Lab11-April5-solutions

Elham

2023-04-03

Question 1: Consider the RealEstate.csv data set.

(a) Plot the distribution of the price per unit area variable.

(b) Describe the shape of the distribution.

(c) Compute an estimate of the mean price per unit area.

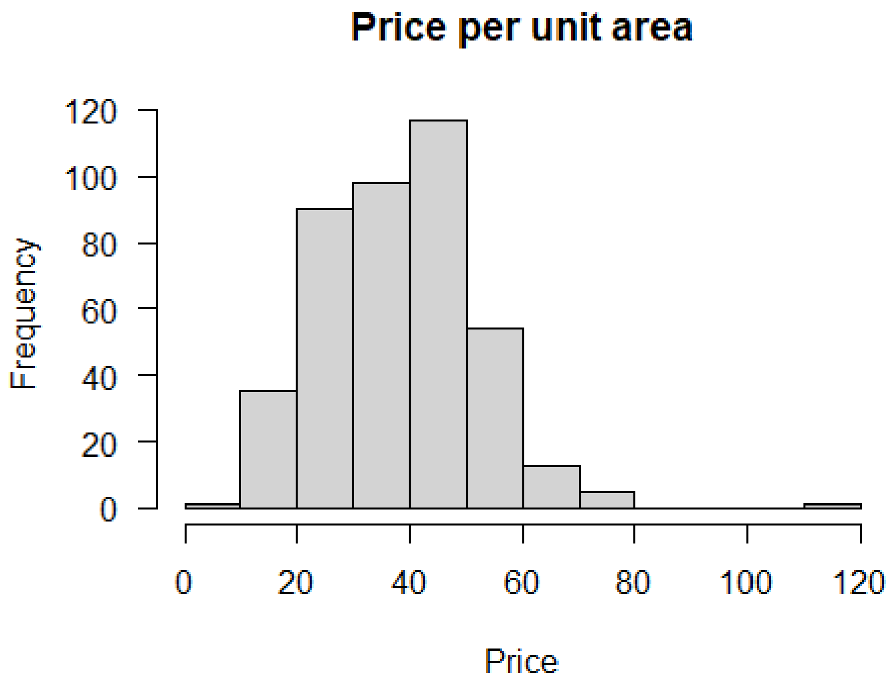
(d) Bootstrap 10000 sample means of the price per unit area. Save the bootstrapped means to a vector called boot_means.

(e) Plot the sampling distribution to see if it appears to be normally distributed.

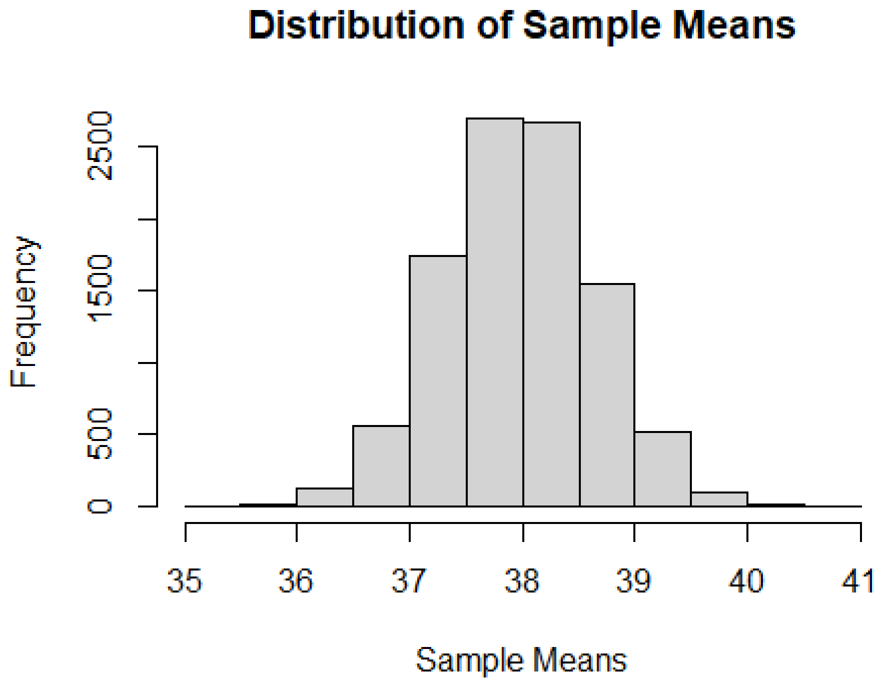
(f) Use the qnorm function (consider the distribution of the sample mean) to compute a 95% confidence interval for the mean.

