Báo cáo SLTK

February 6, 2022

1 Chương 1. Quan sát, xử lý dữ liệu

Đặt lại work directory

[103]:
```
setwd("D:/survey")
getwd()
```

'D:/survey'

Các thư viện cần sử dụng

[104]:
```
library(plyr);
library(dplyr)
library(BSDA)
library(Rmisc)
library(ggpubr)
library(ggplot2)
```

```
Warning message:
"package 'ggpubr' was built under R version 3.6.3"
```

```
Error: package or namespace load failed for 'ggpubr' in loadNamespace(j <-␣
 ↪i[[1L]], c(lib.loc, .libPaths()), versionCheck = vI[[j]]):
 namespace 'broom' 0.5.2 is being loaded, but >= 0.7.4 is required
Traceback:

1. library(ggpubr)
2. tryCatch({
 .     attr(package, "LibPath") <- which.lib.loc
 .     ns <- loadNamespace(package, lib.loc)
 .     env <- attachNamespace(ns, pos = pos, deps, exclude, include.only)
 . }, error = function(e) {
 .     P <- if (!is.null(cc <- conditionCall(e)))
 .         paste(" in", deparse(cc)[1L])
 .     else ""
 .     msg <- gettextf("package or namespace load failed for %s%s:\n %s",
 .         sQuote(package), P, conditionMessage(e))
 .     if (logical.return)
 .         message(paste("Error:", msg), domain = NA)
```

```
 .     else stop(msg, call. = FALSE, domain = NA)
 . })
3. tryCatchList(expr, classes, parentenv, handlers)
4. tryCatchOne(expr, names, parentenv, handlers[[1L]])
5. value[[3L]](cond)
6. stop(msg, call. = FALSE, domain = NA)
```

Nhập dữ liệu từ file CSV

[105]:
```
responses <- read.csv("responses.csv", header = TRUE)
re <- data.frame(responses)
save(re, file = "re.rda")
```

Ta sẽ tách thành hai data frame nhỏ là male cho đối tượng nam, female cho đối tượng nữ. Sau đó ta tạo tiếp một data frame nhỏ cho người là Data Scientist.

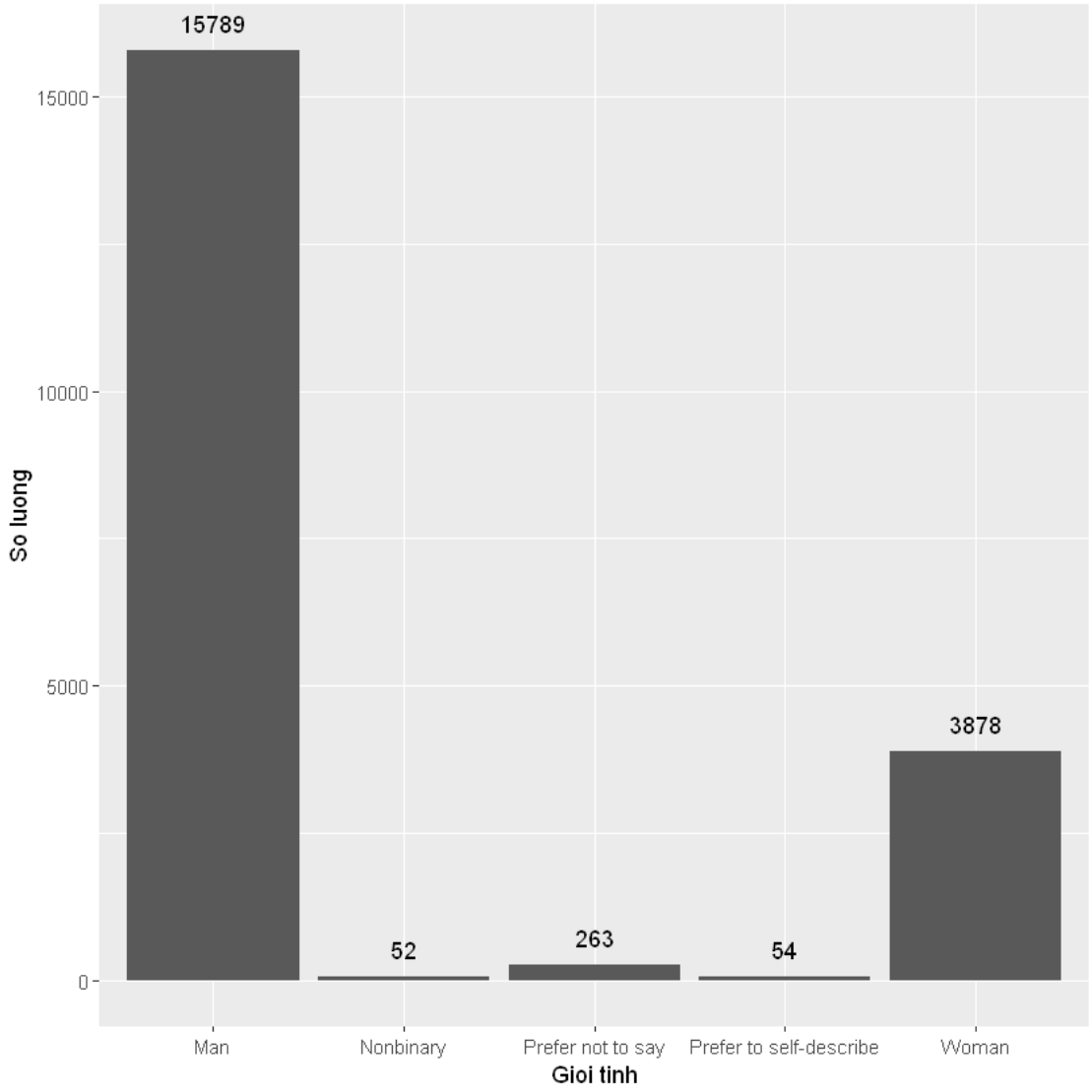
[106]:
```
dasc_all <- filter(re,Q5 =="Data Scientist")
male <- filter(responses,Q2== 'Man')
female <- filter(responses,Q2 == "Woman")
ma_dasc <- filter(male,Q5 =="Data Scientist")
fe_dasc <- filter(female,Q5 =="Data Scientist")
```

