**STAT 350 – Tuo Wang**

**Due Date: Sunday March 27th by the end of the day**

**Data sets for the following problems are posted in our group site in StatCrunch**

**Problem 1: Low Temperature versus Elevation**

Investigate the relationship between the record low temperature and the elevation where it was recorded.

a) First, graph a scatterplot of the 50 states and DC on these two variables (elevation is the explanatory variable). Interpret the graph and provide the correlation.

> elevation = c(760,1100,8180,1250,5532,5920,480,20,410,193,1000,13770,6285,635,785,770,1812,730,194,770,2461,640,785,1460,420,700,5470,3379,5200,6262,70,7350,1720,6525,1929,800,958,4700,1500,425,3115,2277,2471,3275,8092,915,3870,2120,2200,1300,6500)

> temp = c(-27.00,-80.00,-40.00,-29.00,-45.00,-61.00,-32.00,-17.00,-15.00,-2.00,-17.00,12.00,-60.00,-36.00,-36.00,-47.00,-40.00,-37.00,-16.00,-50.00,-40.00,-35.00,-51.00,-60.00,-19.00,-40.00,-70.00,-47.00,-50.00,-47.00,-34.00,-50.00,-52.00,-34.00,-60.00,-39.00,-27.00,-54.00,-42.00,-25.00,-19.00,-58.00,-32.00,-23.00,-69.00,-50.00,-30.00,-48.00,-37.00,-55.00,-66.00)

> data = data.frame(elevation,temp)

> attach(data)

> data

elevation temp

1 760 -27

2 1100 -80

3 8180 -40

4 1250 -29

5 5532 -45

6 5920 -61

7 480 -32

8 20 -17

9 410 -15

10 193 -2

11 1000 -17

12 13770 12

13 6285 -60

14 635 -36

15 785 -36

16 770 -47

17 1812 -40

18 730 -37

19 194 -16

20 770 -50

21 2461 -40

22 640 -35

23 785 -51

24 1460 -60

25 420 -19

26 700 -40

27 5470 -70

28 3379 -47

29 5200 -50

30 6262 -47

31 70 -34

32 7350 -50

33 1720 -52

34 6525 -34

35 1929 -60

36 800 -39

37 958 -27

38 4700 -54

39 1500 -42

40 425 -25

41 3115 -19

42 2277 -58

43 2471 -32

44 3275 -23

45 8092 -69

46 915 -50

47 3870 -30

48 2120 -48

49 2200 -37

50 1300 -55

51 6500 -66

> plot(data$elevation,data$temp,ylab="temperature(F)",xlab="elevation",main="scatter plot")

>

> cor(temp,elevation)

[1] -0.1177765



Form – I can’t see a straight line relationship between the variables. The graph curves down slightly. I can’t see a pattern.

Direction – The data (Except the outlier) runs from upper left to lower right with a negative direction.

Strength – The graph has many scatters with one obvious outlier observed. The relationship between the points is weak. ( correlation is -0.1177765, no linear relationship)

b) Second, delete Hawaii’s observation and produce another scatterplot. Interpret your new scatterplot and provide the new correlation.

> elevation1 = c(760,1100,8180,1250,5532,5920,480,20,410,193,1000,6285,635,785,770,1812,730,194,770,2461,640,785,1460,420,700,5470,3379,5200,6262,70,7350,1720,6525,1929,800,958,4700,1500,425,3115,2277,2471,3275,8092,915,3870,2120,2200,1300,6500)

> temp1 = c(-27.00,-80.00,-40.00,-29.00,-45.00,-61.00,-32.00,-17.00,-15.00,-2.00,-17.00,-60.00,-36.00,-36.00,-47.00,-40.00,-37.00,-16.00,-50.00,-40.00,-35.00,-51.00,-60.00,-19.00,-40.00,-70.00,-47.00,-50.00,-47.00,-34.00,-50.00,-52.00,-34.00,-60.00,-39.00,-27.00,-54.00,-42.00,-25.00,-19.00,-58.00,-32.00,-23.00,-69.00,-50.00,-30.00,-48.00,-37.00,-55.00,-66.00)

> data1 = data.frame(elevation1, temp1)

> attach(data1)

> data1

elevation1 temp1

1 760 -27

2 1100 -80

3 8180 -40

4 1250 -29

5 5532 -45

6 5920 -61

7 480 -32

8 20 -17

9 410 -15

10 193 -2

11 1000 -17

12 6285 -60

13 635 -36

14 785 -36

15 770 -47

16 1812 -40

17 730 -37

18 194 -16

19 770 -50

20 2461 -40

21 640 -35

22 785 -51

23 1460 -60

24 420 -19

25 700 -40

26 5470 -70

27 3379 -47

28 5200 -50

29 6262 -47

30 70 -34

31 7350 -50

32 1720 -52

33 6525 -34

34 1929 -60

35 800 -39

36 958 -27

37 4700 -54

38 1500 -42

39 425 -25

40 3115 -19

41 2277 -58

42 2471 -32

43 3275 -23

44 8092 -69

45 915 -50

46 3870 -30

47 2120 -48

48 2200 -37

49 1300 -55

50 6500 -66

> plot(data1$elevation1,data1$temp1)

> plot(data1$elevation1,data1$temp1,ylab="Temperature(F)",xlab="Elevation",main="scatter plot without Hawai")

> cor(elevtiona1,temp1)

Error in is.data.frame(x) : object 'elevtiona1' not found

> cor(elevation1,temp1)

[1] -0.4599226



Form – I can see a somewhat straight line relationship between the variables. The graph curves down slightly. The pattern is not too strong to observe.

