536 Homework 2

Michael Pena

2024-02-22

Question

Download the dataset hw2.csv from Titanium. In this dataset you will find financial marketing data on 268 start-up companies working in the tech sector. The variables within this dataset are spend and revenue. Spend contains the amount of capital resources that were spent initially on over the first year of each company. Revenue quantifies the amount of revenue generated during the first year of the company.

You have been hired by a new tech start-up who is interested in advertising. Specifically they have the following asks:

1. What is the relationship between advertising spend and revenue within the first year of a start-up.
2. Currently the company is debating between spending $500,000 and $700,000 on advertising, please provide guidance. Please analyze the data and provide relevant output that addresses each of your employers asks.

If you need any clarity on the data feel free to ask Dr. Nichols.

```
# import the data
data <-  read.csv("data.csv")
```

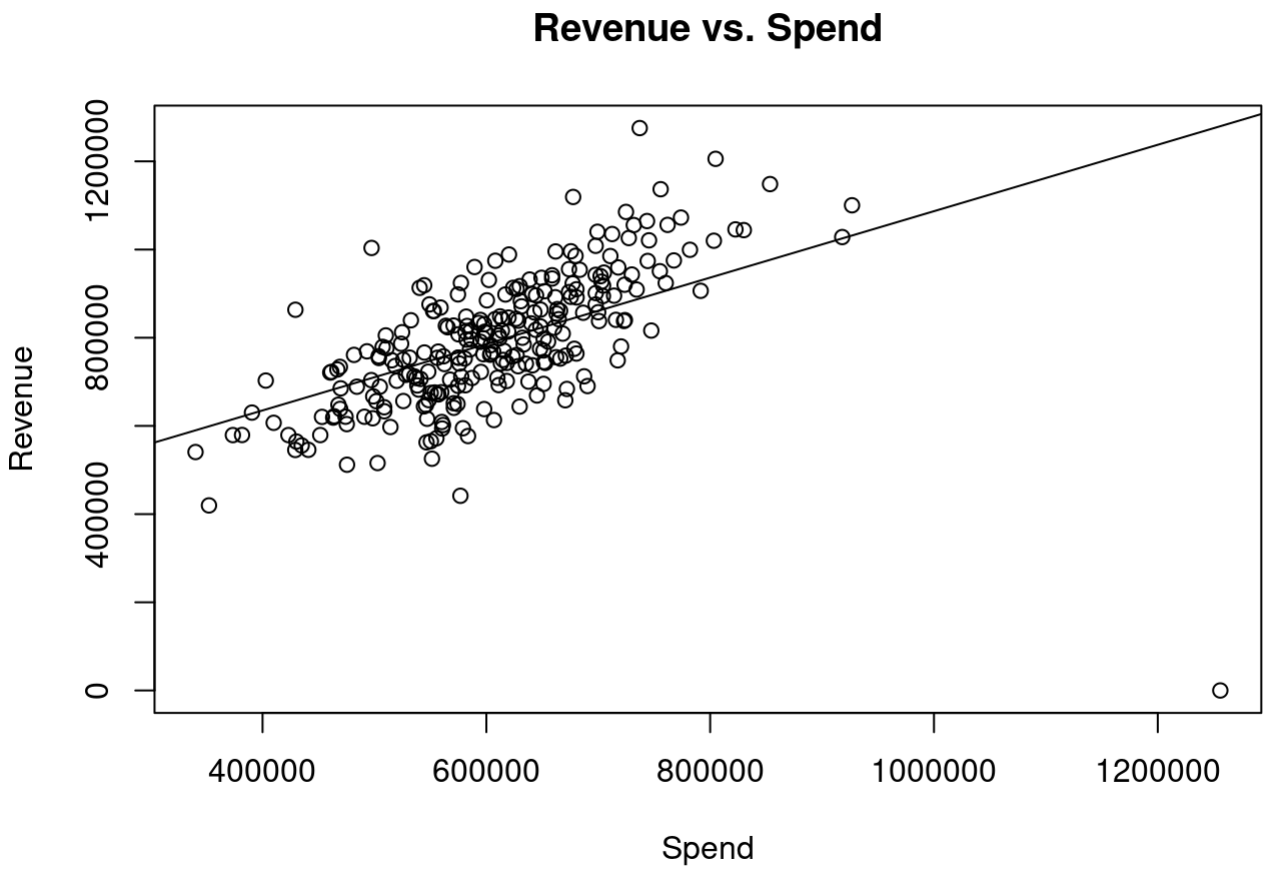
```
# build model

model1 = lm(Revenue ~ Spend, data=data )
model1
```

```
##
## Call:
## lm(formula = Revenue ~ Spend, data = data)
##
## Coefficients:
## (Intercept)        Spend
##   3.339e+05    7.531e-01
```

```
# summary
summary(model1)
```

```
##
## Call:
## lm(formula = Revenue ~ Spend, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1279716   -70754     3568    63913   386477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.339e+05  4.642e+04   7.193 6.49e-12 ***
## Spend       7.531e-01  7.540e-02   9.988  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128200 on 265 degrees of freedom
## Multiple R-squared:  0.2735, Adjusted R-squared:  0.2708
## F-statistic: 99.77 on 1 and 265 DF,  p-value: < 2.2e-16
```

```
# visualization
plot(Revenue ~ Spend,data = data, main = "Revenue vs. Spend")
abline(model1)
```



```r
#remove the outlier in the corner
data <- data %>% filter(Revenue != 0)
#retrain model
model1 = lm(Revenue ~ Spend, data=data)
#revisualise
plot(Revenue ~ Spend,data = data, main = "Revenue vs. Spend")
abline(model1)
```


```r
summary(model1)
```

```
##
## Call:
## lm(formula = Revenue ~ Spend, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -322296  -61120    1873   59363  336863
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.346e+05  3.737e+04    3.60 0.000379 ***
## Spend       1.091e+00  6.106e-02   17.86  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95930 on 264 degrees of freedom
## Multiple R-squared:  0.5473, Adjusted R-squared:  0.5456
## F-statistic: 319.2 on 1 and 264 DF,  p-value: < 2.2e-16
```

Analysis

```
# predictions
newdata.5 = data.frame(Spend = 500000)
predict(model1,newdata.5,interval="predict")
```

```
##        fit      lwr      upr
## 1 680003.4 490347.1 869659.6
```

```
newdata.7 = data.frame(Spend = 700000)
predict(model1,newdata.7,interval="predict")
```

```
##      fit      lwr     upr
## 1 898181 708591.9 1087770
```

```
# let's look at cost effectiveness
# if we spend 500k$
490347.1 - 500000
```

```
## [1] -9652.9
```

```
869659.6 - 500000
```

```
## [1] 369659.6
```

```
# if we spend 700k$
708591.9 - 700000
```

```
## [1] 8591.9
```

```
1087770 - 700000
```

```
## [1] 387770
```

note: I have removed the a single outlying point as it is not representative of the entire data set at hand; standard error is lower in when the outlier is removed making the model more useful for these particular asks.

Conclusion

According to my model, if the company spends $500k on advertising, the lowest predicted revenue is $490347.1 and the highest predicted revenue is $869659.6 If the company rather chooses to spend $700k on advertising, the lowest predicted revenue is $708591.9 and the highest predicted revenue is $1087770 I would argue that because there is more to gain ($18110.4 more) and less to lose ($18244.8 less), the company should pick the $700k spend option. My model implies that there is simply less risk if the company limits their spend to this number.