Ta vẽ biểu đồ tần suất phân phối giới tính để có hình dung ban đầu.

[107]:
```
dasc_summary <-
    re                          %>%
    group_by(Q2)               %>%
    summarise(name_count = n())
dasc_summary
```

Q2	name_count
Man	15789
Nonbinary	52
Prefer not to say	263
Prefer to self-describe	54
Woman	3878

[108]:
```
ggplot(dasc_summary, aes(Q2, name_count)) +
    geom_bar(stat = 'identity') +
    labs(x = 'Gioi tinh', y = 'So luong')+
geom_text(aes(label = name_count), vjust = -1)
```



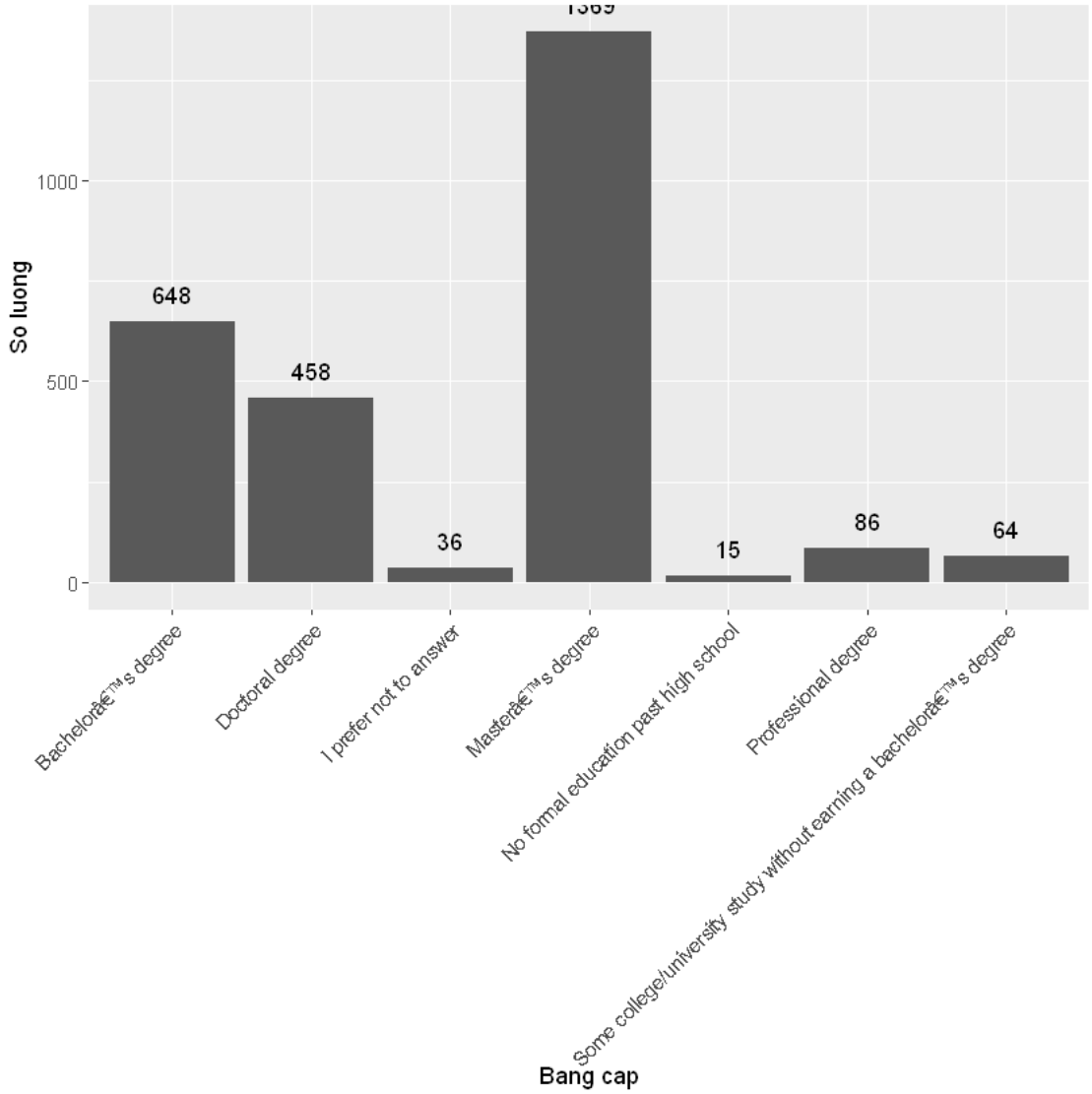
Có thể thấy rõ ràng nam vẫn chiếm đa số trong lĩnh vực công nghệ thông tin.

