

Homework 3

张子乐,3230104237

第一题

数据预处理

对数据的预处理与作业二一致，将字符串格式转为数值格式。观察到 smoke 列和 drunk 列有多种不同的回答，统一改为“是”“否”。使用多重插补的方法，对于连续型变量使用 Predictive Mean Matching(pmm)，对二分类变量使用 Logistic Regression Imputation(logreg)，而对于性别变量则直接删除含缺失值的行。后续需要用到 BMI，因此计算 BMI 作为新的变量，观察到 BMI 出现了异常值 581，将其删除。

```
clean <- function(df){  
  df <- df[ -1, ] # 删除中文描述的行  
  #print(str(df)) # 查看每列的类型  
  # 要转换成数值格式的列  
  numeric_cols <- c("age", "sbp", "dbp", "weight", "height", "FPG", "TG", "HDL-C")  
  df <- df |>  
    mutate(  
      across(all_of(numeric_cols), parse_number)  
    )  
  
  # 观察到 smoke 列有多种回答  
  df <- df %>%  
    mutate(  
      smoke = case_when(  
        smoke == "是" ~ "是",  
        smoke %in% c("否", "已戒烟", "戒烟 3 年", "戒烟 2 个月") ~ "否",  
        TRUE ~ NA_character_  
      ) %>%  
      factor(levels = c("否", "是"))  
    )  
  # 同样，drunk 列有“是”“否”“无”三种回答  
  df <- df %>%  
    mutate(
```

```

drunk = case_when(
  drunk == "是" ~ "是",
  drunk %in% c("否", "无") ~ "否",
  TRUE ~ NA_character_
) %>%
  factor(levels = c("否", "是"))

df <- df %>% filter(!is.na(gender))

meth <- rep("pmm", ncol(df))
names(meth) <- names(df) # 把列名赋给 meth
meth["smoke"] <- "logreg"
meth["drunk"] <- "logreg"
df <- suppressWarnings(mice(df, m = 10, maxit = 50, method =meth, printFlag = FALSE, seed=42)) #
df <- complete(df, 1) # 选取第一个
return(df)
}

df2 <- read_excel("sample.xls", sheet = "Sheet2")
df2 <- clean(df2)
df = df2
df <- df %>%
  mutate(BMI = round(weight / (height / 100)^2, 2)) # BMI = 体重 / (身高的平方)
df <- df[df$BMI <= 50, ]

```

正态性检验

```

X <- c("BMI", "FPG", "sbp", "dbp", "TG", "HDL-C")
tests <- lapply(df[X], shapiro.test)
print(tests)

```

```

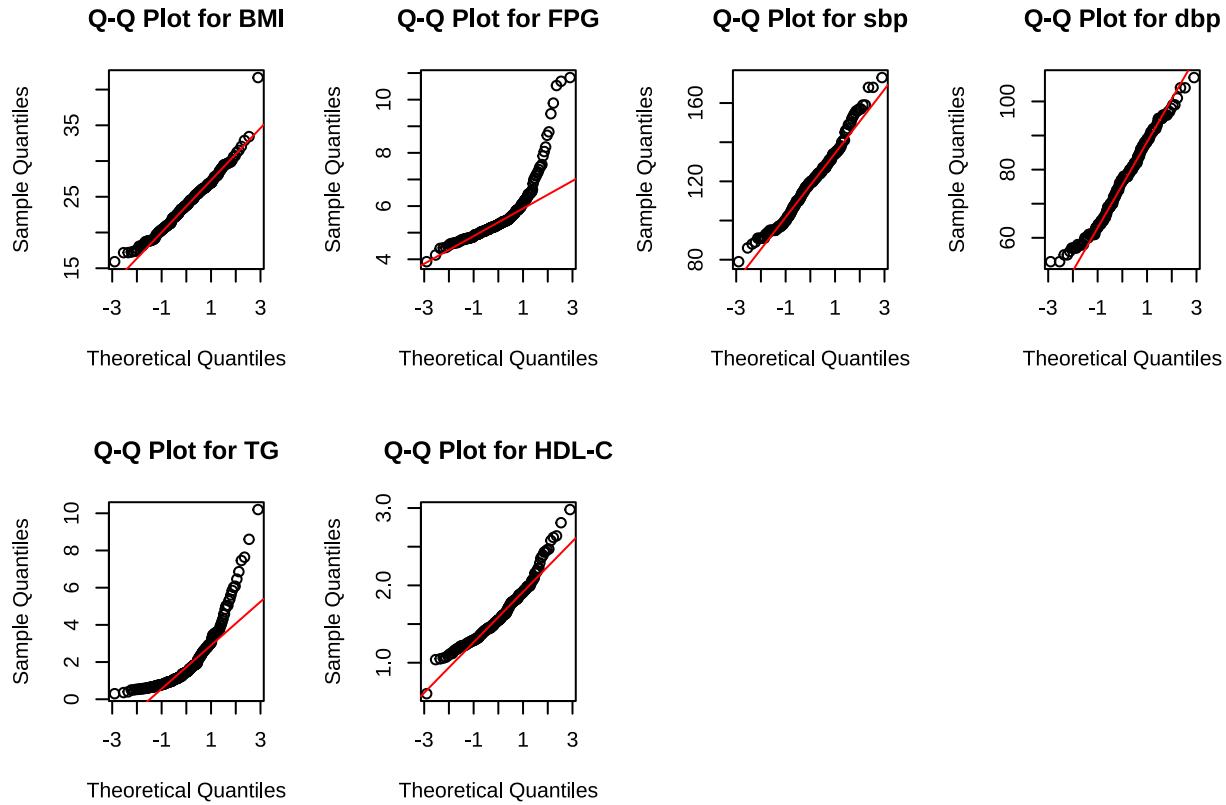
## $BMI
##
## Shapiro-Wilk normality test
##
## data: X[[i]]
## W = 0.97625, p-value = 0.000185
##
```

```
##  
## $FPG  
##  
## Shapiro-Wilk normality test  
##  
## data: X[[i]]  
## W = 0.74028, p-value < 2.2e-16  
##  
##  
## $sbp  
##  
## Shapiro-Wilk normality test  
##  
## data: X[[i]]  
## W = 0.98188, p-value = 0.001705  
##  
##  
## $dbp  
##  
## Shapiro-Wilk normality test  
##  
## data: X[[i]]  
## W = 0.98376, p-value = 0.003756  
##  
##  
## $TG  
##  
## Shapiro-Wilk normality test  
##  
## data: X[[i]]  
## W = 0.79482, p-value < 2.2e-16  
##  
##  
## $`HDL-C`  
##  
## Shapiro-Wilk normality test  
##  
## data: X[[i]]  
## W = 0.94971, p-value = 5.411e-08
```

```

par(mfrow = c(2, 4))
for (var in X) {
  qqnorm(df[[var]], main = paste("Q-Q Plot for", var))
  qqline(df[[var]], col = "red")
}
par(mfrow = c(1, 1))

