**Returns to Education**

The data includes 3,010 men aged between 24 and 34 with 34 variables that is collected from the Young Men Cohort of the National Longitudinal Survey.

Firstly, you should read description data on second spreadsheet. Secondly, you create a random sample with size of 30 from the data. Finally, you answer the following questions.

**Chapter 1: Defining and Collecting Data**

1. How was this data collected? Is the sample representative of the population?

2. Is the data Missing Values of the data? How do you find them? How do you deal with them?

**Chapter 2: Organizing and Visualizing Variables**

3. Classify each of the following variables: nearc4, educ, age, black, wage, IQ, married, exper, lwage and expersq.

4. Use your subsample to graph the variables in exercise 3. Show characteristics of each the variables.

**Chapter 3: Numerical Description Measures**

5. Use your subsample to calculate the statistics for wage, IQ, educ and exper.

6. Use your subsample to calculate the covariance and the coefficient of correlation between wage and IQ. Repeat for wage and educ, wage and exper.

**Chapter 5: Discrete Probability Distributions**

7. Suppose that proportion of blacks in the population is equal to proportion of backs in your subsample.

7.1 How many blacks do you expect in your subsample? How many blacks are in the data?

7.2 Choose a simple random sample with size of 20 from the population. What is the probability that there are 6 blacks.

7.3 Choose a randomly 5 people in your subsample. Find the probability that there is at least one black.

8. Suppose that “educ” is a discrete random variable that has a Poisson distribution such that its mean equals to the sample mean in your subsample. If we randomly choose a person, what is the probability that he/she has 6 years in education.

**Chapter 6: The Normal Distribution and Other Continuous Distributions**

9. Suppose that lwage is a normally distributed random variable with mean μ = 6.26 and standard deviation σ = 0.44. Find the probability that lwage is in the interval (5,6).

10. Assume age is a discrete uniform random variable integers such that 24 ≤ x ≤ 34. Find the mean and the variance of age. Compare these corresponding values to the values in your subsample. Create the bar chart for age. Is age a uniform distribution?

**Chapter 7: Sampling Distributions**

Assume , where and .

11. Find the probability that a random sample n = 20 observations will have a sample mean IQ that falls in the interval from 100 to 103.

12. How large must the random sample be if we want the standard error of the sample mean to be 1?

**Chapter 8: Confidence Interval Estimation**

13. Construct a 97% confidence interval on the true mean lwage based on data in the subsample. Suppose that lwage is a normally distributed random variable with standard deviation σ = 0.44

14. Construct a 94% confidence interval on the true mean IQ based on data in the subsample. Assume that IQ is a normally distributed random variable.

15. If we wanted the error in estimating the mean lwage from the two-size confidence interval to be 0.2 at 95% confidence. What sample size should be used? Assume that lwage is a normally distributed random variable with standard deviation σ = 0.44.

16. Calculate a 99% confidence interval on the true proportion of all people who nearby 4-year college (nearc4) based on the data.

17. If we wanted the error in estimating the proportion of blacks from two-size confidence interval to be 0.01 at 99% confidence. What sample size should be used?

**Chapter 9: Fundamentals of Hypothesis Testing: One-Sample Tests**

18. Use your subsample to test the hypothesis H0: mean(IQ) = 100 against H1: mean(IQ) ≠ 100 at  by three methods. Assume that IQ is a normally distributed random variable with standard deviation σ = 15

19. Use your subsample to test the hypothesis H0: mean(lwage) = 6 against H1: mean(lwage) > 6 at . Assume that lwage is a normally distributed random variable.

20. It is argued that the proportion of people with less than 10 years of work experience is not more than 0.07. Use the given data to test this argument at .

**Chapter 10: Two-Sample Tests**

Assume that the variables are normally distributed and the variance within each population be equal for all populations.

21. Test the difference between the averages wage by variable black at 1% significance level based on the subsample.

22. Test the difference between two independent population proportions that people with low IQ (IQ < 90) by variable black at 5% significance level.

