2) Salary\_hike -> Build a prediction model for Salary\_hike

------------------------------------------------------------

Build a simple linear regression model by performing EDA and do necessary transformations and select the best model using R or Python.

> a=read.csv(file.choose())

> View(a)

> attach(a)

> summary(a)

YearsExperience Salary

Min. : 1.100 Min. : 37731

1st Qu.: 3.200 1st Qu.: 56721

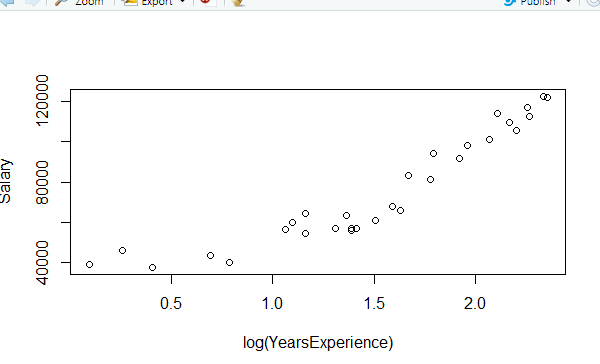
Median : 4.700 Median : 65237

Mean : 5.313 Mean : 76003

3rd Qu.: 7.700 3rd Qu.:100545

Max. :10.500 Max. :122391

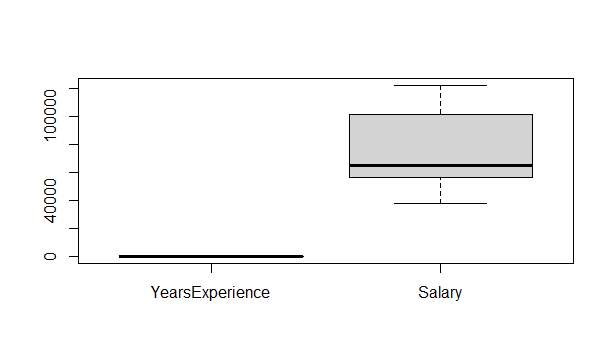
> plot(a$YearsExperience,a$Salary)



> cor(YearsExperience,Salary)

[1] 0.9782416

> boxplot(a)



> reg<-lm(Salary~YearsExperience)

> reg~residuals

reg ~ residuals

> reg$residuals

1 2 3

3155.8412 8127.8488 -2236.1437

4 5 6

-1167.1248 -6691.1173 3444.9091

7 8 9

6007.9128 -1587.0796 8412.9204

10 11 12

-3568.0608 570.9467 -7798.0495

13 14 15

-6635.0495 -7456.0457 -7206.0306

16 17 18

-4159.0156 -7958.0080 7210.9995

19 20 21

-183.9779 11448.0259 1686.0560

22 23 24

5386.0673 855.0975 10530.1088

25 26 27

1424.1276 -5259.8611 1402.1577

28 29 30

-3876.8385 -735.8121 -3144.8046

> confint(reg,level = 0.95)

2.5 % 97.5 %

(Intercept) 21136.061 30448.34

YearsExperience 8674.119 10225.81

> predict(reg,interval = "predict")