```



使用 Shapiro-Wilk 正态性检验，在 95% 的置信水平下， p 值小于 0.05 则拒绝原假设，即不符合正态性。

从结果可以看出，BMI、FPG、SBP、DBP、TG 和 HDL-C 都不具有正态性。画出的 QQ 图也可以验证这一点。

相关性检验

```

cor_matrix <- cor(df[X], use = "complete.obs")
p_matrix <- cor.mtest(df[X], use = "complete.obs")
print(round(cor_matrix, 2))

```

```

##          BMI      FPG      sbp      dbp       TG   HDL-C

```

```

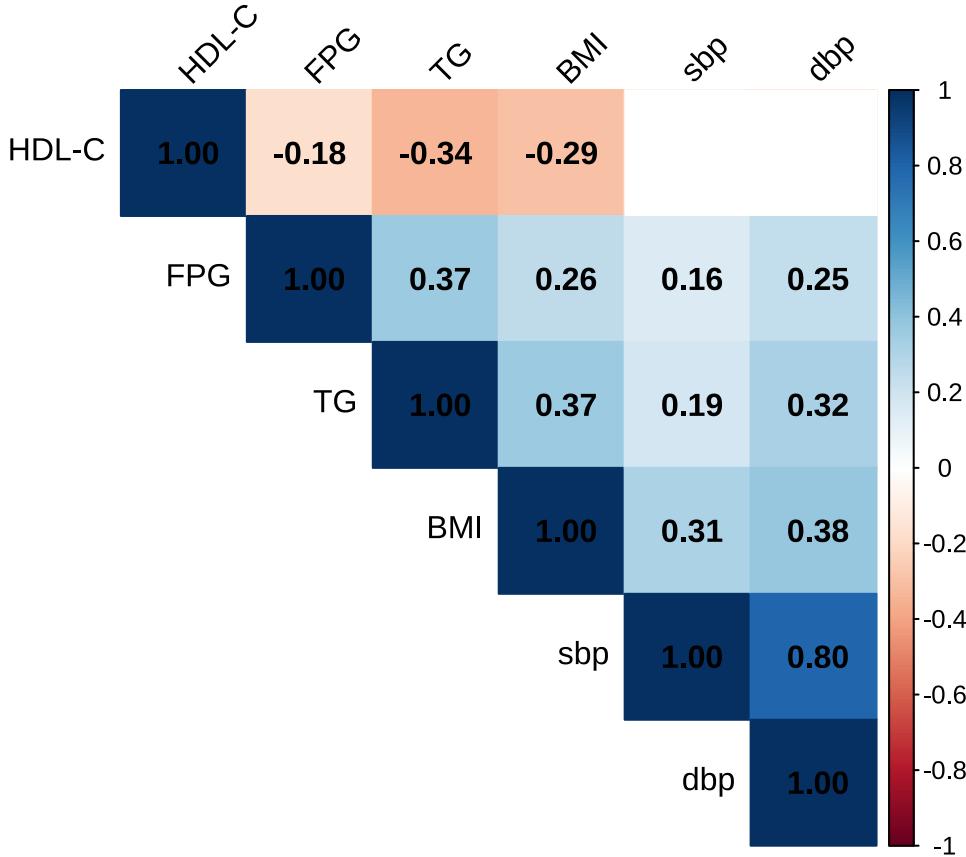
## BMI    1.00  0.26  0.31  0.38  0.37 -0.29
## FPG    0.26  1.00  0.16  0.25  0.37 -0.18
## sbp    0.31  0.16  1.00  0.80  0.19 -0.08
## dbp    0.38  0.25  0.80  1.00  0.32 -0.11
## TG     0.37  0.37  0.19  0.32  1.00 -0.34
## HDL-C -0.29 -0.18 -0.08 -0.11 -0.34  1.00

print(round(p_matrix$p,3))

##      BMI   FPG   sbp   dbp    TG HDL-C
## BMI      0 0.000 0.000 0.000 0.000 0.000
## FPG      0 0.000 0.009 0.000 0.000 0.004
## sbp      0 0.009 0.000 0.000 0.002 0.190
## dbp      0 0.000 0.000 0.000 0.000 0.068
## TG       0 0.000 0.002 0.000 0.000 0.000
## HDL-C    0 0.004 0.190 0.068 0.000 0.000

corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
         addCoef.col = "black",
         tl.col = "black", tl.srt = 45,
         sig.level = 0.05, insig = "blank", p.mat = p_matrix$p)

```



如热力图所示，除了 SBP 与 HDL-C、DBP 与 HDL-C 外，其余变量间的相关性均显著。HDL-C 与其它变量均成负相关关系，其余变量间均成正相关关系。其中 DBP 与 SBP 的正相关性最强，相关系数为 0.8；HDL-C 与 TG 的负相关性最强，相关系数为-0.34。

