Business Analytics- Regression Analytics

Steven Pavliga

2024-10-30

1a: Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X?

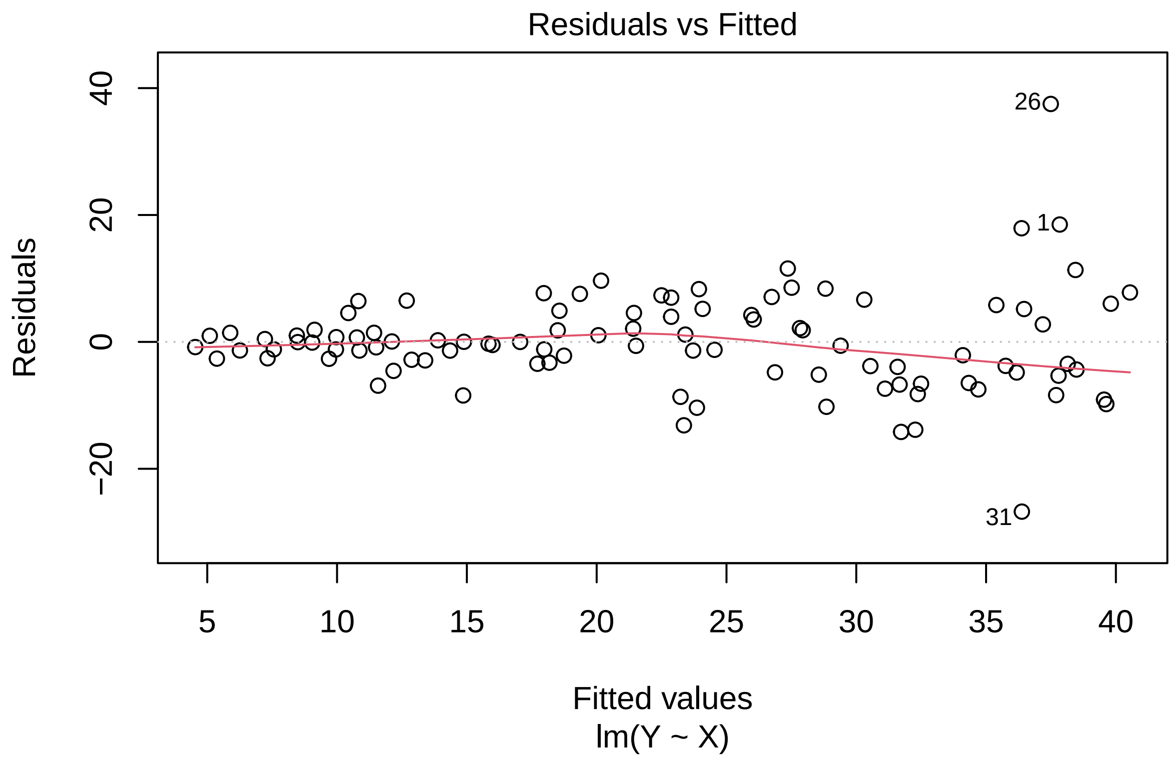
```r
library(dplyr)
```

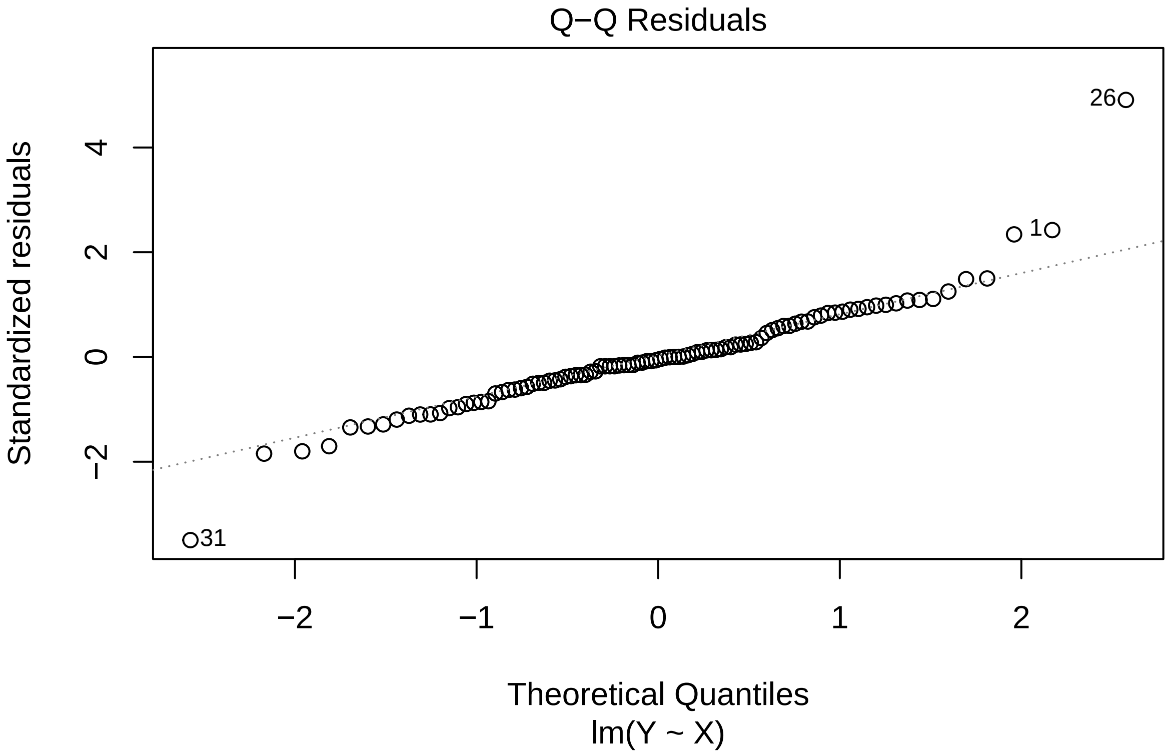
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