23. Test the difference between the variances lwage by variable black at 1% significance level based on the subsample.

**Chapter 11: Analysis of Variance**

Assume that the groups by variable black follow a normal distribution and have equal variances. Using your subsample for testing the hypothesis.

24. Use F-test to test the difference between the means wage by variable black at 5% significance level. Compare the result and T-test.

**Chapter 13: Simple Linea Regression**

Suppose you are interested in studying the effect of education on wages. Use your subsample to calculate the regression equation of educ on wage.

25. Interpret the slope and the intercept for the output.

26. Test the significance of the regression () at 10% significance level.

27. Find the coefficient of determination and explain its meaning.

28. Run regression of educ on lwage. Which model is better? Why?

**Chapter 14-15: Multiple Regression**

29. You are interested in studying factors affecting wages. You use the data to perform multivariate regression according to the equations.

Model 1: 

Model 2: 

Interpret the regression coefficients in the models. Which model is better? Why?

**Instructions**

|  |  |  |
| --- | --- | --- |
| Q | Answers | R-code |
| Preparation:  # install packages (library) if not installed  install.packages("readxl")  install.packages("ggplot2")  library(ggplot2)  library(readxl)  wage <- read\_excel("wage.xlsx")  # copy and paste wage.xlsx file to the Documents Folder or change the path.  id\_number<-101  set.seed(id\_number)  # choice a number to set.seed  yourdata<-wage[sample(1:nrow(wage),30),]  # generate a random sample with size of 30 from the data.  Mỗi sinh viên sử dụng mã id\_number riêng để có được các mẫu khác nhau. | | |
| 1 | Dữ liệu được thu thập bằng cách khảo sát những người đàn ông độ tuổi từ 24-34. Nhiều khả năng một số người không được tiếp cận cho khảo sát nên dễ xảy ra tính chệch (bias). |  |
| 2 | Kết quả cho ra 2059 giá trị khuyết.  Xử lý dữ liệu khuyết bằng 2 cách: Bổ sung hoặc loại bỏ. Bổ sung gồm 2 cách: ghi nhận lại và ước lượng. Tùy từng trường hợp cụ thể mà sẽ có cách xử lý tương ứng. Trường hợp này, ta có thể loại bỏ quan sát có giá trị khuyết. | table(is.na(wage))  # new\_data<-na.omit(wage) |
| 3 | Các biến định tính: black và married. Còn loại các biến định lượng. |  |
| 4 | Các biến định tính có thể sử dụng biểu đồ cột.  Các biến định lượng liên tục dùng histogram, rời rạc dùng biểu đồ boxplot, histogram, dot hoặc stem-leaf. | # boxplot  ggplot(data = yourdata,aes(x=educ))+geom\_boxplot()  # histogram  ggplot(data = yourdata,aes(x=educ))+geom\_histogram(bins = 10)  # dot plot  ggplot(data = yourdata,aes(x=educ))+geom\_dotplot(fill="blue",binwidth = 0.5)  # stem-leaf  stem(yourdata$educ,2)  # bar  ggplot(data = yourdata,aes(x=factor(black),fill=factor(black)))+geom\_bar() |