分析患代谢综合症的比例

首先判断每个样本是否超重、高血糖、高血压、空腹血，从而判断是否患代谢综合症。

```

df$Overweight <- ifelse(df$BMI >= 25, 1, 0) # 超重
df$Hyperglycemia <- ifelse(df$FPG >= 6.1, 1, 0) # 高血糖
df$HTN <- ifelse(df$sbp >= 140 | df$dbp >= 90, 1, 0) # 高血压
df$FBG <- ifelse(df$TG >= 1.7 |
                    (df$`HDL-C` < 0.9 & df$gender == "男") |
                    (df$`HDL-C` < 1 & df$gender == "女"), 1, 0) # 空腹血
df$sick <- ifelse(rowSums(cbind(df$Overweight,
                                    df$Hyperglycemia,
                                    df$HTN,
                                    df$FBG)) >= 3, 1, 0) # 是否患代谢综合征, 1 为患, 0 为不患
df <- df |> mutate(
  sick = factor(sick, levels = c(0,1)),
  
```

```
    gender = as.factor(gender)
)
print(table(df$sick))
```

```
##
##    0     1
## 235   34
```

```
prop <- prop.table(table(df$sick))
print(prop)
```

```
##
##          0           1
## 0.8736059 0.1263941
```

269 个样本中有 34 人患有代谢综合征，占 12.64%。

患代谢综合症的性别差异

判断患代谢综合症的比例有没有性别差异,34 名患者中仅 4 名为女性，仅占 11.77%。

```
table(df$gender, df$sick)
```

```
##
##          0     1
## 男 141  30
## 女  94   4
```

```
contingency_table <- table(df$gender, df$sick)
chisq_test <- chisq.test(contingency_table) # 卡方检验
#fisher.test(contingency_table)
print(chisq_test)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: contingency_table
## X-squared = 9.0421, df = 1, p-value = 0.002638
```

原假设 H0: 患代谢综合征的比例没有显著的性别差异。备择假设 H1: 患代谢综合征的比例有显著的性别差异。

通过卡方检验, $p = 0.002638 < 0.05$, 拒绝原假设, 选择备择假设, 即患代谢综合征的比例有显著的性别差异。

不同性别是否患代谢综合症群体各类指标的均值估计和置信区间

```
df_sick <- df %>% filter(sick == 1)

sick_male <- df_sick %>%
  filter(gender == "男") %>%
  dplyr::select(all_of(X))

sick_female <- df_sick %>%
  filter(gender == "女") %>%
  dplyr::select(all_of(X))

df_healthy <- df %>% filter(sick == 0)

healthy_male <- df_healthy %>%
  filter(gender == "男") %>%
  dplyr::select(all_of(X))

healthy_female <- df_healthy %>%
  filter(gender == "女") %>%
  dplyr::select(all_of(X))

test_sick <- HotellingsT2(sick_male, sick_female)
print(test_sick)

## 
##  Hotelling's two sample T2-test
## 
##  data:  sick_male and sick_female
##  T.2 = 1.698, df1 = 6, df2 = 27, p-value = 0.1599
##  alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0)
```

检测的指标为 BMI、FPG、SBP、DBP、TG 和 HDL-C。

H0: 患病组中检测指标的均值向量没有显著的性别差异。

H1: 患病组中检测指标的均值向量有显著的性别差异。

霍特林统计量 $T^2 = 1.698$, $p = 0.1599 > 0.05$, 无法拒绝原假设, 因此患病组中检测指标的均值向量没有显著的性别差异。

```
test_healthy <- HotellingsT2(healthy_male, healthy_female)
print(test_healthy)
```

```
##
##  Hotelling's two sample T2-test
##
## data: healthy_male and healthy_female
## T.2 = 15.187, df1 = 6, df2 = 228, p-value = 1.288e-14
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0)
```

H0: 未患病组中检测指标的均值向量没有显著的性别差异。

H1: 未患病组中检测指标的均值向量有显著的性别差异。

霍特林统计量 $T^2 = 15.187$, $p < 0.05$, 拒绝原假设, 因此未患病组中检测指标的均值向量有显著的性别差异。

下面计算每个群体均值向量的置信区间, 取置信水平为 95%。

```
CI <- function(df, alpha = 0.05){
  n <- nrow(df)      # 样本量
  p <- ncol(df)      # 变量数量
  X_bar <- colMeans(df) # 样本均值向量
  S <- cov(df)        # 样本协方差矩阵
  sds <- apply(df, 2, sd)

  adjusted_alpha <- alpha / p    # Bonferroni 校正
  t_val <- qt(1 - adjusted_alpha / 2, df = n - 1)
  margin <- t_val * sds / sqrt(n)

  lower_bounds <- X_bar - margin
  upper_bounds <- X_bar + margin

  result <- data.frame(
    Variable = names(X_bar),
    Mean = X_bar,
```

```

    Lower_Bound = lower_bounds,
    Upper_Bound = upper_bounds
)
result[, c("Mean", "Lower_Bound", "Upper_Bound")] <-
  round(result[, c("Mean", "Lower_Bound", "Upper_Bound")], 2)

print(result)
}

```

患病男性群体的均值向量和置信区间：

```
CI(sick_male)
```

| | Variable | Mean | Lower_Bound | Upper_Bound |
|----|----------|--------|-------------|-------------|
| ## | BMI | 27.79 | 26.09 | 29.49 |
| ## | FPG | 6.65 | 5.91 | 7.38 |
| ## | sbp | 132.87 | 125.22 | 140.52 |
| ## | dbp | 88.50 | 83.64 | 93.36 |
| ## | TG | 3.51 | 2.40 | 4.62 |
| ## | HDL-C | 1.45 | 1.27 | 1.63 |

未患病男性群体的均值向量和置信区间：

```
CI(healthy_male)
```

| | Variable | Mean | Lower_Bound | Upper_Bound |
|----|----------|--------|-------------|-------------|
| ## | BMI | 24.33 | 23.62 | 25.04 |
| ## | FPG | 5.51 | 5.33 | 5.69 |
| ## | sbp | 120.16 | 116.95 | 123.37 |
| ## | dbp | 77.07 | 74.86 | 79.29 |
| ## | TG | 2.06 | 1.75 | 2.38 |
| ## | HDL-C | 1.52 | 1.46 | 1.59 |

