**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?

In Figure 1, there is a moderate positive association between Earnings and Hours

( and a weak association between Earnings and Age (). This suggests that a suitable model for Earnings may include Hours but not Age.

The relationship between the two predictors (Age and Hours) appears to be non-linear, so their correlation coefficient is very small (Table 1).

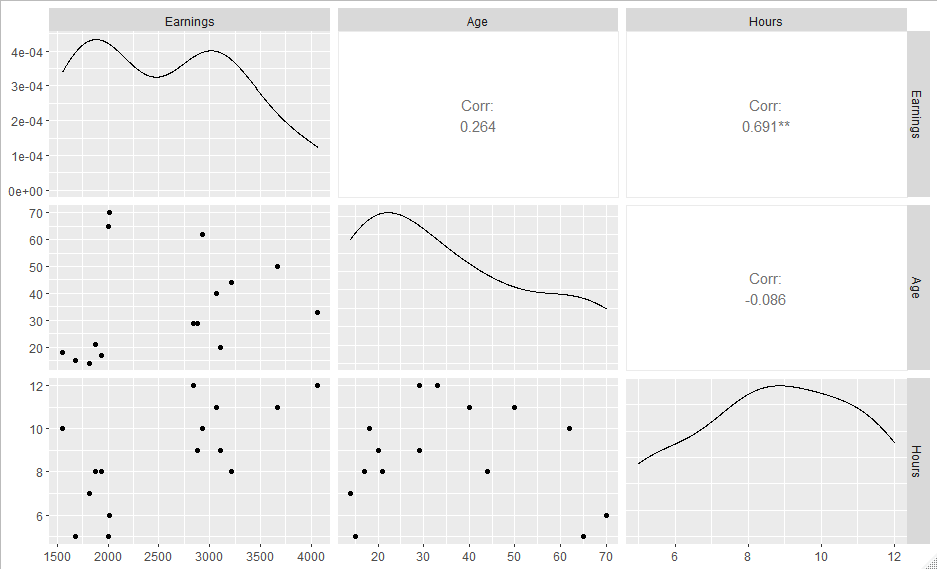


Figure : Street vendors pairs plot

Table 1: Correlations and p-values

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

> 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

Table 2: Summary table for street vendors model

1. Find the regression equation

From Table 2: E(Earnings) = -20.35 + 13.35 \* Age +243.71\*Hours

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

To test the global usefulness of the model, we test the hypothesis:

(all of the predictors are unimportant in predicting Earnings)

at least 1 of the predictors is useful to predict Earnings

From Rstudio output in Table 2, the F statistic is 8.36 on 2 and 12 df, p-value = 0.005 < 0.05.

Therefore, we reject H0 and conclude that at least 1 of the predictors is useful or the model is useful to predict Earnings.

1. Find and interpret the values of

suggests that 51.3% of variability in earnings can be explained by using this model with 2 predictors: Age and Hours.

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

Test using t-test and p-value from the summary table.

T statistic for Age is 1.74 , p-value = 0.107 > 0.05, thus we can’t reject H0. So Age is not a statistically useful predictor when Hours is already in the model.

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

Estimate for betahours = 243.71, 95%CI: (105.3; 382.1) (Table 3)

For each additional hours worked per day, we’d expect the average annual earnings will be increased between $105.3 to $382.1 with 95% confidence.

ALTERNATIVE WORDING:

With 95% confidence, a 1 hour increase in time worked will increase earnings by between $105.33 and $38.09, on average, if Age doesn’t change.

Table 3: CI for the regression coefficients

> confint(mod1)

Estimate Std. Error 2.5 % 97.5 %

(Intercept) -20.35 652.745 -1442.562 1401.86

Age 13.35 7.672 -3.365 30.07

Hours 243.71 63.512 105.334 382.09

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

From the residuals versus fitted value, the residuals are randomly scattered around 0 (constant mean of 0 wrt x), with 1 residual is quite large in magnitude (closer to 1000) which may suggest that this is an outlier. The variability is constant (wrt x), so the conditions of mean=0 and constant variance appear to be met.

From the normal QQ plot, most of the points lie on the straight line so it suggests that the normality assumption for the residuals is reasonable.

From table 4, p-value is 0.4 so we can’t reject the H0: “residuals are normally distributed”. Or the normality assumption is valid.

The condition of independent observations is assessed by considering the design of the study. If the 15 vendors were randomly chosen from a large population of vendors, then this assumption is also met.

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 4: 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.

We are given a confidence interval for the mean earnings for a person with these characteristics, so we can interpret this. The interpretation of a prediction interval would be slightly different. With 95% confidence, average earnings will be between $2721 and $4109 for a 20 year old working 13 hours a day. However, we can see in Figure 3, that this combination of predictor variable values represents an extrapolation of the data from the sample, upon which the model was built, so this confidence interval is unreliable.