Một yếu tố em quan tâm là bằng cấp cao nhất cần có để trở thành Data Scientist

[109]:
```
cer_summary <-
    dasc_all                          %>%
    group_by(Q4)                %>%
    summarise(name_count = n())
cer_summary
```

Q4	name_count
Bachelorâ€™s degree	648
Doctoral degree	458
I prefer not to answer	36
Masterâ€™s degree	1369
No formal education past high school	15
Professional degree	86
Some college/university study without earning a bachelorâ€™s degree	64

[110]:
```
ggplot(cer_summary, aes(Q4, name_count)) +
    geom_bar(stat = 'identity') +
    labs(x = 'Bang cap', y = 'So luong')+
geom_text(aes(label = name_count), vjust = -1)+
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Có thể thấy bằng thạc sĩ chiếm nhiều nhất, sau đó là bằng cử nhân.

2 Chương 2. Ước lượng tham số

2.1 Ước lượng điểm cho sai số tiêu chuẩn khi biết phương sai là $\sigma_{\overline{x}} = \sigma/\sqrt{n}$

[111]:
```
std <- function(x, na.rm = TRUE){
v = var(x, na.rm = TRUE)
sqrt(v / length(x))
}
```

Ước lượng cho sai số tiêu chuẩn của độ tuổi nữ Data Scientist

[112]:
```
std(as.numeric(fe_dasc$Q1))
```

0.395228408471768

2.2 Ước lượng điểm cho tỷ lệ p là $\widehat{P} = \dfrac{x}{n}$

Ước lượng điểm cho tỷ lệ nữ theo nghề Data Scientist

[113]:
```
x = length(na.omit(as.numeric(fe_dasc$Q1)))
n <- length(na.omit(as.numeric(dasc_all$Q1)))
p = x / n
p
```

0.163677130044843

2.3 Ước lượng điểm cho sai số tiêu chuẩn khi chưa biết phương sai $\sigma_{\widehat{\Theta}}$ là $\widehat{\sigma}_{\overline{x}} = \dfrac{s}{\sqrt{n}}$.

[114]:
```
sd.p = function(x){sd(x, na.rm = TRUE)*sqrt((length(x[!is.na(x)])-1)/length(x[!
 ↪is.na(x)]))}
sigma_x = sd.p(as.numeric(fe_dasc$Q1)) / sqrt(length(as.numeric(na.
 ↪omit(fe_dasc$Q1))))
sigma_x
```

0.394776976675706

- Bây giờ, ta cần ước lượng khoảng tin cậy của thu nhập trung bình và phương sai σ^2 chưa biết với độ tin cậy 95%.

2.4 Ước lượng điểm và khoảng tin cậy cho kỳ vọng, phương sai chưa biết, cỡ mẫu $n \geq 30$

Ước lượng điểm cho thu nhập trung bình của nữ ở vị trí Data Scientist.

[115]:
```
mean(fe_dasc$Q24, na.rm = TRUE)
```

44232.0583333333

Viết hàm tìm độ lệch chuẩn tổng thể

[116]:
```
sd.p = function(x){sd(x, na.rm = TRUE)*sqrt((length(x[!is.na(x)])-1)/length(x[!
 ↪is.na(x)]))}
```

Viết hàm tìm khoảng tin cậy

[117]:
```
mean.interval = function(data, conf = 0.95){
  mean <- mean(data, na.rm = TRUE)
  c(mean)

  #Tim co mau, sai so chuan SE
  n <- length(na.omit(data))
  sd <- sd(na.omit(data)) #Do lech chuan
  se <- sd/sqrt(n)
  c(se)

  #Tim so alpha, bac tu do va t
  alpha = 1 - conf
  d.f = n - 1
  t.score = qt(p = alpha/2, d.f,lower.tail = F)
  t.score

  #Tim can tren duoi cua khoang tin cay doi xung
  margin.error <- t.score * se
  lower <- mean - margin.error
  upper <- mean + margin.error
  c(lower,upper)
}
mean.interval(fe_dasc$Q24)
```

1. 38102.4661438231 2. 50361.6505228436

Hoặc sử dụng z.test()

[118]:
```
z.test(fe_dasc$Q24, NULL, mu = mean(fe_dasc$Q24, na.rm = TRUE), alt = "t",
       sigma.x = sd.p(fe_dasc$Q24), sigma.y = NULL, conf = 0.95 )
```