未患病女性群体的均值向量和置信区间：

```
CI(healthy_female)
```

| | Variable | Mean | Lower_Bound | Upper_Bound |
|----|----------|-------|-------------|-------------|
| ## | BMI | 21.96 | 21.18 | 22.74 |
| ## | FPG | 5.18 | 5.00 | 5.35 |

```

## sbp      sbp 112.06      107.29      116.84
## dbp      dbp  70.59      67.52       73.65
## TG       TG   1.23       1.02       1.44
## HDL-C    HDL-C  1.79       1.69       1.89

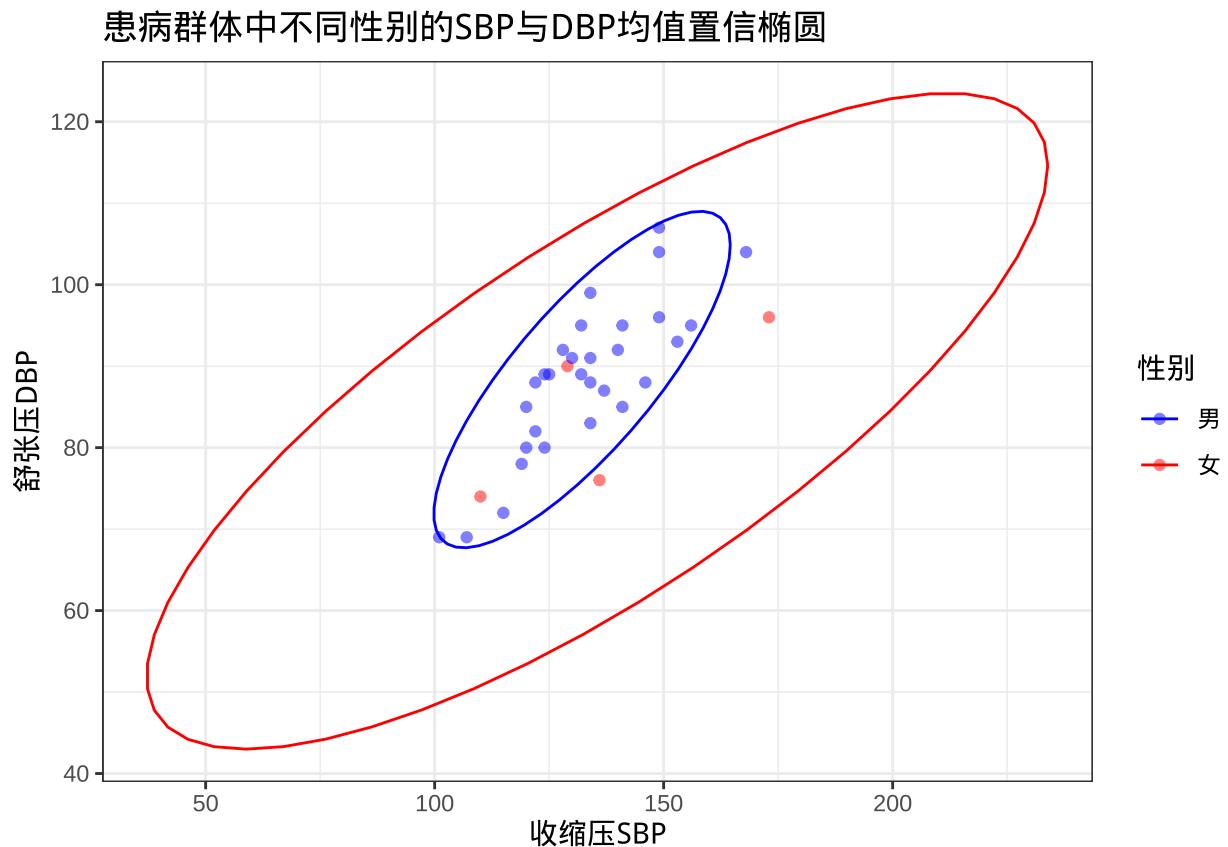
```

我们可以画出置信椭球可视化上面的结果。这里选取收缩压 SBP 和舒张压 DBP 这两个变量。

```

ggplot(df_sick, aes(x = sbp, y = dbp, color = gender)) +
  geom_point(alpha = 0.5) +
  stat_ellipse(type = "t", level = 0.95) +
  labs(
    title = "患病群体中不同性别的 SBP 与 DBP 均值置信椭圆",
    x = "收缩压 SBP",
    y = "舒张压 DBP",
    color = "性别"
  ) +
  theme_bw() +
  scale_color_manual(values = c("男" = "blue", "女" = "red"))

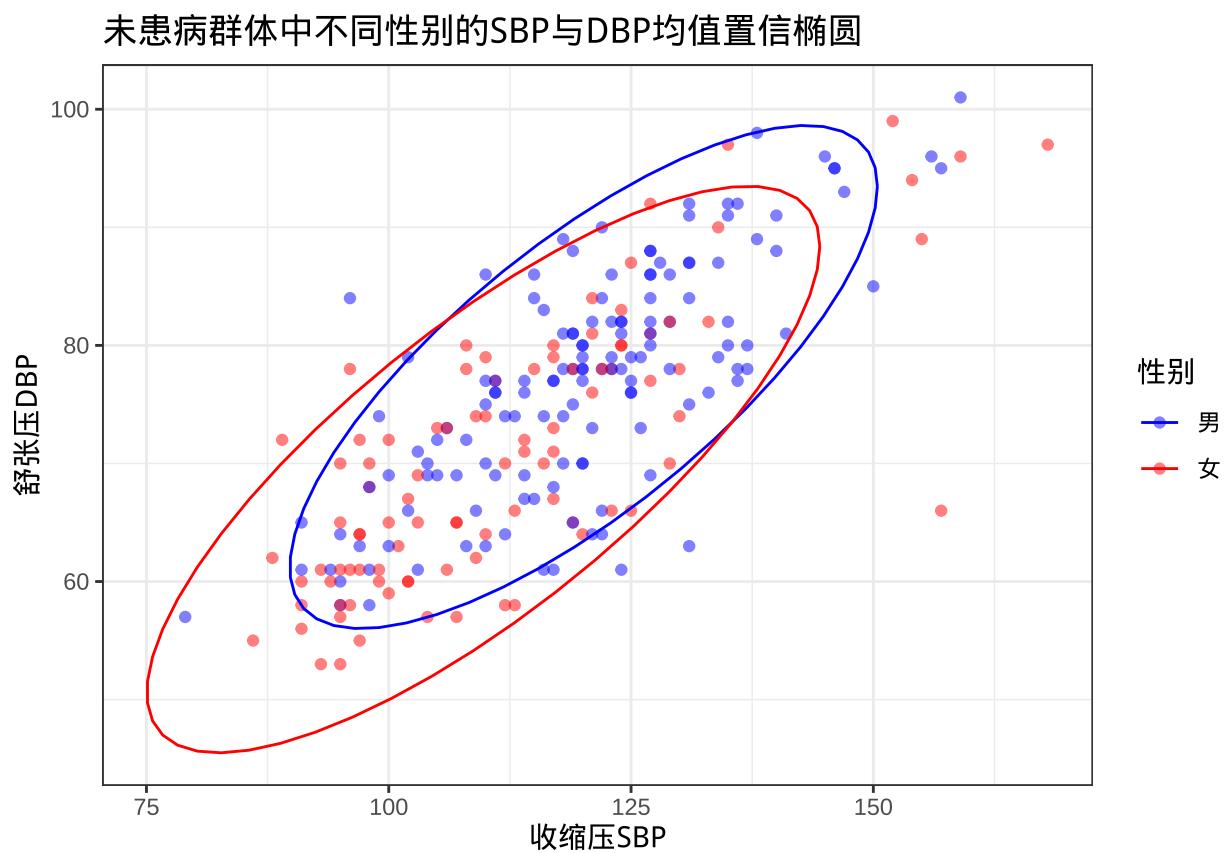
```



```

ggplot(df_healthy, aes(x = sbp, y = dbp, color = gender)) +
  geom_point(alpha = 0.5) +
  stat_ellipse(type = "t", level = 0.95) +
  labs(
    title = "未患病群体中不同性别的 SBP 与 DBP 均值置信椭圆",
    x = "收缩压 SBP",
    y = "舒张压 DBP",
    color = "性别"
  ) +
  theme_bw() +
  scale_color_manual(values = c("男" = "blue", "女" = "red"))