| 5 |  | summary(yourdata[,c("wage", "IQ", "educ", "exper")])  # mean IQ, var and stanrd deviation  mean(yourdata$IQ,na.rm = T)  var(yourdata$IQ,na.rm = T)  sd(yourdata$IQ,na.rm = T)  # mean wage, var and stanrd deviation  mean(yourdata$wage)  var(yourdata$wage)  sd(yourdata$wage) |
| 6 | Tính cov, r cho các cặp biến khi có mặt dữ liệu khuyết. | cov(na.omit(yourdata[,c("wage","IQ","educ","exper")]))  cor(na.omit(yourdata[,c("wage","IQ","educ","exper")])) |
| 7 | 7.1 Số lượng người da đen là 7, trong khi số lượng trong data là 703  Tỉ lệ black trong tổng thể là: p = 7/30  7.2    7.3 | table(yourdata$black)  table(wage$black)  choose(20,6)\*(7/30)^6\*(23/30)^14  # dbinom(6,20,7/30)  1-choose(23,5)/choose(30,5)  #1-dhyper(0,7,23,5)  #sum(dhyper(1:5,7,23,5)) |
| 8 |  | lambda<-mean(yourdata$educ)  # 13.73333  exp(-lambda)\*lambda^6/factorial(6)  # dpois(6,lambda)  # 0.01011608 |
| 9 |  | pnorm((6-6.26)/0.44)-pnorm((5-6.26)/0.44)  #pnorm(6,6.26,0.44)-pnorm(5,6.26,0.44)  # 0.2751966 |
| 10 | Đồ thị cho thấy age *có khả năng* phân phối đều.  # có thể sử dụng package swfscMisc để kiểm định tính đều của histogram | (24+34)/2  # 29  ((34-24+1)^2-1)/12  # 10  mean(yourdata$age)  # 28.23333  var(yourdata$age)  # 10.59885  hist(yourdata$educ)  install.packages("swfscMisc")  library(swfscMisc)  uniform.test(hist(yourdata$educ)) |
| 11 |  | mu<-mean(wage$IQ,na.rm = T)  #102.4498  sig<-sd(wage$IQ,na.rm = T)  # 15.42376  pnorm((103-mu)\*sqrt(20)/sig)-pnorm((100-mu)\*sqrt(20)/sig)  # 0.3246231  # pnorm(103,mu,sig/sqrt(20))-pnorm(100,mu,sig/sqrt(20)) |
| 12 |  | (sig/1)^2  # 237.8923 |
| 13 |  | z.alpha.2<-qnorm(1.5/100,lower.tail = FALSE)  # 2.17009  n<-length(yourdata$lwage)  # 30  mu<-mean(yourdata$lwage)  # 6.289448  CI<-c(mu-z.alpha.2\*0.44/sqrt(n),mu+z.alpha.2\*0.44/sqrt(n))  # 6.115119 6.463777 |
| 14 |  | IQ<-na.omit(wage$IQ)  n<-length(IQ)  mu<-mean(IQ)  sig<-sd(IQ)  z<-qnorm(0.03,lower.tail = FALSE)  CI<-c(mu-z\*sig/sqrt(n),mu+z\*sig/sqrt(n))  # 101.8108 103.0888 |
| 15 |  | n<-(1.96\*0.44/0.2)^2  # 18.59334 |
| 16 |  | n<-length(wage$nearc4)  f<-length(wage$nearc4[wage$nearc4==1])/n  # table(wage$nearc4)  CI<-c(f-2.576\*sqrt(f\*(1-f)/n),f+2.576\*sqrt(f\*(1-f)/n))  # 0.6601949 0.7039247 |
| 17 |  | f<-length(black[black==1])/length(black)  # table(wage$black)  n<-(2.756/0.01)^2\*f\*(1-f)  # 13596.54 |
| 18 | Giá trị cần kiểm định nằm trong miền tin cậy nên chấp nhận H0    Vì nên ta chấp nhận H0    Vì p-value > α nên ta chấp nhận H0 | IQ<-na.omit(yourdata$IQ)  n<-length(IQ)  mu<-mean(IQ)  z.alpha.2<-qnorm(0.01,lower.tail = FALSE)  CI<-c(mu-z\*15/sqrt(n),mu+z\*15/sqrt(n))  # 98.75204 112.99796  z.test<-(mu-100)\*sqrt(n)/15  # 1.918767  abs(z.test)<z.alpha.2  # TRUE  2\*pnorm(-abs(z.test))  # 0.05501383 |
