**Simple Linear Regression Assignment**

### 4 - Salary\_hike -> Build a prediction model for Salary\_hike

> # Load Salary\_hike.csv dataset

> library(readr)

> Salary\_hike <- read.csv("E:/Data Science Asignments/Simple regression/Salary\_data.csv")

>

> View(Salary\_hike)

>

> # Exploratory data analysis

> summary(Salary\_hike)

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

> # x= YearsExperience y= Salary

>

> var(Salary\_hike$Salary)

[1] 751550960

> sd(Salary\_hike$Salary)

[1] 27414.43

> var(Salary\_hike$YearsExperience)

[1] 8.053609

> sd(Salary\_hike$YearsExperience)

[1] 2.837888

>

> #Scatter plot

> plot(Salary\_hike$YearsExperience, Salary\_hike$Salary) # plot(X,Y)

>

> ?plot



attach(Salary\_hike)

> attach(Salary\_hike)

The following objects are masked from Salary\_hike (pos = 4):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 5):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 6):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 7):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 8):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 9):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 10):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 11):

Salary, YearsExperience

The following objects are masked from Salary\_hike (pos = 12):

Salary, YearsExperience

> #Correlation Coefficient (r)

> cor(YearsExperience,Salary) # cor(X,Y)

[1] 0.9782416

>

> # Simple Linear Regression model

> reg <- lm(Salary ~YearsExperience) # lm(Y ~ X)

>

> summary(reg)

Call:

lm(formula = Salary ~ YearsExperience)

Residuals:

Min 1Q Median 3Q Max

-7958.0 -4088.5 -459.9 3372.6 11448.0

Coefficients:

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

(Intercept) 25792.2 2273.1 11.35 5.51e-12 \*\*\*

YearsExperience 9450.0 378.8 24.95 < 2e-16 \*\*\*

---

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

Residual standard error: 5788 on 28 degrees of freedom

Multiple R-squared: 0.957, Adjusted R-squared: 0.9554

F-statistic: 622.5 on 1 and 28 DF, p-value: < 2.2e-16

>

> pred <- predict(reg)

>

> reg$YearsExperience

NULL

> sum(reg$YearsExperience)

[1] 0

>

> mean(reg$YearsExperience)

[1] NA

Warning message:

In mean.default(reg$YearsExperience) :

argument is not numeric or logical: returning NA

> sqrt(sum(reg$YearsExperience^2)/nrow(Salary)) #RMSE

numeric(0)

>

> sqrt(mean(reg$YearsExperience^2))

[1] NaN

>

> 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

>

> # ggplot for adding regresion line for data

> library(ggplot2)

>

> ?ggplot2

>

> ggplot(data = Salary\_hike, aes(x = YearsExperience, y = Salary)) +

+ geom\_point(color='blue') +

+ geom\_line(color='red',data = Salary\_hike, aes(x=YearsExperience, y=pred))

>

> ?ggplot2

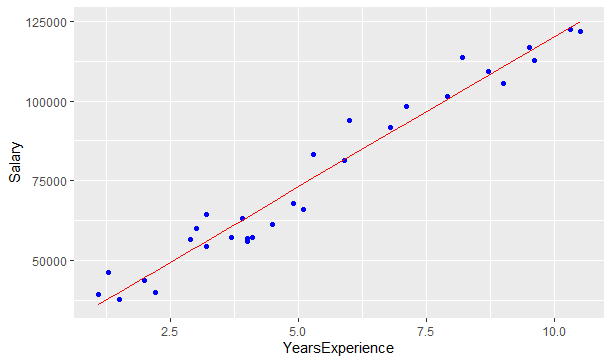
>

> ########################

> # A simple ggplot code for directly showing the line

>

> # ggplot(Salary\_hike,aes(YearsExperience,Salary))+stat\_summary(fun.data=mean\_cl\_normal) + geom\_smooth(method='lm')



# Logarithmic Model

>

> # x = log(YearsExperience); y = Salary

>

> plot(log(YearsExperience), Salary)

> cor(log(YearsExperience), Salary)

[1] 0.9240611

>

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

>

> summary(reg\_log)

Call:

lm(formula = Salary ~ log(YearsExperience))

Residuals:

Min 1Q Median 3Q Max

-15392.6 -7523.0 559.7 6336.1 20629.8

Coefficients:

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

(Intercept) 14928 5156 2.895 0.00727 \*\*

log(YearsExperience) 40582 3172 12.792 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 5 6 7 8

18795.85 25575.24 31382.55 43057.26 46925.14 58136.05 59511.84 62130.94

9 10 11 12 13 14 15 16

62130.94 68022.72 70159.11 71186.55 71186.55 72188.63 75966.42 79422.30

17 18 19 20 21 22 23 24

81045.79 82606.83 86959.07 87641.13 92720.50 94472.51 98805.37 100317.92

25 26 27 28 29 30

102719.92 104095.71 106289.87 106714.81 109571.01 110351.45

>

> reg\_log$YearsExperience

NULL

> sqrt(sum(reg\_log$YearsExperience^2)/nrow(Salary\_hike)) #RMSE

[1] 0

>

> confint(reg\_log,level=0.95)

2.5 % 97.5 %

(Intercept) 4365.921 25490.02

log(YearsExperience) 34083.512 47080.46

> predict(reg\_log,interval="confidence")

fit lwr upr

1 18795.85 8804.575 28787.12

2 25575.24 16568.881 34581.59

3 31382.55 23199.611 39565.49

4 43057.26 36443.652 49670.87

5 46925.14 40794.211 53056.07

6 58136.05 53227.598 63044.50

7 59511.84 54728.455 64295.23

8 62130.94 57565.682 66696.21

9 62130.94 57565.682 66696.21

10 68022.72 63834.617 72210.82

11 70159.11 66062.414 74255.80

12 71186.55 67124.284 75248.82

13 71186.55 67124.284 75248.82

14 72188.63 68153.749 76223.51

15 75966.42 71978.039 79954.81

16 79422.30 75396.508 83448.08

17 81045.79 76976.486 85115.10

18 82606.83 78480.640 86733.02

19 86959.07 82601.871 91316.26

20 87641.13 83238.825 92043.44

21 92720.50 87917.010 97523.99

22 94472.51 89507.201 99437.83

23 98805.37 93397.986 104212.76

24 100317.92 94744.116 105891.72

25 102719.92 96870.949 108568.89

26 104095.71 98083.720 110107.71

27 106289.87 100010.642 112569.09

28 106714.81 100382.883 113046.75

29 109571.01 102877.640 116264.37

30 110351.45 103557.314 117145.59

>

> ######################

>

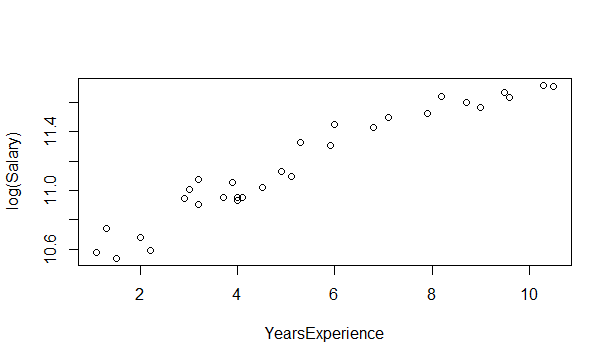
> # Exponential Model

>

> # x = YearsExperience and y = log(Salary)

>

> plot(YearsExperience, log(Salary))

> 

cor(YearsExperience, log(Salary))

[1] 0.9653844

>

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

>

> summary(reg\_exp)

Call:

lm(formula = log(Salary) ~ YearsExperience)

Residuals:

Min 1Q Median 3Q Max

-0.18949 -0.06946 -0.01068 0.06932 0.19029

Coefficients:

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

(Intercept) 10.507402 0.038443 273.33 <2e-16 \*\*\*

YearsExperience 0.125453 0.006406 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

>

> reg\_exp$residuals

1 2 3 4 5 6

-0.065326727 0.070352644 -0.157343914 -0.077216910 -0.189492239 0.073290763

7 8 9 10 11 12

0.120836157 -0.003904845 0.164716284 -0.017460736 0.057676185 -0.079791838

13 14 15 16 17 18

-0.059161577 -0.069532152 -0.051492737 0.004229750 -0.049362313 0.155353356

19 20 21 22 23 24

0.059101960 0.190292327 0.066210424 0.097387189 0.027381712 0.106187655

25 26 27 28 29 30

0.004207459 -0.069234719 -0.029540123 -0.079841852 -0.084590539 -0.113930641

>

> sqrt(mean(reg\_exp$residuals^2))