```



从置信椭球的结果可以看出，患病群体不同性别的置信椭圆完全重叠，而未患病群体的置信椭圆存在更多未重叠的部分，一定程度上可以验证未患病群体的均值向量具有显著的性别差异。

第二题

多元回归

建立回归方程如下：

```

df2 <- read_excel("ex2.1.xls")
model <- lm(y ~ x1+x2+x3+x4, data = df2)
summary(model)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = df2)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -3.6268 -1.2004 -0.2276  1.5389  4.4467 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 5.9433    2.8286   2.101   0.0473 *  
## x1          0.1424    0.3657   0.390   0.7006    
## x2          0.3515    0.2042   1.721   0.0993 .    
## x3         -0.2706    0.1214  -2.229   0.0363 *  
## x4          0.6382    0.2433   2.623   0.0155 *  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 2.01 on 22 degrees of freedom
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.5282 
## F-statistic: 8.278 on 4 and 22 DF,  p-value: 0.0003121

```

以上构建的多元线性回归模型具有统计学意义 ($F = 8.278, P = 0.0003 < 0.05$)，因变量空腹血糖变异的 60.08% 可由血清总胆固醇、甘油、空腹胰岛素和糖化血红蛋白来解释 ($R^2 = 0.6008$, 校正的 $R^2 = 0.5282$)。

空腹胰岛素和糖化血红蛋白的偏回归系数检验的 $p < 0.05$, 在 $\alpha = 0.05$ 的检验水准下有统计学显著性。其中空腹胰岛素的回归系数为-0.2706, 表明对空腹血糖有显著的负向影响；糖化血红蛋白的回归系数为 0.6382, 说明对空腹血糖有显著的正向影响。

