1 a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R. How many rows are in this data set? How many columns? What do the rows and columns represent?

Ans- crim zn indus chas nox rm age dis rad tax ptratio black lstat

1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98

2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14

3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03

4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94

5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33

6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21

medv

1 24.0

2 21.6

3 34.7

4 33.4

5 36.2

6 28.7

[1] 506 14

As we can see each row will represent the set of the predictor for an observation given neighborhood in Boston. And the column represents each predictors variable that was made in 506 neighbor Boston.

1 b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings:

Ans - '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 : int 0 0 0 0 0 0 0 0 0 0 ...

$ 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 : int 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 ...

$ black : 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 ...

Diagram

Description automatically generated with medium confidence

As we can see that some variables appear to be correlated.

1 c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

Ans - ## i j cor p

## 1 crim zn -0.20046922 5.506472e-06

## 2 crim indus 0.40658341 0.000000e+00

## 4 crim chas -0.05589158 2.094345e-01

## 7 crim nox 0.42097171 0.000000e+00

## 11 crim rm -0.21924670 6.346703e-07

## 16 crim age 0.35273425 2.220446e-16

## 22 crim dis -0.37967009 0.000000e+00

## 29 crim rad 0.62550515 0.000000e+00

## 37 crim tax 0.58276431 0.000000e+00

## 46 crim ptratio 0.28994558 2.942924e-11

## 56 crim black -0.38506394 0.000000e+00

## 67 crim lstat 0.45562148 0.000000e+00

## 79 crim medv -0.38830461 0.000000e+00

As we can see based on the correlation coefficients and their corresponding p-values, there is indeed an association between the per capita crime rate (crim) and the other predictors.

1 d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

Ans - As we can see for each their summary and their ggplot with the number of suburbs for the crime rate, Tax rates, Pupil-teacher ratios? The crime rate summary shown below with its plot.

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00632 0.08204 0.25651 3.61352 3.67708 88.97620

Histogram

Description automatically generated with medium confidence

This is for the Boston tax

Min. 1st Qu. Median Mean 3rd Qu. Max.

187.0 279.0 330.0 408.2 666.0 711.0

Chart, histogram

Description automatically generated

And lastly, we have the Boston pupil teacher ration

Min. 1st Qu. Median Mean 3rd Qu. Max.

12.60 17.40 19.05 18.46 20.20 22.00

Graphical user interface, chart

Description automatically generated

But considering that the median and maximum crime rate values are respectively about 0.26% and 89%, there are indeed some neighborhoods where the crime rate is alarmingly high which brings me o the crime rate above 10 percent. As we can see below 11 percent of the neighbor have a crime rate above 10%

[1] 0.1067194

[1] 0.007905138

0.8% of the neighborhoods have crim rates above 50%

Based on the histogram of the Tax rates, they are few neighborhoods where rates are relative higher. The median and average tax amount are $330 and $408.20 (per Full-value property-tax rate per $10,000) respectively.

[1] 0.729249

73% of the neighborhood pay under $600.

[1] 0.270751

27% of the neighborhood pay over $600.

1 e) How many of the suburbs in this data set bound the Charles River?

Ans - [1] 35

1 f) What is the median pupil-teacher ratio among the towns in this data set?

Ans - Min. 1st Qu. Median Mean 3rd Qu. Max.

12.60 17.40 19.05 18.46 20.20 22.00

The median pupil-teacher ratio is 19 pupils for each teacher.

1 g) Which suburb of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors?

Ans - crim zn indus chas nox rm age dis rad tax ptratio black

399 38.3518 0 18.1 0 0.693 5.453 100 1.4896 24 666 20.2 396.9

lstat medv

399 30.59 5

Suburb #399 with a median value of $5000. Suburb 399 can be classified as one of the least desirable places to live in Boston.

1 h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

Ans - [1] 64 There are 64 suburbs with more than 7 rooms per dwelling.