| 19 | Vì tα < t nên ta bác bỏ H0 | n<-length(yourdata$lwage)  mean.x<-mean(yourdata$lwage)  s<-sd(yourdata$lwage)  t<-(mean.x-6)\*sqrt(n)/s  # 3.194853  t.alpha<-qt(0.1,n-1,lower.tail = FALSE)  # 1.311434 |
| 20 | Vì z > zα nên ta bác bỏ H0, tức là có đủ cơ sở để bác bỏ ý kiến trên. | n<-length(wage$educ)  f<-length(wage$educ[wage$educ<10])/length(wage$educ)  z.alpha<-qnorm(0.02,lower.tail = FALSE)  # 2.053749  z<-(f-0.07)\*sqrt(n)/sqrt(0.07\*(1-0.07))  # 9.093994 |
| 21 | Vì |t| > tα/2 nên ta chấp nhận H0 | x1<-subset(yourdata,black==1)$wage  x2<-subset(yourdata,black==0)$wage  n1<-length(x1)  n2<-length(x2)  mean.x1<-mean(x1)  mean.x2<-mean(x2)  s<-sqrt(((n1-1)\*var(x1)+(n2-1)\*var(x2))/(n1+n2-2))  t<-(mean.x1-mean.x2)/(s\*sqrt(1/n1+1/n2))  # -2.509706  t.alpha.2<-qt(0.005,28,lower.tail = FALSE)  # 2.763262 |
| 22 | Vì |z| > zα/2 nên ta bác bỏ H0 | newdata<-na.omit(wage[,c("IQ","black")])  x1<-subset(newdata,black==0)$IQ  x2<-subset(newdata,black==1)$IQ  n1<-length(x1)  n2<-length(x2)  p1<-length(x1[x1<90])/n1  p2<-length(x2[x2<90])/n2  p.bar<-length(newdata$IQ[newdata$IQ<90])/length(newdata$IQ)  # (n1\*f1+n2\*f2)/(n1+n2)  z<-(p1-p2)/sqrt(p.bar\*(1-p.bar)\*(1/n1+1/n2))  # -18.66723 |
| 23 | Vì nên ta chấp nhận H0 | x1<-yourdata$wage[yourdata$black==0]  x2<-yourdata$wage[yourdata$black==1]  n1<-length(x1)  n2<-length(x2)  s1.2<-var(x1)  s2.2<-var(x2)  f<-s1.2/s2.2  # 6.529647  f.u<-qf(0.005,22,6,lower.tail = FALSE)  # 9.526445  f.l<-qf(1-0.005,22,6,lower.tail = FALSE)  # 0.2313489 |
| 24 | Sử dụng phân tích ANOVA 1 nhân tố                Kết quả ta bác bỏ H0.  So sách với t-test ở bài 21 ta có nhận xét sau: bài 21 ta chấp nhận ở mức 1% nhưng lại bác bỏ ở mức 5%. Kết quả này cũng phù hợp với ANOVA one-way. Tuy nhiên, khi t-test dễ sử dụng và linh hoạt hơn ANOVA khi chỉ có 2 nhóm. | x1<-yourdata$wage[yourdata$black==0]  x2<-yourdata$wage[yourdata$black==1]  x.mean<-mean(yourdata$wage)  n1<-length(x1)  n2<-length(x2)  SST<-sum((yourdata$wage-x.mean)^2)  SSA<-n1\*(mean(x1)-x.mean)^2+n2\*(mean(x2)-x.mean)^2  SSE<-sum((x1-mean(x1))^2)+sum((x2-mean(x2))^2)  MSA<-SSA/(2-1)  MSE<-SSE/(30-2)  f<-MSA/MSE  # 6.298623  f.low<-qf(0.025,1,28,lower.tail = TRUE)  # 0.0009997776  f.up<-qf(0.025,1,28,lower.tail = FALSE)  # 5.609564  # ANV<-lm(data = yourdata,wage~factor(black))  # anova(ANV) |
