Linear Regression Project

Jacob Kramp

2/25/2021

Introduction

We have a data set that contains data for 2 random variables, Spending and Revenue, for several company's first year of spending. We would like to evaluate and provide any statistical evidence of a significant relationship between these two variables, then attempt to make a prediction using some sort of a model (probabaly linear) for a hypothetical company .They are requesting a prediction of what potential revenue they would obtain if they spent/invested $500k or 700k and we would like to provide any insights we can given the data that we have.

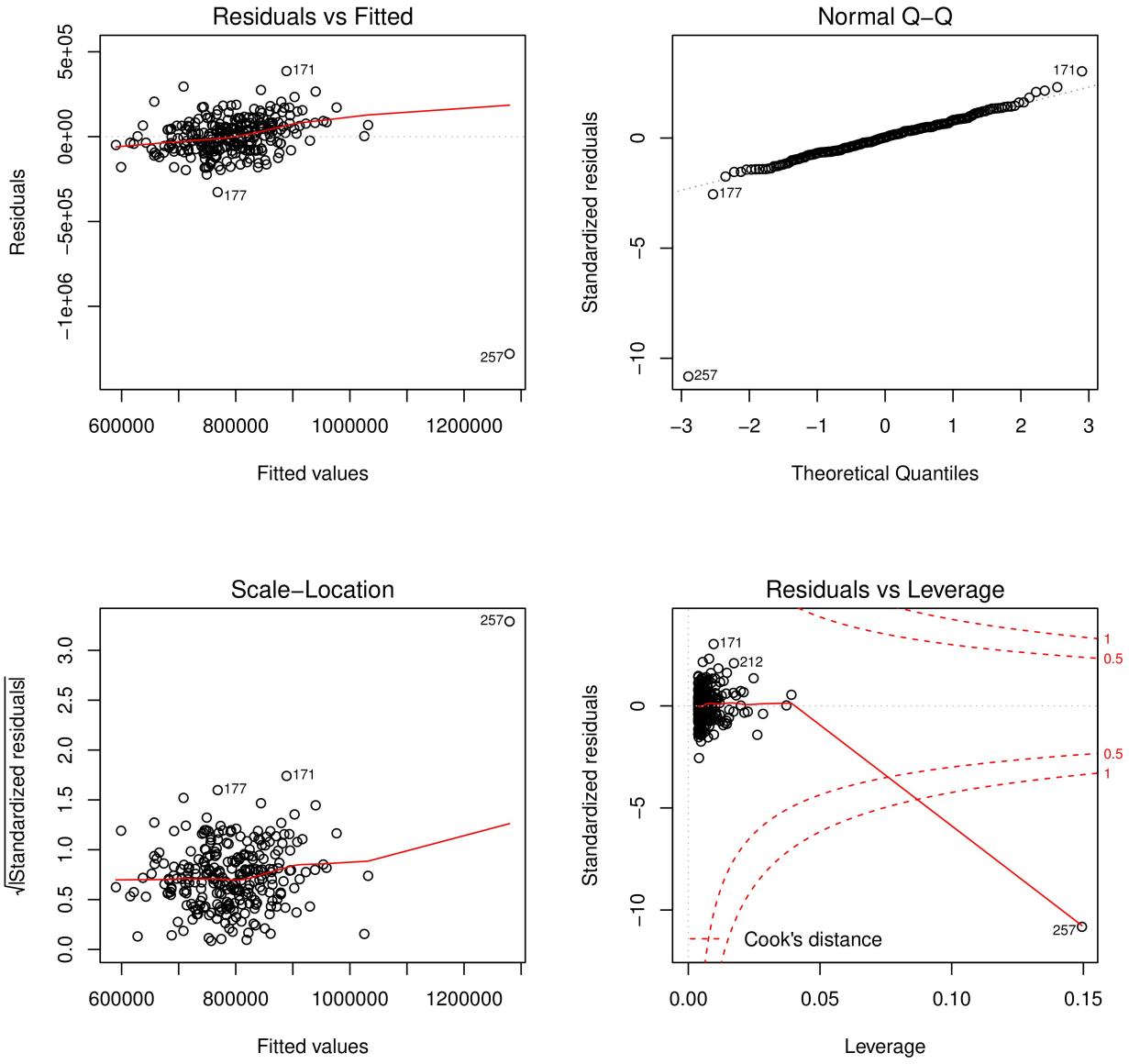
Data Exploration