[1] 13 There are 13 suburbs with more than 7 rooms per dwelling.

crim zn indus chas

Min. :0.02009 Min. : 0.00 Min. : 2.680 Min. :0.0000

1st Qu.:0.33147 1st Qu.: 0.00 1st Qu.: 3.970 1st Qu.:0.0000

Median :0.52014 Median : 0.00 Median : 6.200 Median :0.0000

Mean :0.71879 Mean :13.62 Mean : 7.078 Mean :0.1538

3rd Qu.:0.57834 3rd Qu.:20.00 3rd Qu.: 6.200 3rd Qu.:0.0000

Max. :3.47428 Max. :95.00 Max. :19.580 Max. :1.0000

nox rm age dis

Min. :0.4161 Min. :8.034 Min. : 8.40 Min. :1.801

1st Qu.:0.5040 1st Qu.:8.247 1st Qu.:70.40 1st Qu.:2.288

Median :0.5070 Median :8.297 Median :78.30 Median :2.894

Mean :0.5392 Mean :8.349 Mean :71.54 Mean :3.430

3rd Qu.:0.6050 3rd Qu.:8.398 3rd Qu.:86.50 3rd Qu.:3.652

Max. :0.7180 Max. :8.780 Max. :93.90 Max. :8.907

rad tax ptratio black

Min. : 2.000 Min. :224.0 Min. :13.00 Min. :354.6

1st Qu.: 5.000 1st Qu.:264.0 1st Qu.:14.70 1st Qu.:384.5

Median : 7.000 Median :307.0 Median :17.40 Median :386.9

Mean : 7.462 Mean :325.1 Mean :16.36 Mean :385.2

3rd Qu.: 8.000 3rd Qu.:307.0 3rd Qu.:17.40 3rd Qu.:389.7

Max. :24.000 Max. :666.0 Max. :20.20 Max. :396.9

lstat medv

Min. :2.47 Min. :21.9

1st Qu.:3.32 1st Qu.:41.7

Median :4.14 Median :48.3

Mean :4.31 Mean :44.2

3rd Qu.:5.12 3rd Qu.:50.0

Max. :7.44 Max. :50.0

2 a) Using the rnorm () function, create a vector, x, containing 100 observations drawn from a N (0, 1) distribution. This represents a feature, X.

Ans - x = rnorm (100, mean= 0, sd =1)

2 b*) Using* the rnorm() function, create a vector, eps, containing 100 observations drawn from a N(0, 0.25) distribution i.e. a normal distribution with mean zero and variance 0.25.

Ans - eps = rnorm(100, mean =0, sd = 0.25)

2 c) Using x and eps, generate a vector y according to the model Y = −1+0.5X + eps. What is the length of the vector y? What are the values of βo and β1 in this linear model?

Ans - [1] 100

2 d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

Ans – As shown below we cn see that this is a linear relationship going on in the plot.

*Chart, scatter chart

Description automatically generated*

2 e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do β^o and β^1 compare to βo and β1?

Ans – As we can see below the β^o and β^1 are practically equal to βo and β1. But as expected, β^o and β^1 are statistically significant and R-square = 0.84 which shows that the model fits the data quite well.

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-0.53703 -0.14803 0.02733 0.15956 0.62366

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.03074 0.02409 -42.79 <2e-16 \*\*\*

x 0.47094 0.02544 18.51 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2401 on 98 degrees of freedom

Multiple R-squared: 0.7776, Adjusted R-squared: 0.7754

F-statistic: 342.7 on 1 and 98 DF, p-value: < 2.2e-16

2 f) Display the least squares line on the scatterplot obtained in (d).Draw the population regression line on the plot, in a different color. Use the legend () command to create an appropriate legend.

Ans - *Chart, scatter chart

Description automatically generated*

2 g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

Ans – As we can see below, we can see that the regression coefficient of the quadratic term is not statistically significant; hence that, there is no evidence that the quadratic term improves the model.

Call:

lm(formula = y ~ poly(x, 2))

Residuals:

Min 1Q Median 3Q Max

-0.63559 -0.17951 -0.01978 0.15510 0.77204

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.03425 0.02398 -43.136 <2e-16 \*\*\*

poly(x, 2)1 5.72722 0.23976 23.887 <2e-16 \*\*\*

poly(x, 2)2 0.45334 0.23976 1.891 0.0616 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2398 on 97 degrees of freedom

Multiple R-squared: 0.8555, Adjusted R-squared: 0.8525

F-statistic: 287.1 on 2 and 97 DF, p-value: < 2.2e-16

2 h) Repeat (a)–(f) after modifying the data generation process in such a way that there is less noise in the data. The model (part-c) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term in (b). Describe your results.

