**Question 1 (Ex 6 page 186): MLR with 2 quantitative predictors**

The researchers collected data on age, hours worked per day from 15 vendors in Mexico to study the factors influencing vendors’ incomes. The data set is saved as *STREETVN.txt.*

Read the data into RStudio

> streetvn.df <- read.csv("data sets/STREETVN.csv", header = T)

> library(GGally)

> ggpairs(streetvn.df, columns = c("Earnings","Age","Hours"))

1. Examine the pairs plot (Figure 1). What does this plot suggest?

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

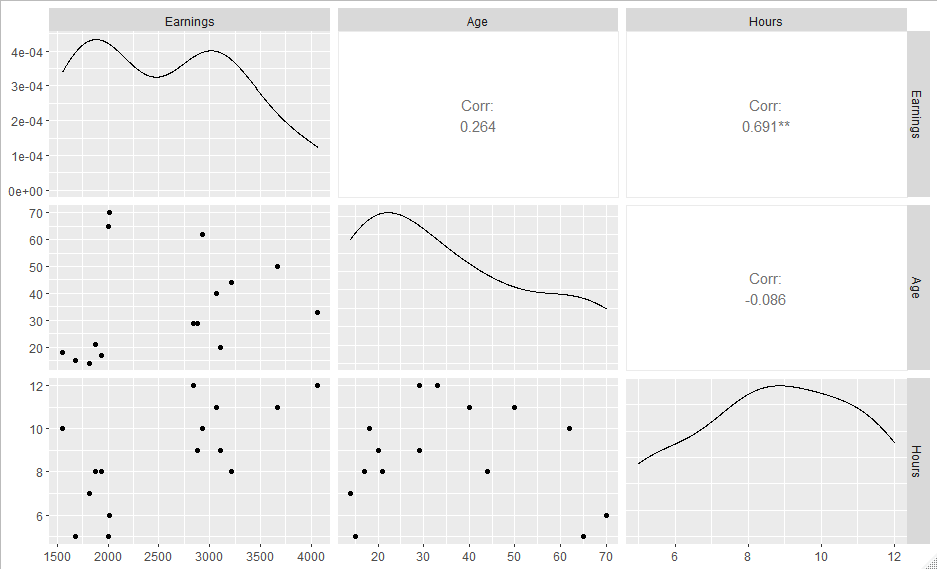


Figure : Street vendors pairs plot

A model was fitted using both predictors, and the summary table produced below.

Table : Summary table for street vendors model

> mod1<-lm(Earnings ~ Age + Hours, data = streetvn.df)

> summary(mod1)

Call:

lm(formula = Earnings ~ Age + Hours, data = streetvn.df)

Residuals:

Min 1Q Median 3Q Max

-1105 -322 -61 332 721

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -20.35 652.75 -0.03 0.9756

Age 13.35 7.67 1.74 0.1074

Hours 243.71 63.51 3.84 0.0024

Residual standard error: 548 on 12 degrees of freedom

Multiple R-squared: 0.582, Adjusted R-squared: 0.513

F-statistic: 8.36 on 2 and 12 DF, p-value: 0.00531

1. Find the regression equation

---------------------------------------------------------------------------------------------------------------------

1. Conduct a test of the global utility of the model (at α = 0.05)

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. Find and interpret the values of

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. Is age a statistically useful predictor of annual earnings? Test using α = 0.05.

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. Find a 95% CI for the slope for hours worked per day. Interpret the interval in the context of the problem.

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table : CI for the regression coefficients

> confint(mod1)

2.5 % 97.5 %

(Intercept) -1442.562 1401.86

Age -3.365 30.07

Hours 105.334 382.09

1. With reference to the following outputs (Table 4 and Figure 2), check the model assumptions

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

A comparison of a plot and a sample quantity

AI-generated content may be incorrect.

Figure : street vendors - Residual plots

> # residuals vs fitted and q-q normal

> library(car)

> library(ggResidpanel)

> resid\_panel(mod1, plots=c(“resid”, “qq”))

> shapiro.test(mod1$residuals)

> ncvTest(mod1)

Shapiro-Wilk normality test

data: mod1$residuals

W = 0.94, p-value = 0.4

Non-constant Variance Score Test

Variance formula: ~ fitted.values

Chisquare = 0.4723515, Df = 1, p = 0.49191

Table : Street vendors - Residuals normality test

1. Use the fitted model to predict annual earning for a 20-year-old person who worked 13h per day.

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

A graph with red dots and numbers

AI-generated content may be incorrect.

> predict(mod1,new=data.frame(Age=20, Hours=13), interval="confidence", se=T)

$fit

fit lwr upr

1 3415 2721 4109

> library(ggplot2)

> ggplot(data=streetvn.df, aes(x=Age, y=Hours))+

geom\_point(col="salmon")+

geom\_point(x=20, y=13, col="purple")+

expand\_limits(y=c(4,14))

Figure : Scatter plot of Hours and Age

**Question 2 (Ex 15 page 191) Modelling IQ**

Because the coefficient of determination R2 always increases when a new independent variable is added to the model, it’s tempting to include many variables in the model to force R2 to be near 1. However, doing so reduce the degree of freedom available for estimating, which adversely affects our ability to make reliable inferences.