fit lwr upr

1 36187.16 23698.92 48675.40

2 38077.15 25628.63 50525.67

3 39967.14 27556.52 52377.76

4 44692.12 32368.22 57016.03

5 46582.12 34289.64 58874.59

6 53197.09 40999.70 65394.48

7 54142.09 41956.37 66327.80

8 56032.08 43868.25 68195.91

9 56032.08 43868.25 68195.91

10 60757.06 48639.42 72874.70

11 62647.05 50544.46 74749.65

12 63592.05 51496.24 75687.86

13 63592.05 51496.24 75687.86

14 64537.05 52447.52 76626.57

15 68317.03 56247.70 80386.36

16 72097.02 60039.93 84154.10

17 73987.01 61933.05 86040.96

18 75877.00 63824.18 87929.82

19 81546.98 69485.57 93608.39

20 82491.97 70427.39 94556.56

21 90051.94 77944.06 102159.83

22 92886.93 80754.66 105019.20

23 100446.90 88228.15 112665.65

24 103281.89 91022.76 115541.02

25 108006.87 95670.98 120342.77

26 110841.86 98454.30 123229.42

27 115566.84 103084.00 128049.68

28 116511.84 104008.59 129015.09

29 123126.81 110468.27 135785.35

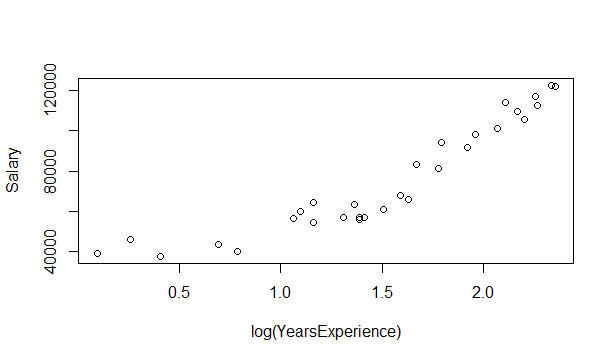
30 125016.80 112309.98 137723.63

Warning message:

In predict.lm(reg, interval = "predict") :

predictions on current data refer to \_future\_ responses

> plot(log(YearsExperience),Salary)



> cor(log(YearsExperience),Salary)

[1] 0.9240611

> reg\_log <- lm(Salary~log(YearsExperience))

> summary(reg\_log)

Call:

lm(formula = Salary ~ log(YearsExperience))

Residuals:

Min 1Q Median 3Q

-15392.6 -7523.0 559.7 6336.1

Max

20629.8

Coefficients:

Estimate

(Intercept) 14928

log(YearsExperience) 40582

Std. Error t value

(Intercept) 5156 2.895

log(YearsExperience) 3172 12.792

Pr(>|t|)

(Intercept) 0.00727 \*\*

log(YearsExperience) 3.25e-13 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05

‘.’ 0.1 ‘ ’ 1

Residual standard error: 10660 on 28 degrees of freedom

Multiple R-squared: 0.8539, Adjusted R-squared: 0.8487

F-statistic: 163.6 on 1 and 28 DF, p-value: 3.25e-13

> predict(reg\_log)

1 2 3 4

18795.85 25575.24 31382.55 43057.26

5 6 7 8

46925.14 58136.05 59511.84 62130.94

9 10 11 12

62130.94 68022.72 70159.11 71186.55

13 14 15 16

71186.55 72188.63 75966.42 79422.30

17 18 19 20

81045.79 82606.83 86959.07 87641.13

21 22 23 24

92720.50 94472.51 98805.37 100317.92

25 26 27 28

102719.92 104095.71 106289.87 106714.81

29 30

109571.01 110351.45

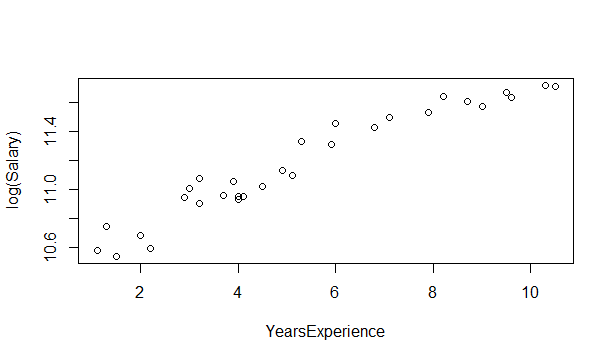
> confint(reg\_log,level = 0.95)

2.5 % 97.5 %

(Intercept) 4365.921 25490.02

log(YearsExperience) 34083.512 47080.46

> plot(YearsExperience,log(Salary))



> cor(YearsExperience,log(Salary))

[1] 0.9653844

> reg\_exp <- lm(log(Salary)~YearsExperience)

> summary(reg\_exp)

Call:

lm(formula = log(Salary) ~ YearsExperience)

Residuals:

Min 1Q Median 3Q

-0.18949 -0.06946 -0.01068 0.06932

Max

0.19029

Coefficients:

Estimate Std. Error

(Intercept) 10.507402 0.038443

YearsExperience 0.125453 0.006406

t value Pr(>|t|)

(Intercept) 273.33 <2e-16 \*\*\*

YearsExperience 19.59 <2e-16 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05

‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09789 on 28 degrees of freedom

Multiple R-squared: 0.932, Adjusted R-squared: 0.9295

F-statistic: 383.6 on 1 and 28 DF, p-value: < 2.2e-16

> predict(reg\_exp)

1 2 3 4

10.64540 10.67049 10.69558 10.75831

5 6 7 8

10.78340 10.87122 10.88376 10.90885

9 10 11 12

10.90885 10.97158 10.99667 11.00921

13 14 15 16

11.00921 11.02176 11.07194 11.12212

17 18 19 20

11.14721 11.17230 11.24757 11.26012

21 22 23 24

11.36048 11.39812 11.49848 11.53612

25 26 27 28

11.59884 11.63648 11.69920 11.71175

29 30

11.79957 11.82466

> confint(reg\_exp,level = 0.95)

2.5 % 97.5 %

(Intercept) 10.4286558 10.5861480

YearsExperience 0.1123316 0.1385742

> predict(reg\_exp,interval = "predict")

fit lwr upr

1 10.64540 10.43420 10.85661

2 10.67049 10.45996 10.88102

3 10.69558 10.48569 10.90547

4 10.75831 10.54988 10.96673

5 10.78340 10.57550 10.99129

6 10.87122 10.66493 11.07750

7 10.88376 10.67767 11.08985

8 10.90885 10.70313 11.11457

9 10.90885 10.70313 11.11457

10 10.97158 10.76664 11.17651

11 10.99667 10.79199 11.20135

12 11.00921 10.80465 11.21378

13 11.00921 10.80465 11.21378

14 11.02176 10.81730 11.22622

15 11.07194 10.86782 11.27606

16 11.12212 10.91821 11.32603

17 11.14721 10.94335 11.35107

18 11.17230 10.96846 11.37614

19 11.24757 11.04359 11.45156

20 11.26012 11.05608 11.46416

21 11.36048 11.15571 11.56525

22 11.39812 11.19293 11.60330

23 11.49848 11.29183 11.70513

24 11.53612 11.32879 11.74345

25 11.59884 11.39021 11.80747

26 11.63648 11.42698 11.84598

27 11.69920 11.48809 11.91032

28 11.71175 11.50029 11.92321

29 11.79957 11.58548 12.01365

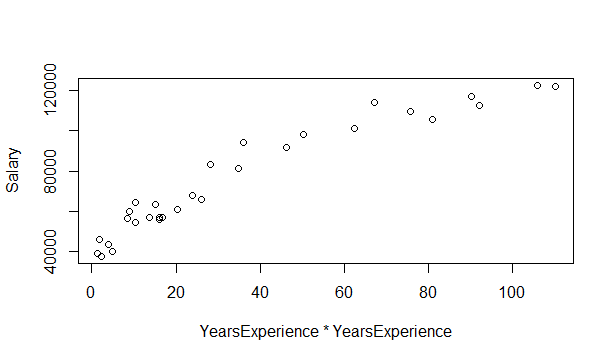
30 11.82466 11.60976 12.03956

Warning message:

In predict.lm(reg\_exp, interval = "predict") :

predictions on current data refer to \_future\_ responses

> plot(YearsExperience\*YearsExperience,Salary)



> cor(YearsExperience\*YearsExperience,Salary)

[1] 0.9567235

> reg\_poly <- lm(Salary~YearsExperience+I(YearsExperience^2))

> summary(reg\_poly)

Call:

lm(formula = Salary ~ YearsExperience + I(YearsExperience^2))

Residuals:

Min 1Q Median 3Q Max

-7835 -4026 -493 3309 11579

Coefficients:

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

(Intercept) 26214.93 4554.67 5.756 4.04e-06 \*\*\*

YearsExperience 9259.28 1811.01 5.113 2.25e-05 \*\*\*

I(YearsExperience^2) 16.39 152.12 0.108 0.915

---

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

Residual standard error: 5893 on 27 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9538

F-statistic: 300.3 on 2 and 27 DF, p-value: < 2.2e-16

> predict(reg\_poly)

1 2 3 4 5 6 7 8

36419.98 38279.71 40140.74 44799.07 46664.70 53204.72 54140.32 56012.50

9 10 11 12 13 14 15 16

56012.50 60698.70 62575.47 63514.35 63514.35 64453.56 68213.66 71979.01

17 18 19 20 21 22 23 24

73863.65 75749.60 81415.33 82360.77 89936.06 92782.20 100386.34 103243.30

25 26 27 28 29 30

108011.46 110876.29 115657.56 116614.80 123324.64 125244.69

> reg\_poly$residuals

1 2 3 4 5 6 7

2923.02004 7925.29483 -2409.74178 -1274.07072 -6773.69725 3437.28256 6009.68256

8 9 10 11 12 13 14

-1567.50100 8432.49900 -3509.69730 642.52923 -7720.34929 -6557.34929 -7372.55566

15 16 17 18 19 20 21

-7102.65964 -4041.00925 -7834.65116 7338.39553 -52.33285 11579.23161 1801.94462

22 23 24 25 26 27 28

5490.80245 915.66454 10568.70328 1419.54415 -5294.28554 1311.44127 -3979.79692

29 30

-933.64409 -3372.69394

> confint(reg\_poly,level = 0.95)

2.5 % 97.5 %

(Intercept) 16869.5127 35560.3527

YearsExperience 5543.4050 12975.1628

I(YearsExperience^2) -295.7344 328.5195

> predict(reg\_poly,interval = "predict")

fit lwr upr

1 36419.98 22934.59 49905.37

2 38279.71 25011.39 51548.02

3 40140.74 27059.48 53222.00

4 44799.07 32066.85 57531.29

5 46664.70 34030.20 59299.19

6 53204.72 40764.58 65644.86

7 54140.32 41712.88 66567.76

8 56012.50 43601.83 68423.18

9 56012.50 43601.83 68423.18

10 60698.70 48290.87 73106.53

11 62575.47 50157.82 74993.12

12 63514.35 51090.24 75938.46

13 63514.35 51090.24 75938.46

14 64453.56 52022.20 76884.91

15 68213.66 55748.59 80678.73

16 71979.01 59479.19 84478.83

17 73863.65 61348.25 86379.05

18 75749.60 63220.68 88278.53

19 81415.33 68861.91 93968.76

20 82360.77 69805.84 94915.70

21 89936.06 77392.43 102479.68

22 92782.20 80249.63 105314.77

23 100386.34 87872.00 112900.67

24 103243.30 90719.45 115767.14

25 108011.46 95430.60 120592.31

26 110876.29 98226.05 123526.52

27 115657.56 102810.50 128504.62

28 116614.80 103713.75 129515.84

29 123324.64 109876.71 136772.58

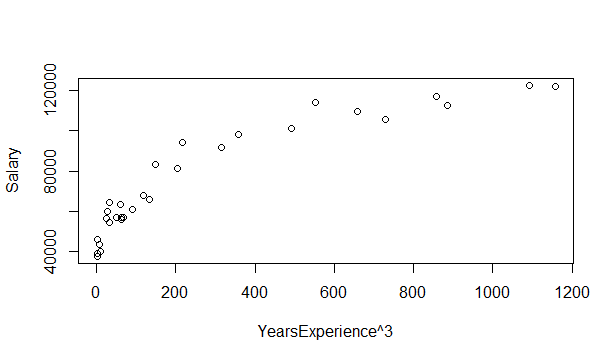
30 125244.69 111578.68 138910.71

Warning message:

In predict.lm(reg\_poly, interval = "predict") :

predictions on current data refer to \_future\_ responses

> plot(YearsExperience^3,Salary)



> cor(YearsExperience^3,Salary)

