ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

——————oOo——————

ĐỒ ÁN MÔN HỌC

MÔ HÌNH HÓA THỐNG KÊ

Học viên thực hiện:

BÙI TẤT HIỆP 22C01007

Ngành: Khoa học dữ liệu - K32

TP. HỒ CHÍ MINH - 2024
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

——————oOo——————

ĐỒ ÁN MÔN HỌC

MÔ HÌNH HÓA THỐNG KÊ

Học viên thực hiện:

BÙI TẤT HIỆP 22C01007

Ngành Khoa học dữ liệu - K32

TP. HỒ CHÍ MINH - 2024
ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

Mục lục

Phần mở đầu

Bài báo cáo này tập trung vào việc diễn giải và báo cáo việc thực hiện mô hình hồi quy tuyến tính, phân tích phương sai ANOVA, đây là những công cụ quan trọng trong lĩnh vực máy học và thống kê. Mô hình hồi quy tuyến tính cho phép mô hình hóa mối quan hệ giữa các biến đầu vào và đầu ra trong dữ liệu. Bằng cách chọn tìm mô hình tối ưu, học viên đã trình bày các nội dung chính yếu nhất của môn học. Vì đồ án được thực hiện dựa trên sự tìm hiểu và hiểu biết của học viên trong quá trình học, vì vậy không thể tránh khỏi việc sai sót, học viên mong người đọc có thể chỉ ra sai sót để có thể hiểu hơn về kiến thức. Qua đó, học viên xin chân thành gửi lời cám ơn cô Nguyễn Thị Mộng Ngọc đã hỗ trợ và truyền dạy kiến thức trong môn học này.
Xin chân thành cám ơn!

Chương 1

Hoạt động 1

1.1 Bài tập 1

Giới thiệu tập dữ liệu và các biến Tập dữ liệu Conventional and Social Media Movies (CSM) cung cấp một số thuộc tính của phim ảnh lấy từ nguồn UCI Machine Learning Repository.
Bộ dữ liệu gồm 231 quan trắc trên 14 biến:

- "Movie": tên phim;
- "Year": năm phát hành;
- "Ratings": điểm đánh giá;
- "Genre": thể loại phim;
- "Gross": tổng doanh thu;
- "Budget": tổng chi phí;
- "Screens": số rạp chiều;
- "Sequel": phần phim;
- "Sentiment": ý kiến khán giả;
- "Views": số lượt xem;
- "Likes": số lượt thích;
- "Dislikes": số lượt không thích;
- "Comments": số bình luận;
- "Aggregate_Followers": số người theo dõi.

1.1.1 Tiền xử lí dữ liệu

Các bước đã thực hiện:
- Loại bỏ dữ liệu bị trùng lặp (nếu có)
- Bỏ đi biến Movie
- Xử lí dữ liệu bị khuyết ở biến Screens và biến Aggregate_Followers

```r
## Kiểm tra dữ liệu khuyết
data1[duplicated(data1),]

## Data profiling
skimr::skim(data1)

## Điền dữ liệu khuyết với thuật toán MICE
imputed_Data <- mice(data1, method = 'cart', seed = 42, print=FALSE)
data1 <- complete(imputed_Data,action=5)
colSums(is.na(data1))

summary(data1)
```

```
> summary(data1)
      Year         Ratings          Genre            Gross              Budget            Screens          Sequel
 Min.   :2014   Min.   :3.100   Min.   : 1.000   Min.   :     2470   Min.   :7.00e+04   Min.   :   2.0   Min.   :1.000
 1st Qu.:2014   1st Qu.:5.800   1st Qu.: 1.000   1st Qu.: 10300000   1st Qu.:9.00e+06   1st Qu.: 372.5   1st Qu.:1.000
 Median :2014   Median :6.500   Median : 3.000   Median : 37400000   Median :2.80e+07   Median :2766.0   Median :1.000
 Mean   :2014   Mean   :6.442   Mean   : 5.359   Mean   : 68066033   Mean   :4.78e+07   Mean   :2144.6   Mean   :1.359
 3rd Qu.:2015   3rd Qu.:7.100   3rd Qu.: 8.000   3rd Qu.: 89350000   3rd Qu.:6.50e+07   3rd Qu.:3360.5   3rd Qu.:1.000
 Max.   :2015   Max.   :8.700   Max.   :15.000   Max.   :643000000   Max.   :2.50e+08   Max.   :4324.0   Max.   :7.000
   Sentiment          Views              Likes            Dislikes          Comments       Aggregate_Followers
 Min.   :-38.00   Min.   :     698   Min.   :     1   Min.   :    0.0   Min.   :    0.0   Min.   :    1066
 1st Qu.:  0.00   1st Qu.:  623302   1st Qu.:  1776   1st Qu.:  105.5   1st Qu.:  248.5   1st Qu.:  147000
 Median :  0.00   Median : 2409338   Median :  6096   Median :  341.0   Median :  837.0   Median :  888000
 Mean   :  2.81   Mean   : 3712851   Mean   : 12732   Mean   :  679.1   Mean   : 1825.7   Mean   : 2983497
 3rd Qu.:  5.50   3rd Qu.: 5217380   3rd Qu.: 15248   3rd Qu.:  697.5   3rd Qu.: 2137.0   3rd Qu.: 3406000
 Max.   : 29.00   Max.   :32626778   Max.   :370552   Max.   :13960.0   Max.   :38363.0   Max.   :31030000
```

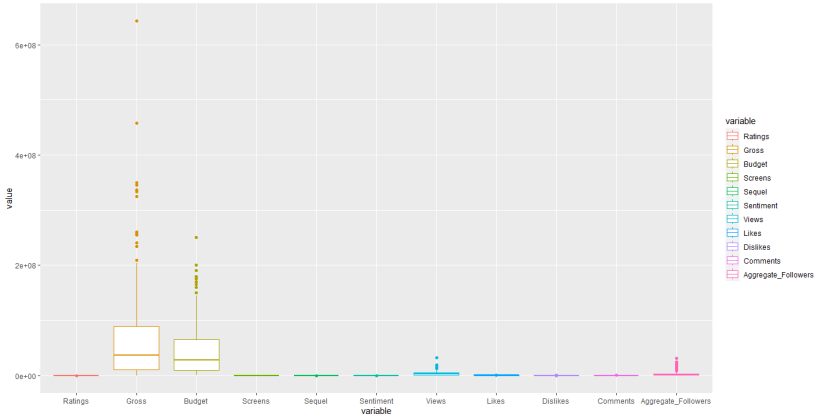
Hình 1.1: Dữ liệu sau khi làm sạch

1.1.2 Chia tập train - tập validation

Các bước đã thực hiện:
- Tạo biến giả cho biến Genre và bỏ đi biến Genre gốc sau khi thực hiện
- Neo lần chạy ngẫu nhiên (chọn cách gieo ngẫu nhiên số 42)
- Thực hiện chia tách tập dữ liệu theo tỉ lệ 80-20

```r
data1 <- dummy_cols(data1, select_columns = 'Genre')
# Loại bỏ biến Genre sau khi tạo biến giả
data1 <- subset(data1, select = -Genre)

set.seed(42)
#Tạo id column
data1$id <- 1:nrow(data1)

#sử dụng 80% cho tập training và 20% cho tập validation
train <- data1 %>% dplyr::sample_frac(0.8)
val  <- dplyr::anti_join(data1, train, by = 'id')
train <- subset(train, select = -id)
val <- subset(val, select = -id)
```

Vì dữ liệu xuất hiện nhiều trường hợp outlier, học viên quyết định không bỏ đi các giá trị outlier vì khi loại bỏ outlier sẽ làm mất đi nhiều thông tin của dữ liệu cung cấp. Ví dụ vẽ boxplot ở tập train với hình 1.2



Hình 1.2: Biểu đồ hộp các biến trong tập dữ liệu

1.1.3 Chọn mô hình

Trước hết học viên em tiến hành xây dựng một mô hình hồi quy tuyến tính, sau đó sẽ tiến hành kiểm tra các giả định của mô hình, ví dụ như kiểm tra tính chuẩn, trung bình và phương sai,...

