基于教育数据的高等教育入学率及失业率预测模型分析报告

完成人：李凌朗

引言

项目背景

该项目依托一个全球范围内的教育数据集（内容涵盖辍学率、出生率、和高等教育入学统计等），通过统计分析方法对各个国家地区的教育现状进行分析，并提供建议。该项目旨在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。

数据介绍

该数据集包含29列数据，数据介绍如下：

1. 数据预处理
2. 数据分析&结论
3. 数据建模及模型质量评估
    a. 建立高等教育总入学人数预测模型（Y：Gross_Tertiary_Education_Enrollment）

b. 建立失业率预测模型（Y：Unemployment_Rate）

4. 模型解读（影响因素分析）
5. 得到教育改进启示和建议

1. **Countries and Areas**: Name of the countries and areas.
2. **Latitude**: Latitude coordinates of the geographical location.
3. **Longitude**: Longitude coordinates of the geographical location.
4. **OOSR_Pre0Primary_Age_Male**: Out-of-school rate for pre-primary age males.
5. **OOSR_Pre0Primary_Age_Female**: Out-of-school rate for pre-primary age females.
6. **OOSR_Primary_Age_Male**: Out-of-school rate for primary age males.
7. **OOSR_Primary_Age_Female**: Out-of-school rate for primary age females.
8. **OOSR_Lower_Secondary_Age_Male**: Out-of-school rate for lower secondary age males.
9. **OOSR_Lower_Secondary_Age_Female**: Out-of-school rate for lower secondary age females.
10. **OOSR_Upper_Secondary_Age_Male**: Out-of-school rate for upper secondary age males.
11. **OOSR_Upper_Secondary_Age_Female**: Out-of-school rate for upper secondary age females.
12. **Completion_Rate_Primary_Male**: Completion rate for primary education among males.
13. **Completion_Rate_Primary_Female**: Completion rate for primary education among females.
14. **Completion_Rate_Lower_Secondary_Male**: Completion rate for lower secondary education among males.
15. **Completion_Rate_Lower_Secondary_Female**: Completion rate for lower secondary education among females.
16. **Completion_Rate_Upper_Secondary_Male**: Completion rate for upper secondary education among males.
17. **Completion_Rate_Upper_Secondary_Female**: Completion rate for upper secondary education among females.
18. **Grade_2_3_Proficiency_Reading**: Proficiency in reading for grade 2-3 students.
19. **Grade_2_3_Proficiency_Math**: Proficiency in math for grade 2-3 students.
20. **Primary_End_Proficiency_Reading**: Proficiency in reading at the end of primary

education.

21. **Primary_End_Proficiency_Math**: Proficiency in math at the end of primary education.
22. **Lower_Secondary_End_Proficiency_Reading**: Proficiency in reading at the end of lower secondary education.
23. **Lower_Secondary_End_Proficiency_Math**: Proficiency in math at the end of lower secondary education.
24. **Youth_15_24_Literacy_Rate_Male**: Literacy rate among male youths aged 15-24.
25. **Youth_15_24_Literacy_Rate_Female**: Literacy rate among female youths aged 15-24.
26. **Birth_Rate**: Birth rate in the respective countries/areas.
27. **Gross_Primary_Education_Enrollment**: Gross enrollment in primary education.
28. **Gross_Tertiary_Education_Enrollment**: Gross enrollment in tertiary education.
29. **Unemployment_Rate**: Unemployment rate in the respective countries/areas.

一、数据导入与预处理

（1）导入数据

数据集包括202个国家和地区的教育信息，包括各阶段教育完成率、辍学率等。首先导入必要的包和数据集：

```
# 导入必要的包
library(ggplot2)
library(caret)
library(tidyverse)
# 导入数据
edu_data <- read.csv("education.csv")
#设置随机种子
set.seed(123)
```

接下来对数据进行预处理。

（2）查找空值及缺失值

```
# 查找缺失值
sapply(edu_data, function(x) sum(is.na(x)))
# 查找空值
sapply(edu_data, function(x) sum(x==""))
summary(edu_data)
```

我们首先查找了缺失值与空值，发现数据集中不存在空值及缺失值。

通过summary，我们发现数据集中存在入学率超过100%的异常值以及大量数值为0的异常值。下一步我们对这些异常值进行处理。

（3）处理异常值

我们对入学率超过100%的异常值采取的处理方法是直接删除，避免异常值影响模型准确性。

```r
# 去除入学率超过100%异常值
edu_data <- edu_data[edu_data$Gross_Primary_Education_Enrollment <= 100 &
edu_data$Gross_Tertiary_Education_Enrollment <=
edu_data$Gross_Primary_Education_Enrollment,]
```

我们对数值为0的异常值采取的处理方法是将其值替换为该列所有非零值的平均值，这样能够确保数据的连续性。

```r
# 处理数值为0的异常值
# 定义函数来处理每一列的零值
replace_zero_with_mean <- function(column) {
  mean_value <- mean(column[column != 0])  # 计算该列非零值的平均值
  column[column == 0] <- mean_value  # 将为0的值替换为平均值
  return(column)
}

# 对数据集除了地理坐标的每一列应用上述函数
for (i in 4:29) {
  edu_data[, i] <- replace_zero_with_mean(edu_data[, i])
}

summary(edu_data)
```

通过summary，我们可以看出，异常值已经被处理完成。

（4）数据计算

我们发现数据集中存在大量数据只存在男女性的区别，为了方便后续分析，我们假设每个国家或地区的男女比例为1：1，该数据的值即为男女性均值。（例如：高中辍学率=(高中男生辍学率+高中女生辍学率)/2）

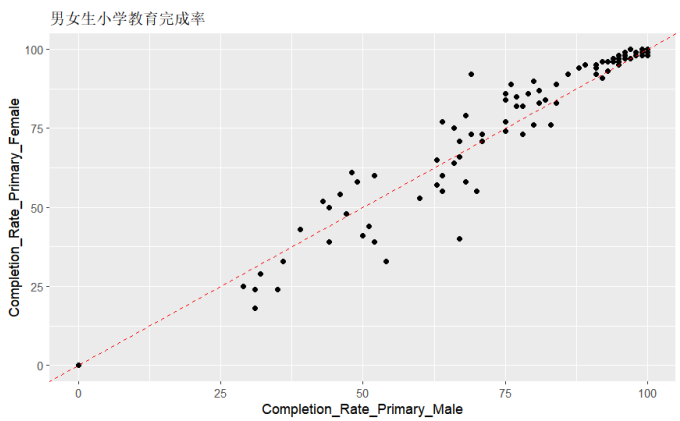
```r
# 后面操作均为计算平均值
edu_data$Average_OOSR_Pre0Primary_Age <- (edu_data$OOSR_Pre0Primary_Age_Male + edu_data$OOSR_Pre0Primary_Age_Female) / 2
```

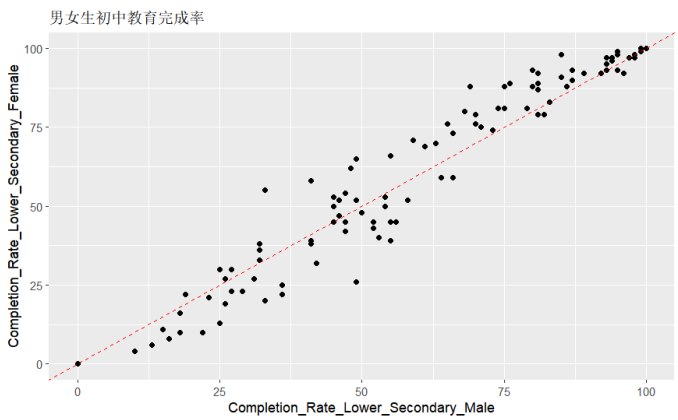
二、数据分析

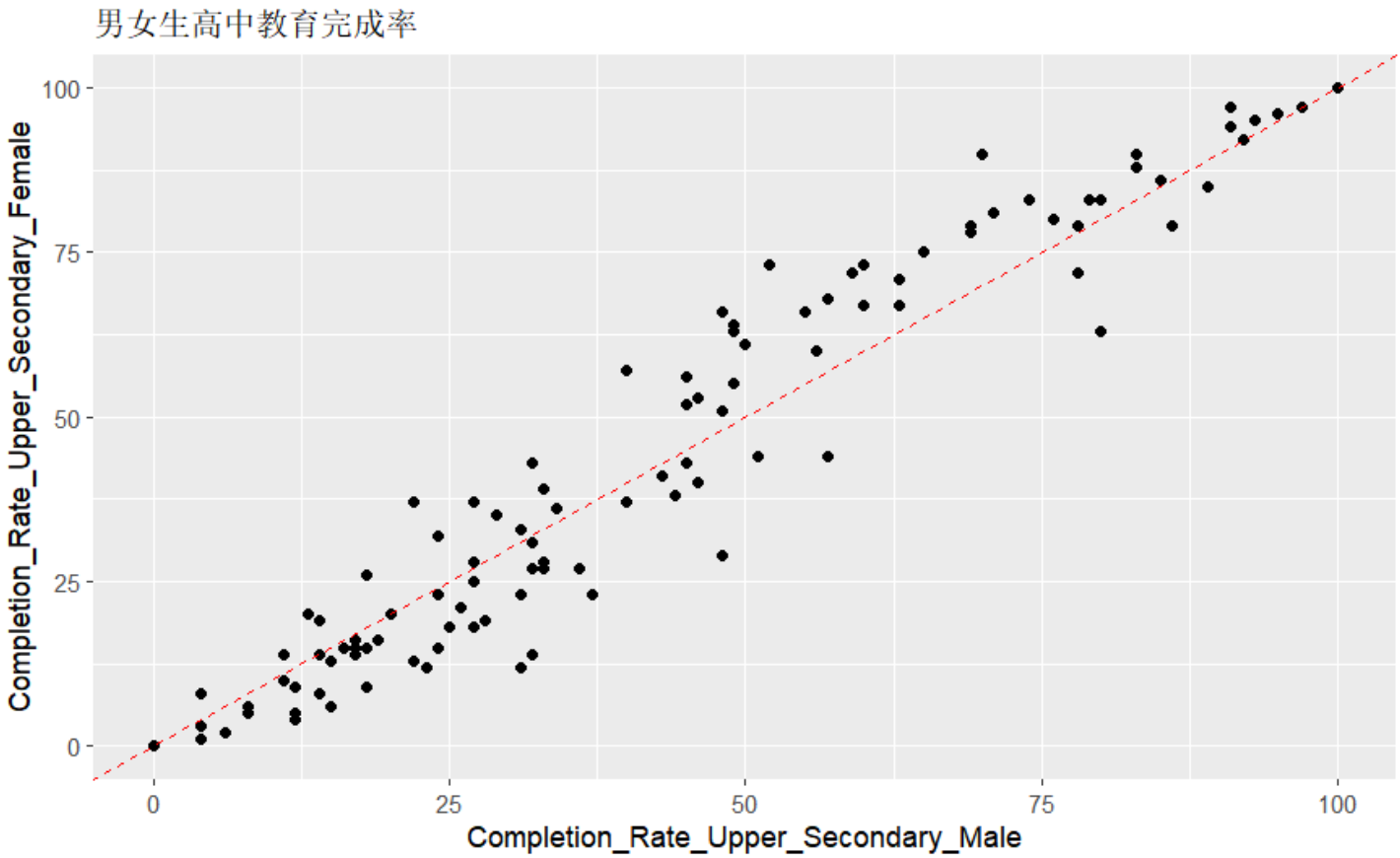
（1）性别对教育完成率的影响

这里我们先创建了一个基础散点图函数，随后调用该函数得到三幅散点图，分别为男女生小学、初中、高中教育完成率散点图。

```r
# 创建基础散点图函数
plot_base <- function(data, x_var, y_var, title_text) {
  ggplot(data, aes(x = !!sym(x_var), y = !!sym(y_var))) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
    labs(title = title_text,
         x = x_var,
         y = y_var)
}

# 依次创建男女生小学、初中、高中教育完成率散点图
plot_base(edu_data, "Completion_Rate_Primary_Male", "Completion_Rate_Primary_Female", "男女生小学教育完成率")

plot_base(edu_data, "Completion_Rate_Lower_Secondary_Male", "Completion_Rate_Lower_Secondary_Female", "男女生初中教育完成率")

plot_base(edu_data, "Completion_Rate_Upper_Secondary_Male", "Completion_Rate_Upper_Secondary_Female", "男女生高中教育完成率")
```







