**Simple Linear Regression Assignment**

### 3 - Emp\_data -> Build a prediction model for Churn\_out\_rate

> library(readr)

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

>

> View(Emp\_data)

>

> # Exploratory data analysis

> summary(Emp\_data)

Salary\_hike Churn\_out\_rate

Min. :1580 Min. :60.00

1st Qu.:1618 1st Qu.:65.75

Median :1675 Median :71.00

Mean :1689 Mean :72.90

3rd Qu.:1724 3rd Qu.:78.75

Max. :1870 Max. :92.00

# x= Salary\_hike y= Churn\_out\_rate

>

> var(Emp\_data$Churn\_out\_rate)

[1] 105.2111

> sd(Emp\_data$Churn\_out\_rate)

[1] 10.25725

> var(Emp\_data$Salary\_hike)

[1] 8481.822

> sd(Emp\_data$Salary\_hike)

[1] 92.09681

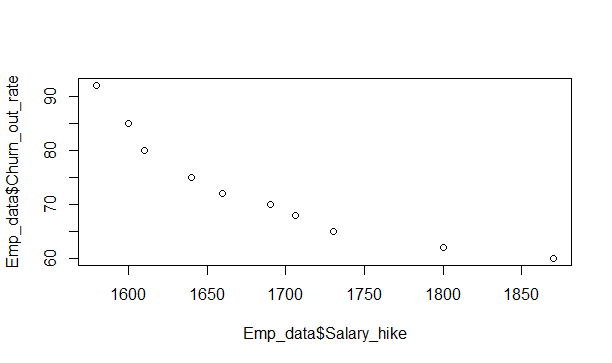
>

> #Scatter plot

> plot(Emp\_data$Salary\_hike, Emp\_data$Churn\_out\_rate) # plot(X,Y)

>

> ?plot



attach(Emp\_data)

> #Correlation Coefficient (r)

> cor(Salary\_hike,Churn\_out\_rate) # cor(X,Y)

[1] -0.9117216

>

> # Simple Linear Regression model

> reg <- lm(Churn\_out\_rate ~Salary\_hike) # lm(Y ~ X)

>

> summary(reg)

Call:

lm(formula = Churn\_out\_rate ~ Salary\_hike)

Residuals:

Min 1Q Median 3Q Max

-3.804 -3.059 -1.819 2.430 8.072

Coefficients:

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

(Intercept) 244.36491 27.35194 8.934 1.96e-05 \*\*\*

Salary\_hike -0.10154 0.01618 -6.277 0.000239 \*\*\*

---

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

Residual standard error: 4.469 on 8 degrees of freedom

Multiple R-squared: 0.8312, Adjusted R-squared: 0.8101

F-statistic: 39.4 on 1 and 8 DF, p-value: 0.0002386

>

> pred <- predict(reg)

>

> reg$Salary\_hike

NULL

> sum(reg$Salary\_hike)

[1] 0

>

> mean(reg$Salary\_hike)

[1] NA

Warning message:

In mean.default(reg$Salary\_hike) :

argument is not numeric or logical: returning NA

> sqrt(sum(reg$Salary\_hike^2)/nrow(Churn\_out\_rate)) #RMSE

numeric(0)

>

> sqrt(mean(reg$Salary\_hike^2))

[1] NaN

>

> confint(reg,level=0.95)

2.5 % 97.5 %

(Intercept) 181.2912317 307.4385905

Salary\_hike -0.1388454 -0.0642399

> predict(reg,interval="predict")

fit lwr upr

1 83.92753 72.38391 95.47115

2 81.89668 70.59327 93.20009

3 80.88125 69.68123 92.08127

4 77.83497 66.87456 88.79538

5 75.80412 64.94216 86.66607

6 72.75784 61.94828 83.56740

7 71.13316 60.30425 81.96206

8 68.69613 57.77694 79.61533

9 61.58815 50.00746 73.16884

10 54.48016 41.72742 67.23290

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 = Emp\_data, aes(x = Salary\_hike, y = Churn\_out\_rate)) +