Using the summary function in R, we can generate some information on our variables, Spend and Revenue:

```
##      Spend            Revenue
##  Min.   : 340098   Min.   :      0
##  1st Qu.: 546323   1st Qu.: 692392
##  Median : 603883   Median : 780149
##  Mean   : 606846   Mean   : 790925
##  3rd Qu.: 667138   3rd Qu.: 893622
##  Max.   :1255897   Max.   :1275447
```

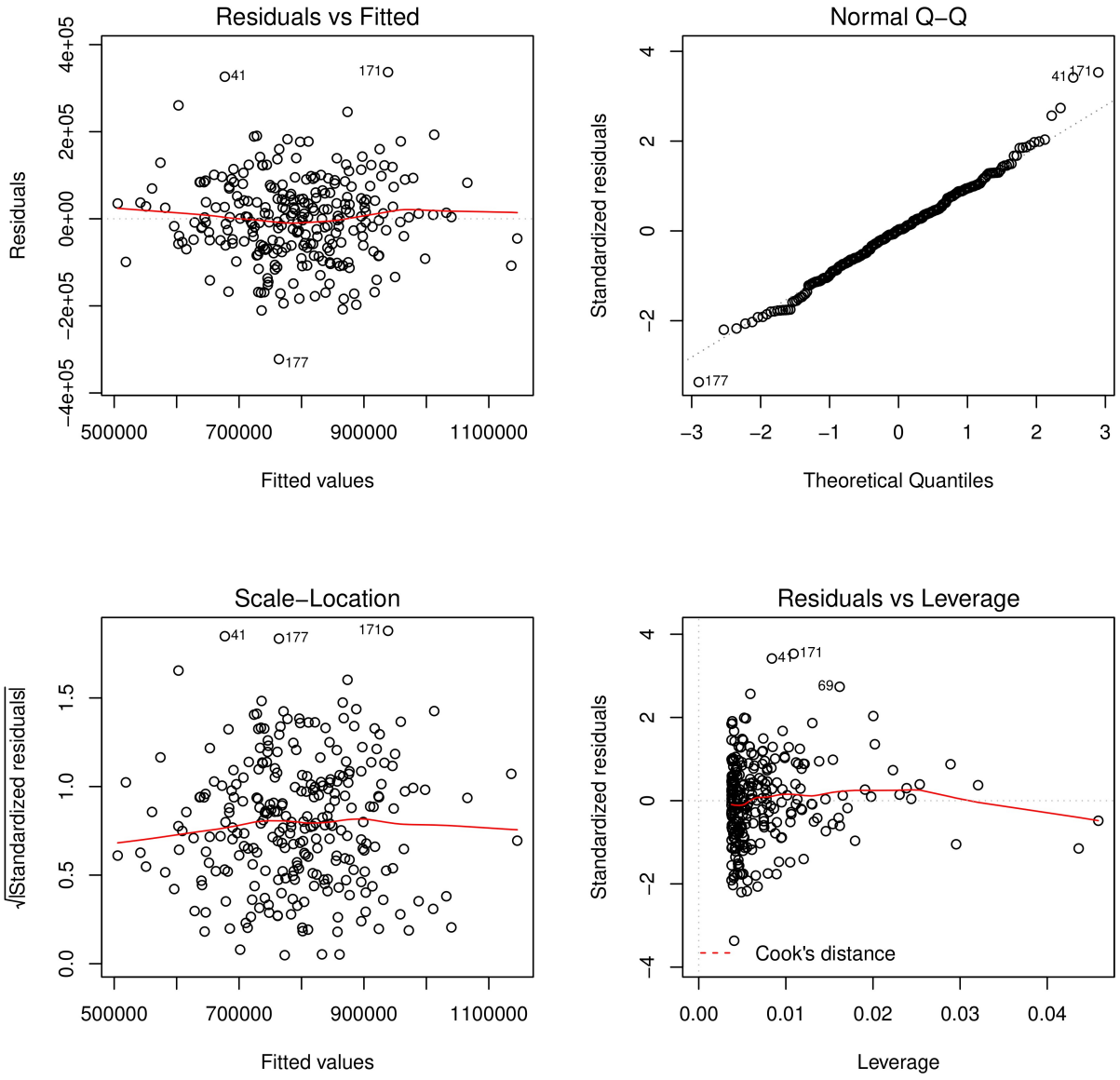
This information provides us with a domain for which we can make predictions. We cannot make a prediction outside of the min or max of either of these variables if they're the predictor.Luckily, 500k and 700k both appear within the domains of each of our variables based on this dataset and we may be able to make reasonable predictions if we can create a model.