As an example, suppose you want to use the responses to a survey consisting of 18 demographic, social, and economic questions to model a college’s student intelligence quotient (IQ). You fit the model

Where y = IQ, and x1, x2,…, x18 are the 18 independent variables. The data has 20 students (n = 20). The fitted model has R2 = 0.95.

1. Test to see whether the model is useful.

Hint:

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

> pf(1.056,18,1,lower.tail = F)

[1] 0.6566

1. Calculate the adjusted R2, denoted as Ra2. Interpret this value.

Hint:

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Question 3**: **Framingham Heart Study** (https://www.framinghamheartstudy.org/)

The Framingham Heart Study (FHS) is dedicated to identify common factors or characteristics that contribute to cardiovascular disease (CVD). High systolic blood pressure predicts the risk of cardiovascular disease, so the study was designed to investigate a number of common factors that contribute to increasing systolic blood pressure.

In 1948, an original cohort of 5,209 men and women between 30 and 62 years old were recruited from Framingham, Massachusetts, USA. The data include biological specimens, molecular genetic data, phenotype data, images, participant vascular functioning data, physiological and demographic etc.

A subset of the FHS data is given in *framingham.csv.*

There are 8 variables: **randid**(random id), **hyperten** (history of hypertension, yes/no), **age**, **sysbp** (systolic blood pressure mmHg), **bmi** (Body Mass Index), **glucose** (casual serum glucose mg/dL), **cigpday** (number of cigarettes smoked each day) and **totchol** (Serum Total Cholesterol).

1. Examine the pairs plot (Figure 4). Write a brief summary for the exploratory analysis.

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

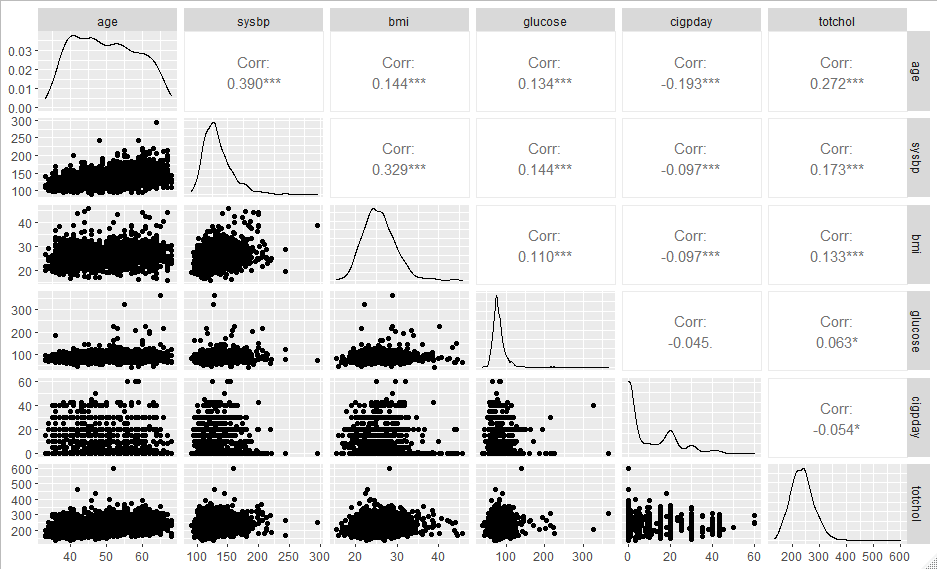


Figure : pairs plot for FHS

A model is fitted using all 5 predictors, and the summary table is given in Table 6.

1. write down the regression equation

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. Interpret the global F-test

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

> FHS.df<-read.csv(“data sets/framingham.csv”)

> FHS.mod<-lm(sysbp~age+bmi+glucose+totchol+cigpday, data=FHS.df)

> summary(FHS.mod)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 41.02136 4.94108 8.30 2.4e-16

age 0.84323 0.06332 13.32 < 2e-16

bmi 1.48986 0.13172 11.31 < 2e-16

glucose 0.07538 0.02632 2.86 0.0042

totchol 0.02214 0.01242 1.78 0.0748

cigpday -0.00335 0.04466 -0.08 0.9402

Residual standard error: 19.7 on 1423 degrees of freedom

Multiple R-squared: 0.234, Adjusted R-squared: 0.232

F-statistic: 87.1 on 5 and 1423 DF, p-value: <2e-16

Table : Summary table for FHS (full model)

1. Should all predictors be retained in the model? If not, which ones would you keep? Justify your answer.

-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

1. Refit the model, remove the insignificant terms, print out the summary table and check the model assumptions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. error | t-value | Pr (>|t|) |
| (intercept) |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| Residual standard error: | | | | |
| Adjusted R-squared: | | | | |
| F-statistic: | | | | |