[1] 0.09457437

>

> logat <- predict(reg\_exp)

> at <- exp(logat)

>

> error =Salary\_hike $Salary - at

> error

1 2 3 4 5 6 7

-2655.9575 3138.9332 -6429.2892 -3494.0292 -8322.6885 4002.8575 6846.3246

8 9 10 11 12 13 14

-213.0149 9786.9851 -1007.3308 3543.0171 -4634.3385 -3471.3385 -4110.2046

15 16 17 18 19 20 21

-3229.1993 286.7539 -3341.1287 11955.3154 4669.3692 16278.1650 5877.2958

22 23 24 25 26 27 28

9119.2746 2736.1904 11465.8895 459.4592 -7568.9334 -3506.8196 -9361.7438

29 30

-10803.6213 -14706.8297

>

> sqrt(sum(error^2)/nrow(Salary\_hike)) #RMSE

[1] 7213.235

>

> confint(reg\_exp,level=0.95)

2.5 % 97.5 %

(Intercept) 10.4286558 10.5861480

YearsExperience 0.1123316 0.1385742

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

fit lwr upr

1 10.64540 10.57909 10.71171

2 10.67049 10.60635 10.73463

3 10.69558 10.63358 10.75758

4 10.75831 10.70147 10.81514

5 10.78340 10.72854 10.83825

6 10.87122 10.82281 10.91962

7 10.88376 10.83620 10.93132

8 10.90885 10.86292 10.95478

9 10.90885 10.86292 10.95478

10 10.97158 10.92929 11.01387

11 10.99667 10.95563 11.03771

12 11.00921 10.96875 11.04968

13 11.00921 10.96875 11.04968

14 11.02176 10.98184 11.06168

15 11.07194 11.03381 11.11007

16 11.12212 11.08511 11.15913

17 11.14721 11.11049 11.18393

18 11.17230 11.13569 11.20891

19 11.24757 11.21016 11.28499

20 11.26012 11.22242 11.29782

21 11.36048 11.31900 11.40197

22 11.39812 11.35464 11.44159

23 11.49848 11.44856 11.54840

24 11.53612 11.48344 11.58879

25 11.59884 11.54127 11.65642

26 11.63648 11.57581 11.69714

27 11.69920 11.63319 11.76522

28 11.71175 11.64464 11.77886

29 11.79957 11.72459 11.87454

30 11.82466 11.74738 11.90194

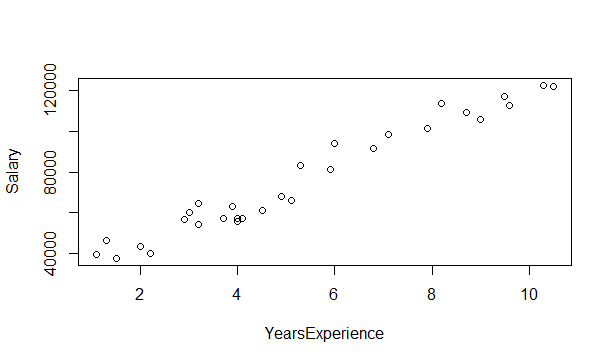
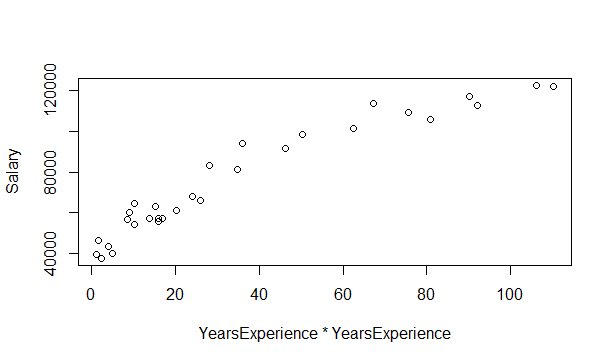
>

> ##############################

> # Polynomial model with 2 degree (quadratic model)

>

> plot(YearsExperience,Salary)