最终的多元回归方程可以写成：

$$y = 5.9433 + 0.1424 * x_1 + 0.3515 * x_2 - 0.2706 * x_3 + 0.6382 * x_4$$

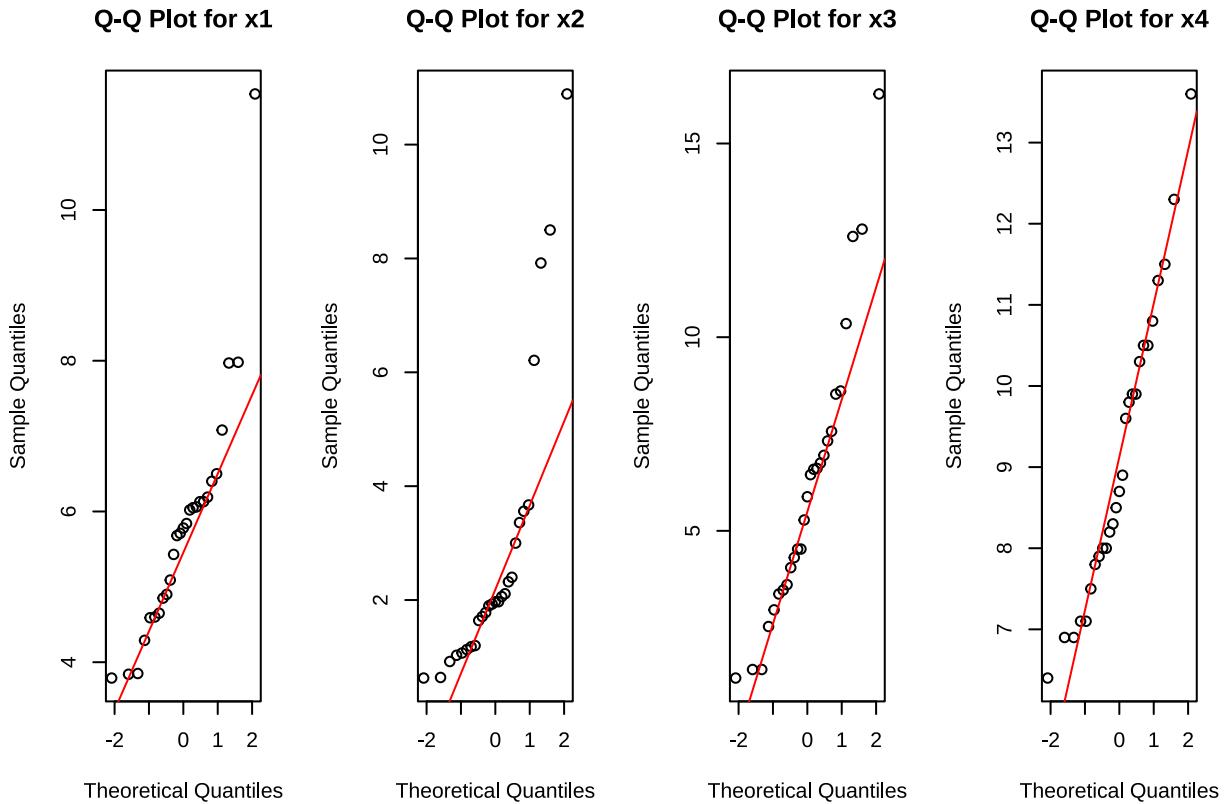
多元正态

对 (X_1, X_2, X_3, X_4) 做多元正态检验，先画出每个单变量的 QQ 图。

```

X2 <- c("x1", "x2", "x3", "x4")
par(mfrow = c(1, 4))
for (var in X2) {
  qqnorm(df2[[var]], main = paste("Q-Q Plot for", var))
  qqline(df2[[var]], col = "red")
}

```



```

par(mfrow = c(1, 1))

```

然后使用 mvn 包进行多元正态性检验，其中对于每个单变量使用 Shapiro-Wilk 检验判断正态性，从检验结果和上面的 QQ 图可以看出，单变量 x1 和 x2 不具有显著的正态性。然后使用 Henze-Zirkler 检验进行多元正态检验，结果显示 (X1,X2,X3,X4) 不具有多元正态性。这和单变量的结果相符，如果单变量不符合正态分布，则它们的组合也不会服从多元正态分布。

```

normal_test <- mvn(df2[X2], descriptives = FALSE, univariate_test = "SW")
print(normal_test$univariate_normality)

```

| ## | Test | Variable | Statistic | p.value | Normality |
|------|--------------|----------|-----------|---------|------------|
| ## 1 | Shapiro-Wilk | x1 | 0.848 | 0.001 | Not normal |

```

## 2 Shapiro-Wilk      x2      0.732 <0.001  Not normal
## 3 Shapiro-Wilk      x3      0.926   0.054      Normal
## 4 Shapiro-Wilk      x4      0.956   0.292      Normal

print(normal_test$multivariate_normality)

##                   Test Statistic p.value      Method      MVN
## 1 Henze-Zirkler     1.167 <0.001 asymptotic  Not normal

```

也可以使用课本上的主成分分析法进行多元正态检验。

```

test_pca <- function(df) {

  pca <- prcomp(df, scale. = TRUE) # 标准化并进行 pca

  score <- pca$x
  n_features <- ncol(score)

  for (i in 1:n_features) {
    pc <- score[, i]

    cat(paste(" 检验主成分", i, "\n"))

    # 使用 Shapiro-Wilk 检验
    shapiro_test <- shapiro.test(pc)
    print(shapiro_test)

    par(mfrow = c(1, 2))
    # 绘制 Q-Q 图
    qqnorm(pc, main = paste("Q-Q Plot of ", i))
    qqline(pc, col = "red", lwd = 2)

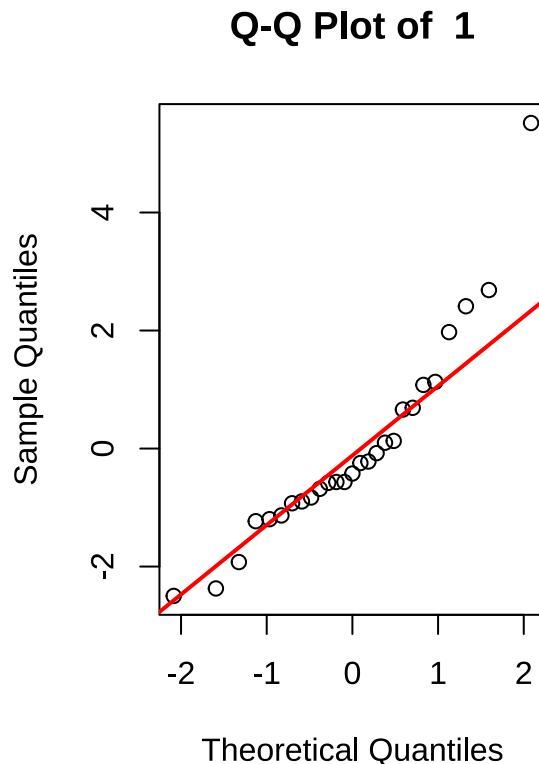
    Sys.sleep(1)
  }
  par(mfrow = c(1, 1))
}

test_pca(df2)

```

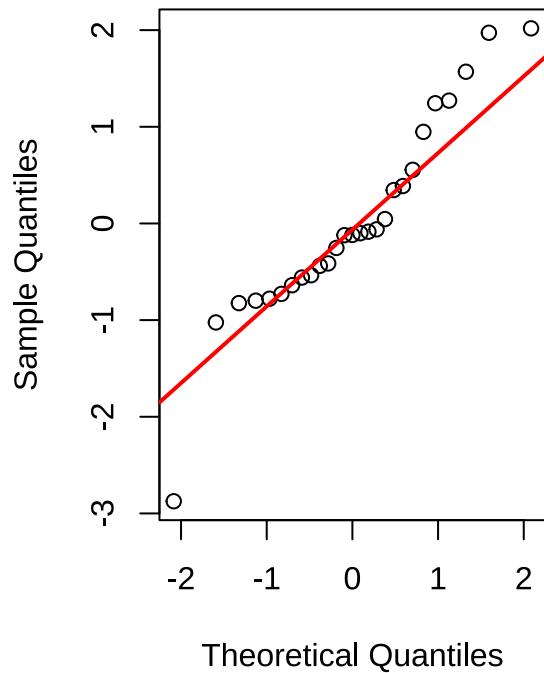
检验主成分 1

```
##  
## Shapiro-Wilk normality test  
##  
## data: pc  
## W = 0.89353, p-value = 0.00954
```