Để xây dựng mô hình hồi quy tuyến tính giữa Gross và tất cả các biến. Đặt giả thuyết:
- H0: $\beta_i = 0, \forall i$
- H1: $\exists \beta_i \neq 0$

Với p-value: <0.05 ở hình 1.3. Ta bác bỏ H0, tồn tại ít nhất 1 $\beta_i \neq 0$. Vì vậy, ta xây dựng mô hình thứ 2, lần này chỉ có các biến có ý nghĩa với p_value < 0.05: 'Ratings' 'Budget' 'Screens' 'Dislikes' 'Aggregate_Followers'

```
## Chạy mô hình hồi quy tuyến tính với tất cả các biến
mod <- lm(Gross ~ ., data = train)
summary(mod)
## Chạy mô hình tuyến tính với các biến Ratings, Budget, Screens, Dislikes, Aggregate_Followers
mod2 <- lm(Gross ~ Ratings + Budget + Screens + Dislikes + Aggregate_Followers, data = train)
summary(mod2)
```

Tuy mô hình thứ 2 cho được kết quả F-statistic lớn hơn ($52.69 > 13.18$), cho thấy mô hình 2 phù hợp hơn mô hình 1. Tuy nhiên, p_value ở biến 'Dislike' có giá trị p_value > 0.05. Học viên tiếp tục loại bỏ biến này khỏi mô hình và chạy lại và gọi đây là mô hình 3. Mô hình 3 cho kết quả tốt khi các biến độc lập đều có ý nghĩa trong việc giải thích cho biến phụ thuộc Gross. Kiểm tra với hệ số tương quan Pearson với những biến đã chọn trên, nhận thấy các biến độc

```
Call:
lm(formula = Gross ~ ., data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-129566584  -30389057   -4381397   20855519  411235192

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -5.256e+09  2.358e+10  -0.223 0.823923
Year                 2.544e+06  1.171e+07   0.217 0.828242
Ratings              2.090e+07  5.607e+06   3.728 0.000266 ***
Budget               7.226e-01  1.290e-01   5.600 8.94e-08 ***
Screens              1.233e+04  4.223e+03   2.920 0.003996 **
Sequel               1.028e+07  5.383e+06   1.910 0.057922 .
Sentiment           -5.884e+05  7.021e+05  -0.838 0.403248
Views               -2.401e+00  1.949e+00  -1.232 0.219729
Likes                6.554e+02  4.320e+02   1.517 0.131242
Dislikes             1.364e+04  6.179e+03   2.208 0.028663 *
Comments            -4.411e+03  3.720e+03  -1.186 0.237462
Aggregate_Followers  2.238e+00  9.924e-01   2.255 0.025474 *
Genre_1             -4.218e+05  2.747e+07  -0.015 0.987766
Genre_2             -2.499e+07  3.259e+07  -0.767 0.444308
Genre_3             -2.566e+07  2.744e+07  -0.935 0.351106
Genre_4             -6.465e+07  6.731e+07  -0.961 0.338191
Genre_6             -3.849e+07  4.350e+07  -0.885 0.377544
Genre_7              8.084e+06  4.983e+07   0.162 0.871320
Genre_8             -2.009e+07  2.667e+07  -0.753 0.452429
Genre_9             -1.582e+07  3.340e+07  -0.474 0.636405
Genre_10            -3.202e+07  3.298e+07  -0.971 0.333143
Genre_12            -1.773e+07  3.386e+07  -0.524 0.601334
Genre_15                    NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61060000 on 163 degrees of freedom
Multiple R-squared:  0.6293,    Adjusted R-squared:  0.5815
F-statistic: 13.18 on 21 and 163 DF,  p-value: < 2.2e-16
```

Hình 1.3: Xây dựng mô hình hồi quy giữa biến Phụ thuộc Gross với tất cả các biến độc lập

```
Call:
lm(formula = Gross ~ Ratings + Budget + Screens + Dislikes +
    Aggregate_Followers, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-136114719  -28220540   -6345914   18750011  434188530

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.243e+08  3.125e+07  -3.977 0.000101 ***
Ratings              1.737e+07  4.787e+06   3.628 0.000373 ***
Budget               8.257e-01  1.065e-01   7.752 6.56e-13 ***
Screens              1.448e+04  3.927e+03   3.688 0.000299 ***
Dislikes             5.370e+03  3.528e+03   1.522 0.129694
Aggregate_Followers  2.390e+00  9.024e-01   2.648 0.008810 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60870000 on 179 degrees of freedom
Multiple R-squared:  0.5954,    Adjusted R-squared:  0.5841
F-statistic: 52.69 on 5 and 179 DF,  p-value: < 2.2e-16
```

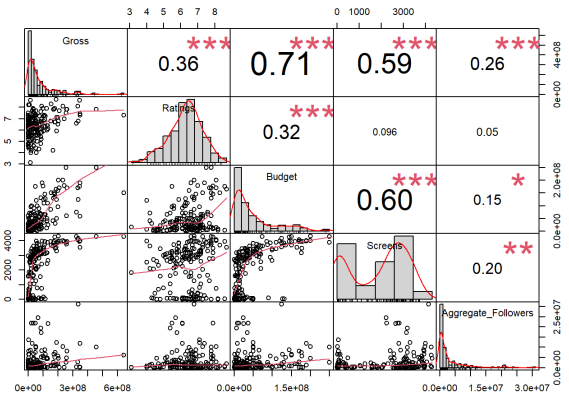
Hình 1.4: Xây dựng mô hình hồi quy giữa biến Phụ thuộc Gross với các biến 'Ratings' 'Budget' 'Screens' 'Dislikes' 'Aggregate_Followers'

lập đều có hiện tượng phụ thuộc tuyến tính với biến phụ thuộc; trong đó có 'Budget' và 'Screen' cho hệ số tương quan khá cao. Vì vậy học viên quyết định chọn 2 biến này để xây dựng mô hình hồi quy tuyến tính.

```
## Chạy mô hình hồi quy tuyến tính lần 3
mod3 <- lm(Gross ~ Ratings + Budget + Screens + Aggregate_Followers, data = train)
summary(mod3)
```

```
Call:
lm(formula = Gross ~ Ratings + Budget + Screens + Aggregate_Followers,
    data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-139692835  -29482749   -6592906   17609432  434853020

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.130e+08  3.046e+07  -3.708 0.000278 ***
Ratings              1.577e+07  4.688e+06   3.364 0.000938 ***
Budget               8.243e-01  1.069e-01   7.710 8.22e-13 ***
Screens              1.575e+04  3.852e+03   4.087 6.57e-05 ***
Aggregate_Followers  2.433e+00  9.053e-01   2.688 0.007865 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 61100000 on 180 degrees of freedom
Multiple R-squared:  0.5902,    Adjusted R-squared:  0.5811
F-statistic:  64.8 on 4 and 180 DF,  p-value: < 2.2e-16
```

Hình 1.5: mod3



Hình 1.6: Hệ số tương quan

```
Call:
lm(formula = Gross ~ Budget + Screens, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-159537837  -28791195   -3559172   12148488  443921008

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.031e+07  8.246e+06  -1.250 0.212967
Budget       9.541e-01  1.054e-01   9.049  < 2e-16 ***
Screens      1.551e+04  3.956e+03   3.920 0.000125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 63870000 on 182 degrees of freedom
Multiple R-squared:  0.5472,    Adjusted R-squared:  0.5422
F-statistic:   110 on 2 and 182 DF,  p-value: < 2.2e-16
```

Hình 1.7: Mô hình tuyến tính được chọn

Từ kết quả tại hình 1.7, ta xây dựng được mô hình như sau:

$$\hat{y} = -1.031e^{+07} + 9.541e^{-01}X_1 + 1.551e^{+04}X_2 \tag{1.1}$$

Với X_1 là Budget và X_2 là biến Screens. Kiểm tra đa cộng tuyến, ta nhận được VIF của 2 biến trên lần lượt là: 1.570738 1.570738. **Như vậy mô hình không xảy ra đa cộng tuyến**.