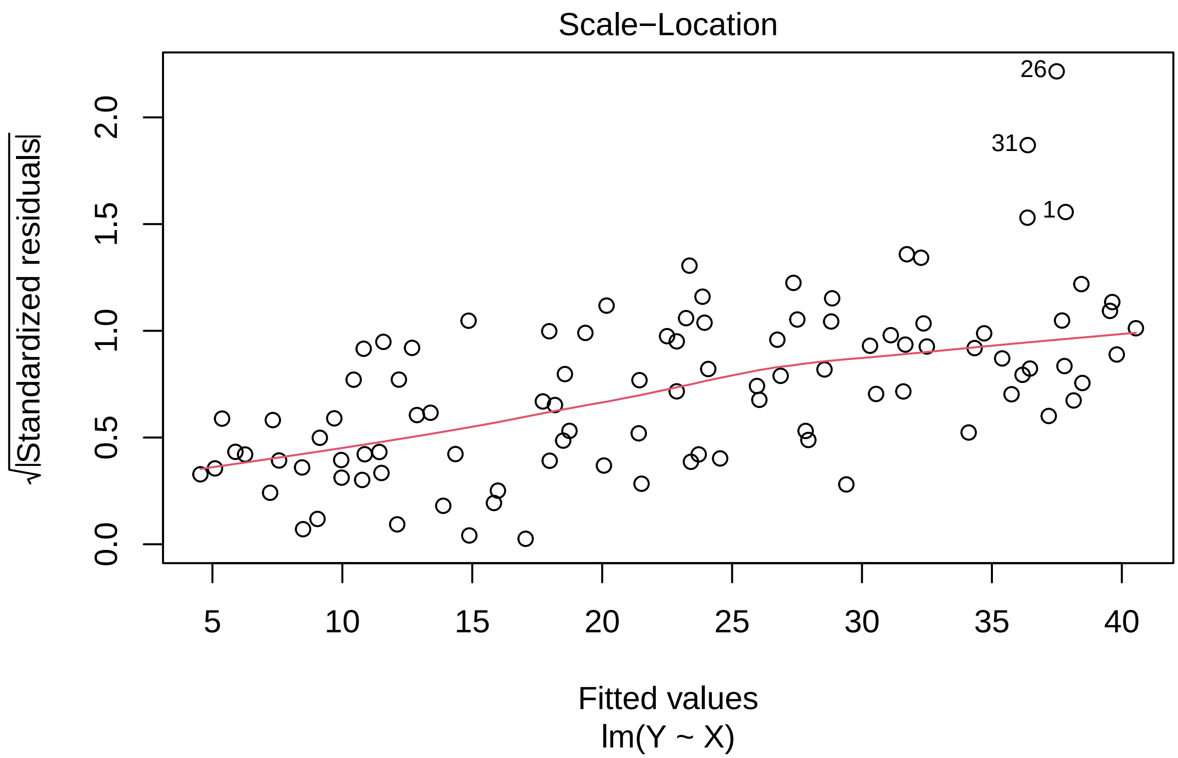
```r
library(ISLR)


set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y

Plot1 <- lm(Y ~ X)
Plot1
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

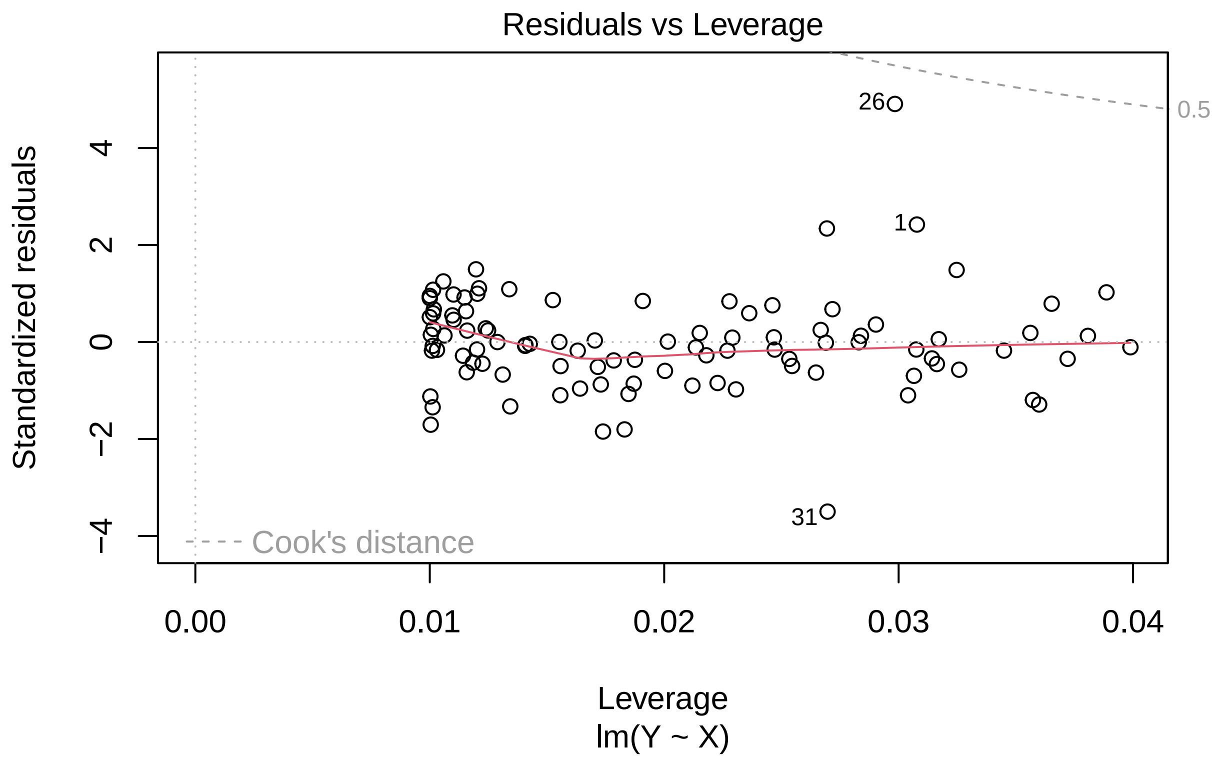
```r
plot(Plot1)
```


Residuals vs Fitted
Residuals
40
20
0
−20
26
1
31
5
10
15
20
25
30
35
40
Fitted values
lm(Y ~ X)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Y ~ X)
26
1
31
−2
0
2
4
−2
−1
0
1
2



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
26
31
1
5
10
15
20
25
30
35
40
Fitted values
lm(Y ~ X)



Yes, aside from a few outliers the linear model appears to fairly accurately predict Y from X.

1b, c: Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model?

How the Coefficient of Determination, R^2, of the model above is related to the correlation coefficient of X and Y?