```
        One-sample z-Test

data:  fe_dasc$Q24
z = 0, p-value = 1
alternative hypothesis: true mean is not equal to 44232.06
95 percent confidence interval:
 38131.62 50332.50
```

```
sample estimates:
mean of x
 44232.06
```

Hoặc sử dụng hàm CI()

[119]:
```
CI(na.omit(fe_dasc$Q24), 0.95)
```

upper 50361.6505228436 **mean** 44232.0583333333 **lower** 38102.4661438231

2.5 Ước lượng điểm $\widehat{\sigma^2} = s^2$ và khoảng tin cậy cho phương sai mẫu σ^2

[120]:
```
var.interval = function(data, conf = 0.95) {
  df = length(na.omit(data)) - 1
  df
  v = var(data, na.rm = TRUE)
  lower = v * df / qchisq((1 - conf) / 2, df, lower.tail = FALSE)
  upper = v * df / qchisq(1 - (1 - conf)/2, df, lower.tail = FALSE)
  c(lower = lower, variance = v, upper = upper)
}
var.interval(fe_dasc$Q24, conf = 0.95)
```

lower 3037209746.68527 **variance** 3497327244.7626 **upper** 4071130670.16062

2.5.1 Ước lượng điểm s và khoảng tin cậy cho độ lệch chuẩn σ

[121]:
```
sd.interval = function(data, conf = 0.95) {
  df = length(na.omit(data)) - 1
  df
  v = var(data, na.rm = TRUE)
  lower = v * df / qchisq((1 - conf) / 2, df, lower.tail = FALSE)
  upper = v * df / qchisq(1 - (1 - conf)/2, df, lower.tail = FALSE)
  c(lower = sqrt(lower), sd = sqrt(v), upper = sqrt(upper))
}
sd.interval(fe_dasc$Q24, conf = 0.95)
```

lower 55110.8859181674 **sd** 59138.2046122691 **upper** 63805.4125459637

3 Chương 3. Kiểm định giả thuyết thống kê

3.1 Viết hàm kiểm định giả thuyết của giá trị trung bình

[122]:
```
kiemdinh <- function(data, mu0, n, alpha, doithuyet,delta){
    x_bar <- mean(data, na.rm = TRUE)
    se <- sd(data, na.rm = TRUE)/sqrt(n)
    u <- (x_bar - mu0)/se
    w1 <- qnorm(1 - alpha/2)
```

```
    w2 <- qnorm(1 - alpha)
    if (doithuyet == "="){
      if (abs(u) > w1) print("Bac bo H_0")
      else print("Chua co co so de bac bo H_0")
    }
    if (doithuyet == ">") {
      if (u > w2) print("Bac bo H_0")
      else print("Chua co co so de bac bo H_0")
    }
    if (doithuyet == "<") {
      if (u < -w2) print("Bac bo H_0")
      else print("Chua co co so de bac bo H_0")
    }
    if(delta != 0){
      be_ta <- pnorm(w1 - delta*sqrt(n)/sigma) - pnorm(-w1 - delta*sqrt(n)/
 ↪sigma)
      m <- ((qnorm(w1)+qnorm(1-be_ta)**2)*(sigma**2))/(delta**2)
      cat("Sai lam loai II la: ", be_ta)
      cat("co mau toi thieu la: ", m)}
}
```

- Bài toán: Kiểm định giả thuyết lương trung bình của nữ Data Scientist là $\mu_0 = 139,542$ với độ tin cậy 95%, $\sigma = 0, \delta = 0$

[123]:
```
fe_dasc1 <- na.omit(fe_dasc)
#fe_dasc1
kiemdinh(fe_dasc1$Q24, 139542, length(fe_dasc1$Q24), 0.05,"=", 0)
```

```
[1] "Bac bo H_0"
```

3.2 Viết hàm kiểm định phương sai và độ lệch chuẩn tổng thể

[124]:
```
sd.p = function(x){sd(x, na.rm = TRUE)*sqrt((length(x[!is.na(x)])-1)/length(x[!
 ↪is.na(x)]))}
p_sai <- function(n, s, sigma_0, alpha, doithuyet){
  x <- (n-1)*(s**2)/(sigma_0**2)
  w1 <- qchisq(alpha/2, df = n-1)
  w2 <- qchisq(1-alpha/2, df = n-1)
  w3 <- qchisq(alpha, df = n-1)
  w4 <- qchisq(1-alpha, df = n-1)
  if (doithuyet == "="){
    if (x<w1 && x>w2) print("bac bo H_0")
    else print("Chua co co so de bac bo H_0")
  }
  if (doithuyet == ">"){
    if (x>w4) print("bac bo H_0")
    else print("Chua co co so de bac bo H_0")
```