Kiểm tra giả định mô hình Bao gồm các giả định sau:
+ Biến phụ thuộc Y và các biến độc lập X có mối quan hệ tuyến tính (thỏa)
+ Không xảy ra đa cộng tuyến (thỏa)
+ Sai số có phân phối chuẩn với trung bình sai số = 0 và phương sai không thay đổi.

```r
## Chạy mô hình hồi quy tuyến tính
mod3 <- lm(Gross ~ Budget + Screens, data = train)

## Kiểm tra sai số có phân phối chuẩn
shapiro.test(resid(mod3))

## Kiểm tra trung bình sai số mu=0
t.test(resid(mod3), mu = 0)

## Kiểm tra tính ổn định của phương sai
ncvTest(mod3)
```

```
    Shapiro-Wilk normality test

data:  resid(mod3)
W = 0.80995, p-value = 2.956e-14
```

1. Kiểm tra sai số có tuân theo phân phối chuẩn hay không

Đặt giả thuyết:

H0: ε_i có phân phối chuẩn
H1: ε_i không có phân phối chuẩn

Ta bác bỏ H0, p-value < 0.05 **sai số của mô hình không có phân phối chuẩn**

```
    One Sample t-test

data:  resid(mod3)
t = 1.4892e-15, df = 184, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -9213740  9213740
sample estimates:
   mean of x
6.954537e-09
```

2. Kiểm tra giả định trung bình sai số $\mu = 0$

Đặt giả thuyết:

H0: $\mathrm{E}(\varepsilon_i) = 0$
H1: $\mathrm{E}(\varepsilon_i) \neq 0$

p_value = 1, không bác bỏ H0. **Trung bình sai số thỏa mãn** $\mathrm{E}(\varepsilon_i) = 0$

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 127.2971, Df = 1, p = < 2.22e-16
```

3. Kiểm tra tính ổn định của phương sai

Đặt giả thuyết

H0: phương sai không thay đổi
H1: phương sai thay đổi

Ta bác bỏ H0 với p_value < 2.22e-16 < 0.05. **Mô hình không thỏa mãn tính ổn định của phương sai**, phương sai của sai số thay đổi.

Như vậy, mô hình không thỏa mãn tính ổn định của phương sai cũng như không có phân phối chuẩn. Việc không thỏa các giả định khiến cho mô hình không đáng tin cậy cho việc ước lượng. Vì vậy cần thực hiện các phép biến đổi biến (trasformation) với phương pháp Box-Cox. Vì cả biến Gross có 14 giá trị outlier, học viên quyết định kiểm tra và cân nhắc biến đổi cả biến phụ thuộc và biến độc lập.

```
a <- powerTransform(cbind(Gross, Budget, Screens) ~ 1, data = train)
summary(a)
```

Dựa theo kiểm định trên, ta rút được kết luận: cần thực hiện ít nhất một phép biến đổi biến và không cần thực hiện phép log-transformation với tất cả các biến. Như vậy dự định biến đổi biến phụ thuộc và độc lập sẽ dựa trên giá trị 'Rounded Pwr'. Kiểm tra lại các giả định, kết quả ở hình 1.1.3

```
bcPower Transformations to Multinormality
        Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Gross      0.2544        0.25       0.2065       0.3023
Budget     0.2047        0.20       0.1281       0.2813
Screens    0.6245        0.62       0.5062       0.7428

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                              LRT df       pval
LR test, lambda = (0 0 0) 277.5498  3 < 2.22e-16

Likelihood ratio test that no transformations are needed
                              LRT df       pval
LR test, lambda = (1 1 1) 684.7745  3 < 2.22e-16
```

Likelihood ratio test that transformation parameters are equal to 0

Để kiểm định việc có nên thực hiện phép biến đổi log-transformation với tất cả các biến hay không. Đặt giả thuyết:

H0: Cần thực hiện phép biến đổi log-transformation cho tất cả các biến
H1: Không cần thực hiện phép biến đổi log-transformation cho tất cả các biến
Với p_value < 2.22e-16. Bác bỏ H0, ta **không cần thực hiện phép biến đổi log-transformation cho tất cả các biến**.

Likelihood ratio test that no transformations are needed

Để kiểm định việc có nên thực hiện phép biến đổi biến hay không. Đặt giả thuyết:

H0: Không cần thực hiện phép transformation
H1: Cần thực hiện tối thiểu một phép transformation
Với p_value < 2.22e-16. Bác bỏ H0, ta **cần thực hiện tối thiểu một phép transformation**. Như vậy em quyết định **thực hiện biến đổi biến với tất cả các giá trị lambda được đề xuất (Rounded Pwr)**

```r
# Rút ra các cột trong tập train và biến đổi biến
subset <- train %>% select(Gross, Budget, Screens)
subset$Gross <- subset$Gross^(0.25)
subset$Budget <- subset$Budget^(0.2)
subset$Screens <- subset$Screens^(0.62)

mod3 <- lm(Gross ~ Budget + Screens, data = subset)
summary(mod3)
```

Như vậy, sau khi thực hiện biến đổi biến ở các biến độc lập cũng như biến phụ thuộc. Học viên đã khắc phục được giả định về tính ổn định của phương sai sai số. Tuy nhiên giả định về tính chuẩn của sai số vẫn chưa khắc phục được mặc dù đã có sự cải thiện đáng kể về giá trị của p_value. Với R^2 hiệu chỉnh = 64.85%, mô hình có các biến độc lập giải thích được cho 64.85% sự biến thiên của biến phụ thuộc. Mô hình không xảy ra hiện tượng đa cộng tuyến giữa các biến độc lập. Tuy nhiên, mô hình vẫn chưa thỏa mãn giả định về tính chuẩn của sai số. Mô hình cuối cùng được xây dựng có dạng như sau:

$$\widehat{y^{0.25}} = -5.63919 + 1.77492{X_1}^{0.2} + 0.24795{X_2}^{0.62} \tag{1.2}$$

```
Call:
lm(formula = Gross ~ Budget + Screens, data = subset)

Residuals:
   Min     1Q Median     3Q    Max
-81.15 -11.84  -0.96  12.79  46.23

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.63919    5.77782  -0.976     0.33
Budget       1.77492    0.22655   7.834 3.79e-13 ***
Screens      0.24795    0.03026   8.193 4.38e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.89 on 182 degrees of freedom
Multiple R-squared:  0.6523,    Adjusted R-squared:  0.6485
F-statistic: 170.7 on 2 and 182 DF,  p-value: < 2.2e-16
```

```
    Shapiro-Wilk normality test

data:  resid(mod3)
W = 0.9843, p-value = 0.03625
```

1. Giả thiết:

H0: ε_i có phân phối chuẩn
H1: ε_i không có phân phối chuẩn

Ta bác bỏ H0, p-value = 0.03625 < 0.05 **sai số của mô hình không có phân phối chuẩn**. Tuy nhiên p_value ở mô hình này đã cải thiện từ 2.956e-14 ở mô hình cũ lên p_value = 0.03625. Có thể thấy mô hình đã cải thiện đáng kể về giả định về tính chuẩn của sai số.

```
    One Sample t-test

data:  resid(mod3)
t = -1.0525e-15, df = 184, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.869418  2.869418
sample estimates:
    mean of x
-1.530804e-15
```

2. Kiểm tra giả định trung bình sai số $\mu = 0$.

H0: $\mathrm{E}(\varepsilon_i) = 0$
H1: $\mathrm{E}(\varepsilon_i) \neq 0$

p_value = 1, không bác bỏ H0. Trung bình sai số thỏa mãn $\mathrm{E}(\varepsilon_i) = 0$

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1531024, Df = 1, p = 0.69559
```

3. Kiểm tra tính ổn định của phương sai

H0: phương sai không thay đổi
H1: phương sai thay đổi

Chấp nhận H0 với p = 0.695 > 0.05. Mô hình thỏa mãn tính ổn định của phương sai.

