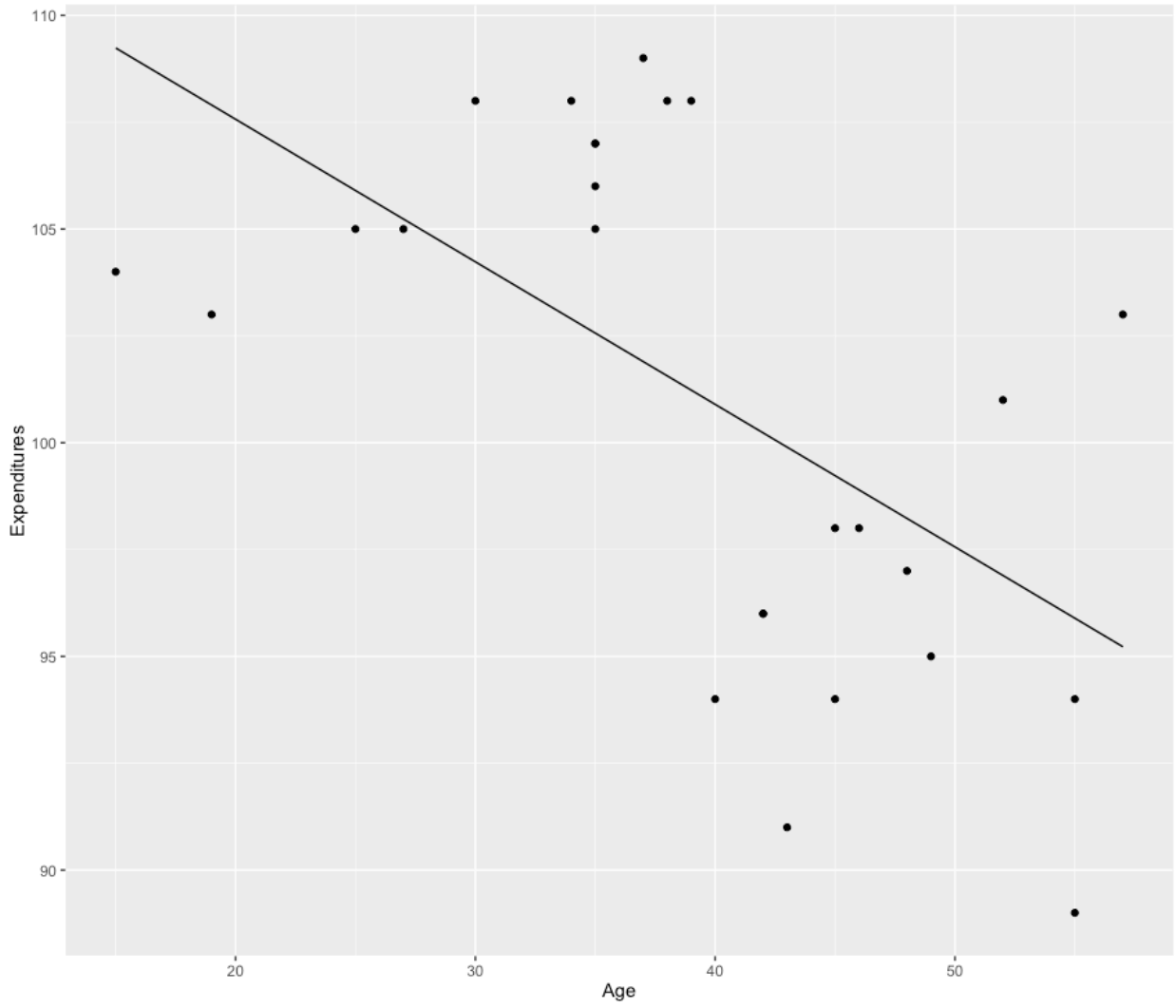
Greg Faletto

Test Exercise 1 Answers

(a) Regression line:
b = -0.333596096606276
a = 114.241107954932

Standard error of b: 0.0953691827886391
t-value of b: -3.49794437628353

(b) The conclusion I draw from this diagram is that Age provides some predictive value in predicting Expenditures, but it seems as though people under 40 have significantly different spending habits as a function of age than people over 40. Splitting the data into two groups might yield better results.