(g) Now, imagine you are not sure the distribution is normal. Use the quantile function to compute a 95% confidence interval for the mean price per unit area.

```
# (a) Plot the distribution of the price per unit area variable.

# Answer (a) below:

RE = read.csv(file.choose())
# Since the price per unit area variable is numerical, we will plot
# it using a histogram.
hist(RE$Y.house.price.of.unit.area,main="Price per unit area",
xlab="Price",las=1)
```



```
# (b) Describe the shape of the distribution.

# Answer (b) below:
# The distribution is skewed to the right (the tail is on the right)

# (c) Compute an estimate of the mean price per unit area.

# Answer (c) below:

# An appropriate estimate would be using the sample mean which gives
# us an estimated mean price per unit area of $37.98
xbar = mean(RE$Y.house.price.of.unit.area)
xbar

## [1] 37.98019

# (d) Bootstrap 10000 sample means of the price per unit area. Save the
# bootstrapped means to a vector called boot_means.

# Answer (d) below:

n = length(RE$Y.house.price.of.unit.area)
boot_means = numeric()
for(i in 1:10000){
temp_sample = sample(RE$Y.house.price.of.unit.area,n,replace=TRUE)
boot_means[i] = mean(temp_sample)
}

# (e) Plot the sampling distribution to see if it appears to be normally
distributed.

# Answer (e) below:
hist(boot_means,xlab="Sample Means",
main="Distribution of Sample Means")
```



```
# (f) Use the qnorm function (consider the distribution of the sample mean)
to compute a 95% confidence interval for the mean.
# Since the sample mean is normally distributed, we can also get a
# confidence interval by computing the critical value and the estimated
# standard error:
s = sd(RE$Y.house.price.of.unit.area)
ese = s/sqrt(n)
crit_val = qnorm(0.975)
lower_bound = xbar - ese*crit_val
upper_bound = xbar + ese*crit_val
lower_bound

## [1] 36.66952

# Now, imagine you are not sure the distribution is normal. Use the quantile
function to
# compute a 95% confidence interval for the mean price per unit area.
quantile(boot_means, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 36.67290 39.30946
```

Question 2: Load the media spend.csv dataset into R and save it to df. Dataset contains information about a fictitious company that's trying to determine how much money to spend on various types of advertising for the coming year. They have historical data showing sales (in millions of dollars) and the amount they spent on TV, Radio, and Newspaper advertising that year (in thousands of dollars). The goal is to determine which of the types of advertising effects sales the most. We are trying to see how the advertising effects sales, therefore, sales is our response

(a) Plot the response variable (as the y-axis) against each of the regressor variables (one plot for each regressor). Use the par(mfrow = c()) function so that all the plots are displayed at once.

(b) Looking only at the plots, which type of advertising do you think will have the largest effect on sales?

(c) Perform a linear regression for each form of advertising vs the response variable, sales.

