|  |
| --- |
| > library(readr)  > salary\_data <- read.csv(file.choose())  > View(salary\_data)  > salary\_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  7 3.0 60150  8 3.2 54445  9 3.2 64445  10 3.7 57189  11 3.9 63218  12 4.0 55794  13 4.0 56957  14 4.1 57081  15 4.5 61111  16 4.9 67938  17 5.1 66029  18 5.3 83088  19 5.9 81363  20 6.0 93940  21 6.8 91738  22 7.1 98273  23 7.9 101302  24 8.2 113812  25 8.7 109431  26 9.0 105582  27 9.5 116969  28 9.6 112635  29 10.3 122391  30 10.5 121872  > # Exploratory data analysis #  > summary(salary\_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  > # Scatter plot #  > plot(salary\_data$YearsExperience, salary\_data$Salary) # plot(X,Y)  > attach(salary\_data)  > # Correlation Coefficient (r) #  > cor(YearsExperience, Salary)  [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)  > pred  1 2 3 4 5 6 7 8 9  36187.16 38077.15 39967.14 44692.12 46582.12 53197.09 54142.09 56032.08 56032.08  10 11 12 13 14 15 16 17 18  60757.06 62647.05 63592.05 63592.05 64537.05 68317.03 72097.02 73987.01 75877.00  19 20 21 22 23 24 25 26 27  81546.98 82491.97 90051.94 92886.93 100446.90 103281.89 108006.87 110841.86 115566.84  28 29 30  116511.84 123126.81 125016.80  > reg$residuals  1 2 3 4 5 6 7 8  3155.8412 8127.8488 -2236.1437 -1167.1248 -6691.1173 3444.9091 6007.9128 -1587.0796  9 10 11 12 13 14 15 16  8412.9204 -3568.0608 570.9467 -7798.0495 -6635.0495 -7456.0457 -7206.0306 -4159.0156  17 18 19 20 21 22 23 24  -7958.0080 7210.9995 -183.9779 11448.0259 1686.0560 5386.0673 855.0975 10530.1088  25 26 27 28 29 30  1424.1276 -5259.8611 1402.1577 -3876.8385 -735.8121 -3144.8046  > sum(reg$residuals)  [1] -7.844392e-12  > mean(reg$residuals)  [1] -2.615537e-13  > sqrt(sum(reg$residuals^2)/nrow(salary\_data)) #RMSE  [1] 5592.044  > sqrt(mean(reg$residuals^2))  [1] 5592.044  > 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  > 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 for adding regresion line for data ##  > library(ggplot2)  > ggplot(data = salary\_data, aes(x = YearsExperience, y = Salary)) +  + geom\_point(color='blue') +  + geom\_line(color='red',data = salary\_data, aes(x=YearsExperience, y=pred))  > ## Logrithamic Model/ Transformation ##  > # x = log(YearsExperience); y = Salary  > plot(log(YearsExperience), Salary)  > cor(log(YearsExperience), Salary)  [1] 0.9240611  > reg\_log <- lm(Salary ~ log(YearsExperience)) ## lm(Y ~ 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 9  18795.85 25575.24 31382.55 43057.26 46925.14 58136.05 59511.84 62130.94 62130.94  10 11 12 13 14 15 16 17 18  68022.72 70159.11 71186.55 71186.55 72188.63 75966.42 79422.30 81045.79 82606.83  19 20 21 22 23 24 25 26 27  86959.07 87641.13 92720.50 94472.51 98805.37 100317.92 102719.92 104095.71 106289.87  28 29 30  106714.81 109571.01 110351.45  > reg\_log$residuals  1 2 3 4 5 6 7  20547.1517 20629.7648 6348.4481 467.7377 -7034.1389 -1494.0501 638.1576  8 9 10 11 12 13 14  -7685.9439 2314.0561 -10833.7185 -6941.1059 -15392.5528 -14229.5528 -15107.6281  15 16 17 18 19 20 21  -14855.4226 -11484.2957 -15016.7917 481.1701 -5596.0667 6298.8670 -982.5021  22 23 24 25 26 27 28  3800.4853 2496.6286 13494.0813 6711.0792 1486.2869 10679.1316 5920.1854  29 30  12819.9928 11520.5459  > sqrt(sum(reg\_log$residuals^2)/nrow(salary\_data)) ###RMSE  [1] 10302.89  > 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 |
|  |
| |  | | --- | | > | |