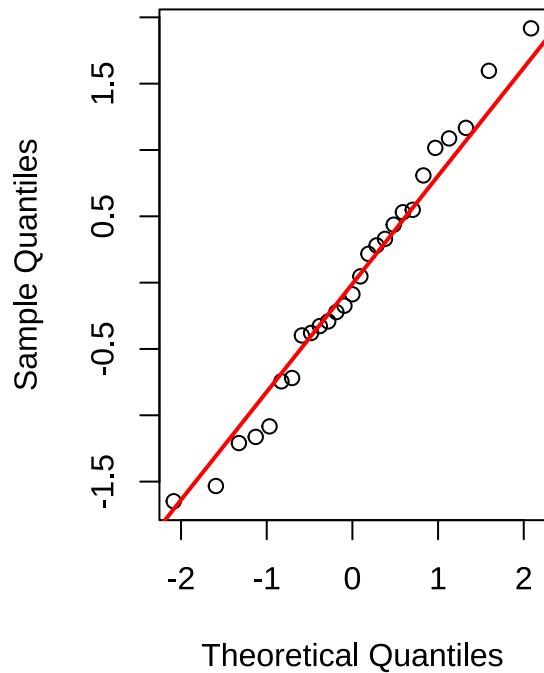
```
## 检验主成分 2  
##  
## Shapiro-Wilk normality test  
##  
## data: pc  
## W = 0.93097, p-value = 0.073
```

Q-Q Plot of 2



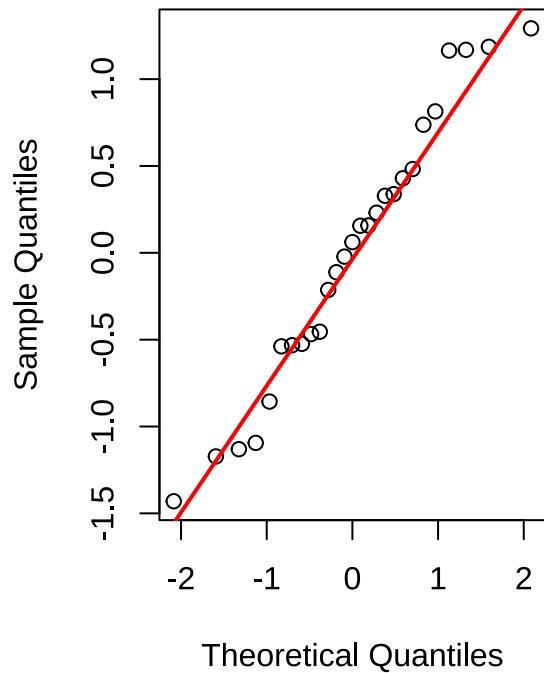
```
## 检验主成分 3
##
## Shapiro-Wilk normality test
##
## data: pc
## W = 0.98447, p-value = 0.9463
```

Q-Q Plot of 3

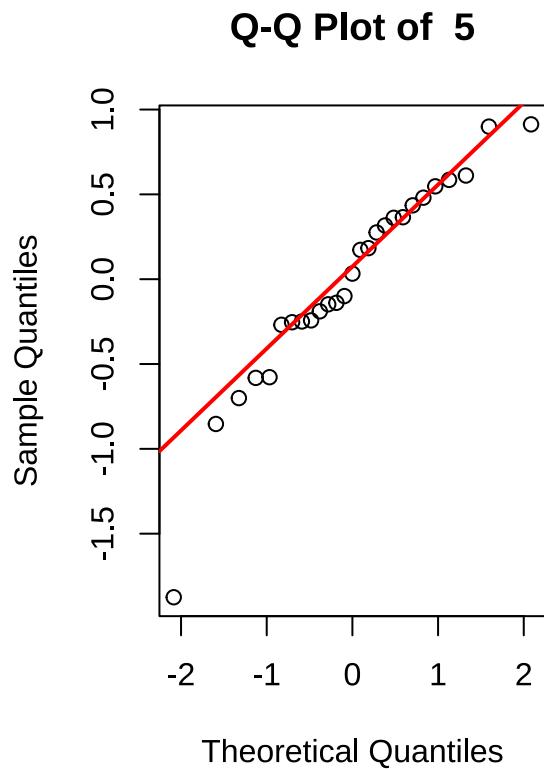


```
## 检验主成分 4
##
## Shapiro-Wilk normality test
##
## data: pc
## W = 0.96513, p-value = 0.4797
```

Q-Q Plot of 4

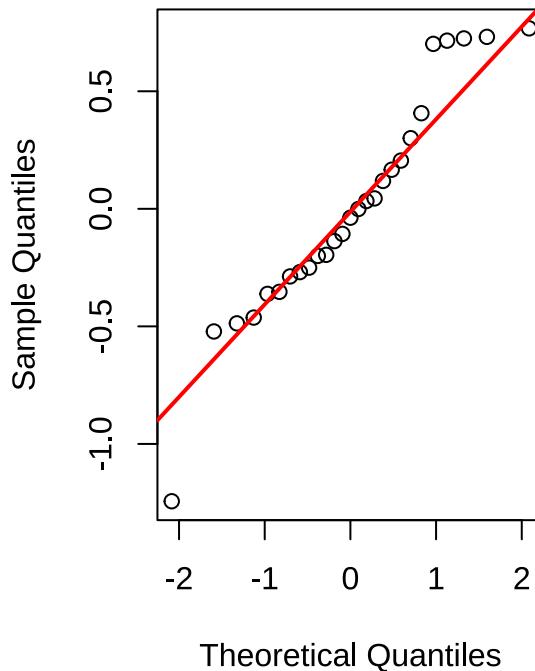


```
## 检验主成分 5
##
## Shapiro-Wilk normality test
##
## data: pc
## W = 0.93026, p-value = 0.07014
```



```
## 检验主成分 6
##
## Shapiro-Wilk normality test
##
## data: pc
## W = 0.94172, p-value = 0.1343
```