Direction – The pattern runs from upper left to lower right with a negative direction

Strength – The graph has medium level scatter with no obvious outlier. The relationship is stronger than in Question 1 but still not very strong with correlation = -0.4599226.

c) Discuss your findings in one to two sentences. Would you consider any other observations as outliers (use your second scatterplot to help).

When we trim the outlier, the pattern is stronger and clearer to observe. The strength becomes stronger in graph 2. The direction remains the same.

I would consider the point at around -80 and 1100 (Alaska) another outlier as I can tell the point is at the left corner on the graph and farm away from other points.

**Problem 2: Always plot your data!**

The data provided for this question present four sets of generated data to illustrate the dangers of conducting a regression analysis without first plotting the data.

**a) Without making scatterplots**, find the correlation and least squares regression line for all four data sets. What do you notice? Use the regression line to predict y for x = 10.

> xa = c(10,8,13,9,11,14,6,4,12,7,5)

> ya = c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68)

> xb = c(10,8,13,9,11,14,6,4,12,7,5)

> yb = c(9.14,8.14,8.74,8.77,9.26,8.1,6.13,3.1,9.13,7.26,4.74)

> xc = c(10,8,13,9,11,14,6,4,12,7,5)

> yc = c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73)

> xd = c(8,8,8,8,8,8,8,8,8,8,19)

> yd = c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,5.56,7.91,6.89,12.5)

> resulta = lm(ya~xa)

> resulta

Call:

lm(formula = ya ~ xa)

Coefficients:

(Intercept) xa

3.0001 0.5001

> cor(xa,ya)

[1] 0.8164205

> resultb = lm(yb~xb)

> resultb

Call:

lm(formula = yb ~ xb)

Coefficients:

(Intercept) xb

3.001 0.500

> cor(xb,yb)

[1] 0.8162365

> resultc = lm(yc~xc)

> resultc

Call:

lm(formula = yc ~ xc)

Coefficients:

(Intercept) xc

3.0025 0.4997

> cor(xc,yc)

[1] 0.8162867

> resultd = lm(yd~xd)

> resultd

> cor(xd,yd)

[1] 0.8165214

Call:

lm(formula = yd ~ xd)

Coefficients:

(Intercept) xd

3.0017 0.4999

For x= 10

* Ya = 3.0001+0.5001\*10 = 8.0011
* Yb = 3.001+0.500\*10 = 8.001
* Yc = 3.0025+0.4997\*10= 7.9995
* Yd= 3.0017+0.4999\*10 = 8.0007

I noticed that the *b0* and *b1*for those 4 data sets are almost identical and the correlation is also identical. When x = 10 the result is also almost identical = 8.

1. Now make a scatterplot for each of the data sets and add the regression line to each plot.

Ya = 3.0001+0.5001xa





Yb = 3.001+0.500\*xb





Yc = 3.0025+0.4997\*xc





Yd= 3.0017+0.4999\*xd





1. In which of the four cases would you be willing to use the regression line to describe the dependence of y on x? Explain your answer in each case.

* I would only use regression line to describe the relationship between X and Y in data set a). We can observe that from in scatter plot that the only the graph from data set a) has a normal straight-line relationship and the line fits the point appropriately.
* In b), it is more likely a polynomial relationship.
* In c), there is an obvious outlier that affects the regression line so it does not describe the relationship between X and Y properly. If we can trim the outlier and get a new straight line, that line would be a better fit.
* In d), the relationship between X an Y are similar in c), and we can also observe that there is no normal straight-line relationship but rather a vertical-line relationship between two variables in the scatter plot.

**Problem 3:**  ***Air Pollution in U.S. Cities***

For 41 cities in the United States, the following 3 variables were recorded:

Y: Sulphur dioxide content of air in micrograms per cubic meter (SO2)

X1: Number of manufacturing enterprises employing 20 or more workers (Manuf)

X2: Population size in thousands (Population)

This data has been collected to investigate the determinants of air pollution.

