library(readr)

> emp\_data <- read.csv(file.choose())

> View(emp\_data)

> emp\_data

Salary\_hike Churn\_out\_rate

1 1580 92

2 1600 85

3 1610 80

4 1640 75

5 1660 72

6 1690 70

7 1706 68

8 1730 65

9 1800 62

10 1870 60

> # 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

> # Scatter plot #

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

> attach(emp\_data)

> # Correlation Coefficient (r) #

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

[1] -0.9117216

> # Simple Linear Regression Model #

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

> summary(reg)

Call:

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

Residuals:

Min 1Q Median 3Q Max

-35.97 -23.13 -21.41 19.24 75.80

Coefficients:

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

(Intercept) 2285.365 95.912 23.828 1.02e-08 \*\*\*

Churn\_out\_rate -8.186 1.304 -6.277 0.000239 \*\*\*

---

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

Residual standard error: 40.13 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)

> pred

1 2 3 4 5 6 7 8 9 10

1532.246 1589.548 1630.479 1671.409 1695.967 1712.340 1728.712 1753.270 1777.828 1794.200

> reg$residuals

1 2 3 4 5 6 7 8 9

47.75415 10.45158 -20.47883 -31.40923 -35.96747 -22.33963 -22.71180 -23.27004 22.17172

10

75.79956

> sum(reg$residuals)

[1] 3.552714e-15

> mean(reg$residuals)

[1] 3.559653e-16

> sqrt(sum(reg$residuals^2)/nrow(emp\_data)) #RMSE

[1] 35.89264

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

[1] 35.89264

> confint(reg,level = 0.95)

2.5 % 97.5 %

(Intercept) 2064.19292 2506.537671

Churn\_out\_rate -11.19332 -5.178839

> predict(reg,interval="predict")

fit lwr upr

1 1532.246 1419.468 1645.023

2 1589.548 1485.897 1693.200

3 1630.479 1531.103 1729.854

4 1671.409 1574.149 1768.669

5 1695.967 1598.875 1793.060

6 1712.340 1614.894 1809.785

7 1728.712 1630.545 1826.879

8 1753.270 1653.350 1853.190

9 1777.828 1675.388 1880.269

10 1794.200 1689.680 1898.721

Warning message:

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

predictions on current data refer to \_future\_ responses

> predict

standardGeneric for "predict" defined from package "stats"

function (object, ...)

standardGeneric("predict")

<environment: 0x000001923d025518>

Methods may be defined for arguments: object

Use showMethods("predict") for currently available ones.

> 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))

> ## Logrithamic Model/ Transformation ##

> # 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 ~ 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 10

84.37627 82.16102 81.06376 77.81241 75.67773 72.52344 70.86397 68.40372 61.41829 54.69939

> reg\_log$residuals

1 2 3 4 5 6 7 8

7.6237297 2.8389757 -1.0637639 -2.8124149 -3.6777253 -2.5234367 -2.8639688 -3.4037229

9 10

0.5817149 5.3006123

> sqrt(sum(reg\_log$residuals^2)/nrow(emp\_data)) ###RMSE

[1] 3.786004

> 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 Transformation ##

> ## 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 7

0.089697629 0.038487005 -0.008174268 -0.030822745 -0.043718043 -0.029998876 -0.036645056

8 9 10

-0.048253455 0.002237097 0.067190712

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

[1] 0.04641748

> logat <- predict(reg\_exp)

> logat

1 2 3 4 5 6 7 8 9 10

4.432091 4.404164 4.390201 4.348311 4.320384 4.278494 4.256153 4.222641 4.124897 4.027154

> at <- exp(logat)

> error = emp\_data$Churn\_out\_rate -Churn\_out\_rate

> error

[1] 0 0 0 0 0 0 0 0 0 0

> sqrt(sum(error^2)/nrow(emp\_data))

[1] 0

> 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

> 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)

> expy

1 2 3 4 5 6 7 8 9 10

89.47028 84.50273 82.26156 76.40060 73.13594 69.07699 67.28213 65.02585 61.07971 60.61269

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

> err

1 2 3 4 5 6 7

2.52971779 0.49727467 -2.26156093 -1.40059507 -1.13594176 0.92300864 0.71787175

8 9 10

-0.02585393 0.92029172 -0.61268579

> 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

> 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))