> 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=streetnv.df, aes(x=Age, y=Hours))+

geom\_point(col="salmon")+

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

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

A graph with red dots and numbers

AI-generated content may be incorrect.

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.

To test the overall utility of the model, we use the Global F-test

Null hypothesis: (all predictors are unimportant)

Test statistics: F =

with k=18 and n – (k+1) = 1 degree of freedom.

p-value = 0.66

Thus, we can’t reject the null hypothesis and so the model is not useful.

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

[1] 0.6566

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

Formula for adjusted R2 is:

Conclusion:

* R2 increases with every predictor added to a model. As R2 always increases and never decreases, it can appear to be a better fit with the more terms you add to the model. This can be completely misleading.
* If your model has too many terms you can run into the problem of over-fitting the data. When you over-fit data, a misleadingly high R2 value can lead to misleading projections.

**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). 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.txt.*

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.

From Figure 4, sysbp appears to have a moderate positive relationship with age and bmi (r = 0.39 and r = 0.33 respectively), and has a very weak correlation with glucose, cigpday and total cholesterol. This suggests that the model for sysbp may include age and bmi but not the other 3 predictors.

Between the predictors, only age and totchol appear to have the largest correlation of 0.27. The rest of the pairwise correlations between predictors is negligible.

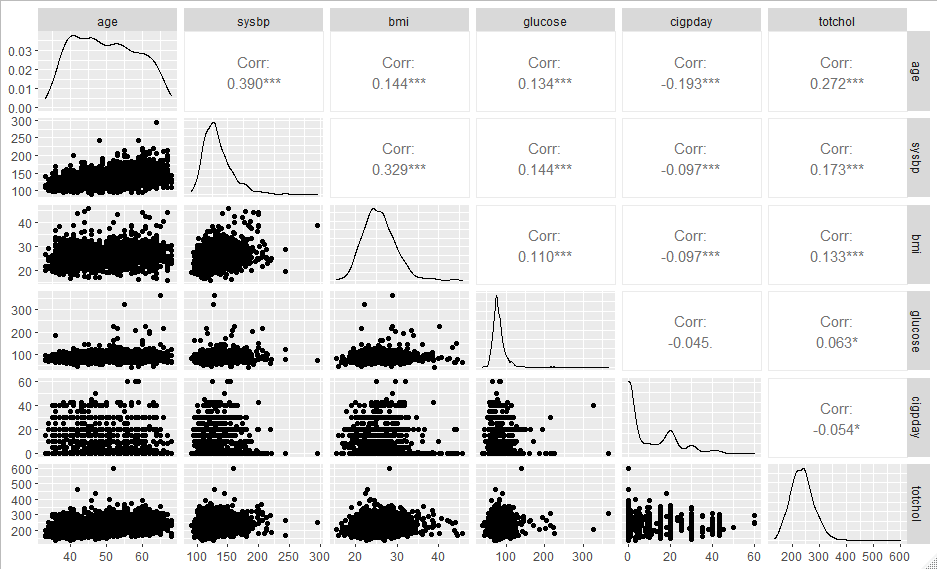


Figure : pairs plot for FHS

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

1. write down the regression equation
2. Interpret the global F-test

Null hypothesis:

Test statistic F = 87.1 on 5 and 1423 degrees of freedom

Thus we reject the null hypothesis and conclude that at least 1 of the predictors is useful, or the model is useful to predict the sysbp.

> 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

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.

**Age**

or age is not a useful predictor, given bmi, glucose, totchol and cigpday are already in the model

T statistic = 13.32 , . Thus, it’s significant, we reject H0 .

**bmi**

or age is not a useful predictor, given age, glucose, totchol and cigpday are already in the model

T statistic = 13.31 , . Thus, it’s significant, we reject H0 .

**Glucose**

or age is not a useful predictor, given age, bmi, totchol and cigpday are already in the model

T statistic = 2.86 , . Thus, it’s significant, we reject H0 .

**Totchol**

or age is not a useful predictor, given age, bmi, glucose and cigpday are already in the model

T statistic = 1.78 , . This p-value is higher than the threshold 0.05, thus it’s not significant, fail to reject H0 .

**Cigpday**

or age is not a useful predictor, given age, bmi, glucose and totchol are already in the model

T statistic = -0.08 , . This p-value is higher than the threshold 0.05, thus it’s not significant, fail to reject H0 .