图中的直线为y=x线，直线下方的点表示男性教育完成率比女性高，直线上方的点表示男性教育完成率比女性低。

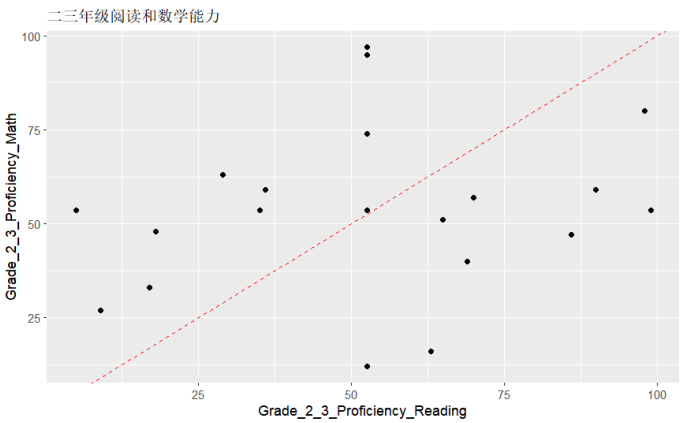
从图中我们可以看出，男女性在不同教育阶段的教育完成率整体上围绕着y=x分布，二者在整体上没有显著差异。不同国家男女生之间的教育完成率存在相当大的差异，许多国家教育完成率依旧处于较低水平，需要增加教育投入。

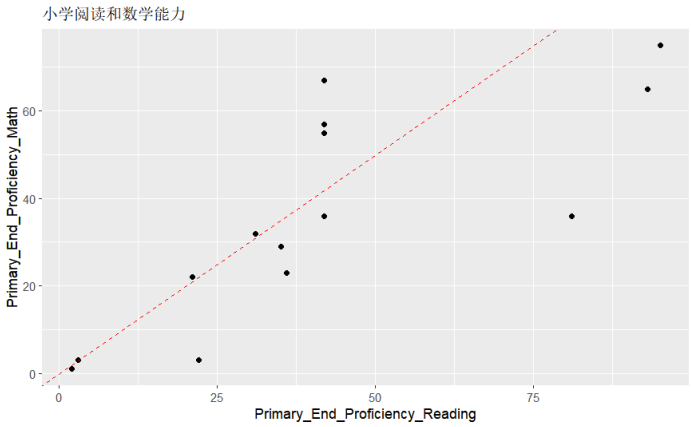
（2）不同年龄段学生的阅读能力和数学能力差异分析

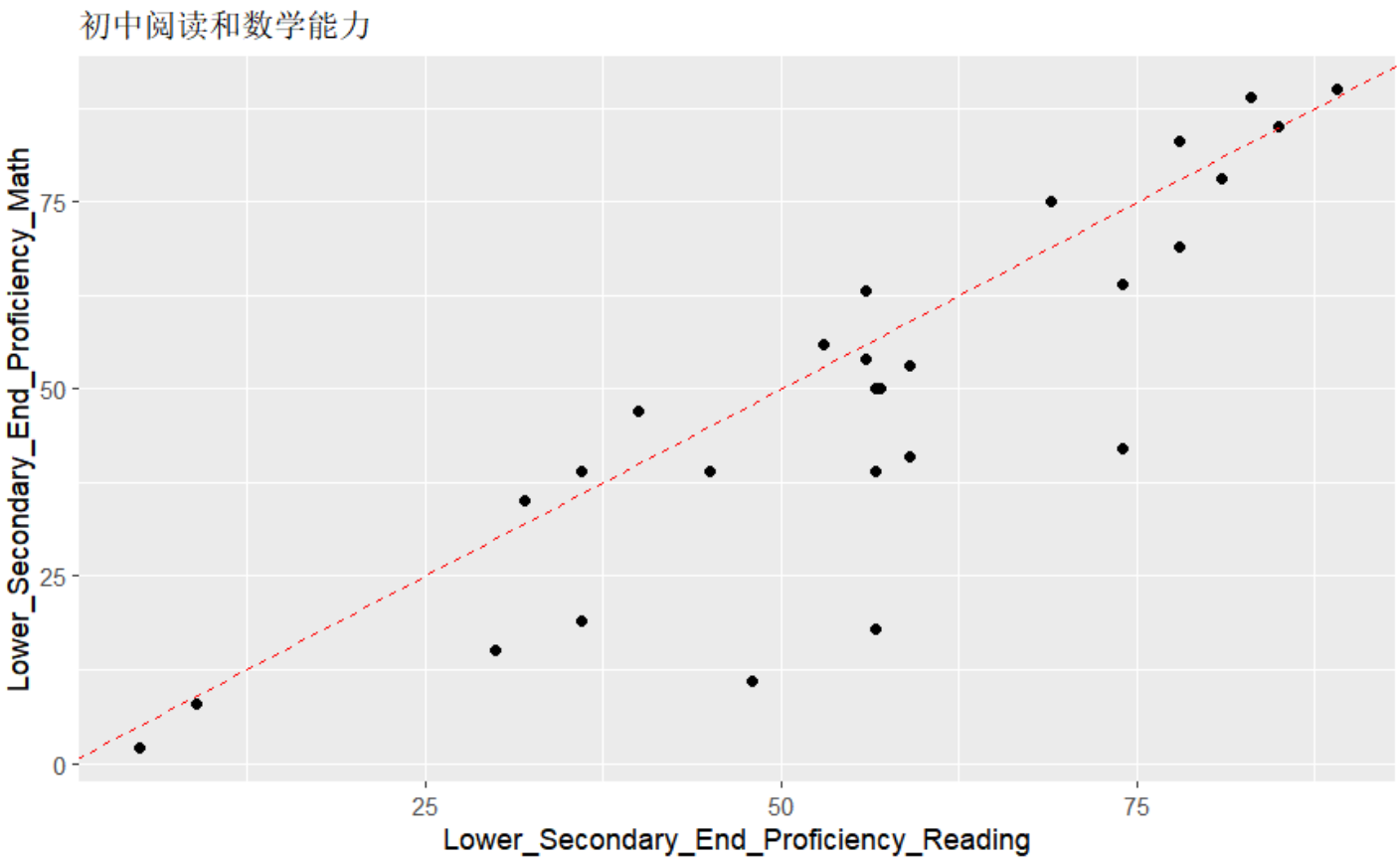
这里我们先创建了一个基础散点图函数，随后调用该函数得到三幅散点图，分别为二三年级、小学、初中阅读和数学能力分析散点图。

```
# 创建基础散点图函数
plot_base <- function(data, x_var, y_var, title_text) {
  ggplot(data, aes(x = !!sym(x_var), y = !!sym(y_var))) +
    geom_point() +
    geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
    labs(title = title_text,
         x = x_var,
         y = y_var)
}

# 依次创建二三年级、小学、初中阅读和数学能力分析散点图
plot_base(edu_data, "Grade_2_3_Proficiency_Reading",
"Grade_2_3_Proficiency_Math", "二三年级阅读和数学能力")
```

```
13
14 plot_base(edu_data, "Primary_End_Proficiency_Reading",
   "Primary_End_Proficiency_Math", "小学阅读和数学能力")
15
16 plot_base(edu_data, "Lower_Secondary_End_Proficiency_Reading",
   "Lower_Secondary_End_Proficiency_Math", "初中阅读和数学能力")
```







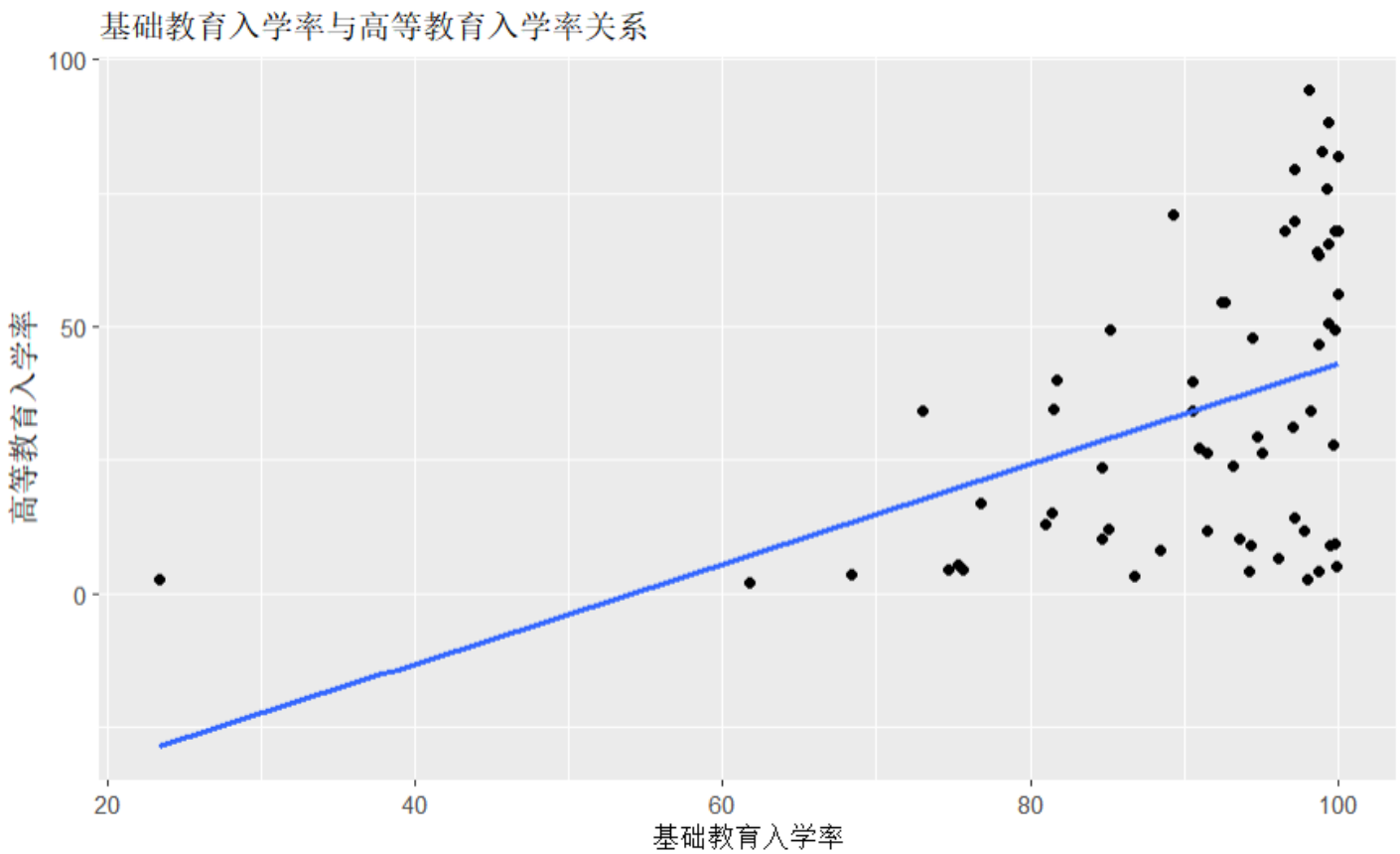
图中的直线为y=x线，直线下方的点表示阅读能力比数学能力高，直线上方的点表示阅读能力比数学能力低。