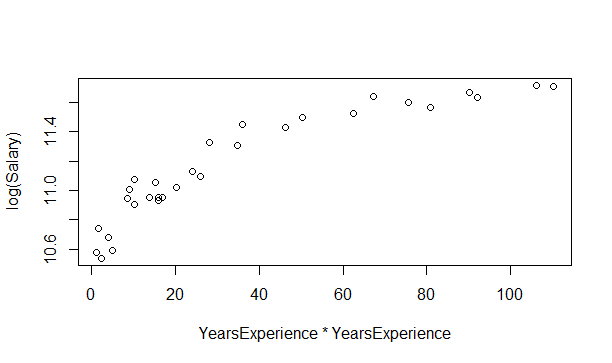
 plot(YearsExperience\*YearsExperience, Salary) 

cor(YearsExperience\*YearsExperience, Salary)

[1] 0.9567235

>

> plot(YearsExperience\*YearsExperience, log(Salary))

 cor(YearsExperience, log(Salary))

[1] 0.9653844

> cor(YearsExperience\*YearsExperience, log(Salary))

[1] 0.9157747

>

> # lm(Y ~ X + I(X\*X) +...+ I(X\*X\*X...))

>

> reg2degree <- lm(log(Salary) ~ YearsExperience + I(YearsExperience\*YearsExperience))

>

> summary(reg2degree)

Call:

lm(formula = log(Salary) ~ YearsExperience + I(YearsExperience \*

YearsExperience))

Residuals:

Min 1Q Median 3Q Max

-0.156176 -0.052355 -0.000915 0.048548 0.156817

Coefficients:

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

(Intercept) 10.336852 0.066962 154.368 < 2e-16 \*\*\*

YearsExperience 0.202382 0.026625 7.601 3.55e-08 \*\*\*

I(YearsExperience \* YearsExperience) -0.006614 0.002236 -2.957 0.00638 \*\*

---

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

Residual standard error: 0.08664 on 27 degrees of freedom

Multiple R-squared: 0.9486, Adjusted R-squared: 0.9448

F-statistic: 249.2 on 2 and 27 DF, p-value: < 2.2e-16

>

> logpol <- predict(reg2degree)

> expy <- exp(logpol)

>

> err = Salary\_hike$Salary - expy

>

> sqrt(sum(err^2)/nrow(Salary\_hike)) #RMSE

[1] 5391.082

>

> confint(reg2degree,level=0.95)

2.5 % 97.5 %

(Intercept) 10.19945640 10.474246649

YearsExperience 0.14775127 0.257011990

I(YearsExperience \* YearsExperience) -0.01120239 -0.002024692

> predict(reg2degree,interval="confidence")

fit lwr upr

1 10.55147 10.46370 10.63924

2 10.58877 10.50847 10.66907

3 10.62554 10.55218 10.69891

4 10.71516 10.65655 10.77377

5 10.75008 10.69623 10.80393

6 10.86814 10.82517 10.91110

7 10.88447 10.84231 10.92664

8 10.91675 10.87567 10.95783

9 10.91675 10.87567 10.95783

10 10.99512 10.95423 11.03602

11 11.02555 10.98401 11.06708

12 11.04056 10.99861 11.08251

13 11.04056 10.99861 11.08251

14 11.05544 11.01303 11.09786

15 11.11364 11.06914 11.15815

16 11.16973 11.12317 11.21629

17 11.19698 11.14952 11.24444

18 11.22370 11.17548 11.27192

19 11.30069 11.25111 11.35027

20 11.31305 11.26339 11.36272

21 11.40724 11.35820 11.45628

22 11.44037 11.39195 11.48880

23 11.52292 11.47552 11.57031

24 11.55169 11.50375 11.59962

25 11.59699 11.54593 11.64805

26 11.62259 11.56795 11.67723

27 11.66261 11.59881 11.72640

28 11.67021 11.60410 11.73632

29 11.71975 11.63323 11.80627

30 11.73272 11.63910 11.82633

>

> # visualization

> ggplot(data = Salary\_hike, aes(x = YearsExperience + I(YearsExperience^2), y = log(Salary))) +

+ geom\_point(color='blue') +

+ geom\_line(color='red',data = Salary\_hike, aes(x=YearsExperience+I(YearsExperience^2), y=logpol))