Dự báo

Copy tập validation và biến đổi biến trên tập này để thực hiện dự báo.

```
val_subset <- val %>% select(Budget, Screens)
val_subset$Budget <- val_subset$Budget^(0.2)
val_subset$Screens <- val_subset$Screens^(0.62)

## Thực hiện dự báo giá trị Gross
y_hat <- predict(mod3, newdata = val_subset)
y_hat <- y_hat^(1/0.25)
MSE <- mean((val$Gross - y_hat)^2)
```

1.1.4 So sánh kết quả

Kiểm tra kết quả của tập validation, ta có $MSE = 1.335894e + 15$

1.1.5 Đề xuất cải tiến/phân tích khác

Mặc dù đã cố gắng sử dụng các phép biến đổi biến, tuy nhiên mô hình hồi quy tuyến tính mà học viên xây dựng vẫn chưa đáp ứng được giả định về tính chuẩn của sai số.
Theo học viên tìm hiểu, một cách tiếp cận khác đó chính là sử dụng **Thống kê phi tham số**, vì lúc này ta không cần đưa ra các giả định về phân phối tổng thể của tập dữ liệu.
Điểm mạnh của thống kê phi tham số là chống nhiễu tốt (outlier- tập dữ liệu này có nhiều giá trị outlier), giảm thời gian xây dựng mô hình khi không cần kiểm tra các giả định phân phối của dữ liệu, thực hiện biến đổi biến... Thực hiện thử nghiệm mô hình cây quyết định (Decision Tree) và mô hình XGBoots- một mô hình Gradient boosting có cấu trúc tree based (và được tối ưu tuning parameter với Grid-search) [4]. Các mô hình được sử dụng với package 'tidyverse', 'caret'. Lưu ý: lệnh train sử dụng từ package 'parsnip'. Ta có kết quả như sau:

```
[1] "Linear Regression: MSE"
[1] 1.335894e+15
[1] "Decision Tree: MSE"
[1] 2.561262e+15
[1] "XGboost: MSE"
[1] 1.630633e+15
```

Dựa trên kết quả MSE, học viên thấy được ở tập dữ liệu này Mô hình hồi quy tuyến tính cho kết quả tốt nhất với MSE = 1.336e+15 dù rằng không thỏa giả định về tính chuẩn của sai số. Trong khi đó mô hình cây quyết định cho kết quả MSE = 2.561e+15 và mô hình XGBoots cho kết quả MSE = 1.631e+15, kết quả 2 mô hình sau tệ hơn MSE của mô hình hồi quy tuyến tính. Tuy nhiên, học viên cho rằng nếu thu thập nhiều dữ liệu hơn, thì các mô hình Decision Tree và XGBoots có thể sẽ cho kết quả tốt hơn.

1.2 Bài tập 2

Giới thiệu tập dữ liệu và các biến : insurance.csv chứa các thông tin về số tiền bảo hiểm dùng trong y tế cho những người dân ở Mỹ. Bộ dữ liệu gồm 1338 quan trắc và 7 biến sau:

- "age": độ tuổi của người sử dụng quyền lợi bảo hiểm y tế;
- "sex": giới tính của người sử dụng bảo hiểm y tế (male/female);
- "bmi": chỉ số BMI;
- "children": số người phụ thuộc (con);
- "smoker": người sử dụng bảo hiểm có hút thuốc hay không (yes/no);
- "region": vùng sinh sống của người sở hữu bảo hiểm y tế;
- "charges": chi phí bảo hiểm y tế chi trả cho người được khảo sát.

1.2.1 Tiền xử lí dữ liệu

Các bước thực hiện:
- Kiểm tra dữ liệu bị duplicate (có 1 trường hợp bị duplicate).

```
## Kiểm tra duplicate
data2[duplicated(data2),]

## Loại bỏ biến duplicate
data2 <- data2 %>% distinct()
skimr::skim(data2)

## Biến đổi biến age
data2$age <- cut(data2$age, breaks = c(0, 30, 60, Inf),
                 labels = c("age < 30", "30 <= age < 60", "age >= 60"))
## Biến đổi biến bmi
data2$bmi <- cut(data2$bmi, breaks = c(0, 20, 35, Inf),
                 labels = c("bmi < 25", "25 <= bmi < 35", "bmi >= 35"))
## Biến đổi biến children
```

```r
data2$children <- cut(data2$children, breaks = c(-1, 1.5, 2.5, Inf),
                      labels = c("0 or 1 child", "2 childs", "> 2 childs"))

## Thực hiện biến đổi factor với các biến.
data2$sex = factor(data2$sex)
data2$smoker = factor(data2$smoker)
data2$region = factor(data2$region)
data2$age <- factor(data2$age)
data2$bmi <- factor(data2$bmi)
data2$children <- factor(data2$children)
```

Sau khi loại bỏ quan sát bị trùng, có 1337 quan sát với 7 biến độc lập. Dữ liệu không có missing value, để thực hiện phân tích ANOVA k nhân tố. Vì mục tiêu là khai thác hết mọi thông tin mà tập dữ liệu cung cấp, học viên quyết định thực hiện chuyển đổi một số biến.

```
tibble [1,337 × 7] (S3: tbl_df/tbl/data.frame)
 $ age     : Factor w/ 3 levels "age < 30","30 <= age < 60",..: 1 1 1 2 2 2 2 2 2 2 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : Factor w/ 3 levels "bmi < 25","25 <= bmi < 35",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ children: Factor w/ 3 levels "0 or 1 child",..: 1 1 3 1 1 1 1 3 2 1 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num [1:1337] 16885 1726 4449 21984 3867 ...
```

Hình 1.8: Tập dữ liệu sau khi được làm sạch

1.2.2 Phân tích phương sai ANOVA k nhân tố

Trải qua quá trình tiền xử lí dữ liệu trên, em có 6 biến cần thực hiện phân tích ANOVA nhiều nhân tố bao gồm [sex, smoker region age bmi children] (lí do chọn 6 biến: để có thể phân tích, khai thác tối đa các thông tin từ bộ dữ liệu đem lại). Vì vậy trước khi thực hiện ANOVA, tiến hành các giả định của 1 mô hình ANOVA nhiều nhân tố gồm:

- Các mẫu phải độc lập: trong dữ liệu này thì các quan sát đều được ghi nhận độc lập.
- Biến phụ thuộc charges là biến liên tục.
- Các nhóm phải có phân phối chuẩn hoặc gần chuẩn.
- Các nhóm có phương sai đồng nhất (thực hiện kiểm định Levene's [5]).

```r
av_residual <- (rstandard(aov(data2$charges ~ data2$age + data2$sex + data2$bmi
                              + data2$children + data2$smoker + data2$region)))

## Giả định phần dư có pp chuẩn hoặc gần chuẩn
shapiro.test(av_residual)

## Giả định các nhóm có phương sai đồng nhất
leveneTest(charges ~ interaction(sex,age,bmi,children,smoker,region), data = data2)
```

```
    Shapiro-Wilk normality test

data:  av_residual
W = 0.92949, p-value < 2.2e-16
```

Đặt giả thuyết:

H0: Dữ liệu có phân phối chuẩn

H1: Dữ liệu không có phân phối chuẩn

Vì p_value < 2.2e-16, ta bác bỏ H0, dữ liệu không có phân phối chuẩn.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group  218  1.4416 0.0001271 ***
      1118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Đặt Giả thuyết kiểm định thống kê ở các biến như sau:

H0: Các nhóm có phương sai đồng nhất

H1: Các nhóm có phương sai không đồng nhất

p_value = 0.0001271 < 0.05. Như vậy ta bác bỏ H0, **phương sai ở các nhóm không đồng nhất**.

Thực hiện biến đổi biến, sau đó thực hiện phân tích ANOVA Vì biến 'charges' là số không âm, ta có thể thực hiện biến đổi biến theo phương pháp Box-Cox.

```r
summary(model <- powerTransform(charges ~ ., data = data2))
```

```r
data2$charges <- data2$charges^(0.17)

## Thực hiện ANOVA
dependent_variable = c("charges")
factor_list = c("age", "sex", "bmi", "children", "smoker", "region")
n_way = nway_aov(dependent_variable, factor_list, data=data2)
print(n_way)

```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.1748        0.17       0.1247       0.2249

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 46.19943  1 1.0681e-11

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 992.2174  1 < 2.22e-16
```

Likelihood ratio test that transformation parameters are equal to 0

Để kiểm định việc có nên thực hiện phép biến đổi log-transformation với tất cả các biến hay không. Đặt giả thuyết:

H0: Cần thực hiện phép biến đổi log-transformation cho tất cả các biến.
H1: Không cần thực hiện phép biến đổi log-transformation cho tất cả các biến.
Với p_value < 2.5535e-15. Bác bỏ H0, **ta không cần thực hiện phép biến đổi log-transformation cho tất cả các biến**.

Likelihood ratio test that no transformations are needed

Để kiểm định việc có nên thực hiện phép biến đổi biến hay không. Đặt giả thuyết:

H0: Không cần thực hiện phép transformation.
H1: Cần thực hiện tối thiểu một phép transformation.
Với p_value < 2.22e-16. Bác bỏ H0, ta **cần thực hiện tối thiểu một phép transformation**. Như vậy em quyết định **thực hiện biến đổi biến với biến Y** với lambda được đề xuất (Rounded Pwr) = 0.17.