从图中我们可以看出，二三年级与小学学生的阅读和数学能力分化较严重，大部分学生只专注于一项能力，初中学生的阅读与数学能力发展逐渐平衡，散点逐渐靠近y=x。

（3）基础教育入学率对高等教育入学率的影响

我们通过绘制带有线性回归线的散点图来研究基础教育入学率对高等教育入学率的影响。

```r
# 创建散点图
ggplot(edu_data, aes(x = Gross_Primary_Education_Enrollment, y = Gross_Tertiary_Education_Enrollment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "基础教育入学率与高等教育入学率关系",
       x = "基础教育入学率",
       y = "高等教育入学率")
```



从图中可以看出基础教育入学率与高等教育入学率有显著的正相关性，基础教育入学率越高，高等教育入学率越高。

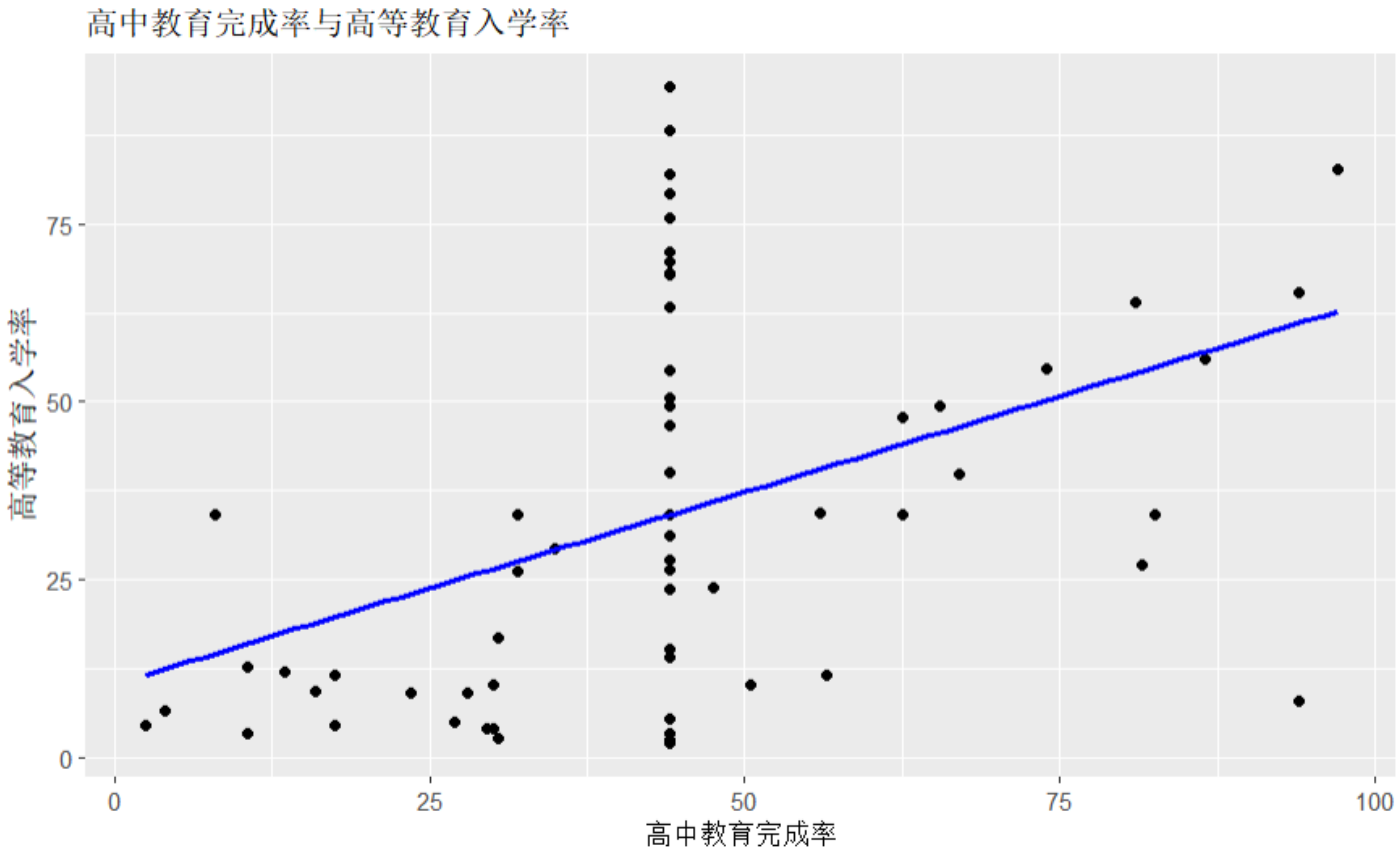
（4）教育完成率对高等教育入学率的影响

同样使用带有线性回归线的散点图来研究教育完成率率对高等教育入学率的影响。

这里我们选取距离高等教育最近的高中教育完成率作为教育完成率的代表。

```r
# 绘制散点图分析高中教育完成率与高等教育入学率的关系
ggplot(edu_data, aes(x = Average_Completion_Rate_Upper_Secondary, y = Gross_Tertiary_Education_Enrollment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "高中教育完成率与高等教育入学率",
```

```
6         x = "高中教育完成率",
7         y = "高等教育入学率")
```



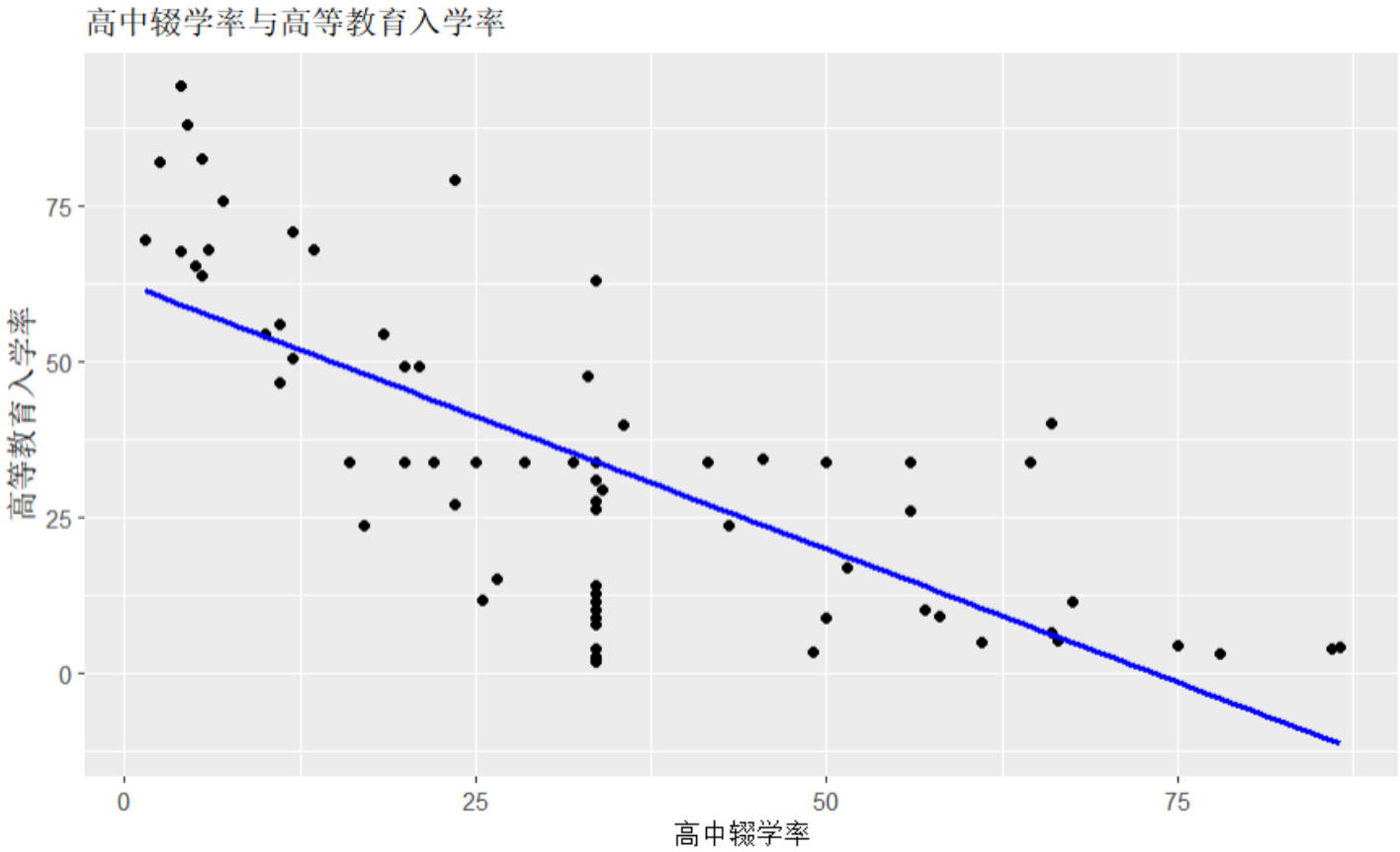
一排竖点的出现是因为在数据预处理时，我们将数据集里的零值替换为了其所在列的平均值。

