prediction model for Salary\_hike

library(readr)  
data<-read.csv("E:\\assignments\\3.simple linear regression\\Salary\_Data.csv")  
View(data)  
attach(data)  
YE<-YearsExperience  
SH<-Salary  
# Exploratory data analysis  
# structure of data  
str(data)

## 'data.frame': 30 obs. of 2 variables:  
## $ YearsExperience: num 1.1 1.3 1.5 2 2.2 2.9 3 3.2 3.2 3.7 ...  
## $ Salary : num 39343 46205 37731 43525 39891 ...

head(data)

## YearsExperience Salary  
## 1 1.1 39343  
## 2 1.3 46205  
## 3 1.5 37731  
## 4 2.0 43525  
## 5 2.2 39891  
## 6 2.9 56642

# Descriptive statistics  
library(psych)  
describe(data)

## vars n mean sd median trimmed mad min  
## YearsExperience 1 30 5.31 2.84 4.7 5.21 2.89 1.1  
## Salary 2 30 76003.00 27414.43 65237.0 75078.88 30203.53 37731.0  
## max range skew kurtosis se  
## YearsExperience 10.5 9.4 0.34 -1.17 0.52  
## Salary 122391.0 84660.0 0.32 -1.40 5005.17

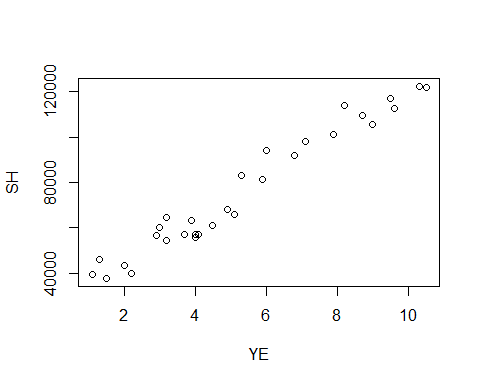
summary(data)

## 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

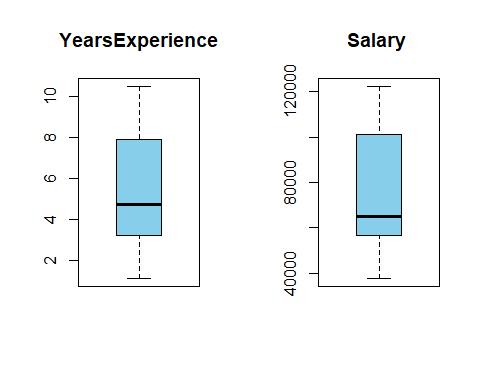
cor(YE,SH)

## [1] 0.9782416

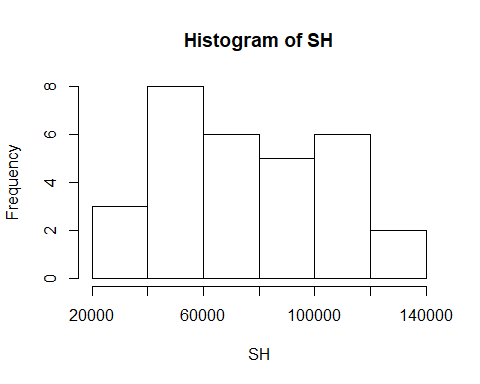
#heighly positively correlated  
plot(YE,SH)



par(mfrow=c(1,2))  
boxplot(YE, main=" YearsExperience ", col="skyblue")  
boxplot(SH, main="Salary ", col="skyblue")



hist(SH)



#Regression Analysis  
reg<-lm(SH ~ YE)  
summary(reg)

##   
## Call:  
## lm(formula = SH ~ YE)  
##   
## 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 \*\*\*  
## YE 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

confint(reg, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 21136.061 30448.34  
## YE 8674.119 10225.81

predict(reg, interval="predict")

## Warning in predict.lm(reg, interval = "predict"): predictions on current data refer to \_future\_ responses

## 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

#p values<0.05 and multiple R-squared value is 0.957   
#logarthmic transformation  
reg\_log<-lm(SH ~ log(YE))  
summary(reg\_log)

##   
## Call:  
## lm(formula = SH ~ log(YE))  
##   
## 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(YE) 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

confint(reg\_log, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 4365.921 25490.02  
## log(YE) 34083.512 47080.46