| 25 | - Ý nghĩa của hệ số góc: (b1=55.341)  Mỗi năm đi học tăng thêm khi cố định các yếu tố khác, ta kỳ vọng tiền công làm việc mỗi giờ tăng thêm 55.34 cents.  - Ý nghĩa hệ số chặn: (b0=-153.92)  Ta kỳ vọng mức tiền công là -153.92 cents tại giá trị educ = 0 khi cố định các nhân tố khác.  Chú ý rằng: giá trị b0 ở một số mô hình đôi khi không có giá trị trong thực tế thực tế. | ols<-lm(data = yourdata,wage~educ)  ols$coefficients  # -153.92036 55.34129 |
| 26 | |t| > tα/2,28 ta bác bỏ H0 | summary(ols)  # t<-summary(ols)$coefficients[2]/summary(ols)$coefficients  # 3.162504  qt(0.05,28,lower.tail = FALSE)  # 1.701131 |
| 27 | Hệ số giải thích    Ý nghĩa: Hệ số xác định r2 đo lường mức đo lường hợp của mô hình. Cụ thể, mô mình giải thích khoảng 26.32% biến động trong wage gây ra bởi educ. | SST<-sum((yourdata$wage-mean(yourdata$wage))^2)  SSR<-sum((-153.92036+55.34129\*yourdata$educ-mean(yourdata$wage))^2)  SSE<-sum((-153.92036+55.34129\*yourdata$educ-yourdata$wage)^2)  SSR/SST  # summary(ols)  # 0.2632  # cor(yourdata$educ,yourdata$wage)^2  # summary(ols)$r.squared |
| 28 | Các biến giải thích đều có ý nghĩa thống kê ở mức 1% ở cả hai mô hình. Tuy nhiên, hệ số xác định ở mô hình 2 cao hơn nên mô hình 2 được xem phù hợp hơn. | reg<-lm(data = yourdata,lwage~educ)  summary(reg)  summary(ols) |
| 29 | Model 1:  Model 2:  Ý nghĩa các hệ số:  - Mô hình 1:  + b1 = 77.51: Mỗi năm đi học tăng them, ta kỳ vọng tiền công tăng thêm 77.51 cents mỗi giờ khi cố định các yếu tố tác động còn lại.  + b2 = 24.95: Mỗi năm kinh nghiệm làm việc tăng thêm, ta kỳ vọng tiền công tăng thêm 77.51 cents mỗi giờ khi cố định các yếu tố tác động còn lại.  - Mô hình 2:  + b1 = 0.146: Mỗi năm đi học tăng thêm, ta kỳ vọng tiền công tăng thêm 0.146 = 14.6% mỗi giờ khi cố định các yếu tố tác động còn lại.  + b2 = 0.158: Mỗi năm kinh nghiệm làm việc tăng thêm, ta kỳ vọng tiền công tăng thêm 0.158 = 15.8% mỗi giờ khi cố định các yếu tố tác động còn lại.  + Biến expersq dùng để kiểm soát phi tuyến. Hệ số b3 có ý nghĩa thống kê hàm ý rằng mô hình tồn tại dạng bậc 2.  Lựa chọn mô hình: Căn cứ vào F-test, Adjusted R-squared và ý nghĩa thực tế.  - F-test: Cả hai mô hình đều phù hợp.  - Adjusted R-squared: Mô hình 2 tốt hơn.  - Ý nghĩa thực tế: kinh nghiệm làm việc và trình độ học vấn (thể hiện bằng số năm đi học) là các yếu tố được kỳ vọng sẽ có tác động đến hiệu quả làm việc. Do đó, ta kỳ vọng các biến này đều có ý nghĩa thống kê trong mô hình hồi quy. Riêng kinh nghiệm làm việc kỳ vọng là một hàm bậc 2 lồi lên, bởi vì một người có kinh nghiệm làm việc ở mức cao đồng nghĩa họ ít đầu tư cho giáo dục hơn. Do vậy, mức thu nhập khi đó sẽ giảm. Căn cứ vào lập luận trên và kết quả của 2 mô hình, mô hình 2 phù hợp hơn.  Như vậy, mô hình 2 giải thích tốt hơn. | model.1<-lm(data = yourdata,wage~educ+exper)  model.2<-lm(data = yourdata,lwage~educ+exper+expersq)  summary(model.1)  summary(model.2) |