从图中可以看出教育完成率与高等教育入学率有显著的正相关性，教育完成率越高，高等教育入学率越高。

（5）辍学率对高等教育入学率的影响

这里我们选取距离高等教育最近的高中辍学率作为辍学率的代表。

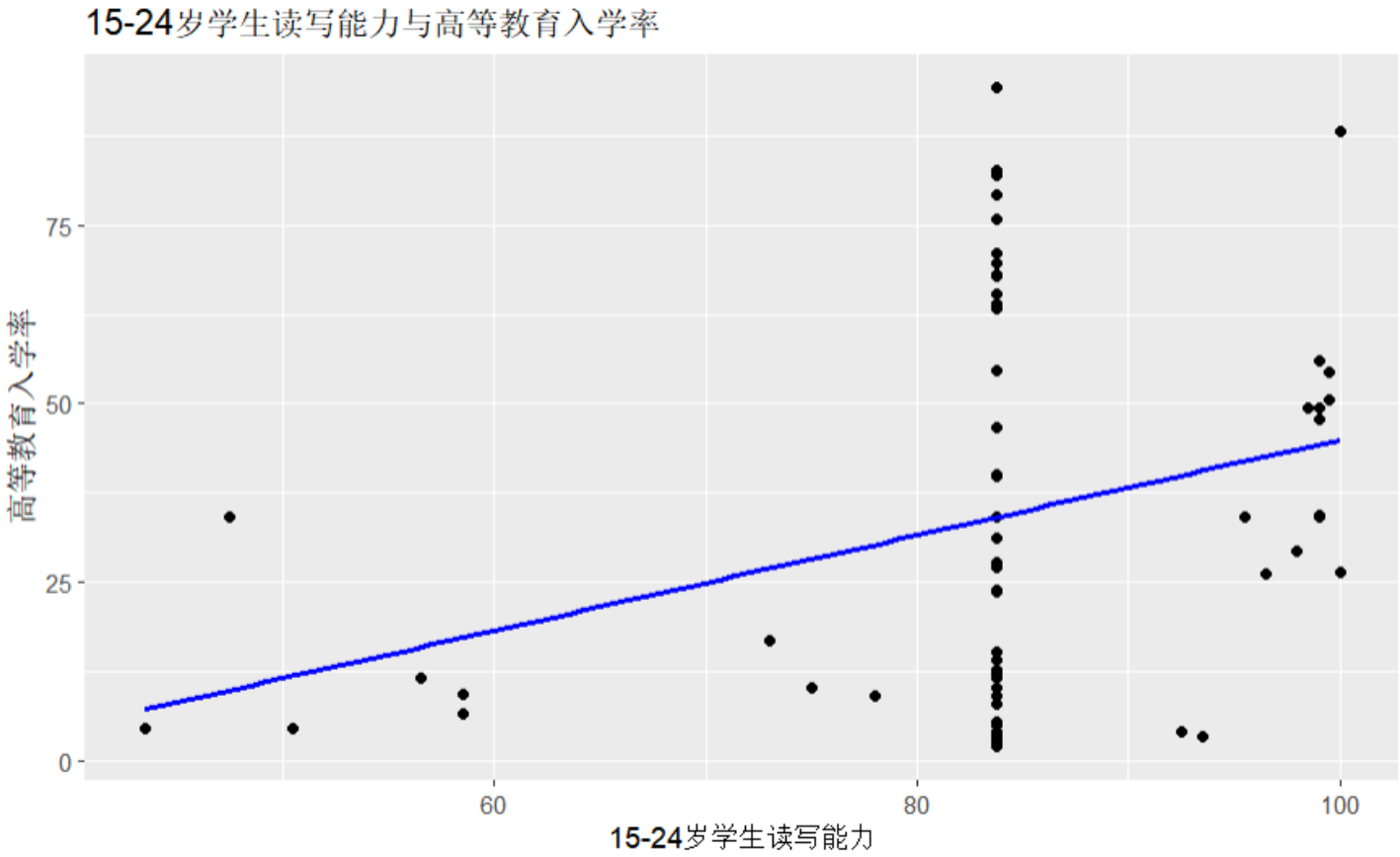
```
# 绘制散点图分析高中辍学率与高等教育入学率的关系
ggplot(edu_data, aes(x = Average_OOSR_Upper_Secondary_Age, y = Gross_Tertiary_Education_Enrollment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "高中辍学率与高等教育入学率",
       x = "高中辍学率",
       y = "高等教育入学率")
```



从图中可以看出辍学率与高等教育入学率有显著的负相关性，辍学率越高，高等教育入学率越低。

（6）15-24岁学生读写能力对高等教育入学率的影响

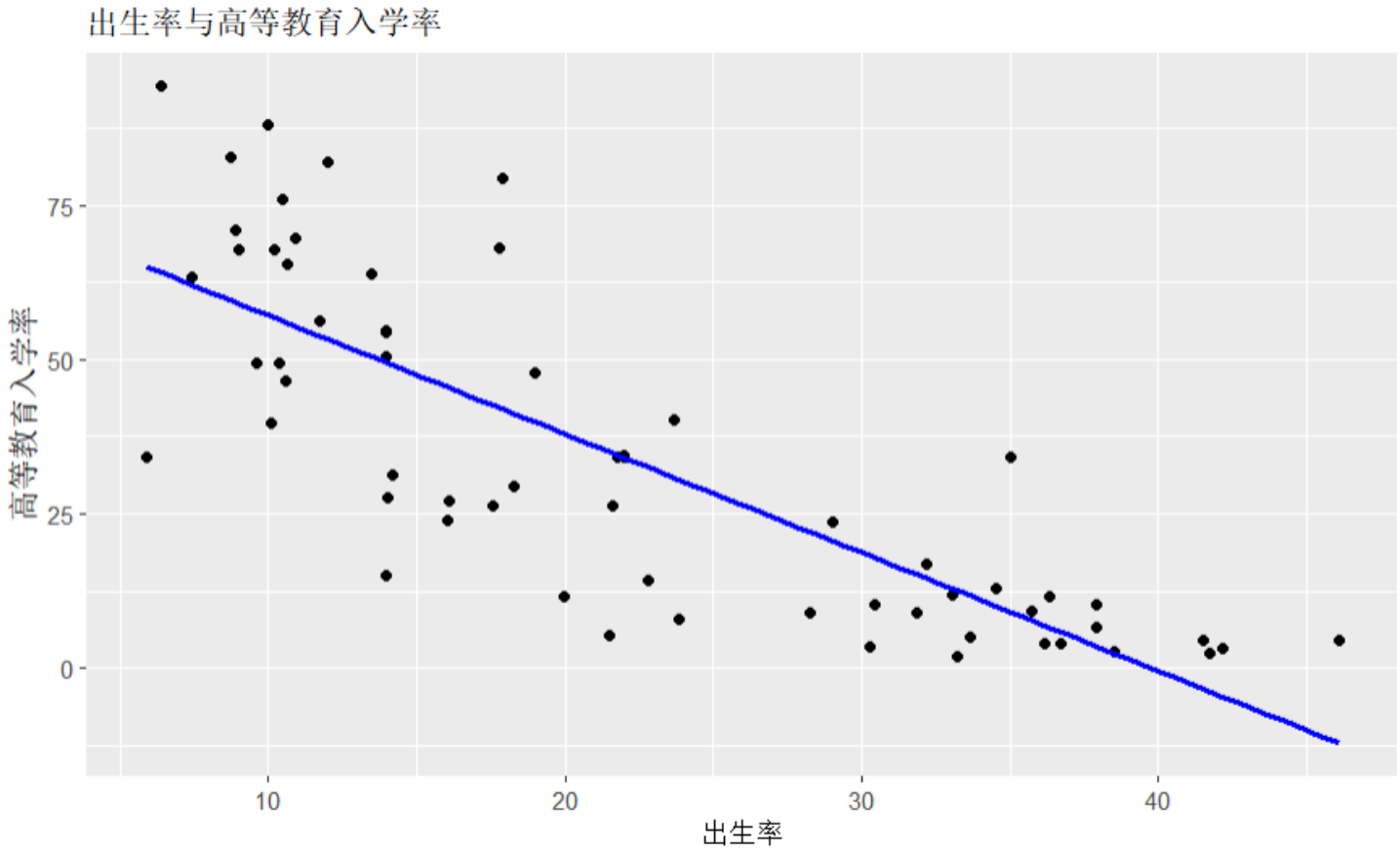
```r
# 绘制散点图分析15-24岁学生读写能力与高等教育入学率的关系
ggplot(edu_data, aes(x = Average_Youth_15_24_Literacy_Rate, y = Gross_Tertiary_Education_Enrollment)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "15-24岁学生读写能力与高等教育入学率",
       x = "15-24岁学生读写能力",
       y = "高等教育入学率")
```



从图中可以看出15-24岁学生读写能力与高等教育入学率有一定的正相关性，15-24岁学生读写能力越高，高等教育入学率越高。

（7）出生率对高等教育入学率的影响

```
ggplot(edu_data, aes(x = Birth_Rate, y = Gross_Tertiary_Education_Enrollment))
+
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "出生率与高等教育入学率",
       x = "出生率",
       y = "高等教育入学率")
```