```r
## Kết quả code trên cho cặp biến tốt nhất gồm: age:children:smoker
## Thực hiện ANOVA với age:children:smoker
dependent_variable = c("charges")
factor_list = c("age", "children", "smoker")
n_way = nway_aov(dependent_variable, factor_list, data=data2)
print(n_way)

## Kiểm tra giả định mô hình
## Các nhóm phải có phân phối chuẩn hoặc gần chuẩn.
av_residual <- (rstandard(aov(data2$charges ~ data2$age + data2$children + data2$smoker)))
shapiro.test(av_residual)
## Kiểm định các nhóm có phương sai đồng nhất.
leveneTest(charges ~ interaction(age,children,smoker), data = data2)
```

1.2.3 Kiểm tra giả định mô hình

```r
av_residual <- (rstandard(aov(data2$charges ~ data2$age
                              + data2$children + data2$smoker)))
shapiro.test(av_residual)
```

```
                     Df Sum Sq Mean Sq  F value   Pr(>F)
age                   2  136.1    68.1  455.474  < 2e-16 ***
children              2    9.7     4.8   32.420 1.80e-14 ***
smoker                1  359.7   359.7 2407.279  < 2e-16 ***
age:children          4    7.9     2.0   13.146 1.67e-10 ***
age:smoker            2   15.1     7.5   50.363  < 2e-16 ***
children:smoker       2    1.2     0.6    4.011   0.0183 *
age:children:smoker   4    1.8     0.5    3.086   0.0153 *
Residuals          1319  197.1     0.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
4
5  leveneTest(charges ~ interaction(age,children,smoker), data = data2)
```

```
    Shapiro-Wilk normality test

data:  av_residual
W = 0.9338, p-value < 2.2e-16
```

Đặt giả thuyết:

H0: Dữ liệu có phân phối chuẩn

H1: Dữ liệu không có phân phối chuẩn

Vì p_value < 2.2e-16, ta bác bỏ H0, **dữ liệu không có phân phối chuẩn**.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group   17   3.092 2.182e-05 ***
      1319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Đặt Giả thuyết kiểm định thống kê ở các biến như sau:

H0: Các nhóm có phương sai đồng nhất

H1: Các nhóm có phương sai không đồng nhất

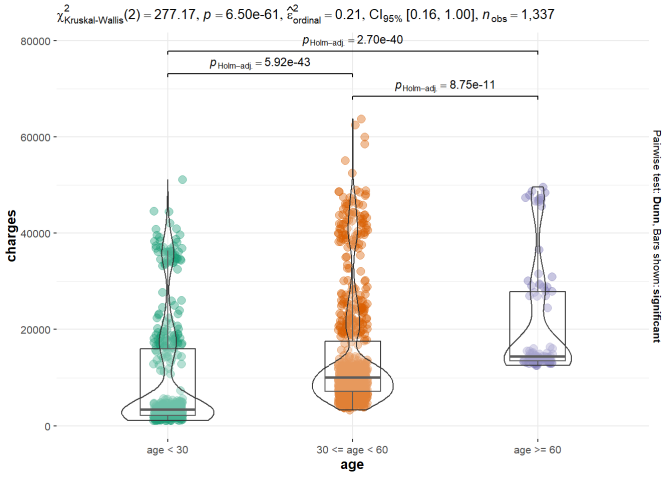
Sau khi thực hiện biến đổi biến Y và thực hiện kiểm định phương sai đồng nhất. **Bác bỏ H0** với p_value = 2.182e-05 < 0.05.

Kết luận: theo học viên, phân tích ANOVA k nhân tố không phù hợp cho mô hình này vì không thỏa giả định phần dư tuân theo phân phối chuẩn cũng như có phương sai đồng nhất giữa các nhóm. Điều này có thể đến từ việc dữ liệu chưa thu thập đủ thông tin, cần phải thu thập nhiều dữ liệu hơn. Một phần kết quả trên của em cũng chịu ảnh hưởng bởi cách phân chia các nhân tố (ở các biến age, children, smoker).

1.2.4 Đề xuất cải tiến/phân tích khác

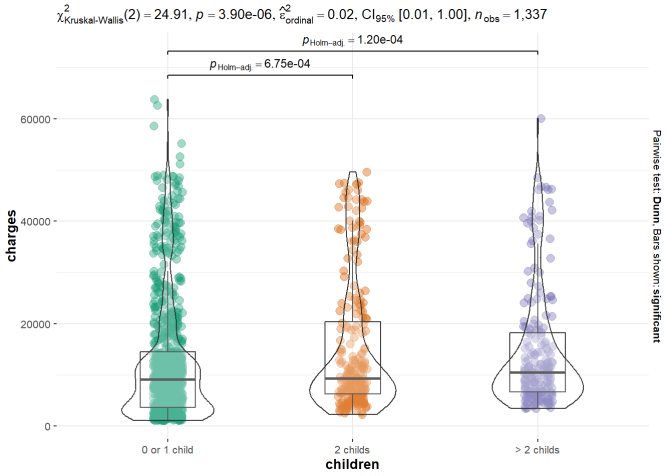
Trong trường hợp với tập dữ liệu này, học viên đề xuất sử dụng thống kê phi tham số cho các nhân tố ở các biến age, children, smoker với kiểm định Wilcox (biến có 2 nhân tố smoker) [1] và Kruskal-Wallis Test (biến có 3 nhân tố- age, children) [2]. Sau đó tiến hành phân tích hậu định (kiểm định Dunn) với package FSA.

- Biến age:
  H0: Không có sự khác biệt giữa 3 nhóm tuổi
  H1: Có ít nhất 1 nhóm tuổi khác với 2 nhóm tuổi còn lại
  Với p-value $<$ 2.2e-16 $<$ 0.05. Bác bỏ H0, có ít nhất 1 nhóm tuổi khác với 2 nhóm tuổi còn lại. Thực hiện Dunn test, học viên thấy được có sự khác biệt có ý nghĩa giữa từng nhóm tuổi với nhau.

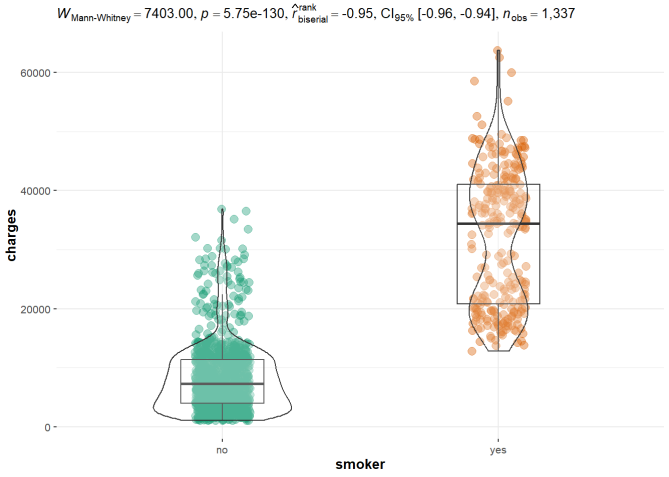


- Biến children:
  H0: Không có sự khác biệt giữa số lượng người phụ thuộc (con)
  H1: Có ít nhất 1 nhóm khác với 2 nhóm còn lại
  Với p-value = 3.899e-06 $<$ 0.05. Bác bỏ H0, có ít nhất 1 nhóm khác với 2 nhóm còn lại.
  Thực hiện Dunn test, học viên thấy được có sự khác biệt có ý nghĩa giữa các nhóm không có hoặc có 1 con với nhóm có 2 con trở lên. Không có sự khác biệt có ý nghĩa giữa nhóm có 2 con và nhóm có 2 con trở lên với biến charges.

- Biến smoker:
  H0: chi phí bảo hiểm của người hút thuốc và không hút thuốc bằng nhau
  H1: chi phí bảo hiểm giữa hai nhóm khác nhau Với p-value = 5.75e-130 $<$



2.2e-16 $<$ 0.05. Bác bỏ H0, kiểm định cho thấy có sự khác biệt giữa nhóm người hút thuốc và không hút thuốc đến biến charges.



Chương 2

Hoạt động 2

2.1 Giới thiệu đề tài

Dữ liệu xếp hàng đại học trên thế giới được nguồn từ Kaggle [3]. Đây là tập dữ liệu xếp hạng các trường đại học trên thế giới.
Mục tiêu thực hiện: Sử dụng tập dữ liệu trên để dự báo điểm tổng xếp hạng (Overall Scores). Theo học viên đánh giá, đây là một tập dữ liệu thú vị và khó để dự đoán điểm tổng Overall, lí do là vì số điểm tổng Overall chính xác như một con số chỉ được quy ước cho các trường trong vòng TOP 200 hoặc khi điểm tổng đạt giá trị $>$ [55,60]. Khi điểm tổng thấp hơn thì bảng xếp hạng sẽ không còn để số điểm chính xác mà chỉ cho khoảng giá trị- Vì cơ bản, bảng xếp hạng đại học này vốn được xây dựng cũng chỉ mang tính tham khảo tổng quát cho người sử dụng là người học, phụ huynh, doanh nghiệp,... Ví dụ: Trường Michigan State University có hạng 50 với điểm tổng là 64.2 trong khi trường University of Bayreuth có hạng 150 sẽ có điểm tổng là trong khoảng 47.4–49.2. Giải thích ý nghĩa các biến như sau:

- rank: Xếp hạng trường/viện
- ranking-institution-title: Tên trường đại học/viện nghiên cứu
- location: Quốc gia
- Overall scores: Điểm tổng
- Research Quality Score: Chỉ số về chất lượng nghiên cứu (ghi nhận chất lượng nghiên cứu như chỉ số trích dẫn, nội lực nghiên cứu và ảnh hưởng nghiên cứu ở trường/viện đến cộng đồng học thuật)
- Industry Score: Chỉ số công nghiệp (ghi nhận tác động, thu nhập từ việc thương mại hóa các nghiên cứu, bằng sáng chế,...)

- International Outlook: Điểm quốc tế hóa (ghi nhận về tỉ lệ sinh viên quốc tế, nhân viên quốc tế và các hoạt động hợp tác quốc tế)
- Research Environment Score: Chỉ số về môi trường nghiên cứu (đo lường và ghi nhận điểm về danh tiếng học thuật, thu nhập từ hoạt động và sản phẩm nghiên cứu).
- Teaching Score: Chỉ số giảng dạy (ghi nhận điểm về danh tiếng giảng dạy, tỉ lệ giảng viên/sinh viên, tỉ lệ tiến sĩ/cử nhân,...)

2.2 Tiền xử lí dữ liệu

Vì dữ liệu được xếp hạng theo thứ tự, nên trước hết, học viên quyết định xáo trộn dataset và đổi tên biến trước. Tập dữ liệu này không có dữ liệu duplicate vì mỗi trường chỉ được xếp hạng một lần.

