male   .
## Completion_Rate_Lower_Secondary_Female *
## Completion_Rate_Upper_Secondary_Female
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.698 on 168 degrees of freedom
## Multiple R-squared:  0.1786, Adjusted R-squared:  0.1444
## F-statistic: 5.218 on 7 and 168 DF,  p-value: 2.106e-05
```

```
backward <- step(unemploy_model2, direction = 'backward', scope = formula(unemploy_model2), trace =
backward$anova
```

```
##                                       Step Df Deviance Resid. Df Resid. Dev
## 1                                         NA       NA       168   3708.442
## 2 - Completion_Rate_Upper_Secondary_Female  1 3.641297       169   3712.084
## 3         - Completion_Rate_Primary_Female  1 8.018438       170   3720.102
## 4           - Completion_Rate_Primary_Male  1 2.027557       171   3722.130
## 5        - OOSR_Upper_Secondary_Age_Female  1 9.731522       172   3731.861
##        AIC
## 1 552.4274
## 2 550.6002
## 3 548.9799
## 4 547.0758
## 5 545.5354
```

```
summary(backward)
```

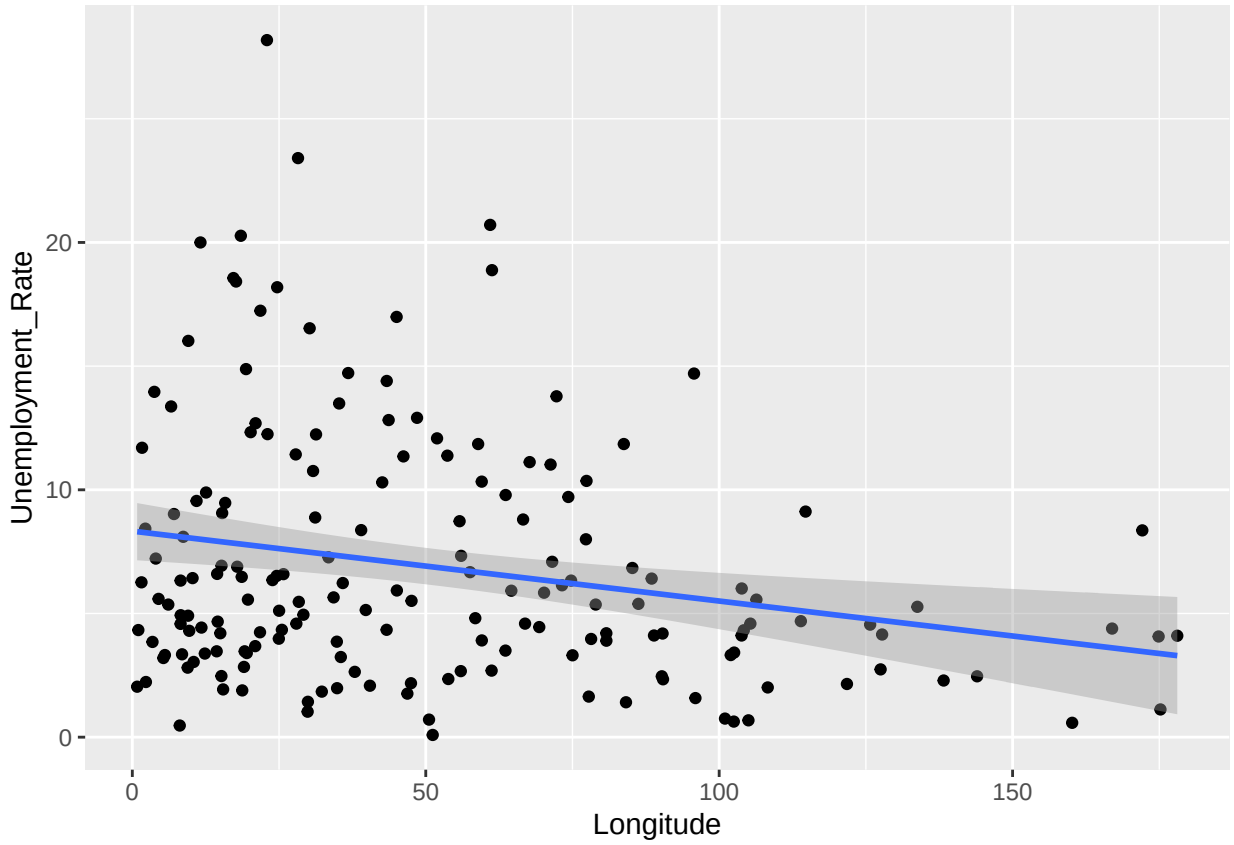
```
##
## Call:
## lm(formula = Unemployment_Rate ~ Longitude + Completion_Rate_Lower_Secondary_Male +
##     Completion_Rate_Lower_Secondary_Female, data = edu.unemploy)
##
## Residuals:
```

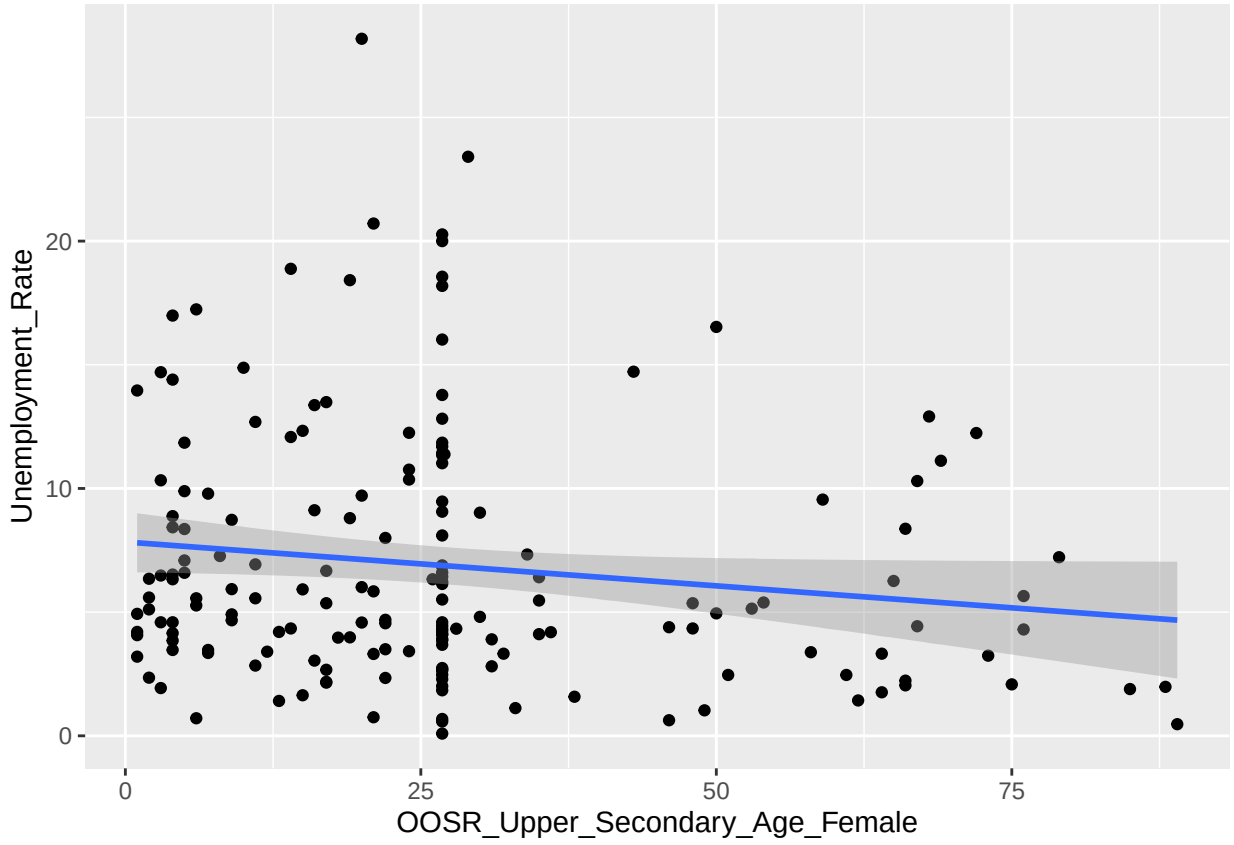
```
##    Min     1Q Median     3Q    Max
## -7.206 -3.378 -1.212  2.280 17.729
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5.390786   1.166498   4.621 7.45e-06
## Longitude                         -0.041195   0.008843  -4.659 6.34e-06
## Completion_Rate_Lower_Secondary_Male  -0.132673   0.068172  -1.946  0.05326