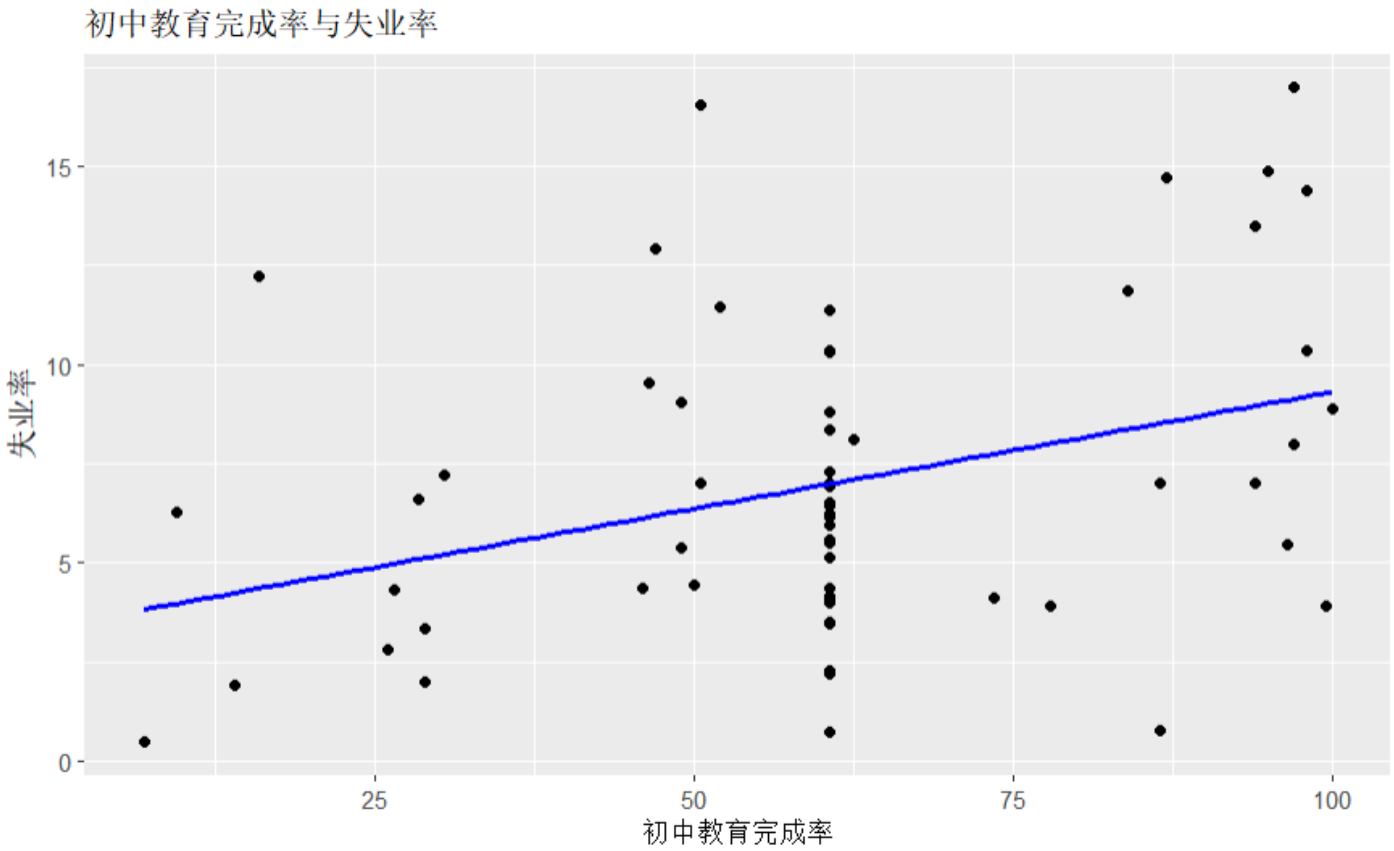


从图中可以看出二者存在显著的负相关关系，出生率越高，高等教育入学率越低。

（8）教育完成率对失业率的影响

这里我们选取初中教育完成率作为教育完成率的代表。

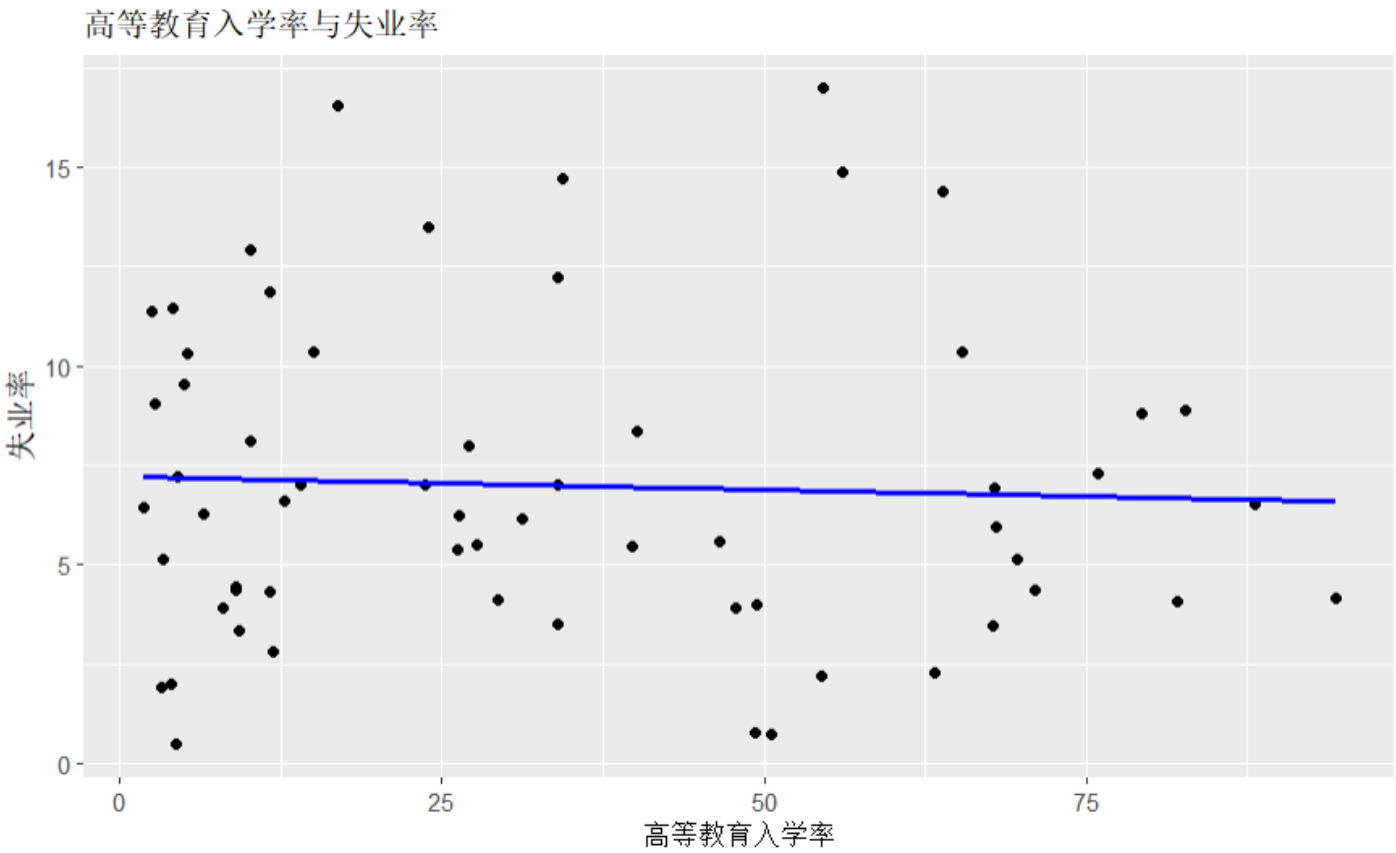
```r
# 绘制散点图分析平均初中教育完成率与失业率的关系
ggplot(edu_data, aes(x = Average_Completion_Rate_Lower_Secondary, y = Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "初中教育完成率与失业率",
       x = "初中教育完成率",
       y = "失业率")
```



可以看出二者存在一定的正相关关系，教育完成率越高，失业率越高。

（9）高等教育入学率对失业率的影响

```
ggplot(edu_data, aes(x = Gross_Tertiary_Education_Enrollment, y =
Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "高等教育入学率与失业率",
       x = "高等教育入学率",
       y = "失业率")
```

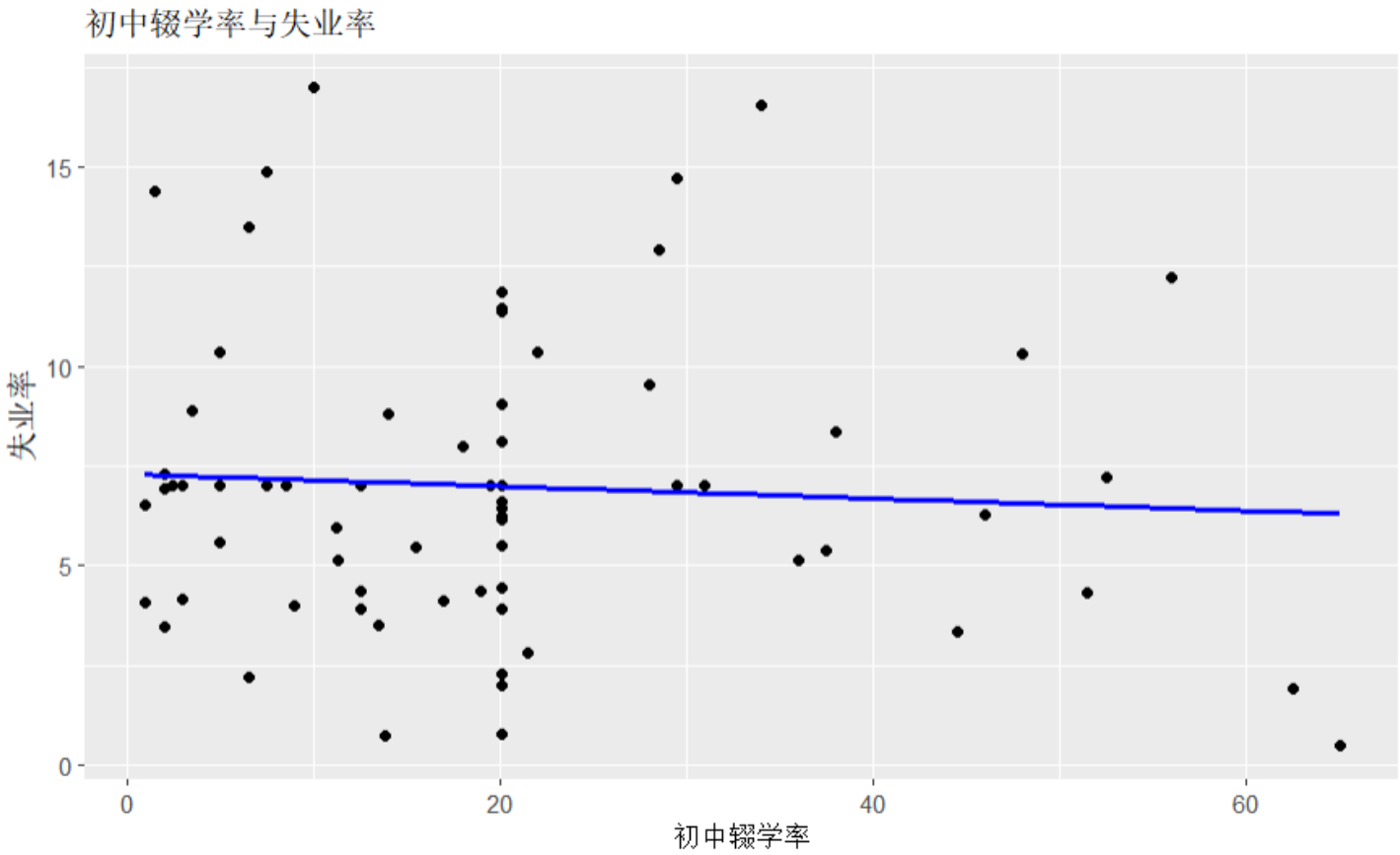


从图中可以看出，二者存在微弱的负相关关系，高等教育入学率越高，失业率越低。

（10）辍学率对失业率的影响

这里我们选取初中辍学率作为辍学率的代表。

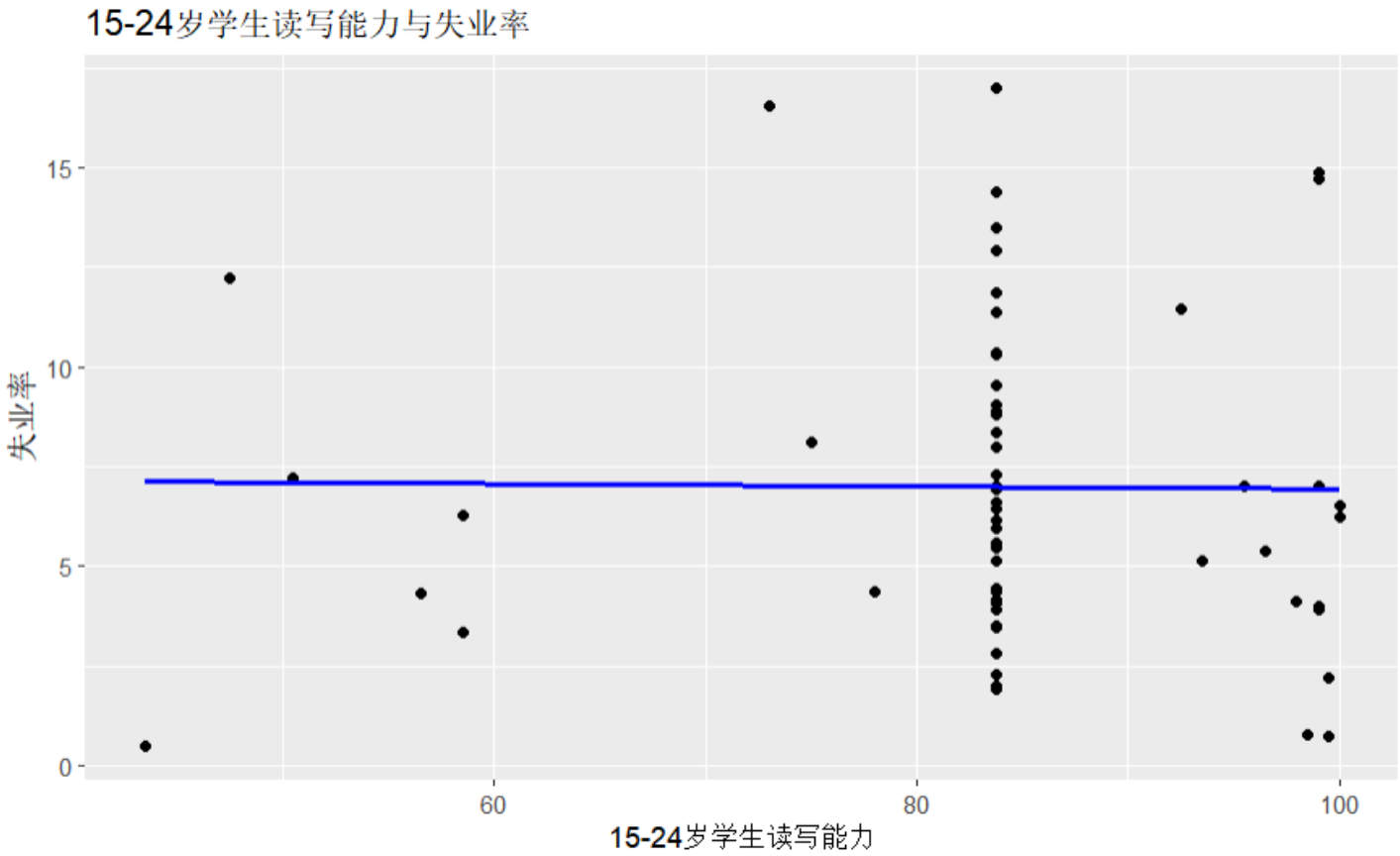
```
# 绘制散点图分析初中辍学率与失业率的关系
ggplot(edu_data, aes(x = Average_OOSR_Lower_Secondary_Age, y =
Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "初中辍学率与失业率",
       x = "初中辍学率",
       y = "失业率")
```