(c) Regression line (40 and over):

b = 0.146470828233374
a = 88.8718890248878
Standard error of b (40 and over): 0.197384418725913
t-value of b (40 and over): 0.742058715570468

Regression line (Under 40):
b = 0.197971278778208
a = 100.232277182585
Standard error of b (under 40): 0.0443836675864513
t-value of b (under 40): 4.46045335015626

(d) In part (c), I found a significantly different way to model the data by splitting the data into two groups and modeling them separately. Based on our limited sample data set, this seems to provide more predictive value than simply grouping all the data together and making one model. We may need a larger sample before we can know for sure if this is the best model for the data. Sometimes it may be better to group the data into different groups and then model the data separately.

R code that I wrote and used to complete the assignment:

```
rm(list=ls())

library("ggplot2")

data <- read.table("TestExer1-holiday expenditures-round2.txt",
        header=T)

n <- nrow(data)

### Part (a)

# Calculate b

A.bar <- mean(data$Age)
E.bar <- mean(data$Expenditures)

b <- (data$Age-A.bar)%*%(data$Expenditures-E.bar)/((data$Age-
        A.bar)%*%(data$Age-A.bar))

# Calculate a

a <- E.bar-b*A.bar

# Display results

print("Regression line:")
print(paste("b=", b))
print(paste("a=", a))
```

```
# Standard error and t-value of b

predictions <- a + b*data$Age

s.squared <- (data$Expenditures-predictions)%*%
(data$Expenditures-predictions)/(n-2)

se.b <- sqrt(s.squared/((data$Age-A.bar)%*%(data$Age-A.bar)))

print(paste("Standard error of b:", se.b))

t.b <- b/se.b

print(paste("t-value of b:", t.b))

# Part (b)

##### Plot of data

df <- data.frame(data)

plot <- ggplot(data=df, aes(x=Age, y=Expenditures)) +
geom_point() + stat_function(fun=function(x)(a+b*x),
          geom="line")

print(plot)

# The conclusion I draw from this diagram is that Age
# provides some predictive value in predicting
# Expenditures, but it seems as though people under
# 40 have significantly different spending habits as
# a function of age than people over 40. Splitting the
# data into two groups might yield better results.


# Part (c)

# Parts (d) and (e): histograms, scatter plot

forty.and.over <- data[which(data$Age>=40), ]
under.forty <- data[which(data$Age<40), ]

# Calculate b for each group

A.bar.over <- mean(forty.and.over$Age)
E.bar.over <- mean(forty.and.over$Expenditures)

b.over <- (forty.and.over$Age-A.bar.over)%*%
```

```
(forty.and.over$Expenditures-E.bar.over)/((forty.and.over$Age-
        A.bar.over)%*%(forty.and.over$Age-A.bar.over))

A.bar.under <- mean(under.forty$Age)
E.bar.under <- mean(under.forty$Expenditures)

b.under <- (under.forty$Age-A.bar.under)%*%
(under.forty$Expenditures-E.bar.under)/((under.forty$Age-
        A.bar.under)%*%(under.forty$Age-A.bar.under))

# Calculate a for each group

a.over <- E.bar.over-b.over*A.bar.over

a.under <- E.bar.under-b.under*A.bar.under

# Standard error and t-value of b for each group

predictions.over <- a.over + b.over*forty.and.over$Age

s.squared.over <- (forty.and.over$Expenditures-
        predictions.over)%*%(forty.and.over$Expenditures-
        predictions.over)/(nrow(forty.and.over)-2)

se.b.over <- sqrt(s.squared.over/((forty.and.over$Age-
        A.bar.over)%*%(forty.and.over$Age-A.bar.over)))

t.b.over <- b.over/se.b.over

#

predictions.under <- a.under + b.under*under.forty$Age

s.squared.under <- (under.forty$Expenditures-
        predictions.under)%*%(under.forty$Expenditures-
        predictions.under)/(nrow(under.forty)-2)

se.b.under <- sqrt(s.squared.under/((under.forty$Age-
        A.bar.under)%*%(under.forty$Age-A.bar.under)))

t.b.under <- b.under/se.b.under

# Display results

print("Regression line (40 and over):")
print(paste("b=", b.over))
print(paste("a=", a.over))

print(paste("Standard error of b (40 and over):", se.b.over))
```

```
print(paste("t-value of b (40 and over):", t.b.over))

print("Regression line (Under 40):")
print(paste("b=", b.under))
print(paste("a=", a.under))

print(paste("Standard error of b (under 40):", se.b.under))

print(paste("t-value of b (under 40):", t.b.under))

# Part (d)

# In part (c), I found a significantly different way to model
# the data by splitting the data into two groups and modeling
# them separately. Based on our limited sample data set, this
# seems to provide more predictive value than simply grouping
# all the data together and making one model. Sometimes it
# may be better to group the data into different groups and
# then model the data separately.
```