```
set.seed(123)
data3 <- data3[sample(nrow(data3)),]
## Đổi tên biến
data3 <- data3 %>%
  rename(nation = location, overall = `Overall scores`,
         research_quality = `Research Quality Score`,
         industry = `Industry Score`, international = `International Outlook`,
         research_environment = `Research Environment Score`, teaching = `Teaching Score`)

## Tìm các giá trị bị thiếu tại biến `nation`
data3 %>% filter(is.na(nation))
```

Thông tin về dataset sau khi kiểm tra data profiling: Dữ liệu có 910 quan sát và 9 biến (3 biến định dạng character và 6 biến dạng numeric). Có 12 value ở biến nation bị khuyết. Xem xét và điền vào giá trị khuyết hợp lí (có thể sử dụng mode cho quốc gia xuất hiện nhiều nhất), dựa vào tên quốc gia mà tạo nên biến mới continent: châu lục nhằm xây dựng mô hình thay vì sử dụng tên quốc gia. Cần loại bỏ đi biến rank và ranking-institution-title. Các biến research_quality, industry, international, research_environment, teaching có giá trị mean và median không quá chênh lệch nhau. Kiểm tra dữ liệu khuyết, nhận thấy đây là trường hợp đặc biệt, việc dữ liệu bị khuyết ở biến nation là có chủ đích chứ không phải lí do khách quan làm dữ liệu bị khuyết (có thể là vì lí do chính trị). Các trường đại học có quốc gia bị khuyết này sau khi kiểm tra đều thuộc Liên Bang Nga. Vì vậy, thay vì tìm giá trị mode để thay thế, học viên quyết định điền giá trị khuyết bằng tên Russian.

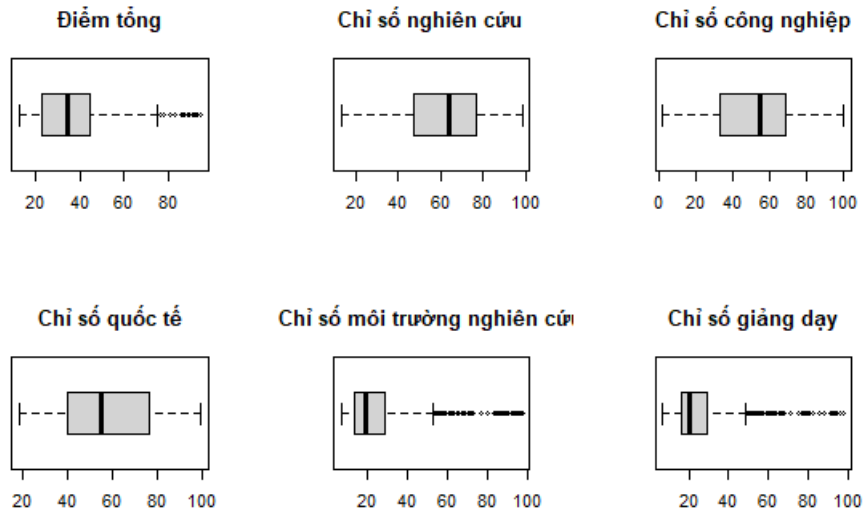
```
# A tibble: 12 × 9
    rank `ranking-institution-title`    nation overall research_quality industry
   <dbl> <chr>                          <chr>  <chr>              <dbl>    <dbl>
 1   890 Southern Federal University    <NA>   12.6-2…             29.9     18.1
 2   580 RUDN University                <NA>   31.3-3…             50.6     19.5
 3   885 Russian Presidential Academy … <NA>   12.6-2…             18.6     24.6
 4   679 ITMO University                <NA>   23.0-3…             56.3     41.8
 5   383 South Ural State University    <NA>   37.4-4…             91.3     19.3
 6   190 Peter the Great St Petersburg… <NA>   46.2-4…             72.1     24.2
 7   173 Ural Federal University        <NA>   47.4-4…             76.3     50
 8    87 HSE University                 <NA>   58.1                54.7     69.7
 9    62 Lomonosov Moscow State Univer… <NA>   61.4                43.3     76.6
10   530 Financial University under th… <NA>   31.3-3…             38       46.2
11   686 Kazan Federal University       <NA>   23.0-3…             42       22.2
12   570 Plekhanov Russian University … <NA>   31.3-3…             35.1     34.3
# i 3 more variables: international <dbl>, research_environment <dbl>,
#   teaching <dbl>
```

```r
## Điền giá trị khuyết với tên Russian
data3 <- data3 %>%
  mutate(nation = ifelse(is.na(nation), 'Russian', nation))

## Tạo biến continent với tên châu lục tương ứng với quốc gia
data3$continent <- countrycode(sourcevar = data3$nation,
                               origin = "country.name",
                               destination = "continent")

## bỏ đi biến rank, ranking-institution-title, nation
data3 <- data3 %>% select(-c(rank, 'ranking-institution-title', nation))
```

Vì overall có định dạng kí tự (character), cần được đổi về dạng số. Tuy nhiên, đối với các trường/viện nghiên cứu thuộc top dưới sẽ gây lỗi khi đổi về định dạng numeric. Một khó khăn khác là không thể để giá trị mean vì nhiều trường có cùng khoảng điểm overall như nhau. Một giải pháp đưa ra đó là: random ngẫu nhiên số điểm overall theo phân phối chuẩn với giá trị random sẽ nằm trong khoảng tương ứng mà khoảng điểm overall của trường/viện nhận được ban đầu (không gieo hạt để tránh các trường có cùng range điểm được gieo giống nhau). **Để đảm bảo thuận tiện cho việc sử dụng lại kết quả, học viên chỉ thực hiện duy nhất một lần và lưu dữ liệu lại với tên là ranking_cleaned.csv**.