## Completion_Rate_Lower_Secondary_Female  0.189913   0.061700   3.078  0.00243
##
## (Intercept)                       ***
## Longitude                         ***
## Completion_Rate_Lower_Secondary_Male  .
## Completion_Rate_Lower_Secondary_Female **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.658 on 172 degrees of freedom
## Multiple R-squared:  0.1734, Adjusted R-squared:  0.159
## F-statistic: 12.03 on 3 and 172 DF,  p-value: 3.466e-07
```

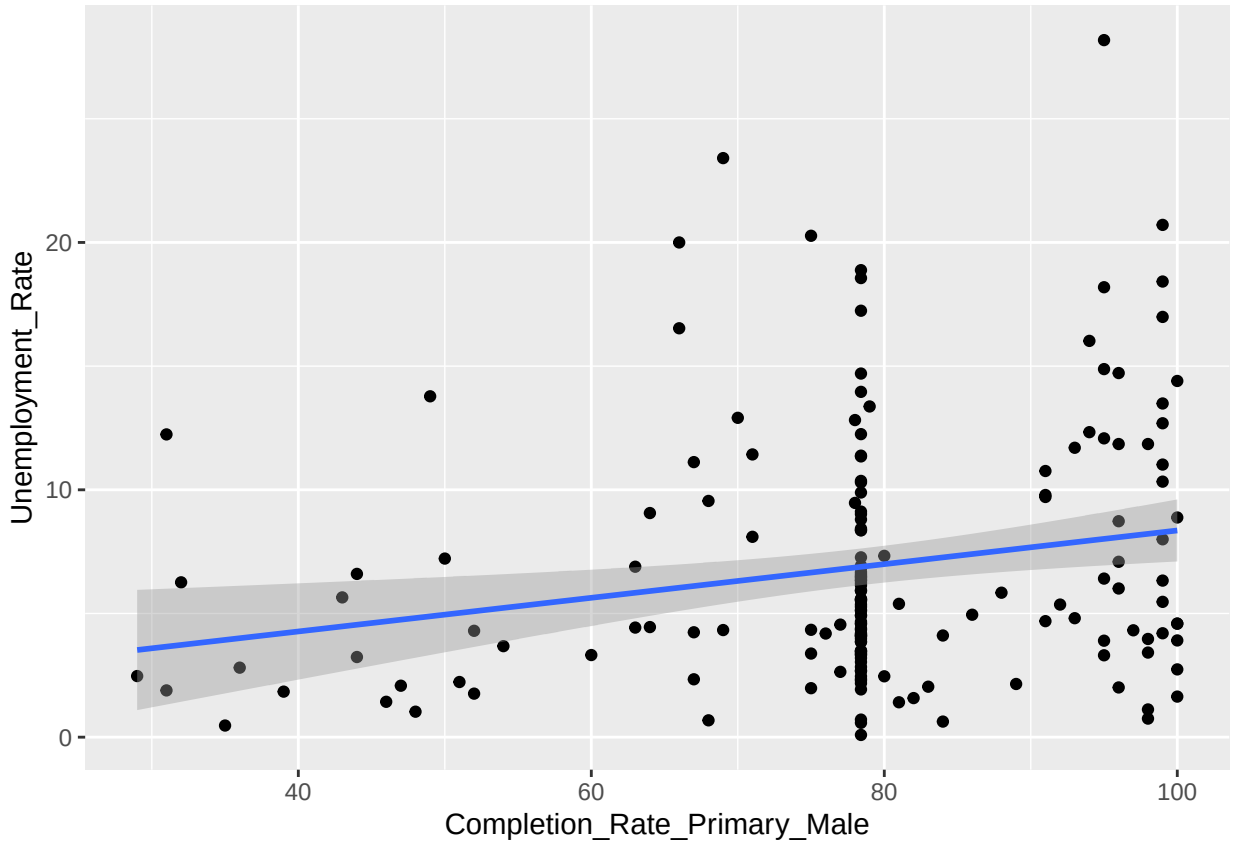
```
unemploy_model2 <- backward
```

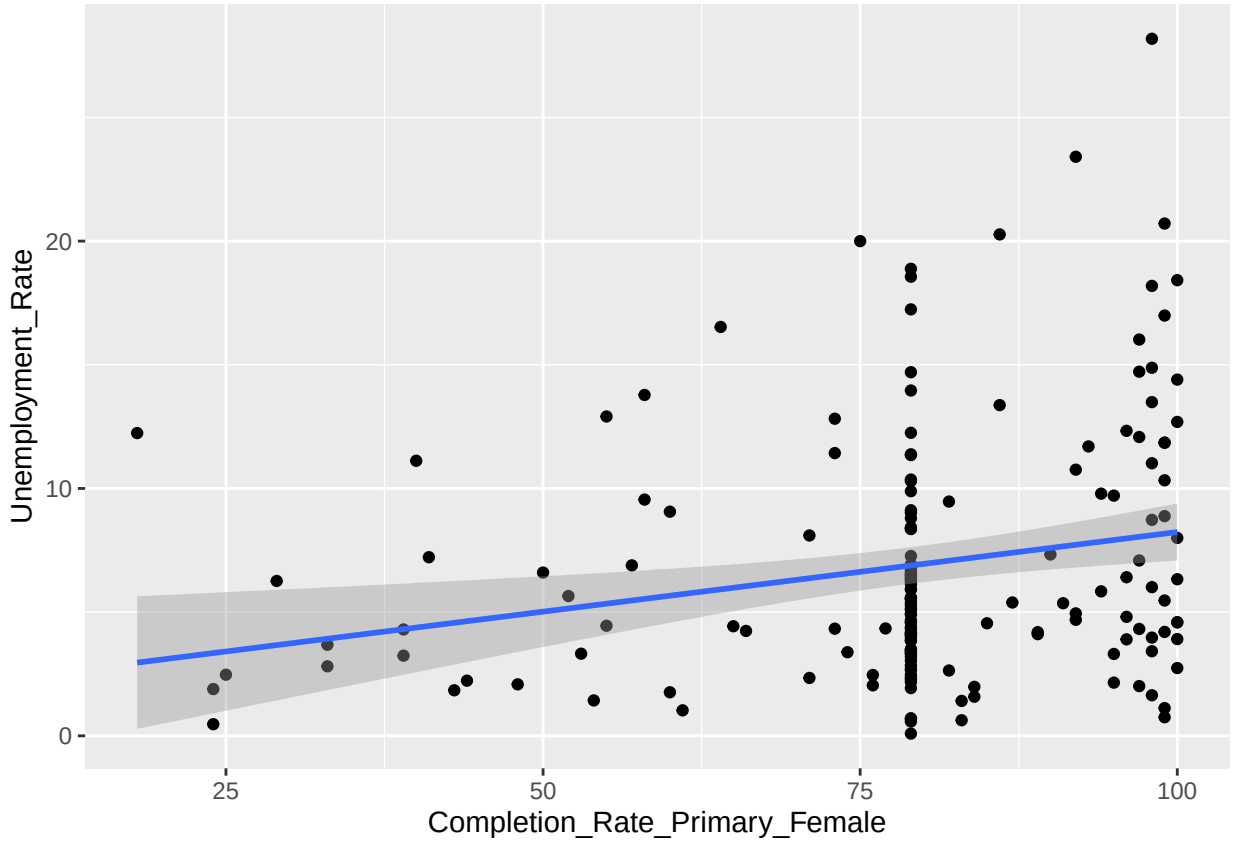
3.4.5 单独影响显著因素散点图

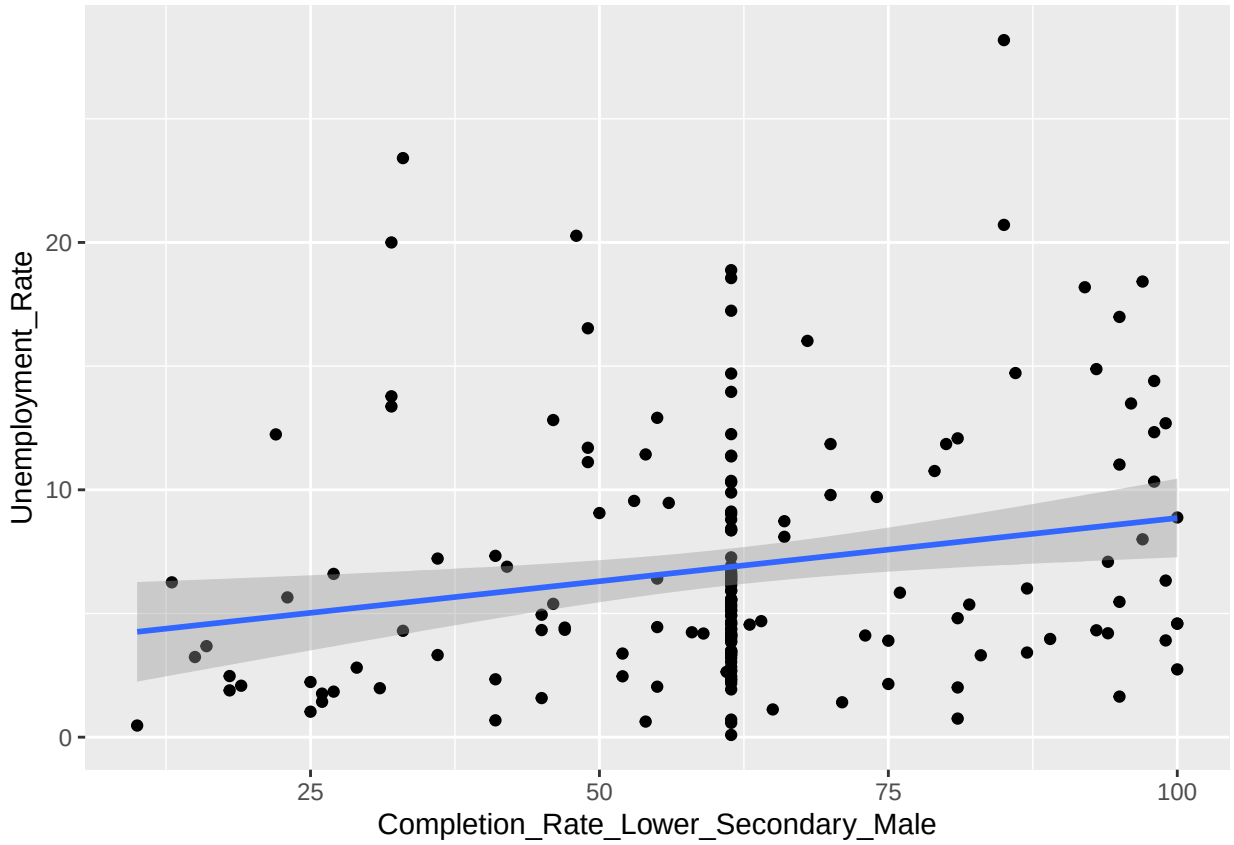
```
visualize <- function(column) {
  plot <- ggplot(data = edu.unemploy, aes(x = !!as.name(column), y = Unemployment_Rate)) +
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm", se = T)