```
  }
  if (doithuyet == "<"){
    if (x<w3) print("bac bo H_0")
    else print("Chua co co so de bac bo H_0")
  }
}
```

- Bài toán: Kiểm định xem phương sai có lớn hơn $\sigma_0^2 = 250000$ không với mức ý nghĩa 0.05

[125]:
```
p_sai(length(fe_dasc1$Q24), sd.p(fe_dasc1$Q24), sqrt(250000), 0.05, ">")
```

```
[1] "bac bo H_0"
```

4 Viết hàm so sánh hai kỳ vọng

[126]:
```
t.test(x, y, alt, conf)
```

```
Error in match.arg(alternative): object 'alt' not found
Traceback:

1. t.test(x, y, alt, conf)
2. t.test.default(x, y, alt, conf)
3. match.arg(alternative)
```

So sánh kỳ vọng của lương Data Scientist nam và nữ với độ tin cậy 95%

[127]:
```
ma_dasc1 <- na.omit(ma_dasc)
t.test(fe_dasc1$Q24,ma_dasc1$Q24, alt = "t", conf = 0.95 )
```

```
        Welch Two Sample t-test

data:  fe_dasc1$Q24 and ma_dasc1$Q24
t = -3.5522, df = 650.96, p-value = 0.0004096
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20066.420  -5779.258
sample estimates:
mean of x mean of y
 44232.06  57154.90
```

4.1 So sánh hai tỷ lệ

So sánh tỷ lệ giữa nam và Data Scientist với độ tin cậy 95%.

[128]:
```
ma_hundred <- ma_dasc1[c(1:200),c(2,11)]
fe_hundred <- fe_dasc1[c(1:200),c(2,11)]
fe_hundred <- fe_hundred %>%
  rename(Q2_1=Q2,
       Q24_1= Q24)
hundred = cbind(ma_hundred,fe_hundred)
size <- length(hundred$Q24_1)
size2 <- length(hundred$Q24)
```

Chạy hàm so sánh

[129]:
```
prop.test(x = c(438,2194), n = c(2676, 2676), alt = "t", conf = 0.95, correct =␣
 ↪FALSE)
```

```
        2-sample test for equality of proportions without continuity
        correction

data:  c(438, 2194) out of c(2676, 2676)
X-squared = 2305.2, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.6764146 -0.6359920
sample estimates:
   prop 1    prop 2
0.1636771 0.8198804
```

4.2 Viết hàm phân tích phương sai một nhân tố

- Kiểm định giả thuyết: Thu nhập của người làm trong lĩnh vực công nghệ thông tin là như nhau ở tất cả các nước. Ta xét thu nhập của người ở ba nước Việt Nam, Indonesia và Phillipines

[130]:
```
country <- read.csv(file.choose(), header = TRUE,stringsAsFactors=T)
```

[131]:
```
levels(country$Country)
country <- country %>%
  rename(Compensation=ï..Compensation)
names(country)
```

1. 'Indonesia' 2. 'Philippines' 3. 'Viet Nam'

1. 'Compensation' 2. 'Country'

[132]:
```
country.aov <- aov(Compensation ~ Country, data=country)
summary(country.aov)
```

```
             Df    Sum Sq   Mean Sq F value Pr(>F)
Country       2 3.450e+09 1.725e+09   0.363  0.696
```

```
Residuals   258 1.225e+12 4.749e+09
275 observations deleted due to missingness
```

[133]:
```r
country.aov <- lm(country$Compensation ~ country$Country)
summary(country.aov)
```

```
Call:
lm(formula = country$Compensation ~ country$Country)

Residuals:
   Min     1Q Median     3Q    Max
-17322 -16256  -9012  -3075 941947

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                    17850       6115   2.919  0.00382 **
country$CountryPhilippines     -2532      11424  -0.222  0.82477
country$CountryViet Nam        -8264       9727  -0.850  0.39631
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68910 on 258 degrees of freedom
  (275 observations deleted due to missingness)
Multiple R-squared:  0.002808,        Adjusted R-squared:  -0.004922
F-statistic: 0.3633 on 2 and 258 DF,  p-value: 0.6957
```

4.3 Kiểm định phi tham số

4.3.1 Kiểm định xem X có phân phối Poisson không

[134]:
```r
ps <- function(alpha, X, N){
  n <- N
  x <- mean(X)
  X_qs <- 0
  p <- vector(length = length(N))
  for (i in 1:N){
    p[i] <- ((2**(X[i]))*exp(-2)/factorial(X[i]))
    X_qs <- X_qs + ((N[i] - n*p[i])**2)/(n*p[i])
  }
  w <- qchisq(1-alpha, length(N) - 2)
  if (X_qs > w){
    cat("Co co so bac bo H0, khong the xem X co phan phoi poisson voi tham so␣
 ↪lamda = ",x)
  }
  else cat("Chua co co so bac bo H0, co the xem X co phan phoi poisson voi tham␣
 ↪so lamda = ",x)
```

```
}
```

4.3.2 Kiểm tra xem X có phân phối chuẩn không

[135]:
```
shapiro.test(ma_dasc1$Q24)
```