Q-Q Plot of 6

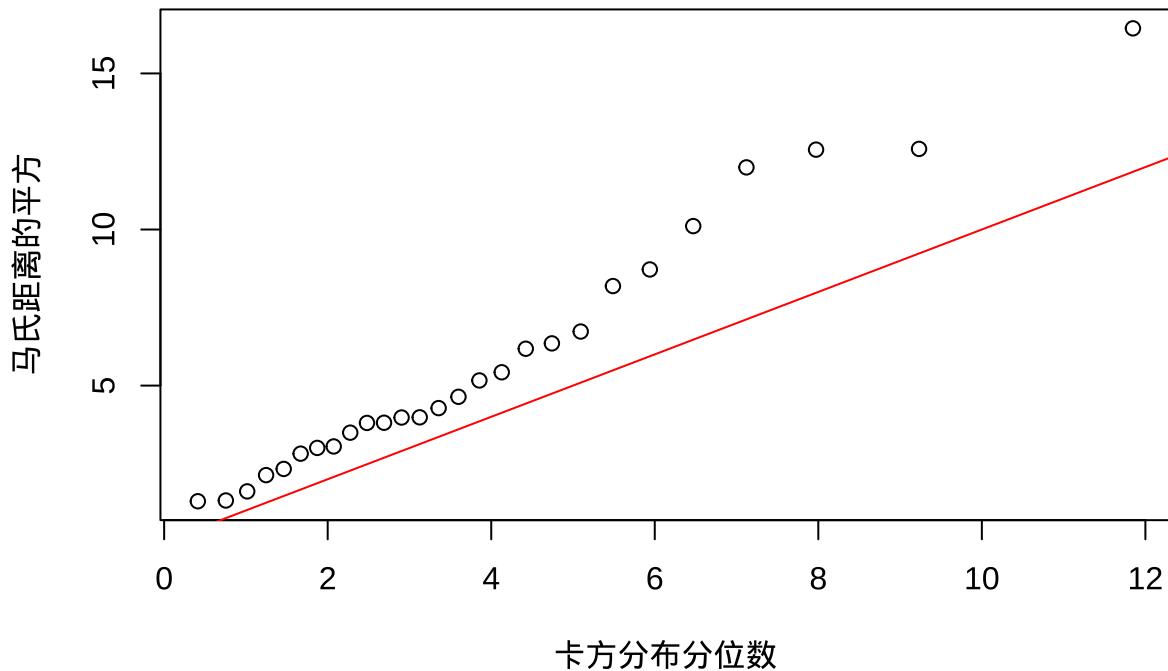


由第一个主成分不具有一元正态性推出 (X_1, X_2, X_3, X_4) 不具有显著的多元正态性。

最后我们可以画出多元正态分布的 QQ 图，可以看出点并不在直线上，与我们上面的分析相符。

```
x_bar <- colMeans(df2)
S <- cov(df2)
d2 <- mahalanobis(df2, center = x_bar, cov = S)

qqplot(qchisq(ppoints(length(d2)), df = 4), d2,
       xlab = "卡方分布分位数",
       ylab = "马氏距离的平方")
abline(0, 1, col = "red")
```



(X1,X2) 和 (X3,X4) 是否相互独立

```
test_independence <- function(df, group_sizes, alpha = 0.05) {
  n <- nrow(df) # 样本量
  p <- ncol(df) # 总变量数

  S <- cov(df) # 协方差阵
  A <- (n - 1) * S # 离差阵
  det_A <- det(A)

  Aii1 <- 1
  start <- 1
  for (size in group_sizes) {
    end <- start + size - 1
    Aii <- A[start:end, start:end, drop = FALSE]
    Aii1 <- Aii1 * det(Aii)
    start <- end + 1
  }
}
```

```

    }

V <- det_A / Aii1
f <- 4
b <- n - 3.5

xi <- -b * log(V)
p_value <- pchisq(xi, df = f, lower.tail = FALSE)
print(p_value)
}

test_independence(df2[X2], group_sizes = c(2,2))

```

[1] 0.1072814

H0: (X1,X2) 和 (X3,X4) 相互独立, H1: (X1,X2) 和 (X3,X4) 不相互独立。

检验的 p > 0.05, 无法拒绝原假设, 即 (X1,X2) 和 (X3,X4) 相互独立。

第三题

多元回归

```

df3 <- read_excel("ex2.5.xls")
model <- lm(y ~ x1+x2+x3+x4+x5+x6, data = df3)
summary(model)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = df3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3078.2  -713.3  -118.6   674.8  2852.2 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 3.716e+04  1.585e+04   2.345 0.028937 *  
## x1          -7.792e-01  3.351e-01  -2.326 0.030138 *  

```

```

## x2      2.308e-01 5.888e-02  3.920 0.000786 ***
## x3      5.425e-01 8.940e-01  0.607 0.550460
## x4     -3.059e-01 1.636e-01 -1.869 0.075580 .
## x5      4.600e-01 1.527e-01  3.012 0.006636 **
## x6     -5.757e-01 6.274e-01 -0.918 0.369255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1428 on 21 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9981
## F-statistic:  2338 on 6 and 21 DF,  p-value: < 2.2e-16

```

以上构建的多元线性回归模型具有统计学意义 ($F = 2338, P < 0.05$), 因变量财政收入变异的 99.85% 可由第一产业增加值、工业增加值、建筑业增加值、年末总人口、社会消费品零售总额和受灾面积来解释 ($R^2 = 0.9985$, 校正的 $R^2 = 0.9981$)。

第一产业增加值、工业增加值和社会消费品零售总额的偏回归系数检验的 $p < 0.05$, 在 $\alpha = 0.05$ 的检验水准下有统计学显著性。其中工业增加值的回归系数为 0.2308, 表明对财政收入有显著的负向影响; 社会消费品零售总额的回归系数为 0.46, 说明对财政收入也有显著的正向影响; 而第一产业增加值的回归系数为-0.78, 说明对财政收入有显著的负向作用。

最终的多元回归方程可以写成:

$$y = 3.7e+04 - 0.78 * x1 + 0.23 * x2 + 0.54 * x3 - 0.31 * x4 + 0.46 * x5 - 0.58 * x6$$