**Conclusion: only age, bmi and glucose are significant predictors and should be remained in the model. The other 2 variables: totchol and cigpday are not useful predictors and can be removed.**

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

Coefficients:

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

(Intercept) 44.1146 4.4493 9.91 <2e-16

age 0.8723 0.0603 14.46 <2e-16

bmi 1.5132 0.1308 11.57 <2e-16

glucose 0.0763 0.0263 2.90 0.0038

Residual standard error: 19.7 on 1425 degrees of freedom

Multiple R-squared: 0.233, Adjusted R-squared: 0.231

F-statistic: 144 on 3 and 1425 DF, p-value: <2e-16

The new model with only 3 predictors: age, bmi and glucose was fitted and the summary table of regression coefficients is given in Table 8.

All predictors are significant, the model explains about 23.1% of the variability in the sysbp.

**Model assumptions:** Residuals are independent, normally distributed about 0 and having constant variance.

From Figure 5, the residuals versus fitted value, the residuals seem to spread out for fitted values of sysbp above 130. This suggests that the residuals do not have constant variance.

There are 3 observations with very large residuals (observation numbers 81, 125 and 453), these can be the outliers and should be investigated further.

From the normal QQ plot, most of the points lie on the straight line however there is a clear bend at one tail, suggesting a deviation from a normal distribution. Three observations (81, 125, 453) have extremely large standardised residuals, more than 4.

From both plots in Figure 5, it appears that two of the assumptions for residuals (normality assumption and constant variance) are violated, so interpreting the outputs of this model and making predictions should be taken with care.

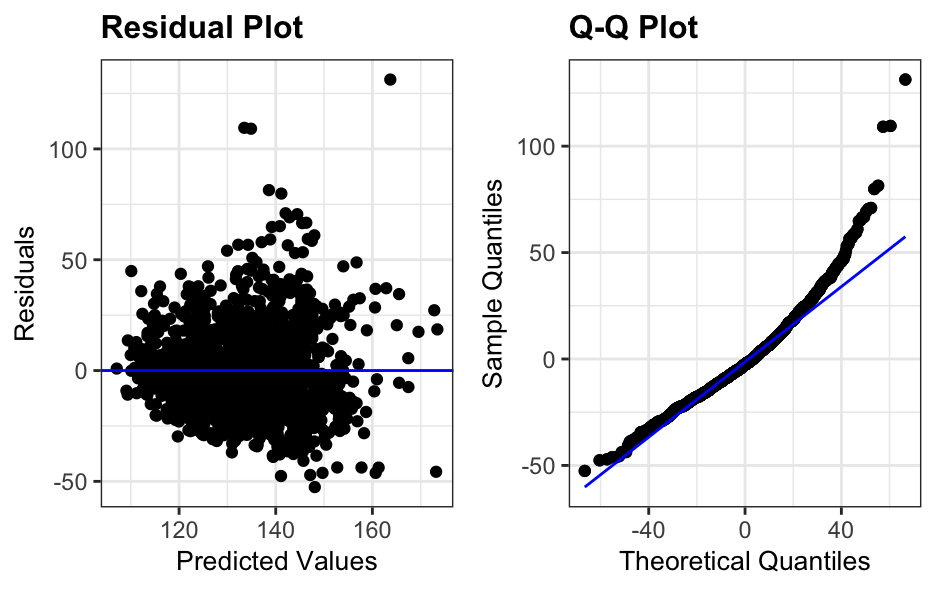


Figure : residuals plots for final model in Q3

**Total Rscript**

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

library(GGally)

library(ggplot2)

library(car)

library(ggResidpanel)

options(digits=3, show.signif.stars=F)

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

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

summary(mod1)

confint(mod1)

# diagnostic plots

# residuals vs fitted and q-q normal

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

shapiro.test(mod1$residuals)

ncvTest(mod1)plot(mod1, which=1)

#make predictions

predict(mod1,new=data.frame(Age=20, Hours=13),

interval = "confidence", se=T)

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))

#Question 2

#Assign the number of variables(k),

#sample size(n) and r2 value(r2)

k<-18

n<-20

r2<-0.95

#Use these values to calculate the

#f value (F) and the DF(df)

f<-(r2/k)/((1-r2)/(n-(k+1)))

f

df<-n-(k+1)

#use these values to find the p-value.

#use lower.tail=FALSE because we want to find

#the righthand side of the f-value

pf(f,k,df,lower.tail = FALSE)

#OR use the values you calculated by hand

pf(1.056,18,1,lower.tail=FALSE)

#Question 3

library(GGally)

FHS.df<-read.csv("data sets/framingham.csv", header=TRUE)

ggpairs(FHS.df,

columns = c("sysbp", "age", "bmi", "glucose", "cigpday",

"totchol")])

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

summary(FHS.mod)

FHS.mod2<-lm(sysbp~age+bmi+glucose, data=FHS.df)

summary(FHS.mod2)

resid\_panel(FHS.mod2, plots=c(”resid”,”qq”))

shapiro.test(mlr2$residuals)