Tutorial Activity 6

Sean Leggett - BDA201 Winter 2020 March 10, 2020

Question

A) The data
B) Which variable would make more sense as response variable? Is this relationship positive or negative?

C) Using the data above create a linear model in R, what is the slope, y-intercept and coefficient of determination.

D) Using the equation for the linear regression that you calculated, estimate the monthly income of an employee at this company who spends 5000 dollars per month on recreation. Is this interpolation or extrapolation?

Answers

A) I want to work where these people are paid these amounts per month. Probably never going to happen but here is the data in a dataframe:

```
spend <- data.frame("Income" = c(41200, 50100, 52000, 66000, 44500, 37700, 73500, 37500, 56700, 35600),
                    "Expenditure" = c(2400, 2650, 2350, 4950, 3100, 2500, 5106, 3100, 2900, 1750)
)
spend
```

```
##    Income Expenditure
## 1   41200        2400
## 2   50100        2650
## 3   52000        2350
## 4   66000        4950
## 5   44500        3100
## 6   37700        2500
## 7   73500        5106
## 8   37500        3100
## 9   56700        2900
## 10  35600        1750
```

B) We chose expenditure as the response variable. A large income obviously provides more potential for discretionary expenditure on entertainment. There are many other variables at play in terms of disposition, how people are with money, etc. But it requires income to spend. Therefore, we anticipate people with more income will spend more on entertainment.

Is the relationship positive or negative? It is positive.

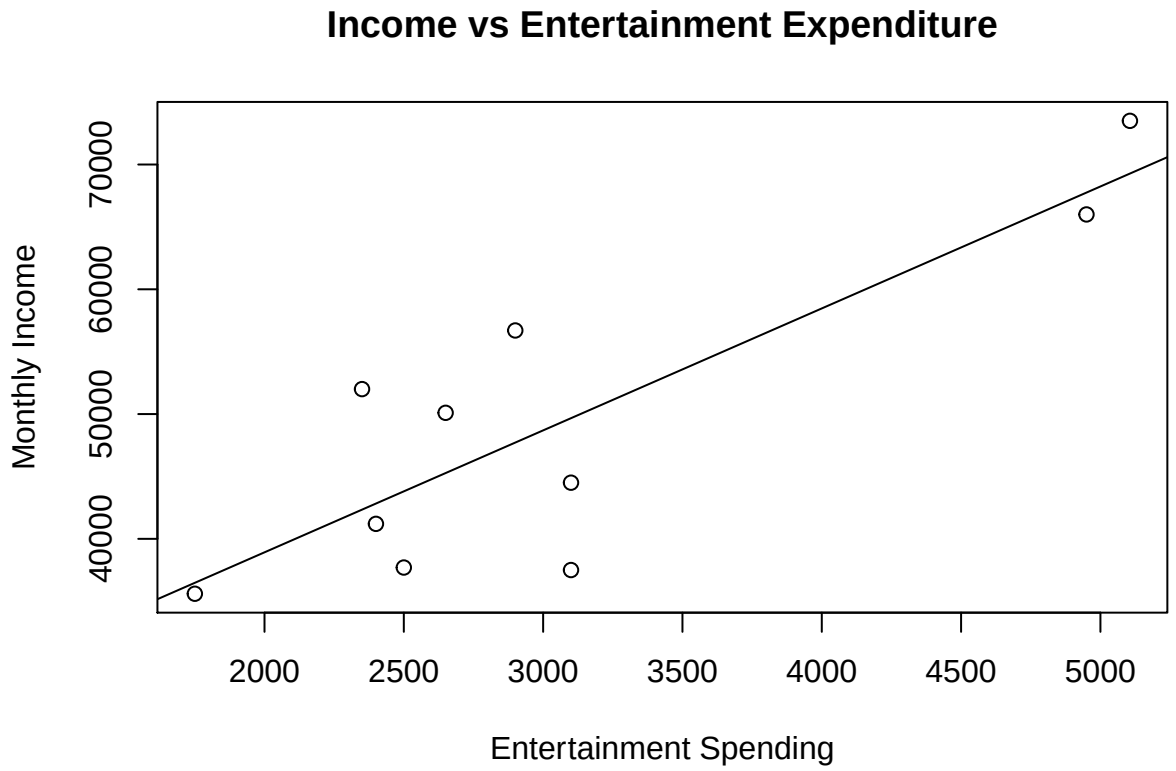
```
cor.test(spend$Income, spend$Expenditure)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  spend$Income and spend$Expenditure
## t = 4.3862, df = 8, p-value = 0.002329
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4476812 0.9613477
## sample estimates:
##       cor
## 0.8404175
```

We see a small P-value and a correlation coefficient of 84% which convinces us there is a positive relationship. Let's confirm via a plot. We will add a regression line to confirm which also anticipates the answer to the next question by establishing an lm.

```
spendlm <- lm(Income ~ Expenditure, data = spend)
spendplot <- plot(spend$Expenditure, spend$Income,
                  main = "Income vs Entertainment Expenditure",
                  xlab = "Entertainment Spending",
                  ylab = "Monthly Income")
abline(lm(spend$Income ~ spend$Expenditure))
```




```
spendplot
```

```
## NULL
```

Scatter plot shows some variability but extreme monthly salary certainly leads to extreme entertainment expenses. However, we belive the expected trend holds true.

C) Linear Model and Data. Model created for previous abline...
Printing the linear model provides the following regression line information including intercept of 19370.109, slope of 9.774

```
spendlm
```

```
##
## Call:
## lm(formula = Income ~ Expenditure, data = spend)
##
## Coefficients:
## (Intercept)  Expenditure
##   19370.109        9.774
```

We can also easily determine coefficient of determination (R-Squared) by using the summary() function for our linear model.
This value is 0.7063

```
summary(spendlm)
```

```
##
## Call:
## lm(formula = Income ~ Expenditure, data = spend)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12170  -4315  -1251   4677   9661
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19370.109   7248.883   2.672  0.02827 *
## Expenditure     9.774      2.228   4.386  0.00233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7364 on 8 degrees of freedom
## Multiple R-squared:  0.7063, Adjusted R-squared:  0.6696
## F-statistic: 19.24 on 1 and 8 DF,  p-value: 0.002329
```

D) Predict Income of a 5000 expenditure...
We can use the predict() function to achieve this. We establish a new dataframe containing the value we wish to test and pass to the predict() function. This produces an answer of 68240.28 which eyeballs to a correct value when checking original dataset with expenditures close to 5000. This answer is interpolated rather than extrapolated. We have not simply added amounts to forecast. We have developed a model that assesses relationships and predicts based on statistical fit with behaviour between two variables.

```r
newexpense <- data.frame("Expenditure" = 5000)
predict(spendlm, newdata = newexpense)
```

```
##        1
## 68240.28
```