```r
summary(Plot1)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
```

```
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

The simple linear model is y = 3.611x + 4.465

The R^2 is 0.6517, meaning 65.17% of variance in y is accounted for by x.

2a: James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question.

```r
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
JamesLM <- lm(mtcars$hp ~ mtcars$wt)
JamesLM
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt)
##
## Coefficients:
## (Intercept)    mtcars$wt
##      -1.821       46.160
```

```r
summary(JamesLM)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## mtcars$wt     46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```r
ChrisLM <- lm(mtcars$hp ~ mtcars$mpg)
ChrisLM
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg)
##
## Coefficients:
## (Intercept)   mtcars$mpg
##      324.08        -8.83
```

```r
summary(ChrisLM)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mtcars$mpg     -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Based on the R^2 values, mpg (Chris) appears to appears to be a better estimator than hp (James).

2b: Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```r
HPModel <- lm(hp ~ cyl + mpg, data = mtcars)
HPModel
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl          mpg
##      54.067       23.979       -2.775
```

```r
TwoBData <- data.frame(cyl = 4, mpg = 22)
TwoBModel <- predict(HPModel, newdata = TwoBData)
TwoBModel
```

```
##        1
## 88.93618
```

The estimated horse power is ~88.9hp.

3a: Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R 2)

```r
library(mlbench)
data(BostonHousing)

str(BostonHousing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
MedValueModel <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)
MedValueModel
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Coefficients:
## (Intercept)         crim           zn      ptratio        chas1
##    49.91868     -0.26018      0.07073     -1.49367      4.58393
```

```r
summary(MedValueModel)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

The R^2 value is around .36, meaning the model does not appear to be very accurate.

3b: Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much?

Based on the standard deviation of chas1, it appears the house that bounds the Charles River would be more expensive by ~$4584.

3c:

Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.

All of the p values are <.05, meaning they are all significant.

3d:

Use the anova analysis and determine the order of importance of these four variables.

```r
anova(MedValueModel)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based off sum of squares, crim would be most important, followed by ptratio, followed zn, followed by chas being least important.