```r
## Bỏ thao tác gieo hạt
set.seed(NULL)
## tạo biến overall_num để lưu các giá trị numeric
data3$overall_num <- as.numeric(data3$overall)

## tạo biến overcheck1 lưu chặng dưới
data3$overcheck1 <- substr(data3$overall, 1, 4)
data3$overcheck1 <- as.numeric(data3$overcheck1)
data3$overcheck1[!is.na(data3$overall_num)] <- NA

```

```r
## tạo biến overcheck2 lưu chặng trên
data3$overcheck2 <- substr(data3$overall, 6, 9)
data3$overcheck2 <- as.numeric(data3$overcheck2)

## tạo giá trị với random có phân phối chuẩn trong khoảng (overcheck1, overcheck2)
data3$overall_check <- round(rtruncnorm(n=1,a=data3$overcheck1, b=data3$overcheck2),2)

## merge overall_num và overall_check gán vào biến overall
data3$overall_num[is.na(data3$overall_num)] <- data3$overall_check[is.na(data3$overall_num)]
data3$overall <- data3$overall_num

## xóa đi các biến đã thực hiện
data3 <- data3 %>%
  select(-c(overall_num, overcheck1, overcheck2, overall_check))

## lưu dataset với tên ranking_cleaned.csv
write_csv(data3, "./data/ranking_cleaned.csv")
data3_clean <- read_csv("./data/ranking_cleaned.csv")

## tạo biến giả cho continent
data3_clean <- dummy_cols(data3_clean, select_columns = 'continent')
# Loại bỏ biến continent sau khi tạo biến giả
data3_clean <- subset(data3_clean, select = -continent)
```



Hình 2.1: Biểu đồ hộp

Kiểm tra boxplot chung ở từng biến dữ liệu ta có được kết luận, về mặt bằng chung trên bảng xếp hạng, các trường đại học/viện nghiên cứu có các hoạt động về nghiên cứu (research_quality), công nghiệp (industry), quốc tế (international) có phân phối chuẩn (có thể lệch nhẹ) nhưng vẫn khá đối xứng khi xét thấy giá trị mean và median tương đối gần nhau- không chênh lệch quá nhiều. Tuy nhiên ta cũng thấy được các dữ liệu ngoại lai chủ yếu xuất hiện trên các biến overall, research_environment, teaching. Điều này là hợp lí khi chỉ có một số ít các trường đại học/viện nghiên cứu là nơi được tập trung nguồn lực để đào tạo và nghiên cứu phát triển công nghệ lõi, công nghệ mới, các yếu tố này đã thể hiện qua giá trị có điểm tổng, điểm môi trường nghiên cứu, giảng dạy vượt trội, outstanding hầu hết các trường/viện còn lại.
Bộ dữ liệu nhìn chung được đánh giá là khách quan và thực tiễn.

2.3 Chia tập train - tập validation

```
set.seed(42)
#Tạo id column
data3_clean$id <- 1:nrow(data3_clean)

#sử dụng 80% cho tập training và 20% cho tập validation
train <- data3_clean %>% dplyr::sample_frac(0.8)
val  <- dplyr::anti_join(data3_clean, train, by = 'id')
train <- subset(train, select = -id)
val <- subset(val, select = -id)

## Xây dựng mô hình hồi quy tuyến tính bội với tất cả các biến
mod <- lm(overall ~ ., data = train)
summary(mod)
```

2.4 Chọn mô hình

```
model_ranking <- lm(overall ~ research_quality + industry
                    + international + research_environment
                    + teaching, data = train)
summary(model_ranking)

## Kiểm tra đa cộng tuyến
vif(model_ranking)
```

```
Call:
lm(formula = overall ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6889 -1.2189  0.2252  1.3449 13.2076

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -8.403188   0.558418 -15.048   <2e-16 ***
research_quality      0.297894   0.005609  53.108   <2e-16 ***
industry              0.057967   0.004660  12.439   <2e-16 ***
international         0.104136   0.005229  19.916   <2e-16 ***
research_environment  0.315540   0.014707  21.455   <2e-16 ***
teaching              0.339164   0.015302  22.165   <2e-16 ***
continent_Africa      0.342816   0.607566   0.564    0.573
continent_Americas    0.260441   0.440495   0.591    0.555
continent_Asia        0.283168   0.437455   0.647    0.518
continent_Europe      0.257279   0.415406   0.619    0.536
continent_Oceania           NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.281 on 718 degrees of freedom
Multiple R-squared:  0.9794,    Adjusted R-squared:  0.9791
F-statistic:  3787 on 9 and 718 DF,  p-value: < 2.2e-16
```

Xây dựng mô hình hồi quy tuyến tính giữa overall và tất cả các biến

Đặt giả thuyết:

- **H0**: $\beta_i = 0, \forall i$
- **H1**: $\exists \beta_i \neq 0$

  Với p-value: < 2.2e-16 ở kết quả trên, em kết luận bác bỏ H0, tồn tại ít nhất 1 $\beta_i \neq 0$. Vì vậy, em xây dựng mô hình thứ 2, lần này chỉ có các biến có ý nghĩa với p_value < 0.05: `research_quality`, `industry`, `international`, `research_environment`, `teaching`

  Ngoài ra, cá nhân em cũng khá bất ngờ khi vị trí địa lí của châu lục (có thể quốc gia sẽ có) lại không có ý nghĩa cho việc xây dựng mô hình dự báo điểm tổng cho chất lượng trường đại học/viện nghiên cứu.

Kiểm tra với mô hình chạy với các biến: research_quality, industry, international, research_environment, teaching, tất cả các biến đều có p_value <2e-16. Tuy nhiên khi kiểm tra đa cộng tuyến nhận được biến research_environment có VIF=7.95, học viên quyết định loại bỏ biến research_environment ra khỏi mô hình và chạy lại.

```
  research_quality              industry         international
          1.598503              1.536825              1.610942
research_environment              teaching
          7.953889              6.809575
```

```
model_ranking <- lm(overall ~ research_quality + industry
                    + international + teaching, data = train)
summary(model_ranking)
```

```
Call:
lm(formula = overall ~ research_quality + industry + international +
    teaching, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3503 -1.8960  0.1448  2.0365 12.2482

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -10.763920   0.396716  -27.13   <2e-16 ***
research_quality   0.315374   0.007105   44.39   <2e-16 ***
industry           0.090551   0.005523   16.39   <2e-16 ***
international      0.108415   0.006308   17.19   <2e-16 ***
teaching           0.629856   0.008903   70.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.934 on 723 degrees of freedom
Multiple R-squared:  0.9656,    Adjusted R-squared:  0.9654
F-statistic:  5078 on 4 and 723 DF,  p-value: < 2.2e-16
```

research_quality	industry	international	teaching
1.566333	1.370252	1.606861	1.460593

Không có hiện tượng đa cộng tuyến nữa, ta chọn mô hình này để dự báo điểm overall.

2.5 Kiểm tra giả định

```r
## 1. Kiểm tra sai số có tuân theo phân phối chuẩn hay không
shapiro.test(resid(model_ranking))

## 2. Kiểm tra giả định trung bình sai số mu=0
t.test(resid(model_ranking), mu = 0)

## 3. Kiểm tra tính ổn định của phương sai
ncvTest(model_ranking)
```

Như vậy, mô hình không thỏa mãn tính ổn định của phương sai. Học viên quyết định cần thực hiện các phép biến đổi biến (trasformation) với phương pháp Box-Cox và vì chỉ có duy nhất điều kiện 3 là tính ổn định của phương sai bị vi phạm, nên học viên chọn chỉ kiểm tra liệu có nên biến đổi đối với biến phụ thuộc overall hay không.