Ans – As we can see below that the *R-square =* 0.95 shows that this model fits the data better. And this observation in (1) is evidenced graphically by how well the regression line fits the data points. While the quadratic term is, again, statistically insignificant.

Call:

lm(formula = y ~ x)

Residuals:

Min 1Q Median 3Q Max

-0.258409 -0.060633 -0.000761 0.064953 0.231338

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.006430 0.009875 -101.92 <2e-16 \*\*\*

x 0.507536 0.009133 55.57 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09832 on 98 degrees of freedom

Multiple R-squared: 0.9692, Adjusted R-squared: 0.9689

F-statistic: 3088 on 1 and 98 DF, p-value: < 2.2e-16

Chart, scatter chart

Description automatically generated

Call:

lm(formula = y ~ poly(x, 2))

Residuals:

Min 1Q Median 3Q Max

-0.25527 -0.06407 -0.00005 0.06457 0.23355

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.057261 0.009872 -107.097 <2e-16 \*\*\*

poly (x, 2)1 5.463915 0.098720 55.348 <2e-16 \*\*\*

poly (x, 2)2 0.045654 0.098720 0.462 0.645

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09872 on 97 degrees of freedom

Multiple R-squared: 0.9693, Adjusted R-squared: 0.9687

F-statistic: 1532 on 2 and 97 DF, p-value: < 2.2e-16

3 a) The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

Ans - The regression coefficients are βo = 2+rnorm (100), β1 = 2, and β3 = 0.3

3 b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

Ans - [1] 0.8351212

Chart, scatter chart

Description automatically generated

1. c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are βˆ0, βˆ1, and βˆ2? How do these relate to the true βo, β1, and β2? Can you reject the null hypothesis Ho: β1 = 0? How about the null hypothesis Ho : β2 = 0?

Ans – As we can see below our coefficient estimates are βo = 2.1305, β1 = 1.4396 , and β3 = 1.0097, meaning this are poor estimates. But for β1, we can only reject the null hypothesis at a 99% level of confidence (not 95%). And for β2, we must reject the null hypothesis.

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*

x1 1.4396 0.7212 1.996 0.0487 \*

x2 1.0097 1.1337 0.891 0.3754

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 97 degrees of freedom

Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925

F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

3 d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis Ho: β1 = 0?

Ans – As we can see below, we have to reject the null hypothesis for the regression coefficient β1 because of how very low the associated p-value is and In addition, per the R-squared value, the predictor x1 can, on its own, explain about 20% of the changes in the response variable y.

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*

x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942

F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

3 e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis Ho: β2 = 0?

Ans – As we can see below, we can reject the null hypothesis for the regression coefficient β2 because of the very low associated p-value. Per the R-squared value, the predictor x2 can, on its own, explain about 17% of the changes in the response variable y.

Call:

lm(formula = y ~ x2)

Residuals:

Min 1Q Median 3Q Max

-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*

x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom

Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679

F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

3f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

Ans – For part C based on our evidence from the multi-linear regression model, we discarded x1 and x2 as significant predictor variables. But in part E, separate linear regression models with each of the predictors show that x1 and x2 are quite significant and can help explain up to 20% and 17% of the changes in y respectively. So, I would say, this is a contradiction.

3 g) Now suppose we obtain one additional observation, which was unfortunately mismeasured-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

Ans - Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-2.73348 -0.69318 -0.05263 0.66385 2.30619

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.2267 0.2314 9.624 7.91e-16 \*\*\*