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

## Warning in predict.lm(reg\_log, interval = "predict"): predictions on current data refer to \_future\_ responses

## fit lwr upr  
## 1 18795.85 -5225.823 42817.52  
## 2 25575.24 1946.237 49204.23  
## 3 31382.55 8054.979 54710.13  
## 4 43057.26 20232.824 65881.70  
## 5 46925.14 24235.859 69614.42  
## 6 58136.05 35746.140 80525.96  
## 7 59511.84 37149.017 81874.67  
## 8 62130.94 39813.758 84448.13  
## 9 62130.94 39813.758 84448.13  
## 10 68022.72 45779.622 90265.82  
## 11 70159.11 47933.039 92385.17  
## 12 71186.55 48966.805 93406.30  
## 13 71186.55 48966.805 93406.30  
## 14 72188.63 49973.872 94403.38  
## 15 75966.42 53760.064 98172.78  
## 16 79422.30 57209.189 101635.40  
## 17 81045.79 58824.757 103266.83  
## 18 82606.83 60375.307 104838.35  
## 19 86959.07 64683.513 109234.62  
## 20 87641.13 65356.711 109925.56  
## 21 92720.50 70353.368 115087.64  
## 22 94472.51 72070.071 116874.96  
## 23 98805.37 76300.817 121309.93  
## 24 100317.92 77772.799 122863.04  
## 25 102719.92 80105.200 125334.64  
## 26 104095.71 81438.281 126753.14  
## 27 106289.87 83560.068 129019.67  
## 28 106714.81 83970.397 129459.23  
## 29 109571.01 86723.331 132418.68  
## 30 110351.45 87474.053 133228.86

#p values <0.05 and multiple r-squared value is 0.8539  
#Exponential Transformation  
reg\_exp<-lm(log(SH) ~ YE)  
summary(reg\_exp)

##   
## Call:  
## lm(formula = log(SH) ~ YE)  
##   
## 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 \*\*\*  
## YE 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

confint(reg\_exp, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 10.4286558 10.5861480  
## YE 0.1123316 0.1385742

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

## Warning in predict.lm(reg\_exp, interval = "predict"): predictions on current data refer to \_future\_ responses

## 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

#p values<0.05 and multiple R-squared value is 0.932  
#polynomial of second degree transformation  
reg\_poly<-lm(SH ~ YE+I(YE^2))  
summary(reg\_poly)

##   
## Call:  
## lm(formula = SH ~ YE + I(YE^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 \*\*\*  
## YE 9259.28 1811.01 5.113 2.25e-05 \*\*\*  
## I(YE^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

confint(reg\_poly, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 16869.5127 35560.3527  
## YE 5543.4050 12975.1628  
## I(YE^2) -295.7344 328.5195

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

## Warning in predict.lm(reg\_poly, interval = "predict"): predictions on current data refer to \_future\_ responses

## 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

#multiple R squared valued is 0.957  
#polynomial of three degree transformation  
reg\_poly1<-lm(SH ~ YE+I(YE^2)+I(YE^3))  
summary(reg\_poly1)

##   
## Call:  
## lm(formula = SH ~ YE + I(YE^2) + I(YE^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 \*\*\*  
## YE -718.71 4892.11 -0.147 0.8843   
## I(YE^2) 2099.35 968.36 2.168 0.0395 \*   
## I(YE^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

confint(reg\_poly1, level=0.95)

## 2.5 % 97.5 %  
## (Intercept) 24032.9470 53693.19673  
## YE -10774.5874 9337.17053  
## I(YE^2) 108.8691 4089.83482  
## I(YE^3) -239.0934 -6.73744

predict(reg\_poly1, interval="predict")

## Warning in predict.lm(reg\_poly1, interval = "predict"): predictions on current data refer to \_future\_ responses

## 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

#multiple R square value is 0.9636  
#Adjusted R squared value is 0.9594  
#polynomial of three degree id best for R-squared values  
predicted\_values<-predict(reg\_poly1)  
predicted\_values

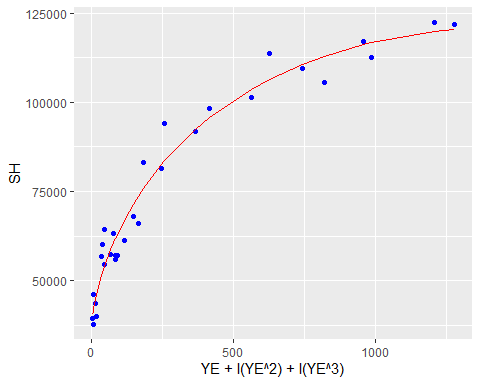
## 1 2 3 4 5 6 7 8   
## 40449.11 41206.61 42093.71 44839.74 46133.97 51436.58 52282.40 54032.88   
## 9 10 11 12 13 14 15 16   
## 54032.88 58717.94 60700.03 61711.28 61711.28 62735.02 66940.09 71285.97   
## 17 18 19 20 21 22 23 24   
## 73496.95 75725.44 82456.89 83577.76 92401.35 95595.79 103603.74 106358.46   
## 25 26 27 28 29 30   
## 110570.09 112836.87 116117.25 116692.05 119867.63 120480.23

library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

ggplot(data=data, aes(x=YE + I(YE^2) + I(YE^3), y=SH))+geom\_point(color="blue")+geom\_line(color="red", data=data, aes(x=YE + I(YE^2) + I(YE^3), y=predicted\_values))



par(mfrow=c(2,2))  
plot(reg\_poly1)