We are going to attempt to fit a simple linear regression model and evaluate the diagnostic plots to determine whether or not we can reasonably use the linear model.



There appears to be a statistical outlier that is seriously breaking our strong linear model. This outlier labeled "257" in our plots is bending our Residuals vs. Fitted plot upwards since our line of best fit will have a large residual for that point. Thankfully, it doesn't seem to be breaking our Q-Q plot too badly, but the scale-location does have a bend on the right edge of our plot. This point also violates our Cook's distance line and is telling us that this point may be significantly altering our line of best fit. We could remove this outlier from our dataset and these problems would likely be solved. Since this is only one point and seems to be a rare case, removing the point may be reasonable.

Upon further investigation, the point in question is the coordinate pair $(1255897, 0)$, a company that invested over a million dollars and didn't see any revenue in the following year. This should be a very rare case and likely a special circumstance and I think it's safe to remove it from our dataset.



The Residuals vs. Fitted values line is relatively flat, the normal Q-Q plot follows a line, the scale location plot is relatively flat with no strong bends, and the Residuals vs. Leverage plot shows no points outside of the Cook's Line plot. This is a strong subjective interpretation that our relationship is linear. The estimated linear coefficent, $\hat{B}_1$ in $y = \hat{B}_0 + \hat{B}_1 x$ has an estimated value of 1.091e+00 with a calculated p-value of $<$ 2e-16 ***. This indicates a statistically significant relationship between x and y using our linear model.We can now conclude based on these diagnostic plots as well as the p-value generated from the following summary of our model fit that this data is linear.

```
##
## Call:
## lm(formula = Revenue ~ Spend, data = comp.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
```

Model Predictions

Now that we have a model that fits our data pretty well, we can generate prediction estimates and intervals given some input of interest. We are asked for our predictions using the model we built for 500k and 700k investments. Using the predict function in R, we are able to generate the following table.

Spend	fit	lwr	upr
5e+05	680003.4	490347.1	869659.6
7e+05	898181.0	708591.9	1087770.0

The model prediction intervals we have generated seem to lean favorable to investing 700k instead of 500k. The prediction interval for 500K includes values below the investment value and therefore, an investment of 500k may lead to a loss over the first year of net income for the company. They are more likely to profit if they invest 700k.

Index (Code)

```
library(data.table)

comp.data <- data.frame(fread('hw2.csv'))
str(comp.data)
summary(comp.data)

fit.data = lm(Revenue ~ Spend,data = comp.data)
summary(fit.data)
par(mfrow = c(2,2))
plot(fit.data)

comp.data <- comp.data[-257,]
#str(comp.data)
s <- summary(comp.data)
fit.data = lm(Revenue ~ Spend,data = comp.data)
sf <- summary(fit.data)
par(mfrow = c(2,2))
plot(fit.data)

predict(fit.data)

new <- data.frame(Spend = c(500000,700000))
cbind(new,predict(fit.data,new,interval = 'prediction'))
```