x1 0.5394 0.5922 0.911 0.36458

x2 2.5146 0.8977 2.801 0.00614 \*\*

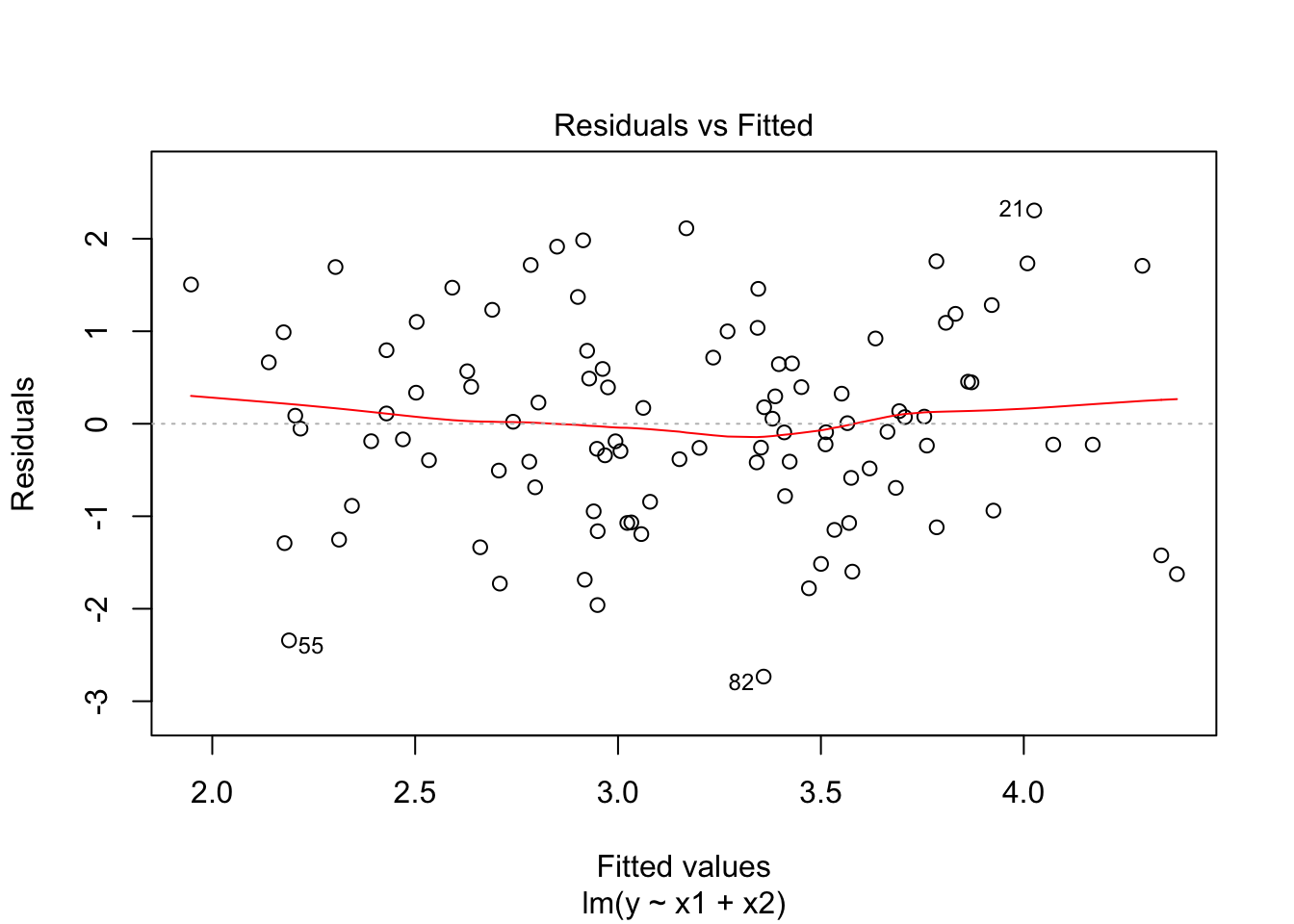
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.075 on 98 degrees of freedom

Multiple R-squared: 0.2188, Adjusted R-squared: 0.2029

F-statistic: 13.72 on 2 and 98 DF, p-value: 5.564e-06



Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

As we can see in both cases, the x1 is statistically significant to the variations in y. But however, we observe a sizeable negative change is the R-squared value. Hence, this model is worse than the former. And for this model, the added data point is a relative outlier. It is not an absolute outlier in the sense that the magnitude of its separation from the “normal” range is quite small.

Call:

lm(formula = y ~ x2)

Residuals:

Min 1Q Median 3Q Max

-2.64729 -0.71021 -0.06899 0.72699 2.38074

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3451 0.1912 12.264 < 2e-16 \*\*\*

x2 3.1190 0.6040 5.164 1.25e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 99 degrees of freedom

Multiple R-squared: 0.2122, Adjusted R-squared: 0.2042

F-statistic: 26.66 on 1 and 99 DF, p-value: 1.253e-06

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Now for x2 in both cases, the x2 is statistically significant to the variations in y. But however, we observe a sizeable positive change is the R-squared value. Hence, this model is better than the former. So, for this model, when using cook’s distance as a reference, the added data point is not an outlier or leverage point.