  print(plot)
}

for (x in unemploy_remark[-1]) {
  visualize(x)
}
```

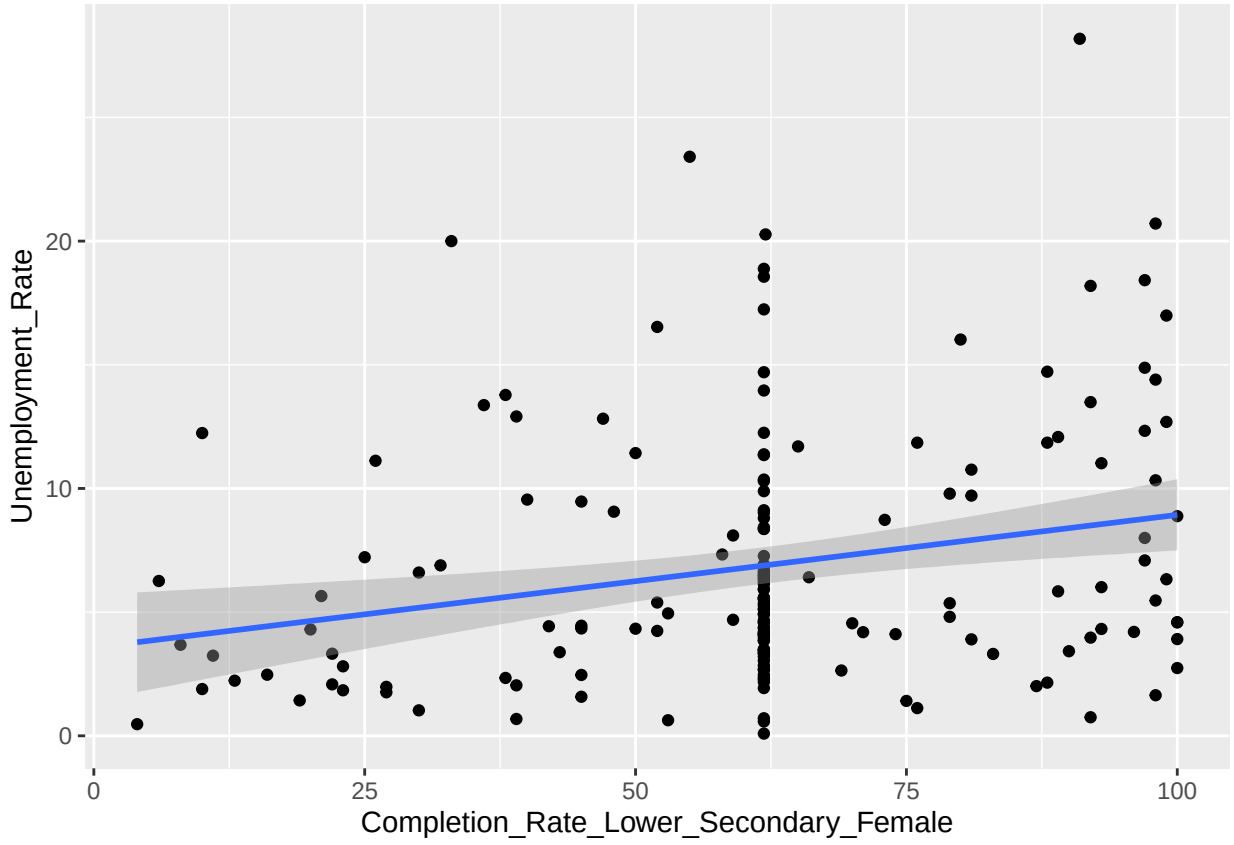

Unemployment_Rate
20
10
0
0
50
100
150
Longitude

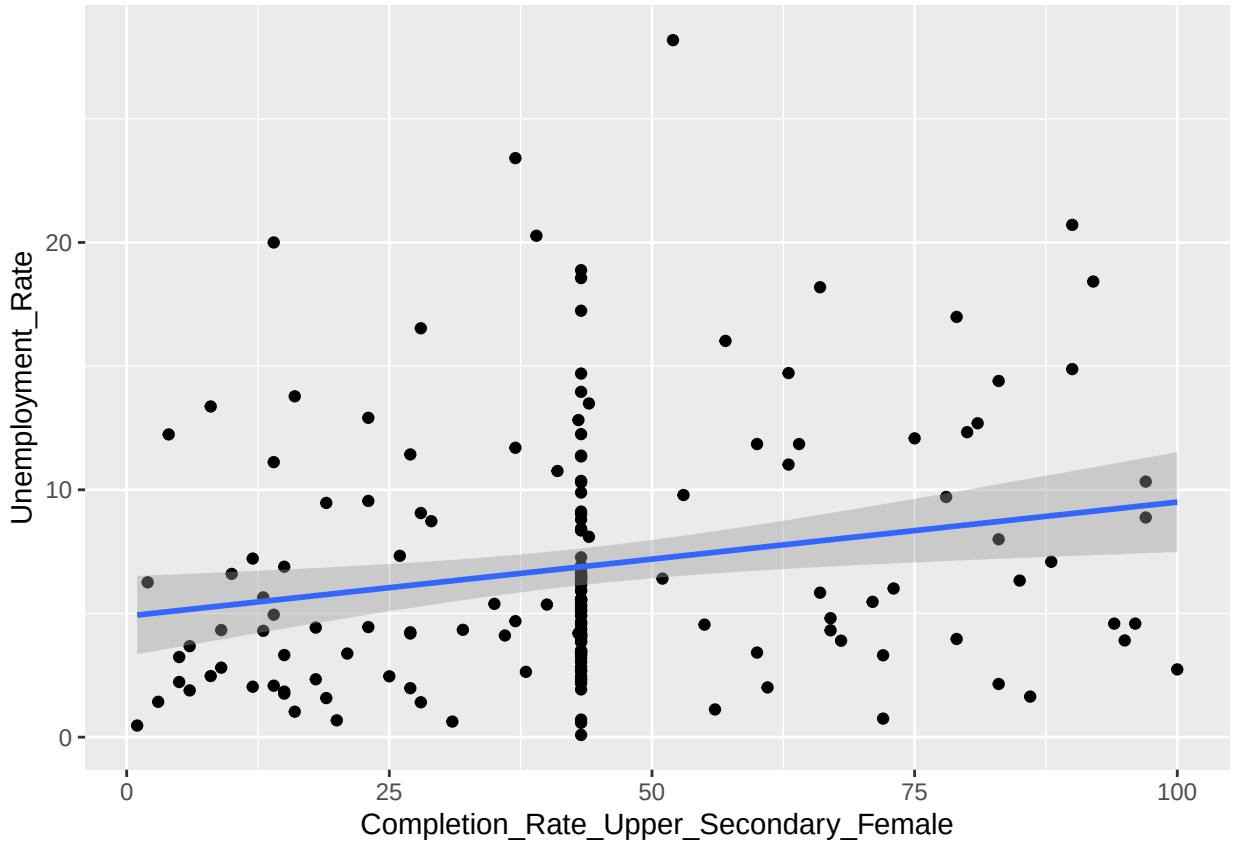

Unemployment_Rate
20
10
0
0
25
50
75
OOSR_Upper_Secondary_Age_Female


Unemployment_Rate
20
10
0
40
60
80
100
Completion_Rate_Primary_Male


Unemployment_Rate