1. Use StatCrunch to conduct a complete regression analysis between variables Y and X1

> y = c(10,13,12,17,56,36,29,14,10,24,110,28,17,8,30,9,47,35,29,14,56,14,11,46,11,23,65,26,69,61,94,10,18,9,10,28,31,26,29,31,16)

> x1 = c(213,91,453,454,412,80,434,136,207,368,3344,361,104,125,291,204,625,1064,699,381,775,181,46,44,391,462,1007,266,1692,347,343,337,275,641,721,137,96,197,379,35,569)

> data = data.frame(x1,y)

> attach(data)

> data

x1 y

1 213 10

2 91 13

3 453 12

4 454 17

5 412 56

6 80 36

7 434 29

8 136 14

9 207 10

10 368 24

11 3344 110

12 361 28

13 104 17

14 125 8

15 291 30

16 204 9

17 625 47

18 1064 35

19 699 29

20 381 14

21 775 56

22 181 14

23 46 11

24 44 46

25 391 11

26 462 23

27 1007 65

28 266 26

29 1692 69

30 347 61

31 343 94

32 337 10

33 275 18

34 641 9

35 721 10

36 137 28

37 96 31

38 197 26

39 379 29

40 35 31

41 569 16

> plot(data$x1,data$y)



> cov(data)

x1 y

x1 317502.89 8527.7201

y 8527.72 550.9476

> cor(data)

x1 y

x1 1.0000000 0.6447687

y 0.6447687 1.0000000

> results=lm(y~x1)

> results

Call:

lm(formula = y ~ x1)

Coefficients:

(Intercept) x1

17.61057 0.02686

> abline(results)

> plot(data$x1,data$y, ylab="SO2 level", xlab="Number of manufacturing enterprises employed",col="blue",main="scatter graph with best fit line")

> abline(results)



> plot(x1,result$res)



> qqnorm(result$res)

> qqline(result$res)



> hist(result$res)



> summary(result)

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-26.976 -12.968 -3.495 6.710 67.177

Coefficients:

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

(Intercept) 17.610574 3.691587 4.770 2.58e-05 \*\*\*

x1 0.026859 0.005099 5.268 5.36e-06 \*\*\*

---

Signif. codes: 0 ‚Äò\*\*\*‚Äô 0.001 ‚Äò\*\*‚Äô 0.01 ‚Äò\*‚Äô 0.05 ‚Äò.‚Äô 0.1 ‚Äò ‚Äô 1

Residual standard error: 18.17 on 39 degrees of freedom

Multiple R-squared: 0.4157, Adjusted R-squared: 0.4007

F-statistic: 27.75 on 1 and 39 DF, p-value: 5.363e-06

> confint(results)

2.5 % 97.5 %

(Intercept) 10.14363536 25.07751339

x1 0.01654568 0.03717175

1. Predict the SO2 level for a city with 400 manufacturing enterprises.

Y =17.610574 + 0.026859\*x1,

when x1 = 400

Y = 17.610574 + 0.026859\*400 = 28.354174

1. You will notice that one city stands out in the model fit as being an outlier and another stands out as being an influential point. Locate these cities and comment.

We noticed that there is an outlier from the scatter plot in a), the outlier has both extreme X and Y value and that is Chicago with 110 = Y and X =3344. The city has extreme so2 level and enterprise numbers. It is not an idea observation in this conducted analysis.

The influential point is the outlier that greatly affects the slope of the regression line. We can observe that in the graph, which is Providence where Y = 94 and X = 343. The city has normal enterprise numbers but extreme high level so2. It is not an idea observation in the analysis maybe the factory is heavy-duty so they produce more SO2 than other cities. Anyway, it affects the slope of regression line and should be eliminated.

1. Use StatCrunch to investigate the correlation between variables X1 and X2

Comment on the value you obtain.

> x1 = c(213,91,453,454,412,80,434,136,207,368,3344,361,104,125,291,204,625,1064,699,381,775,181,46,44,391,462,1007,266,1692,347,343,337,275,641,721,137,96,197,379,35,569)

> x2 = c(582,132,716,515,158,80,757,529,335,497,3369,746,201,277,593,361,905,1513,744,507,622,347,244,116,463,453,751,540,1950,520,179,624,448,844,1233,176,308,299,531,71,717)

> data1 = data.frame(x1,x2)

> attach(data1)

> data1

x1 x2

1 213 582

2 91 132

3 453 716

4 454 515

5 412 158

6 80 80

7 434 757

8 136 529

9 207 335

10 368 497

11 3344 3369

12 361 746

13 104 201

14 125 277

15 291 593

16 204 361

17 625 905

18 1064 1513

19 699 744

20 381 507

21 775 622

22 181 347

23 46 244

24 44 116

25 391 463

26 462 453

27 1007 751

28 266 540

29 1692 1950

30 347 520

31 343 179

32 337 624

33 275 448

34 641 844

35 721 1233

36 137 176

37 96 308

38 197 299

39 379 531

40 35 71

41 569 717

> cor(data1)

x1 x2

x1 1.0000000 0.9552693

x2 0.9552693 1.0000000

>

The Correlation between x1 and x2 is around 0.95, hence we can say that the relationship between x1 and x2 or number of manufacturing enterprises and population is quite strong with an almost perfect positive linear relationship. More population with more enterprise and less population with less enterprise.