从图中可以看出，二者存在微弱的负相关关系，辍学率越高，失业率越低。

（11）15-24岁学生读写能力对失业率的影响

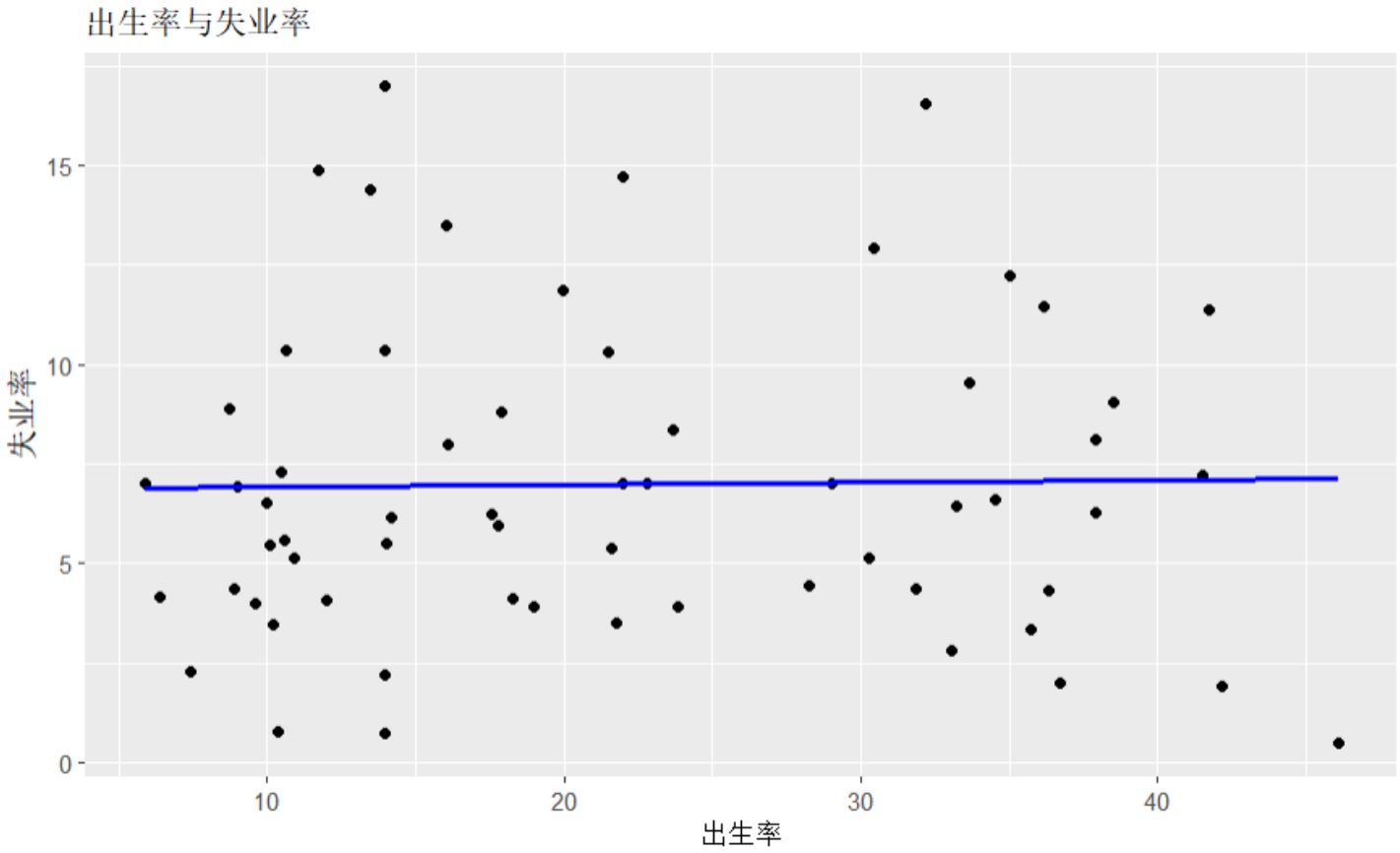
```
ggplot(edu_data, aes(x = Average_Youth_15_24_Literacy_Rate, y =
Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "15-24岁学生读写能力与失业率",
       x = "15-24岁学生读写能力",
       y = "失业率")
```



从图中可以看出，线性回归线几乎为一条直线，15-24岁学生读写能力与失业率二者没有显著的线性相关性。

（12）出生率对失业率的影响

```
ggplot(edu_data, aes(x = Birth_Rate, y = Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "出生率与失业率",
       x = "出生率",
       y = "失业率")
```

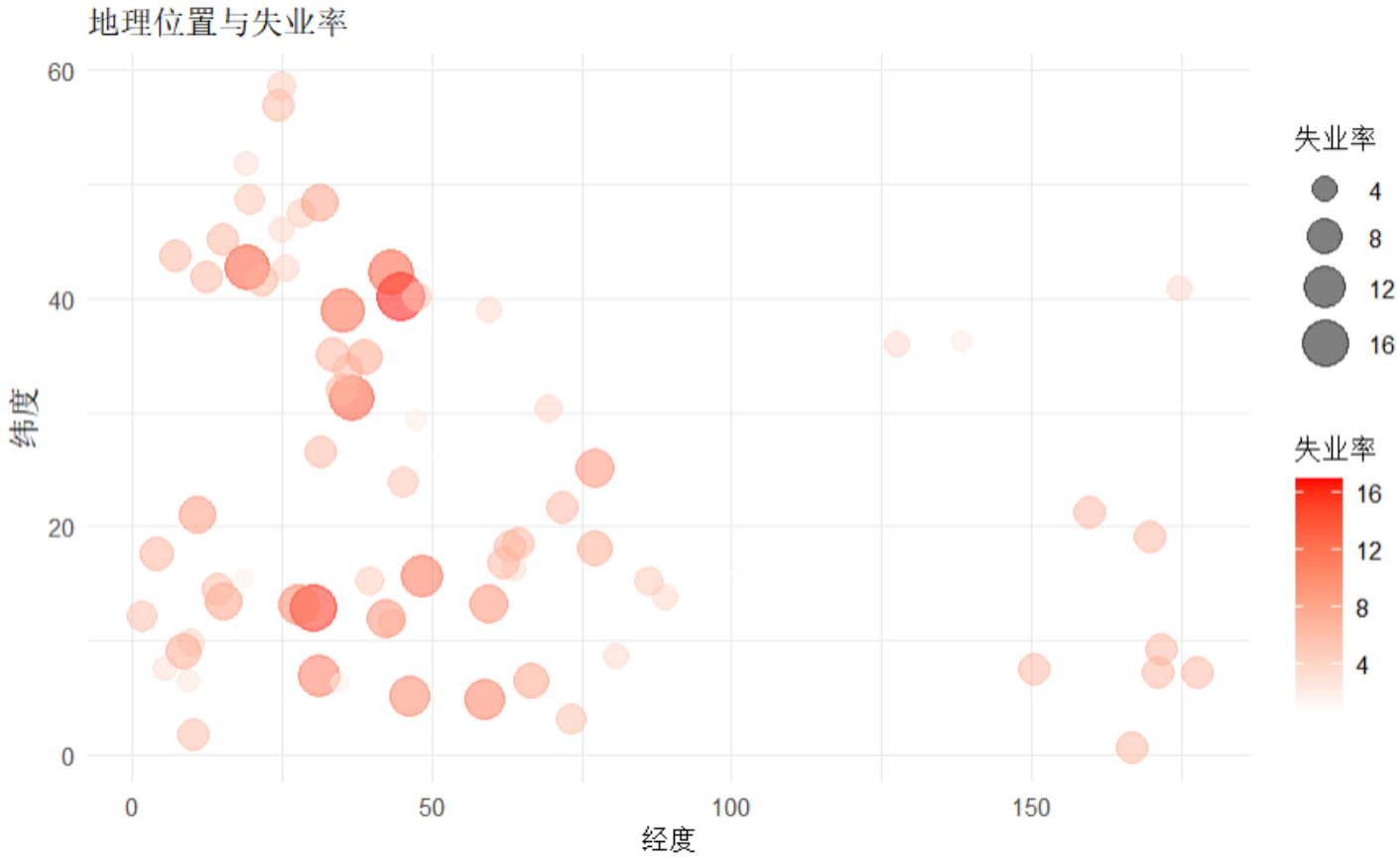


从图中可以看出，线性回归线几乎为一条直线，出生率与失业率二者没有显著的线性相关性。

（13）地理位置对失业率的影响

使用散点图来展示地理位置与失业率的关系。

```
ggplot(edu_data, aes(x = Longitude, y = Latitude)) +
  geom_point(aes(size = Unemployment_Rate, color = Unemployment_Rate), alpha = 0.5) +
  scale_size_continuous(range = c(1, 8)) +  # 调整大小
  scale_color_gradient(low = "white", high = "red") +  # 调整颜色
  labs(title = "地理位置与失业率",
       x = "经度",
       y = "纬度",
       size = "失业率",
       color = "失业率") +
  theme_minimal()
```