```
        Shapiro-Wilk normality test

data:  ma_dasc1$Q24
W = 0.616, p-value < 2.2e-16
```

5 Chương 4. Phân tích tương quan và hồi quy

5.1 Xác định hệ số tương quan và mô hình hồi quy tuyến tính

- Khảo sát 4706 người làm trong lĩnh vực Data Science về độ tuổi Q1, số năm kinh nghiệm lập trình Q6, độ lớn công ty Q20, số người trong công ty làm trong lĩnh vực Data Science Q21, thu nhập hàng năm Q24.

[136]:
```
correlation <- read.csv("responses_correlation.csv", header=TRUE)
correlation
```

ï..Q24	Q1	Q6	Q20	Q21
119026	28	10	93028	21
141224	29	10	382	5
74565	27	7	7355	24
97120	27	0	31	3
725	25	0	76	1
14877	27	0	28	3
16644	25	7	7674	3
4962	24	3	9432	2
721647	19	4	30	22
986	29	9	45	2
69246	28	3	9532	24
73683	27	1	43	3
641	18	0	21	23
762	27	0	33	2
545	29	1	72	2
658	21	0	30	0
559	22	1	636	3
42215	55	12	27	3
86111	25	13	6377	12
13649	23	3	54584	24
93350	47	10	8905	9
559	23	0	35325	0
7721	18	0	7419	4
728	26	8	88111	12
122596	26	5	92271	24
278645	49	11	7273	21
82824	28	8	5602	3
5936	26	1	37938	4
512	20	0	26	4
4528	26	5	40	4
...	...	...	...	...
869	26	1	39	4
842	25	0	42	1
1993	28	0	28	1
8466	44	0	22	2
22447	27	4	420	2
22127	25	3	215	4
13388	28	2	6400	19
85621	29	1	384	7
25330	22	0	48	0
80569	29	20	5772	16
31898	25	18	13957	23
369783	57	5	30	21
112203	29	0	2313	24
1720	28	5	64517	1
5420	21	2	45	4
33155	45	6	33468	17
55836	28	2	77056	23
651	24	4	45	0
50705	40	13	424	14
17818	57	0	3524	4
667	28	0	28	0

[137]:
```
cor(exp, salary)
```

0.340436837821418

Ta ước lượng được hệ số tương quan là 0.3404

[138]:
```
age <- (correlation$Q1) #tuoi
exp <- (correlation$Q6) #so nam kinh nghiem
company <- (correlation$Q20) #so luong nhan vien cong ty
dasc <- (correlation$Q21) #so nguoi lam lien quan Data Science trong cong ty
salary <- (correlation$ï..Q24) #luong hang nam
cor_n <- length(correlation$ï..Q24)
```

[139]:
```
cor <- lm(salary ~ age + exp + company + dasc)
summary(cor)
```

```
Call:
lm(formula = salary ~ age + exp + company + dasc)

Residuals:
    Min      1Q  Median      3Q     Max
-139784  -28814  -14297   15509  955106

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.580e+03  3.283e+03   0.786    0.432
age         4.331e+02  1.104e+02   3.921 8.94e-05 ***
exp         3.061e+03  1.717e+02  17.823  < 2e-16 ***
company     6.088e-03  4.197e-02   0.145    0.885
dasc        1.434e+03  1.262e+02  11.356  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64870 on 4700 degrees of freedom
Multiple R-squared:  0.1461,        Adjusted R-squared:  0.1454
F-statistic:   201 on 4 and 4700 DF,  p-value: < 2.2e-16
```

Kết quả phân tích cho thấy hệ số $\hat{\alpha} = 2580, \hat{\beta}_1 = 433.1, \hat{\beta}_2 = 3061, \hat{\beta}_3 = 0.0061, \hat{\beta}_4 = 1434$

Phương trình hồi quy tuyến tính: $\hat{y} = 2580 + 433.1x_1 + 3061x_2 + 0.0061x_3 + 1434x_4$

5.1.1 Lập bảng phân tích ANOVA

[140]:
```
sse <- sum((fitted(cor) - salary)^2)
ssr <- sum((fitted(cor) - mean(salary))^2)
sst <- sse + ssr
msr <- ssr/4 # k = 4
```

```
mse <- sse/(cor_n-4-1)
f_qs <- msr/mse
cat("SSE:", sse, "  SSR:", ssr, "  SST:", sst, "  MSE:", mse, "  MSR:", msr, "
 ↪f_qs:", f_qs)
```

```
SSE: 1.97804e+13   SSR: 3.384546e+12   SST: 2.316494e+13   MSE: 4208595199
MSR: 846136492716   f_qs: 201.0496
```

Nguồn	Bậc tự do	SS	MS	F
Hồi quy (R)	4	3.38×10^{12}	4208595199	201.0496
Sai số (E)	4700	1.98×10^{12}	846136492716	
Tổng	4704	5.36×10^{12}	850345087915	

5.1.2 Nhận xét về sự phù hợp của mô hình hồi quy

[141]:
```
F <- qf(1-0.05, df1 = 4, df2 = cor_n-4-1)
cat("F:", F)
```