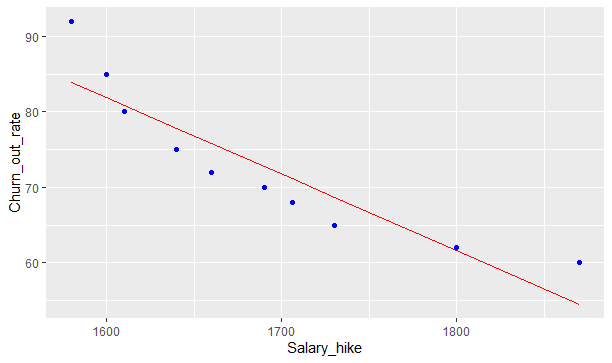
+ geom\_point(color='blue') +

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

>

> ?ggplot2

>



########################

> # A simple ggplot code for directly showing the line

>

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

>

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

>

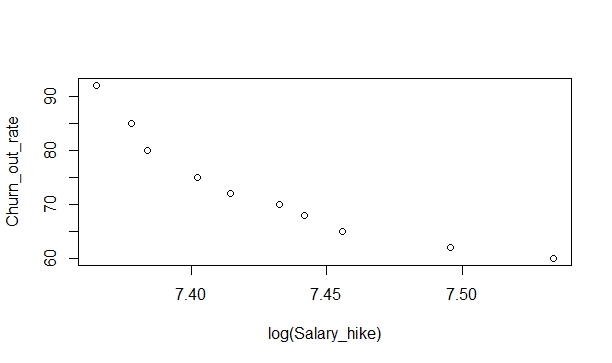
> # Logarithmic Model

>

> # x = log(Salary\_hike); y = Churn\_out\_rate

>

> plot(log(Salary\_hike), Churn\_out\_rate)

>

> cor(log(Salary\_hike), Churn\_out\_rate)

[1] -0.9212077

>

> reg\_log <- lm(Churn\_out\_rate ~ log(Salary\_hike)) # lm(Y ~ log(X))

>

> summary(reg\_log)

Call:

lm(formula = Churn\_out\_rate ~ log(Salary\_hike))

Residuals:

Min 1Q Median 3Q Max

-3.678 -2.851 -1.794 2.275 7.624

Coefficients:

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

(Intercept) 1381.5 195.4 7.070 0.000105 \*\*\*

log(Salary\_hike) -176.1 26.3 -6.697 0.000153 \*\*\*

---

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

Residual standard error: 4.233 on 8 degrees of freedom

Multiple R-squared: 0.8486, Adjusted R-squared: 0.8297

F-statistic: 44.85 on 1 and 8 DF, p-value: 0.0001532

> predict(reg\_log)

1 2 3 4 5 6 7 8 9

84.37627 82.16102 81.06376 77.81241 75.67773 72.52344 70.86397 68.40372 61.41829

10

54.69939

>

> reg\_log$Salary\_hike

NULL

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

[1] 0

>

> confint(reg\_log,level=0.95)

2.5 % 97.5 %

(Intercept) 930.8584 1832.0540

log(Salary\_hike) -236.7512 -115.4682

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

fit lwr upr

1 84.37627 79.36190 89.39065

2 82.16102 77.72288 86.59916

3 81.06376 76.88883 85.23870

4 77.81241 74.29260 81.33223

5 75.67773 72.44622 78.90923

6 72.52344 69.43400 75.61287

7 70.86397 67.69864 74.02930

8 68.40372 64.95049 71.85696

9 61.41829 56.40243 66.43414

10 54.69939 47.71330 61.68547

>

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

>

> # Exponential Model

>

> # x = Salary\_hike and y = log(Churn\_out\_rate)

>

> plot(Salary\_hike, log(Churn\_out\_rate))

>

> cor(Salary\_hike, log(Churn\_out\_rate))

[1] -0.9346361

>

> reg\_exp <- lm(log(Churn\_out\_rate) ~ Salary\_hike) #lm(log(Y) ~ X)

>

> summary(reg\_exp)

Call:

lm(formula = log(Churn\_out\_rate) ~ Salary\_hike)

Residuals:

Min 1Q Median 3Q Max

-0.04825 -0.03519 -0.01909 0.02942 0.08970

Coefficients:

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

(Intercept) 6.6383000 0.3175983 20.902 2.88e-08 \*\*\*

Salary\_hike -0.0013963 0.0001878 -7.434 7.38e-05 \*\*\*

---

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

Residual standard error: 0.0519 on 8 degrees of freedom

Multiple R-squared: 0.8735, Adjusted R-squared: 0.8577

F-statistic: 55.26 on 1 and 8 DF, p-value: 7.377e-05

>

> reg\_exp$residuals

1 2 3 4 5 6

0.089697629 0.038487005 -0.008174268 -0.030822745 -0.043718043 -0.029998876

7 8 9 10

-0.036645056 -0.048253455 0.002237097 0.067190712

>

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

[1] 0.04641748

>

> logat <- predict(reg\_exp)

> at <- exp(logat)

>

> error =Emp\_data $Churn\_out\_rate - at

> error

1 2 3 4 5 6 7

7.8929033 3.2092422 -0.6566215 -2.3477013 -3.2175185 -2.1317363 -2.5380839

8 9 10

-3.2133794 0.1385450 3.8989881

>

> sqrt(sum(error^2)/nrow(Emp\_data)) #RMSE

[1] 3.541549

>

> confint(reg\_exp,level=0.95)