```r
summary(model <- powerTransform(overall ~ research_quality + industry
                    + international + teaching, data = train))
```

```
    Shapiro-Wilk normality test

data:  resid(model_ranking)
W = 0.99692, p-value = 0.1798
```

1. Kiểm tra sai số có tuân theo phân phối chuẩn hay không

Đặt giả thuyết:

H0: ε_i có phân phối chuẩn
H1: ε_i không có phân phối chuẩn

Ta không bác bỏ H0, p-value = 0.1845 > 0.05 **sai số của mô hình có phân phối chuẩn**

```
    One Sample t-test

data:  resid(model_ranking)
t = -2.0954e-16, df = 727, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.2129022  0.2129022
sample estimates:
    mean of x
-2.272357e-17
```

2. Kiểm tra giả định trung bình sai số $\mu = 0$

Đặt giả thuyết:

H0: $\mathrm{E}(\varepsilon_i) = 0$
H1: $\mathrm{E}(\varepsilon_i) \neq 0$

p_value = 1, không bác bỏ H0. **Trung bình sai số thỏa mãn** $\mathrm{E}(\varepsilon_i) = 0$

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 16.37145, Df = 1, p = 5.2063e-05
```

3. Kiểm tra tính ổn định của phương sai

Đặt giả thuyết

H0: phương sai không thay đổi
H1: phương sai thay đổi

Ta bác bỏ H0 với p_value < 5.2324e-05 < 0.05. **Mô hình không thỏa mãn tính ổn định của phương sai**, phương sai của sai số thay đổi.

```r
model_ranking <- lm((overall)^1.08 ~ research_quality + industry
                    + international + teaching, data = train)
summary(model_ranking)

## Kiểm tra lại giả định sau khi biến đổi biến Y
## 1. Kiểm tra sai số có tuân theo phân phối chuẩn hay không
shapiro.test(resid(model_ranking))

## 2. Kiểm tra giả định trung bình sai số mu = 0
t.test(resid(model_ranking), mu = 0)

## 3. Kiểm tra tính ổn định của phương sai
ncvTest(model_ranking)
```

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    1.0846        1.08       1.0281       1.1411

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 755.2951  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                           LRT df      pval
LR test, lambda = (1) 8.365487  1 0.0038241
```

Likelihood ratio test that transformation parameters are equal to 0

Để kiểm định việc có nên thực hiện phép biến đổi log-transformation với tất cả các biến hay không. Đặt giả thuyết:

H0: Cần thực hiện phép biến đổi log-transformation cho tất cả các biến.
H1: Không cần thực hiện phép biến đổi log-transformation cho tất cả các biến.
Với p_value < 2.22e-16. Bác bỏ H0, **ta không cần thực hiện phép biến đổi log-transformation cho tất cả các biến**.

Likelihood ratio test that no transformations are needed

Để kiểm định việc có nên thực hiện phép biến đổi biến hay không. Đặt giả thuyết:

H0: Không cần thực hiện phép transformation.
H1: Cần thực hiện tối thiểu một phép transformation.
Với p_value = 0.004 < 0.05. Bác bỏ H0, ta **cần thực hiện tối thiểu một phép transformation**. Như vậy em quyết định **thực hiện biến đổi biến với biến Y** với lambda được đề xuất (Rounded Pwr) = 1.08. và chọn đây là **mô hình phù hợp nhất**

2.6 So sánh kết quả với biến overall và Đề xuất cải tiến/phân tích khác

```r
y_hat <- predict(model_ranking, newdata = val)
## Trả lại giá trị ban đầu
y_hat <- y_hat^(1/1.08)
MSE <- mean((val$overall - y_hat)^2)

## Tương tự, sử dụng Decision Tree, XGboost
tree_model <- decision_tree(min_n=2) %>%
  set_engine("rpart") %>%
  set_mode(.,"regression")

tree_fit <-
  tree_model %>%
  fit(overall ~ ., data=train)

result <-
  val %>%
  select(overall) %>%
  bind_cols(predict(tree_fit,val))

MSE_tree <- mean((result$overall - result$.pred)^2)
```

```
    Shapiro-Wilk normality test

data:  resid(model_ranking)
W = 0.99752, p-value = 0.3501
```

1. Kiểm tra sai số có tuân theo phân phối chuẩn hay không

Đặt giả thuyết:

H0: ε_i có phân phối chuẩn
H1: ε_i không có phân phối chuẩn

Ta không bác bỏ H0, p-value = 0.3572 > 0.05 **sai số của mô hình có phân phối chuẩn**

```
    One Sample t-test

data:  resid(model_ranking)
t = 8.7e-16, df = 727, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.3014775  0.3014775
sample estimates:
  mean of x
1.33599e-16
```

2. Kiểm tra giả định trung bình sai số $\mu = 0$

Đặt giả thuyết:

H0: $\mathrm{E}(\varepsilon_i) = 0$
H1: $\mathrm{E}(\varepsilon_i) \neq 0$

p_value = 1, không bác bỏ H0. **Trung bình sai số thỏa mãn** $\mathrm{E}(\varepsilon_i) = 0$

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 9.955284, Df = 1, p = 0.0016039
```

3. Kiểm tra tính ổn định của phương sai

Đặt giả thuyết

H0: phương sai không thay đổi
H1: phương sai thay đổi

Ta bác bỏ H0 với p_value = 0.0015884 < 0.05. **Mô hình không thỏa mãn tính ổn định của phương sai**, phương sai của sai số thay đổi.

```
21
22  ## XGboost
23  train_control = trainControl(method = "cv", number = 5, search = "grid")
24  set.seed(42)
25  # Tạo thông số để thực hiện Grid search
26  gbmGrid <-  expand.grid(max_depth = c(3, 5, 7),
27                          nrounds = (1:10)*50,    # số lượng cây
28                          # cài đặt các value mặc định
29                          eta = 0.3,
30                          gamma = 0,
31                          subsample = 1,
32                          min_child_weight = 1,
33                          colsample_bytree = 0.6)
34
35  # huấn luyện mô hình XGboost Regression tree model
36  model <- (train(overall~.,
37                  data = train, method = "xgbTree",
38                  trControl = train_control,
```

```
                tuneGrid = gbmGrid, verbosity = 0))
pred_y = predict(model, val)

test_y = val$overall
MSE_xgboots <- mean((test_y - pred_y)^2)
MSE_xgboots
```

```
[1] "Linear Regression: MSE"
[1] 9.845632
[1] "Decision Tree: MSE"
[1] 32.97056
[1] "XGboost: MSE"
[1] 9.345697
```

Tương tự như bài tập 1 (hoạt động 1), sử dụng mô hình Cây Quyết định (Decision Tree) và XGBoost [4] để thực hiện dự báo Như vậy trong tập dữ liệu này thì mô hình XGboost đã cho kết quả tốt nhất (9.3456975 so với 9.8456322 của mô hình hồi quy tuyến tính).

Kết luận : Theo ý kiến cá nhân học viên, đây là một bài toán khá thú vị và đa dạng cách để xử lí.
Trước tiên, dữ liệu khuyết ở tập dữ liệu này được để khuyết một cách có chủ đích (bị bỏ đi tên quốc gia ở đây cụ thể là Nga) nên không dùng cách thông thường (thay thế giá trị mode). Tuy nhiên nhìn chung dữ liệu vẫn đảm bảo được tính khách quan, các giá trị outlier ở tập dữ liệu cũng đóng vai trò quan trọng khiến chúng không thể bị loại bỏ (vì các outlier này là các chỉ số thuộc các trường đại học/viện nghiên cứu hàng đầu nổi bật trong giảng dạy và nghiên cứu). Biến phụ thuộc mà học viên chọn cũng khá đặc biệt khi một phần là điểm số cụ thể và một phần là các khoảng điểm để xếp hạng ranking gây khó khăn để tìm ra được cách xử lí. Một khó khăn khác với mô hình này đó là khác với bài tập 1 (hoạt động 1) khi tính ổn định phương sai không thỏa mãn và có xuất hiện đa cộng tuyến.

Tài liệu tham khảo

[1] Antoine Soetewey, Wilcoxon test in R: how to compare 2 groups under the non-normality assumption?, 2020, <https://statsandr.com/blog/wilcoxon-test-in-r-how-to-compare-2-groups-under-the-non-normality-assumption/>.

[2] Antoine Soetewey, Kruskal-Wallis test, or the nonparametric version of the ANOVA, 2020, <https://statsandr.com/blog/kruskal-wallis-test-nonparametric-version-anova/>.

[3] Rafsun Ahmad, Kaggle, World All University Ranking Factors, 2023 <https://www.kaggle.com/datasets/rafsunahmad/world-all-university-ranking-factors/data>.

[4] Kuhn, M. and Silge, J., 2022. Tidy modeling with R. " O'Reilly Media, Inc.".

[5] GeeksforGeeks, Levene's Test in R Programming, 2020 <https://www.geeksforgeeks.org/levenes-test-in-r-programming/> .