```
F: 2.373823
```

$f_{qs} = 201.0496 > 2.373823$ nên bác bỏ giả thuyết H_0

[143]:
```
Rsquared <- ssr/sst
R <- sqrt(Rsquared)
cat("Rsquared:", Rsquared, "  R:", R)
```

```
Rsquared: 0.1461064   R: 0.3822386
```

$R = 0.3822$ nên mối tương quan tuyến tính giữa biến phụ thuộc và các biến giải thích là trung bình.

$R^2 = 0.1461$, có nghĩa là 14,61% tổng biến động của thu nhập của người làm việc liên quan đến Data Science đã được giải thích bằng các yếu tố độ tuổi, số năm kinh nghiệm, độ lớn công ty, số người trong công ty làm cùng lĩnh vực. Vậy mô hình hồi quy chưa phù hợp.

5.2 Kiểm định và ước lượng cho hệ số hồi quy, hệ số tương quan

5.2.1 Ước lượng và kiểm định hệ số tương quan

- Ước lượng và kiểm định hệ số tương quan giữa thu nhập và số năm kinh nghiệm của các nhà Data Scientist với độ tin cậy 95%.

[144]:
```
cor.test(x = exp, y = salary, alt = "t", conf = 0.95)
```

```
        Pearson's product-moment correlation

data:  exp and salary
t = 24.83, df = 4703, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
 0.3149253 0.3654568
sample estimates:
      cor
0.3404368
```

Vậy bác bỏ có sự tương quan tuyến tính thuận giữa thu nhập và số năm kinh nghiệm, mức độ tương quan trung bình. khoảng tin cậy của hệ số tương quan là $(0.3149, 0.3655)$.

5.2.2 Kiểm định hệ số hồi quy

- Kiểm định giả thuyết: $H_0 : \beta_j = 0,\ j = 1,2,3,4$ với mức ý nghĩa $\alpha = 0.05$

$H_1 : \exists\ \beta_j \neq 0$

[145]:
```
cor_test <- function(beta_j, SE_beta_j, n, k, alpha){
  t <- qt(1-alpha/2, df = n-k-1)
  t_qs <- beta_j/SE_beta_j
  if(t_qs<-t || t_qs>t){
    "Bac bo gia thuyet H0 : beta_j = 0."
  }
  else{
    "Chua du co so bac bo gia thuyet H0 : beta_j = 0."
  }
}
```

Kiểm định hệ số β_1

[146]:
```
cor_test(433.1,110.4,cor_n,4,0.05)
```

'Bac bo gia thuyet H0 : beta_j = 0.'

Kiểm định hệ số β_2

[147]:
```
cor_test(3061,171.7,cor_n,4,0.05)
```

'Bac bo gia thuyet H0 : beta_j = 0.'

Kiểm định hệ số β_3

[148]:
```
cor_test(0.0061,0.042,cor_n,4,0.05)
```

'Bac bo gia thuyet H0 : beta_j = 0.'

Kiểm định hệ số β_4

[149]:
```
cor_test(1434,126.2,cor_n,4,0.05)
```

'Bac bo gia thuyet H0 : beta_j = 0.'

[150]:
```
summary(cor)
```

```
Call:
lm(formula = salary ~ age + exp + company + dasc)

Residuals:
    Min      1Q  Median      3Q     Max
-139784  -28814  -14297   15509  955106

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.580e+03  3.283e+03   0.786    0.432
age         4.331e+02  1.104e+02   3.921 8.94e-05 ***
exp         3.061e+03  1.717e+02  17.823  < 2e-16 ***
company     6.088e-03  4.197e-02   0.145    0.885
dasc        1.434e+03  1.262e+02  11.356  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64870 on 4700 degrees of freedom
Multiple R-squared:  0.1461,        Adjusted R-squared:  0.1454
F-statistic:   201 on 4 and 4700 DF,  p-value: < 2.2e-16
```

$p-$ giá trị của biến company $(0.885 > 0.05)$ cho thấy có cơ sở để chấp nhận giả thuyết $\beta_3 = 0$, có nghĩa là độ lớn của công ty không ảnh hưởng đến lương.

- Ta thử loại bỏ biến company.

[151]:
```
cor_2 <- lm(salary ~ age + exp + dasc)
summary(cor_2)
```

```
Call:
lm(formula = salary ~ age + exp + dasc)

Residuals:
    Min      1Q  Median      3Q     Max
-139885  -28826  -14324   15577  955080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2609.0     3276.5   0.796    0.426
age            432.5      110.4   3.919 9.03e-05 ***
exp           3061.7      171.6  17.840  < 2e-16 ***
dasc          1441.2      115.1  12.517  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64870 on 4701 degrees of freedom
```

```
Multiple R-squared:  0.1461,        Adjusted R-squared:  0.1456
F-statistic: 268.1 on 3 and 4701 DF,  p-value: < 2.2e-16
```

[152]:
```r
sse2 <- sum((fitted(cor_2) - company)^2)
ssr2 <- sum((fitted(cor_2) - mean(company))^2)
sst2 <- sse + ssr
Rsquared2 <- ssr2/sst2
R2 <- sqrt(Rsquared2)
cat("SSE:", sse2, "  SSR:", ssr2, "  SST:", sst2, "  Rsquared2:", Rsquared2)
```