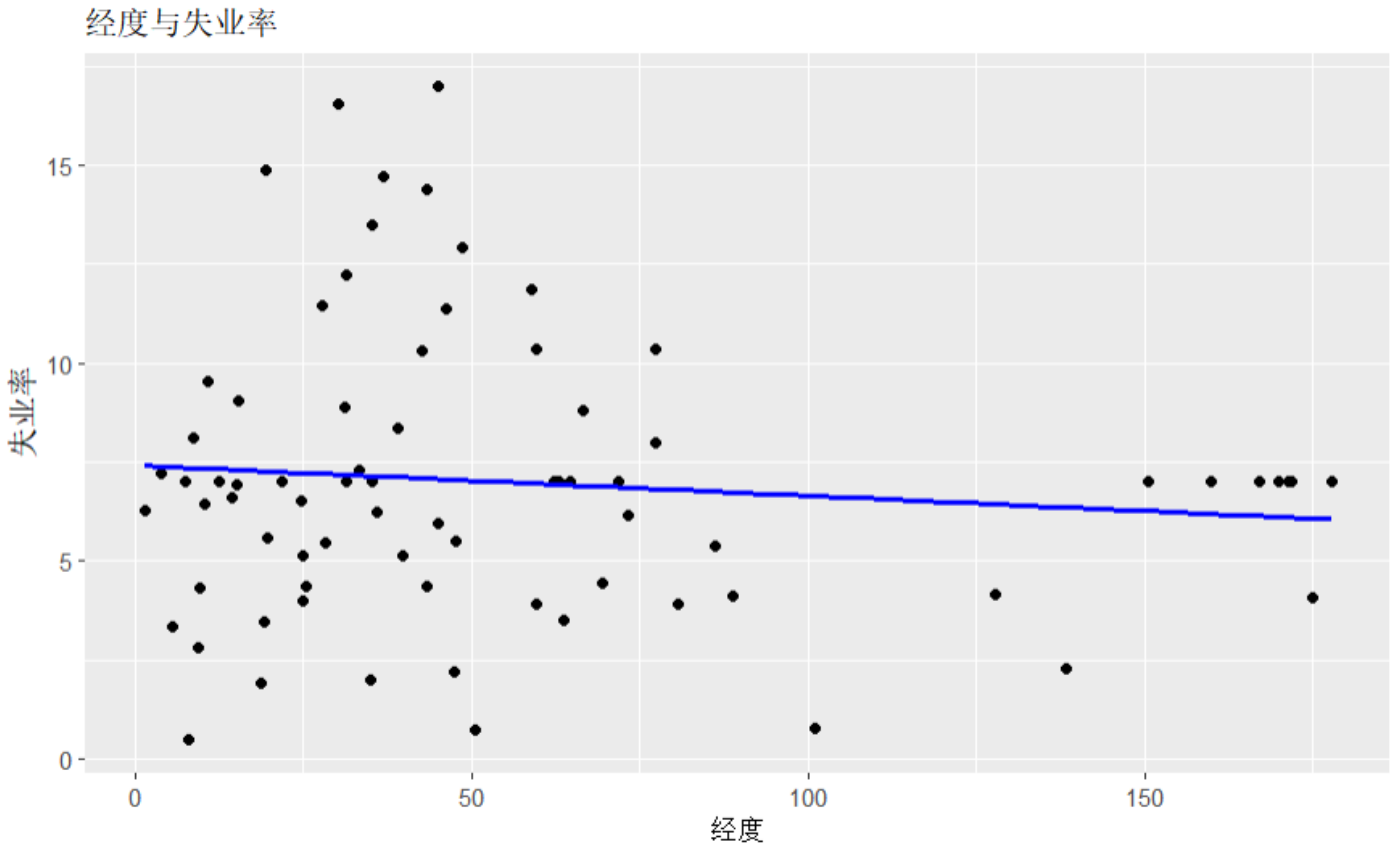


散点的大小与颜色表示失业率高低，散点越大，颜色越深，代表失业率越高；散点越小，颜色越浅，代表失业率越低。

从图中可以看出，经度与纬度较低的国家或地区，失业率较高，失业率与地理位置有一定关系。下面单独对经度和纬度进行分析。

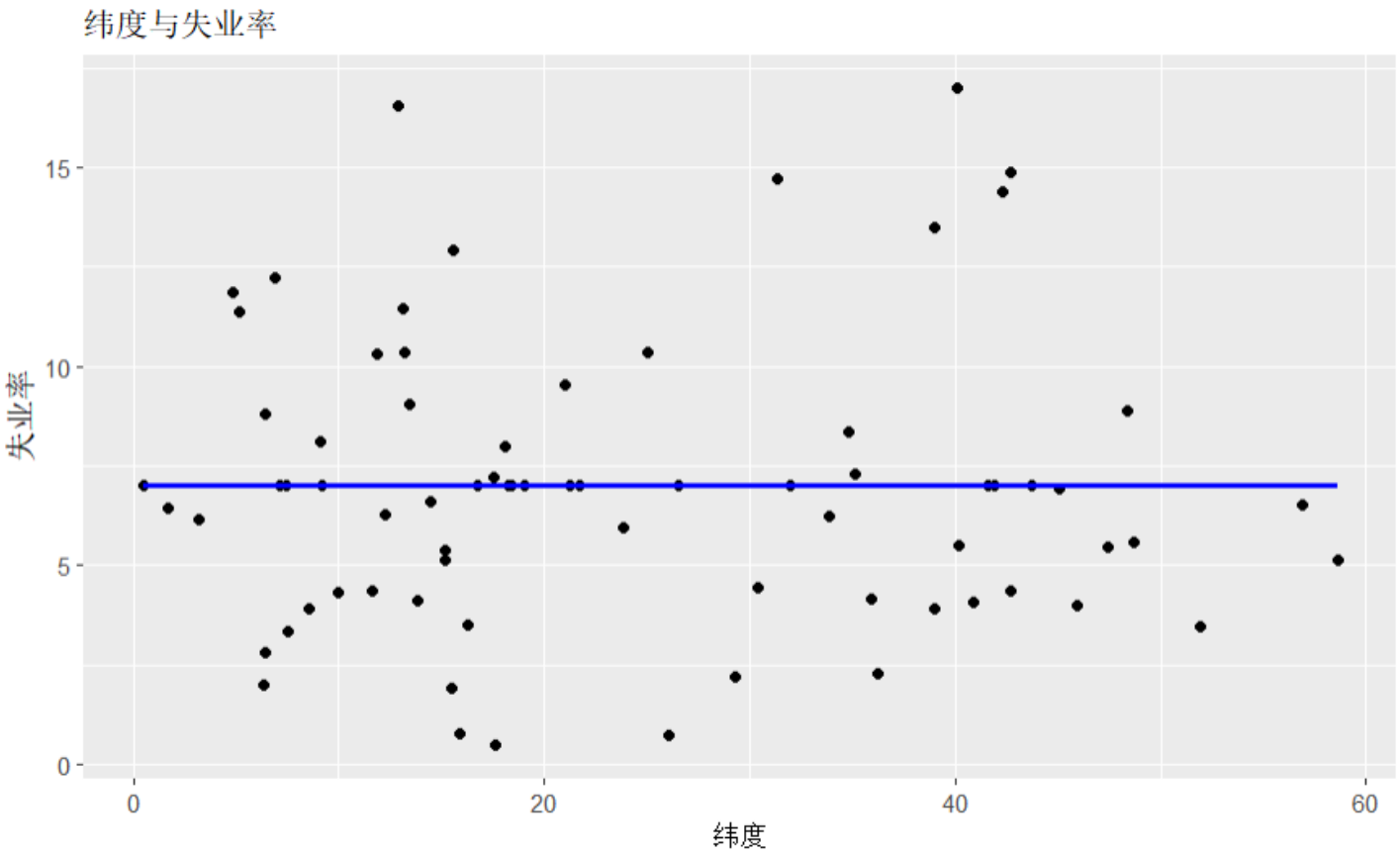
首先是经度：

```
ggplot(edu_data, aes(x = Longitude, y = Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "经度与失业率",
       x = "经度",
       y = "失业率")
```



随后是纬度：

```
ggplot(edu_data, aes(x = Latitude, y = Unemployment_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "纬度与失业率",
       x = "纬度",
       y = "失业率")
```



可以看出，经度对失业率的影响较大，二者存在负相关关系；纬度对失业率几乎没有影响。

三、数据建模及模型质量评估

A. 高等教育总入学人数预测模型

建模

本次建模使用线性回归模型。

```
# 划分训练数据集以及测试数据集
trainIndex1 <- createDataPartition(edu_data$Gross_Tertiary_Education_Enrollment, p = 0.8, list = FALSE)
train_data1 <- edu_data[trainIndex1, ]
test_data1 <- edu_data[-trainIndex1, ]

# 初步建立线性回归模型
model1 <- lm(Gross_Tertiary_Education_Enrollment ~ Latitude + Longitude +
               Average_OOSR_Pre0Primary_Age + Average_OOSR_Primary_Age +
               Average_OOSR_Lower_Secondary_Age + Average_OOSR_Upper_Secondary_Age +
               Gross_Primary_Education_Enrollment + Average_Completion_Rate_Primary +
```

```
              Average_Completion_Rate_Lower_Secondary +
Average_Completion_Rate_Upper_Secondary +
              Average_Youth_15_24_Literacy_Rate + Birth_Rate,
            data = train_data1)

summary(model1)

# model1结果如下
# Coefficients:
#                                     Estimate Std. Error t value
Pr(>|t|)
# (Intercept)                       115.753778  32.495592   3.562
0.000843 ***
# Latitude                            0.186273   0.162077   1.149
0.256132
# Longitude                           0.064225   0.042104   1.525
0.133730
# Average_OOSR_Pre0Primary_Age        0.028706   0.103350   0.278
0.782397
# Average_OOSR_Primary_Age            0.113118   0.237934   0.475
0.636646
# Average_OOSR_Lower_Secondary_Age    0.046103   0.267746   0.172
0.864013
# Average_OOSR_Upper_Secondary_Age   -0.556485   0.196324  -2.835
0.006696 **
# Gross_Primary_Education_Enrollment -0.003769   0.178379  -0.021
0.983231
# Average_Completion_Rate_Primary     0.347720   0.358871   0.969
0.337439
# Average_Completion_Rate_Lower_Secondary  -0.984178   0.350212  -2.810
0.007144 **
# Average_Completion_Rate_Upper_Secondary   0.491966   0.237013   2.076
0.043303 *
# Average_Youth_15_24_Literacy_Rate  -0.313754   0.199061  -1.576
0.121555
# Birth_Rate                         -1.575008   0.340343  -4.628 2.83e-
05 ***
```

查看model1的结果如上，发现一些变量与高等教育总入学率的相关性不高，将其剔除后，再次进行建模。

```
#筛选变量、优化模型
model2 <- lm(Gross_Tertiary_Education_Enrollment ~ Latitude + Longitude +
              Average_OOSR_Upper_Secondary_Age +
Average_Completion_Rate_Lower_Secondary +
```