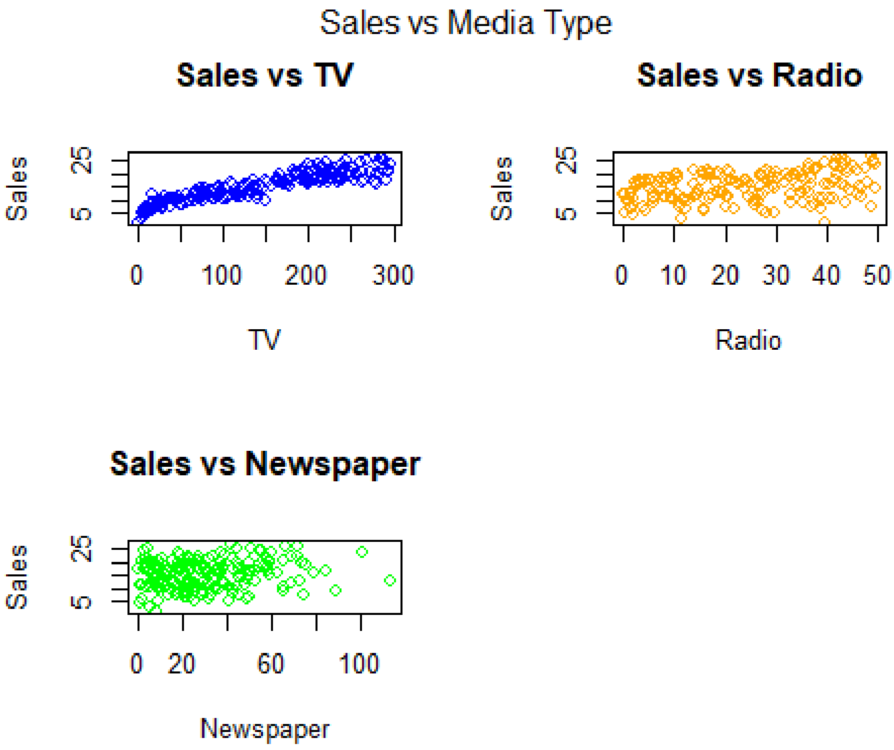
(d) Print out the summary for each of these regressions and take note of the p-value for the t-test on the significance of the coefficient for each.Which of the regressors is the most significant?

(e) Create a vector of these p-values and name each element with the corresponding type of advertising

```
# (a) Plot the response variable (as the y-axis) against each of the
regressor variables
# (one plot for each regressor). Use the par(mfrow = c()) function so that
all the plots
# are displayed at once

# Answer (a) below:
df <- read.csv(file.choose())
head(df)

##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3  12.0
## 4 151.5  41.3      58.5  16.5
## 5 180.8  10.8      58.4  17.9
## 6   8.7  48.9      75.0   7.2

par(mfrow = c(2, 2))
par(mar = c(4.8, 4.5, 5, 2.1))
colours <- c("blue", "orange", "green")
cnames <- c("TV", "Radio", "Newspaper")
```

```
for(i in 1:3){
title <- paste("Sales vs", cnames[i])
plot(df[,i], df[,4], col = colours[i], xlab = cnames[i], ylab = "Sales", main
= title)
}
mtext("Sales vs Media Type", outer = TRUE, line = -1.5)

# (b) Looking only at the plots, which type of advertising do you think will
have
# the largest effect on sales?

# Answer (b) below:
# TV is the most affective advertising about the sales.


# (c) Perform a linear regression for each form of advertising vs the
response variable,
# sales.

# Answer (c) below
fit_news <- lm(df$Sales ~ Newspaper + TV + Radio, df)


# (d) Print out the summary for each of these regressions and take note of
the p-value for # the t-test on the significance of the coefficient for each.
Which of the regressors is
# the most significant?.

# Answer (d) below:


summary(fit_news)

##
## Call:
## lm(formula = df$Sales ~ Newspaper + TV + Radio, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3034 -0.8244 -0.0008  0.8976  3.7473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.6251241  0.3075012  15.041   <2e-16 ***
## Newspaper   0.0003357  0.0057881   0.058    0.954
## TV          0.0544458  0.0013752  39.592   <2e-16 ***
## Radio       0.1070012  0.0084896  12.604   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.662 on 196 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16

# TV and RADIO are significant

# (e) Create a vector of these p-values and name each element with the
corresponding type
# of advertising

# Answer (e) below:
res <- numeric(3)
names(res) <- c("Newspaper", "TV", "Radio")
res[1] <- summary(fit_news)$coefficients[2,4]
res[2] <- summary(lm(Sales ~ TV, data = df))$coefficients[2,4]
res[3] <- summary(lm(Sales ~ Radio, data = df))$coefficients[2,4]

print(res)

##    Newspaper           TV        Radio
## 9.538145e-01 7.927912e-74 3.882892e-07
```


Sales vs Media Type
Sales vs TV
Sales
TV
Sales vs Radio
Sales
Radio
Sales vs Newspaper
Sales
Newspaper