20
10
0
25
50
75
100
Completion_Rate_Primary_Female


Unemployment_Rate
20
10
0
25
50
75
100
Completion_Rate_Lower_Secondary_Male


Unemployment_Rate
20
10
0
0
25
50
75
100
Completion_Rate_Lower_Secondary_Female



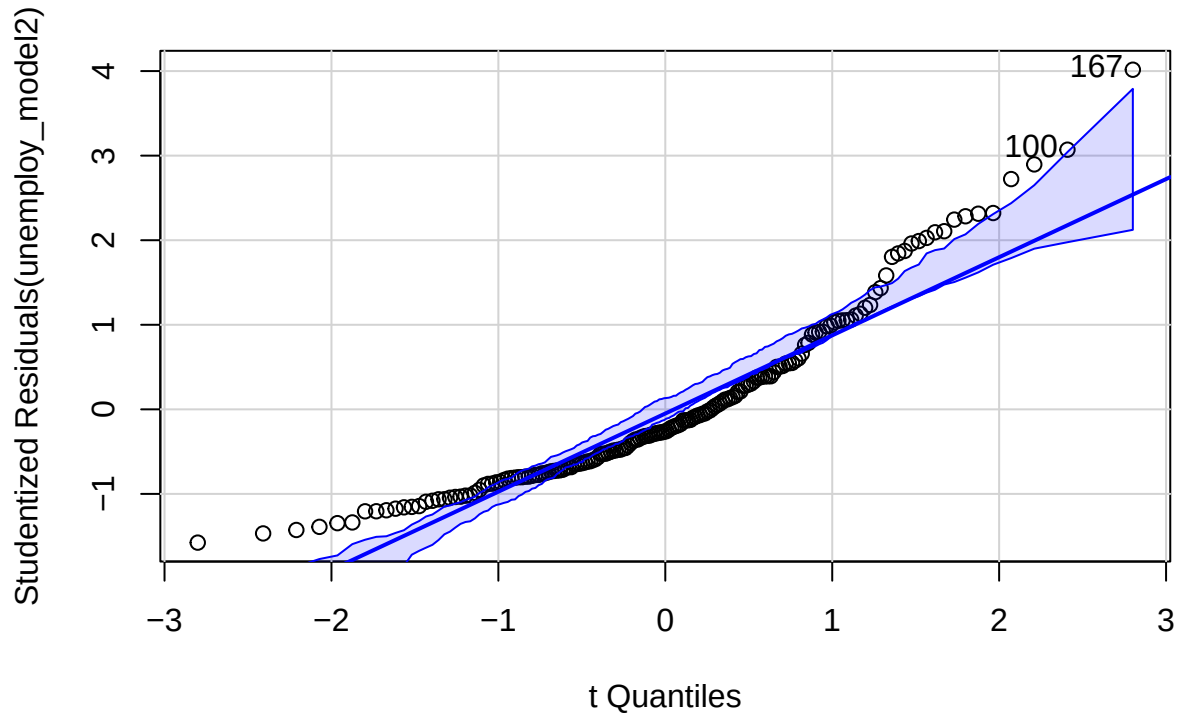
3.4.6 模型评估

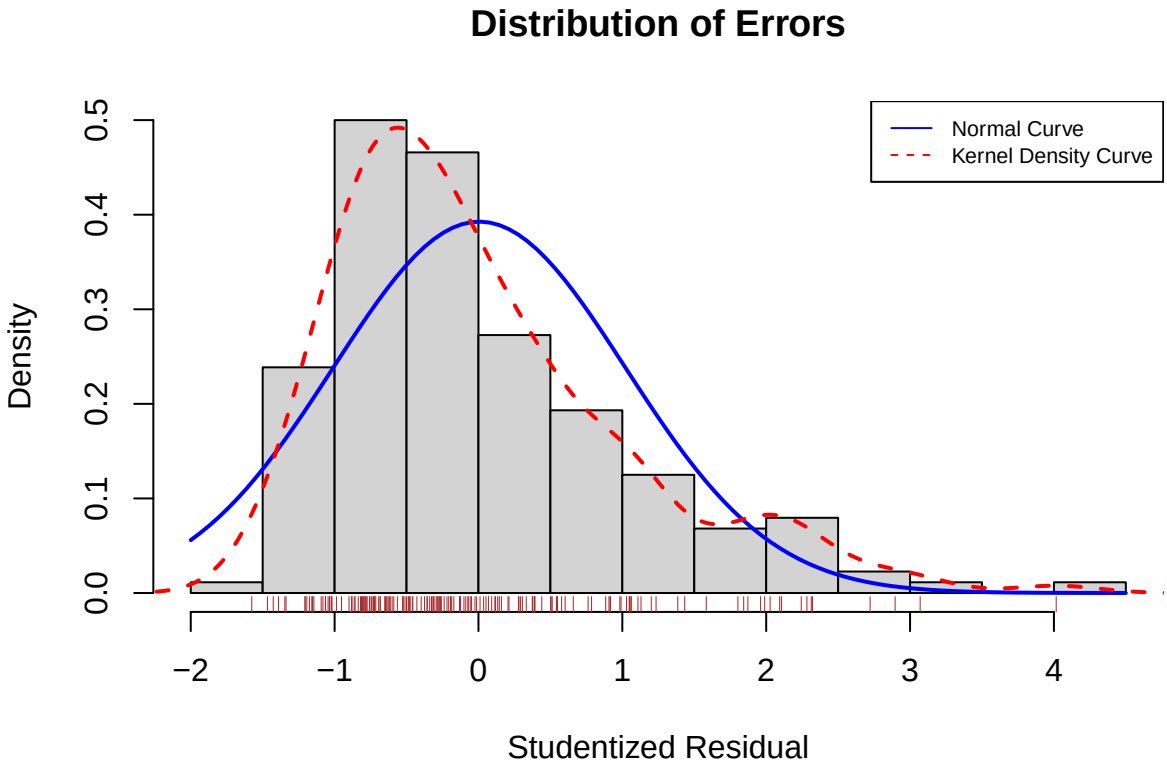
```
# 模型的 QQ 图
qqPlot(unemploy_model2, simulate = T)
```