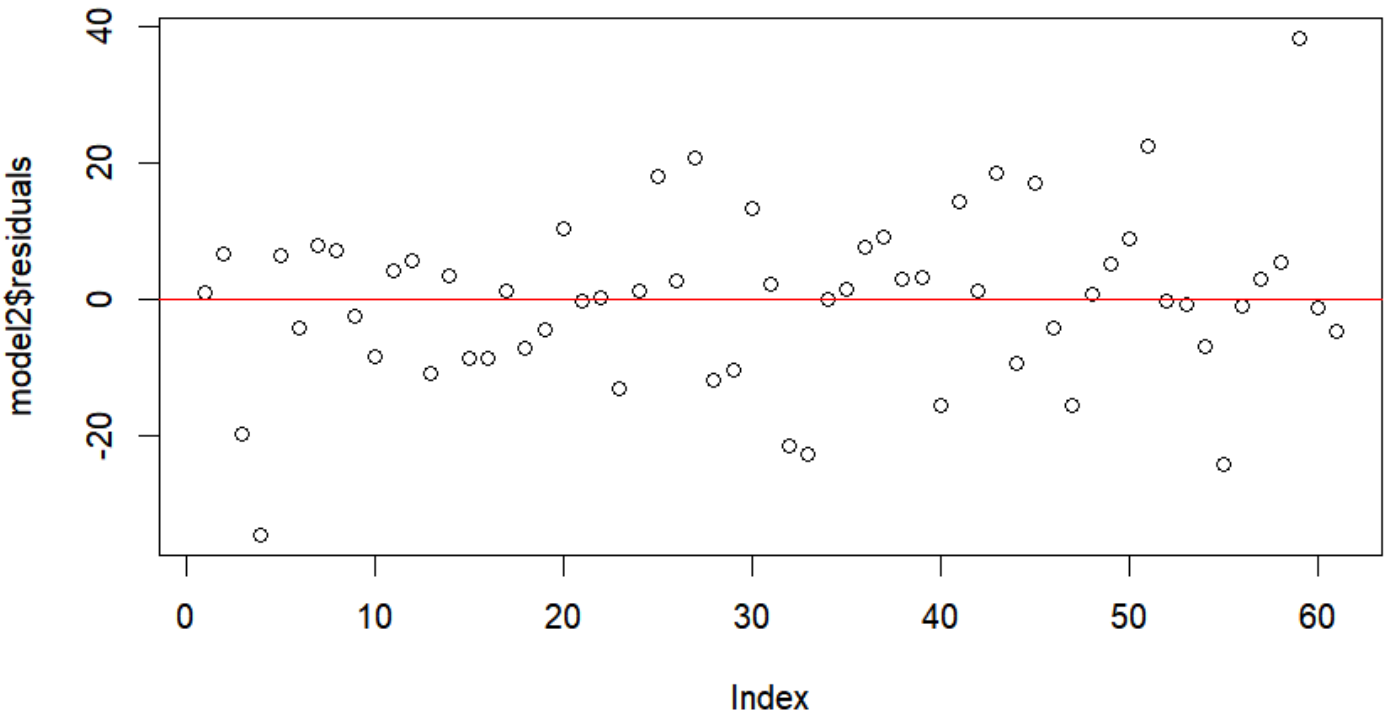
```
              Average_Completion_Rate_Upper_Secondary + Birth_Rate +
              Average_Youth_15_24_Literacy_Rate,
            data = train_data)

summary(model2)

# 优化后的model2结果如下
# Coefficients:
#                                      Estimate Std. Error t value
Pr(>|t|)
# (Intercept)                        122.42903   18.34652   6.673 1.51e-
08 ***
# Latitude                             0.16933    0.15018   1.128
0.264594
# Longitude                            0.05747    0.03868   1.486
0.143239
# Average_OOSR_Upper_Secondary_Age    -0.43669    0.10779  -4.051
0.000167 ***
# Average_Completion_Rate_Lower_Secondary  -0.72516    0.22470  -3.227
0.002146 **
# Average_Completion_Rate_Upper_Secondary   0.48810    0.22621   2.158
0.035500 *
# Birth_Rate                         -1.61857    0.27668  -5.850 3.13e-
07 ***
# Average_Youth_15_24_Literacy_Rate   -0.26139    0.17220  -1.518
0.134962
```

此时的模型与第一个模型相比，解释变量的显著性明显增强。

模型质量评估

建模完成后，使用测试数据集进行预测，绘制残差图并计算均方根误差。

```
# 使用测试集进行预测
predictions1 <- predict(model2, test_data1)

# 绘制残差图
plot(model2$residuals)
abline(h = 0, col = "red")

# 计算均方根误差
test_error1 <- sqrt(mean((test_data1$Gross_Tertiary_Education_Enrollment -
predictions1)^2))
print(test_error1)
# 11.5801
```



从残差图中我们可以看到，大部分点随机分布在0周围，没有明显的趋势，且计算得出的均方根误差为11.5801，R平方为0.7664。考虑到高等教育入学率的取值范围，该模型的预测效果较好。

B. 失业率预测模型

建模

本次建模使用线性回归模型。

```r
# 划分训练数据集以及测试数据集
trainIndex2 <- createDataPartition(edu_data$Unemployment_Rate, p = 0.8, list = FALSE)
train_data2 <- edu_data[trainIndex2, ]
test_data2 <- edu_data[-trainIndex2, ]

# 初步建立线性回归模型
model3 <- lm(Unemployment_Rate ~ Longitude + Average_OOSR_Pre0Primary_Age +
               Average_OOSR_Primary_Age + Average_OOSR_Lower_Secondary_Age +
               Average_OOSR_Upper_Secondary_Age +
Average_Completion_Rate_Primary +
               Average_Completion_Rate_Lower_Secondary +
               Average_Completion_Rate_Upper_Secondary,
             data = train_data2)

summary(model3)

# model3结果如下
```

```
# Coefficients:
#                                      Estimate Std. Error t value
Pr(>|t|)
# (Intercept)                          3.086406   3.109818   0.992
0.32556
# Longitude                           -0.003682   0.009522  -0.387
0.70055
# Average_OOSR_Pre0Primary_Age         0.066356   0.026334   2.520
0.01485 *
# Average_OOSR_Primary_Age             0.024387   0.044836   0.544
0.58883
# Average_OOSR_Lower_Secondary_Age    -0.101207   0.072095  -1.404
0.16632
# Average_OOSR_Upper_Secondary_Age     0.051847   0.050392   1.029
0.30830
# Average_Completion_Rate_Primary     -0.188000   0.092482  -2.033
0.04719 *
# Average_Completion_Rate_Lower_Secondary  0.359427   0.109318   3.288
0.00181 **
# Average_Completion_Rate_Upper_Secondary -0.139600   0.066791  -2.090
0.04152 *
```

查看model3的结果如上，发现一些变量与失业率的相关性不高，将其剔除后，再次进行建模。

```
#筛选变量、优化模型
model4 <- lm(Unemployment_Rate ~ Average_OOSR_Pre0Primary_Age +
               Average_OOSR_Upper_Secondary_Age +
Average_Completion_Rate_Primary +
               Average_Completion_Rate_Lower_Secondary +
Average_OOSR_Lower_Secondary_Age +
               Average_Completion_Rate_Upper_Secondary,
             data = train_data2)

summary(model4)

# 优化后的model4结果如下
# Coefficients:
#                                     Estimate Std. Error t value Pr(>|t|)

# (Intercept)                          3.11449    2.96834   1.049 0.298659

# Average_OOSR_Pre0Primary_Age         0.07489    0.02404   3.116 0.002913
**
# Average_OOSR_Lower_Secondary_Age    -0.03264    0.04334  -0.753 0.454690
```

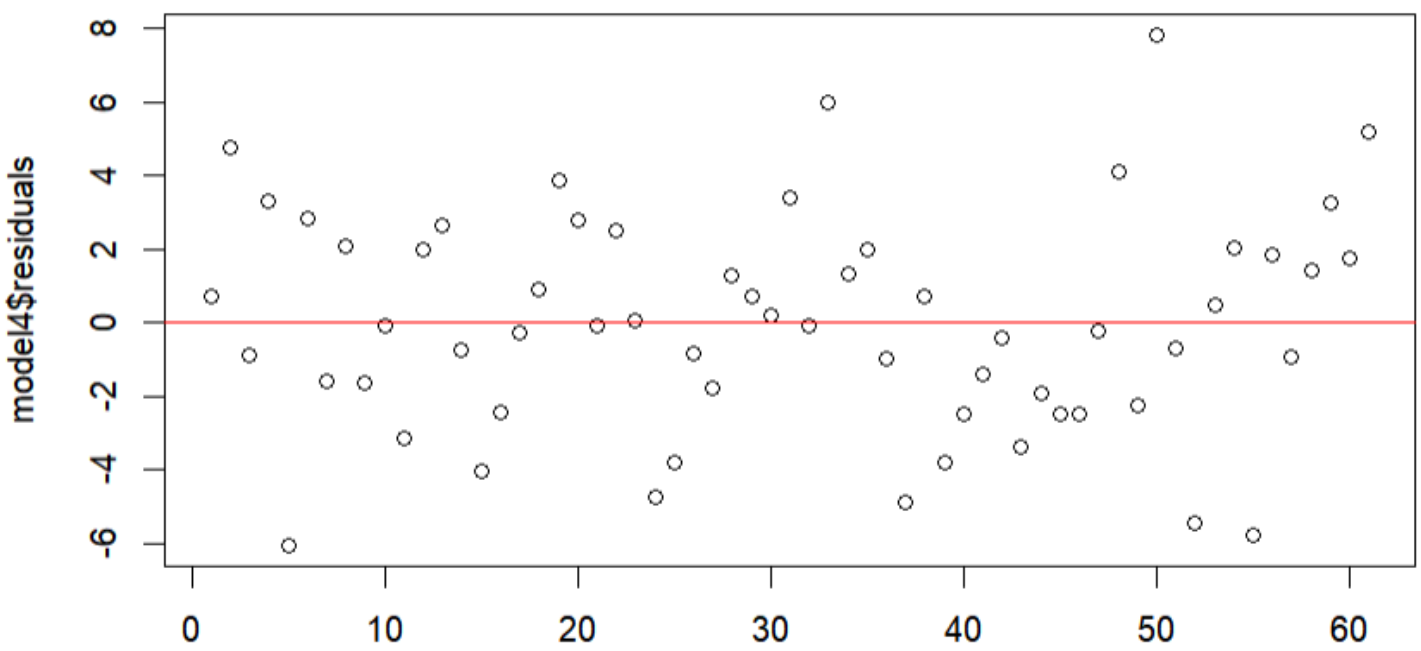
```
# Average_Completion_Rate_Primary          -0.18316    0.08499  -2.155 0.035549
*
# Average_Completion_Rate_Lower_Secondary  0.36914    0.10406   3.547 0.000805
***
# Average_Completion_Rate_Upper_Secondary -0.15690    0.06447  -2.434 0.018224
*
```

此时的模型与第一个模型相比，解释变量的显著性明显增强。

模型质量评估

建模完成后，使用测试数据集进行预测，绘制残差图并计算均方根误差。

```r
# 使用测试集进行预测
predictions2 <- predict(model4, test_data2)

# 绘制残差图
plot(model4$residuals)
abline(h = 0, col = "red")

# 计算均方根误差
test_error2 <- sqrt(mean((test_data2$Unemployment_Rate - predictions2)^2))
print(test_error2)
# 2.9881
```



从残差图中可以看到，点偏离较严重。计算得出的均方根误差为2.9881，R平方为0.3579。考虑到失业率的取值范围，该模型的预测效果较差。出现这种现象原因可能是这里仅仅使用了教育因素对失业率

进行建模，而实际的失业率不仅受教育因素影响，还会受到经济、社会等因素的影响，因此只使用教育因素进行建模的模型的解释力较差。

四、启示和建议

- 不同国家和地区之间的教育水平差距仍较大，许多国家教育水平依旧处于较低状态，需要增加教育投入。
- 尽管男女性受教育率在整体上没有显著差异，一些国家男女性之间的受教育率仍存在相当大的差异，教育公平仍有欠缺。
- 二三年级与小学学生的阅读和数学能力差别较大，大部分学生只专注于一项能力，另一项能力水平较低，直到初中二者的发展才逐渐平衡。出现这种现象的地区需要更加注重综合能力训练。
- 基础教育直接影响到高等教育以及失业率。需要加大投入，确保基础教育入学率以及完成率，尽量减少辍学人数。