2.5 % 97.5 %

(Intercept) 5.905917079 7.3706828388

Salary\_hike -0.001829477 -0.0009631923

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

fit lwr upr

1 4.432091 4.371718 4.492464

2 4.404164 4.350267 4.458062

3 4.390201 4.339297 4.441105

4 4.348311 4.305006 4.391616

5 4.320384 4.280564 4.360204

6 4.278494 4.240645 4.316343

7 4.256153 4.217566 4.294740

8 4.222641 4.180763 4.264518

9 4.124897 4.063575 4.186220

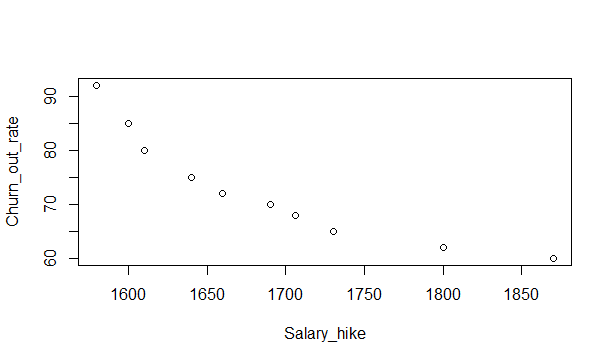
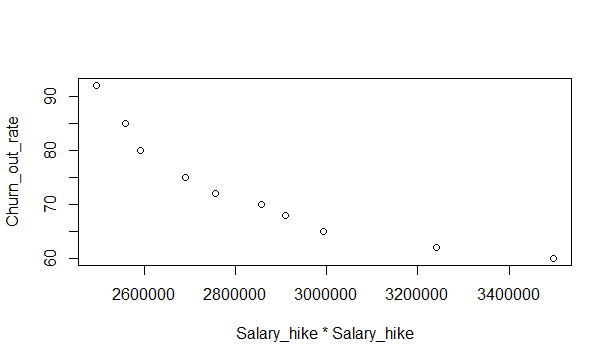
10 4.027154 3.939943 4.114365

>

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

> # Polynomial model with 2 degree (quadratic model)

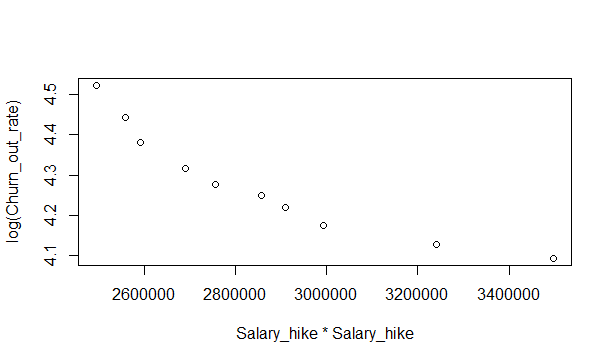
> plot(Salary\_hike,Churn\_out\_rate)

 plot(Salary\_hike\*Salary\_hike, Churn\_out\_rate) > cor(Salary\_hike\*Salary\_hike, Churn\_out\_rate)

[1] -0.9017223

>

> plot(Salary\_hike\*Salary\_hike, log(Churn\_out\_rate))



>> cor(Salary\_hike, log(Churn\_out\_rate))

[1] -0.9346361

> cor(Salary\_hike\*Salary\_hike, log(Churn\_out\_rate))

[1] -0.925803

>

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

>

> reg2degree <- lm(log(Churn\_out\_rate) ~ Salary\_hike + I(Salary\_hike\*Salary\_hike))

>

> summary(reg2degree)

Call:

lm(formula = log(Churn\_out\_rate) ~ Salary\_hike + I(Salary\_hike \*

Salary\_hike))

Residuals:

Min 1Q Median 3Q Max

-0.027877 -0.014280 0.002735 0.012608 0.027882

Coefficients:

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

(Intercept) 2.318e+01 2.415e+00 9.597 2.8e-05 \*\*\*

Salary\_hike -2.068e-02 2.813e-03 -7.351 0.000156 \*\*\*

I(Salary\_hike \* Salary\_hike) 5.605e-06 8.175e-07 6.857 0.000241 \*\*\*

---

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

Residual standard error: 0.01997 on 7 degrees of freedom

Multiple R-squared: 0.9836, Adjusted R-squared: 0.9789

F-statistic: 210.1 on 2 and 7 DF, p-value: 5.634e-07

>

> logpol <- predict(reg2degree)

> expy <- exp(logpol)

>

> err = Emp\_data$Churn\_out\_rate - expy

>

> sqrt(sum(err^2)/nrow(Emp\_data)) #RMSE

[1] 1.32679

>

> confint(reg2degree,level=0.95)

2.5 % 97.5 %

(Intercept) 1.746563e+01 2.888684e+01

Salary\_hike -2.733275e-02 -1.402780e-02

I(Salary\_hike \* Salary\_hike) 3.672104e-06 7.538047e-06

> predict(reg2degree,interval="confidence")

fit lwr upr

1 4.493907 4.461937 4.525876

2 4.436784 4.412723 4.460845

3 4.409904 4.388698 4.431110

4 4.335990 4.318381 4.353600

5 4.292320 4.273865 4.310775

6 4.235222 4.214108 4.256335

7 4.208895 4.186590 4.231199

8 4.174785 4.151429 4.198140

9 4.112180 4.087586 4.136774

10 4.104504 4.060961 4.148048

>

> # visualization

> ggplot(data = Emp\_data, aes(x = Salary\_hike + I(Salary\_hike^2), y = log(Churn\_out\_rate))) +

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

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