```
## 100 167
##  90 145
```

```
# 绘制预测残差分布直方图
residplot <- function(fit, nbreaks=10) {
 z <- rstudent(fit)
 hist(z, breaks=nbreaks, freq=FALSE,
 xlab="Studentized Residual",
 main="Distribution of Errors")
 rug(jitter(z), col="brown")
 curve(dnorm(x, mean=mean(z), sd=sd(z)),
 add=TRUE, col="blue", lwd=2)
 lines(density(z)$x, density(z)$y,
 col="red", lwd=2, lty=2)
 legend("topright",
```

```
 legend = c( "Normal Curve", "Kernel Density Curve"),
 lty=1:2, col=c("blue","red"), cex=.7)
}

residplot(unemploy_model2, 10)
```





4 分析结果解读

4.1 原始数据

- 在预处理时，我们发现数据缺失比较严重，某些列的缺失程度达到 3/4。而且由于原始数据的列较多，我们在预处理时就进行了初步筛选。
- 方法是对于某一列 x 和自变量 y，选出 x 和 y 都不缺失的行，单独建立 y ~ x 的线性回归模型。拟合斜率 p 值小于 0.05 则保留，否则初步剔除。若该列缺失 1/2 则也剔除
- 对原始数据中除 y 以外的所有列进行上述的操作，对保留下来的列进行均值插补后，建立和 y 的多元线性回归模型，称为 model_all
- 同时我们使用了针对 AIC 的模型优化，另外建立了 y 和常量 1 的回归模型 model_1，分别对 model_all 和 model_1 进行 backward 和 forward 优化
- 最后我们根据 model_all 中各因素的影响情况，手动剔除一些影响不显著/拟合结果显然违反常理的因素后，建模并进行 backward 优化，称为 model_2。停止剔除因素的情况是 model_2：
  - AIC 值得到优化或相差不大
  - 与其余两个模型的影响显著的因素相近
  - R 方大小合适
  - 因素普遍影响显著
- 比较 model_1、model_2、model_all，择一采用

4.2 高教入学率模型

4.2.1 模型比较

项目	全因素-backward	全因素-forward	手动选择因素-backward
AIC	1003.214	1003.219	1007.727
残差标准差	15.84	15.96	16.21
R 方	0.6839	0.6732	0.6612
调整后 R 方	0.6691	0.6638	0.6534
因素总数	8	5	4
显著因素总数	7	5	4
模型 p 值	<2.2e-16	<2.2e-16	<2.2e-16

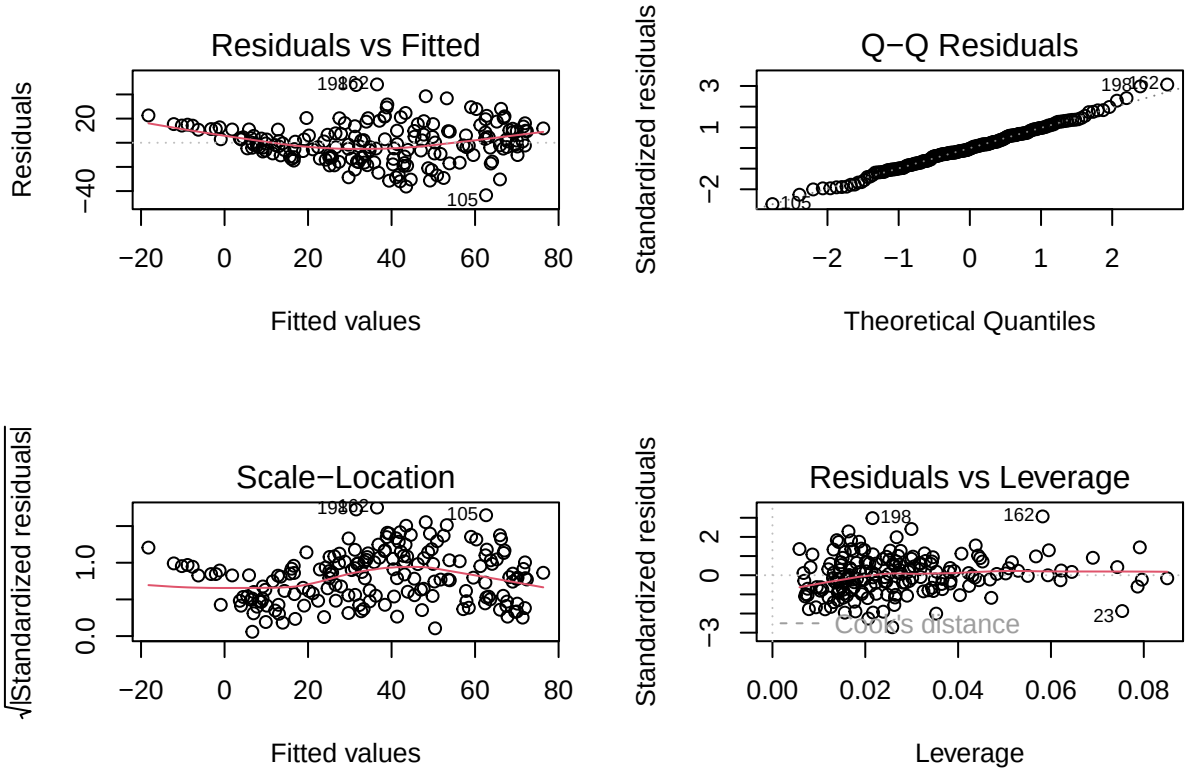
可以发现 AIC 优化后的 model_all、model_1 虽然数据较为漂亮，但有显然违反常理的拟合结果，因此在手动剔除后采用了较为合理的 model_2。

4.2.2 模型分析

```
print(enroll_model2$coefficients)
```

```
##                    (Intercept)                       Latitude
##                     57.3364130                      0.4612649
##     OOSR_PreOPrimary_Age_Male OOSR_Upper_Secondary_Age_Male
##                     -0.1345733                     -0.3475586
##                     Birth_Rate
##                     -0.9606103
```