```
SSE: 9.975349e+12   SSR: 8.575033e+12   SST: 2.316494e+13   Rsquared2: 0.3701729
```

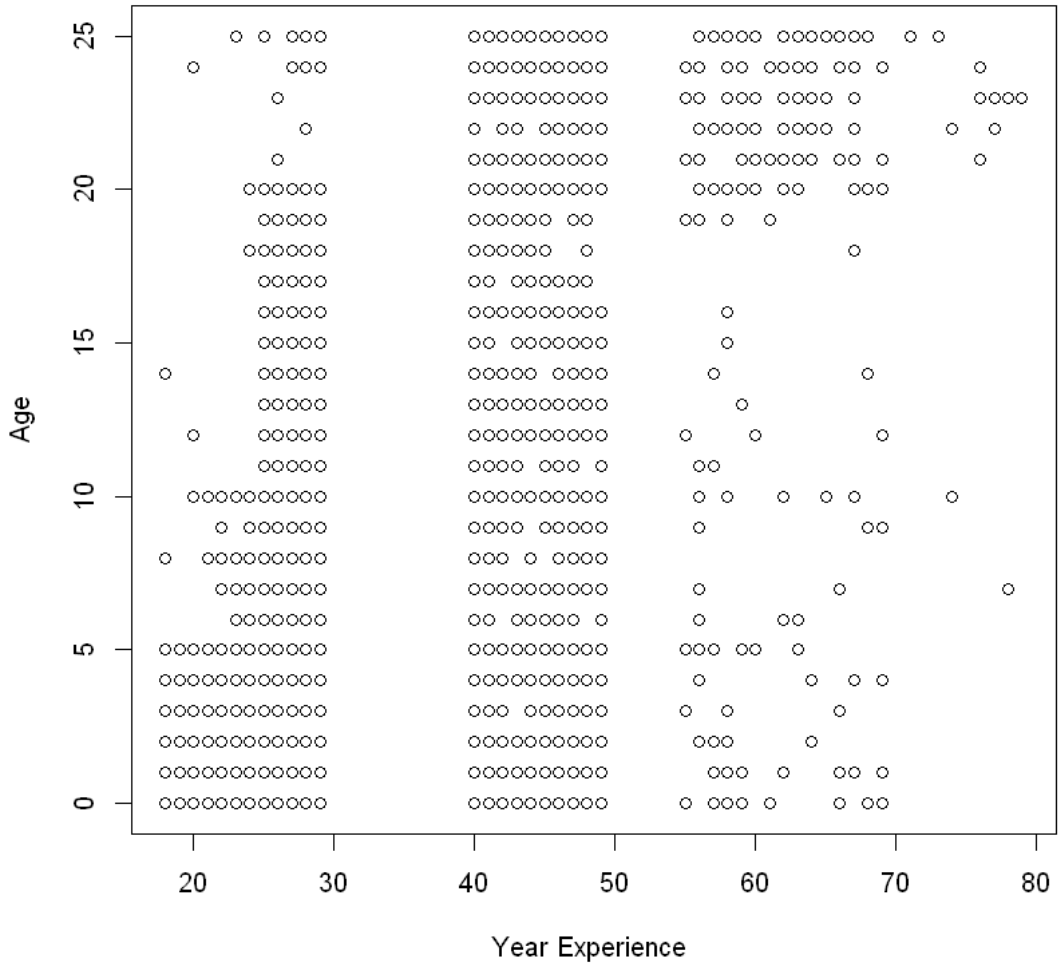
Ta thấy R^2 ở mô hình mới là 0.3702 , có nghĩa là $37,02\%$ sự thay đổi của thu nhập được giải thích bởi các biến phụ thuộc còn lại. Mức độ phù hợp của mô hình gần như tăng gấp đôi so với khi chưa bỏ biến company ($R^2 = 0.1461$)

[153]:
```r
re_summary <-
    re                          %>% # Pipe df into group_by
    group_by(Q1)                %>% # grouping by 'type' column
    summarise(Count = n())      # calculate the name count for each group
## 'df_summary' now contains the summary data for each 'type'
re_summary
```

Q1	Count
18	812
19	871
20	903
21	883
22	1259
23	1314
24	1213
25	1785
26	1790
27	1757
28	1750
29	1731
40	289
41	283
42	298
43	268
44	259
45	344
46	332
47	305
48	348
49	357
55	88
56	85
57	84
58	77
59	77
60	29
61	44
62	32
63	43
64	43
65	41
66	39
67	37
68	50
69	40
70	8
71	6
72	5
73	7
74	8
75	7
76	12
77	12
78	4
79	7

Vẽ thử đám mây điểm

[157]:
```
plot(age, exp,
   xlab="Year Experience ", ylab="Age")
```



[ ]:
```

```