[1] 0.9133658

> reg\_poly3 <- lm(Salary~YearsExperience+I(YearsExperience^2)+I(YearsExperience^3))

> summary(reg\_poly3)

Call:

lm(formula = Salary ~ YearsExperience + I(YearsExperience^2) +

I(YearsExperience^3))

Residuals:

Min 1Q Median 3Q Max

-7468 -4286 -1100 2639 10412

Coefficients:

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

(Intercept) 38863.07 7214.75 5.387 1.21e-05 \*\*\*

YearsExperience -718.71 4892.11 -0.147 0.8843

I(YearsExperience^2) 2099.35 968.36 2.168 0.0395 \*

I(YearsExperience^3) -122.92 56.52 -2.175 0.0389 \*

---

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

Residual standard error: 5524 on 26 degrees of freedom

Multiple R-squared: 0.9636, Adjusted R-squared: 0.9594

F-statistic: 229.4 on 3 and 26 DF, p-value: < 2.2e-16

> predict(reg\_poly3)

1 2 3 4 5 6 7 8

40449.11 41206.61 42093.71 44839.74 46133.97 51436.58 52282.40 54032.88

9 10 11 12 13 14 15 16

54032.88 58717.94 60700.03 61711.28 61711.28 62735.02 66940.09 71285.97

17 18 19 20 21 22 23 24

73496.95 75725.44 82456.89 83577.76 92401.35 95595.79 103603.74 106358.46

25 26 27 28 29 30

110570.09 112836.87 116117.25 116692.05 119867.63 120480.23

> reg\_poly3$residuals

1 2 3 4 5 6 7

-1106.1080 4998.3895 -4362.7116 -1314.7395 -6242.9734 5205.4167 7867.6021

8 9 10 11 12 13 14

412.1234 10412.1234 -1528.9444 2517.9673 -5917.2828 -4754.2828 -5654.0203

15 16 17 18 19 20 21

-5829.0938 -3347.9653 -7467.9504 7362.5647 -1093.8886 10362.2381 -663.3491

22 23 24 25 26 27 28

2677.2051 -2301.7374 7453.5384 -1139.0884 -7254.8667 851.7483 -4057.0544

29 30

2523.3688 1391.7709

> confint(reg\_poly3,level = 0.95)

2.5 % 97.5 %

(Intercept) 24032.9470 53693.19673

YearsExperience -10774.5874 9337.17053

I(YearsExperience^2) 108.8691 4089.83482

I(YearsExperience^3) -239.0934 -6.73744

> predict(reg\_poly3,interval = "predict")

fit lwr upr

1 40449.11 27225.47 53672.74

2 41206.61 28443.63 53969.59

3 42093.71 29671.89 54515.53

4 44839.74 32883.55 56795.93

5 46133.97 34259.01 58008.93

6 51436.58 39635.79 63237.38

7 52282.40 40481.08 64083.71

8 54032.88 42229.44 65836.31

9 54032.88 42229.44 65836.31

10 58717.94 46916.97 70518.91

11 60700.03 48905.33 72494.74

12 61711.28 49920.66 73501.90

13 61711.28 49920.66 73501.90

14 62735.02 50948.94 74521.10

15 66940.09 55173.09 78707.10

16 71285.97 59529.80 83042.13

17 73496.95 61739.31 85254.59

18 75725.44 63960.18 87490.69

19 82456.89 70627.62 94286.16

20 83577.76 71732.13 95423.39

21 92401.35 80394.05 104408.65

22 95595.79 83530.42 107661.17

23 103603.74 91465.11 115742.37

24 106358.46 94235.02 118481.90

25 110570.09 98511.11 122629.07

26 112836.87 100814.05 124859.68

27 116117.25 104045.46 128189.04

28 116692.05 104577.17 128806.94

29 119867.63 106823.53 132911.74

30 120480.23 106880.01 134080.44

Warning message:

In predict.lm(reg\_poly3, interval = "predict") :

predictions on current data refer to \_future\_ responses

> plot(reg\_poly3)