```
par(mfrow=c(2,2))
plot(enroll_model2)
```



以下介绍 model_2 中，对一个国家或地区高教入学率影响显著的因素

4.2.2.1 纬度 拟合结果：每高 1°，高教入学率**增加 0.46%**

我们惊讶地发现三个模型中 Latitude 都有显著的影响，为此还特意进行了可视化，散点图显示的确低纬度地区集中在高教入学率低的区域，高纬度地区集中在高教入学率高的区域。最有可能的原因就是位于各个纬度区间的国家和地区发展情况差异明显。这里为了阐述方便，纬度的高低未采用地理上的约定：

- **低纬度（0-20）**：东南亚、印度、非洲北部和美洲中部（多欠发达国家）
- **中纬度（20-40）**：中东、东亚、美国、非洲和南美洲（多发展中国家）
- **高纬度（40-60）**：欧洲、俄罗斯、加拿大（多发达国家）

这无疑会带来以下结果：

- **人口密度**：高纬度地区的人口密度可能较低，这可能导致更好的教育资源分配和更低的竞争，从而提高高教入学率。
- **经济状况**：高纬度地区可能拥有更发达的经济体系，有更多的资源投入到教育领域，包括高等教育。
- **社会状况**：不同纬度的国家或地区可能对于教育的社会看法不同，这可能对高教入学率产生影响。

4.2.2.2 学前、高中男生失学率 拟合结果：

- 学前每高 1%，高教入学率**降低 0.13%**

- 高中每高 1%，高教入学率**降低 0.35%**

同为失学率，学前和高中可以分开看待。

首先，由于学前教育就学习本身来说竞争压力较小，同时学前教育对于形成价值观、认识世界有重要的影响，一般家庭普遍会让孩子接受学前教育。因此，学前失学最有可能是家庭原因，比如经济问题或家庭背景问题；亦或者是当地学前教育匮乏。无法接受学前教育带来的学习能力、知识缺失，会暗中影响到孩子能否接受高等教育。

高中失学则不同，一方面，就学习来说高中失学意味着为高等教育做的学习准备没有做好，高等教育所需要的一些学习品质没有养成，因此能力不足以接受高等教育，直接影响其高教入学；另一方面，在某些社会中，可能存在对高等教育的不同看法或文化倾向，这些文化中的人可能更倾向于早期就业，提倡职业教育。综上，虽然同为下降趋势，高中失学明显比学前失学影响大。

4.2.2.3 生育率 拟合结果：每高 1%，高教入学率降低 0.96%

表面上，高生育率意味着学生群体有更多的新鲜血液，有更多人可能接受高等教育，可以带动整个国家或地区的高教入学率。但以下原因可以解释该拟合结果：

- **资源竞争**：生育率高导致的人口增长，在高等教育资源增长与人口增长不同步时，就会导致入学率下降。因为资源分配极大激化了有限高等教育资源的竞争。
- **家庭压力**：高生育率意味着家庭面临的经济挑战更多，也就意味着家庭难以支付较高的高等教育费用，提供足够支持鼓励孩子接受高等教育的可能就更大
- **价值观念**：在一些高生育率文化中，人们可能更看重一个人的家庭责任而非受教育水平，导致更多的家庭倾向于让孩子早日进入工作岗位，而不是继续接受高等教育。

4.3 失业率模型

4.3.1 模型比较

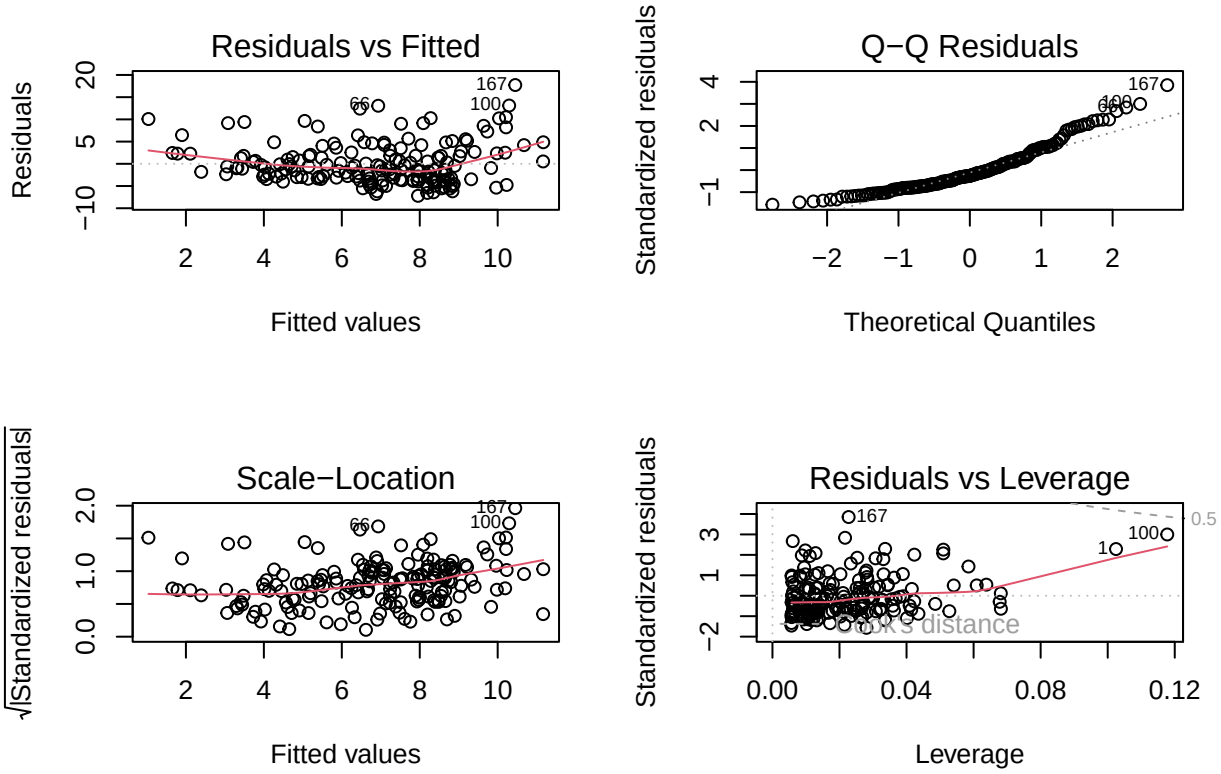
项目	全因素-backward/全因素-forward/手动选择因素-backward
AIC	545.5354
残差标准差	4.658
R 方	0.1734
调整后 R 方	0.159
因素总数	3
显著因素总数	2
模型 p 值	3.466e-07

```
print(unemploy_model2$coefficients)
```

```
##                (Intercept)                  Longitude
```

```
##                                5.39078643                                -0.04119532
##   Completion_Rate_Lower_Secondary_Male Completion_Rate_Lower_Secondary_Female
##                               -0.13267348                                 0.18991261
```

```
par(mfrow=c(2,2))
plot(unemploy_model2)
```



三个模型完全一致，且无论从 R 方、残差分布直方图来看拟合效果都非常一般。2 个显著影响因素还是非常反直觉且难以解释的：

- **经度**：世界范围内国家和地区在经度上的分布毫无规律，同一经度上的国家和地区共同特点也很少
- **女生初中完成率**：完全无法解释为什么女生初中完成率越高，失业率越高（即拟合斜率为正）

参考高教入学率的拟合结果，我们采用的统计方法有一定合理性。因此经过讨论，我们决定放弃以回归模型对失业率进行预测。因为失业率是一个更为复杂，与社会情况关系更密切的因素。它需要考虑到一个国家或地区的政治、经济因素（如经济大萧条引起的美国高失业率），单单从教育因素预测失业率不免过于片面。

4.4 总结

1. 从单因素散点图上看，实际上高教入学率的确和各类失学率成反比，和各类完成率成正比。单单从模型上考虑其实也不完全足够。但无论如何，经过本次分析，我们可以给全球教育工作者的建议如下：

- **关注失学人群：**接受高等教育是一个水到渠成的结果，离不开孩子成长各个阶段的教育。无论是哪一环的缺失，带来的影响都不可估量，因为初等教育对于人的价值观塑造非常重要，没有正确引导，孩子接触不良因素的可能就会极大提升。
- **关注个体差异：**意识到每个学生都是独特的个体，有着不同的学习风格、兴趣和能力。同时这也意味着要考虑学生所处地区的社会情况，因地制宜进行教育引导，高等教育也许真的不是唯一的出路，不必一味强求，帮助孩子培养良好品质和能力才是目标。

2. 失业率虽然与预期不符，但并不代表教育与就业完全无关。孩子通过接受教育培养出的学习能力、人际交往能力在工作中依然有帮助，显然不能因为看似关系不大而放弃对学生的教育，相反还应该帮助学生树立正确的择业观、就业观，这才是教育的意义所在。当然，前面也提到就业还需考虑许多教育以外的因素，因此教育工作者若想在学生就业方面提供帮助，需要以开阔的视野去获取相关信息，以求结合专业知识提供更加合理的建议。