**Simple Linear Regression Assignment**

**1 - Calories\_consumed-> predict weight gained using calories consumed**

> # Load Calories\_consumed.csv dataset

> library(readr)

> Calories\_consumed <- read.csv("E:/Data Science Asignments/Simple regression/calories\_consumed.csv")

> View(Calories\_consumed)

> # Exploratory data analysis

> summary(Calories\_consumed)

Weight.gained..grams. Calories.Consumed

Min. : 62.0 Min. :1400

1st Qu.: 114.5 1st Qu.:1728

Median : 200.0 Median :2250

Mean : 357.7 Mean :2341

3rd Qu.: 537.5 3rd Qu.:2775

Max. :1100.0 Max. :3900

>

> var(Calories\_consumed$Calories.Consumed)

[1] 565668.7

> sd(Calories\_consumed$Calories.Consumed)

[1] 752.1095

> var(Calories\_consumed$Weight.gained..grams.)

[1] 111350.7

> sd(Calories\_consumed$Weight.gained..grams.)

[1] 333.6925

Error: Unable to establish connection with R session

Error: Unable to establish connection with R session

> #Scatter plot

> plot(Calories\_consumed$Calories.Consumed, Calories\_consumed$Weight.gained..grams.) # plot(X,Y)

> attach(Calories\_consumed)

>

>

> #Correlation Coefficient (r)

> cor(Calories.Consumed,Weight.gained..grams.) # cor(X,Y)

[1] 0.946991

>

> # Simple Linear Regression model

> reg <- lm(Weight.gained..grams. ~ Calories.Consumed) # lm(Y ~ X)

>

> summary(reg)

Call:

lm(formula = Weight.gained..grams. ~ Calories.Consumed)

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

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

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

Calories.Consumed 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

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

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

>

> pred <- predict(reg)

>

> reg$Calories.Consumed

NULL

> sum(reg$Calories.Consumed)

[1] 0

Error: Unable to establish connection with R session

Error: Unable to establish connection with R session

> mean(reg$Calories.Consumed)

[1] NA

Warning message:

In mean.default(reg$Calories.Consumed) :

argument is not numeric or logical: returning NA

> sqrt(sum(reg$Calories.Consumed^2)/nrow(Weight.gained..grams.)) #RMSE

numeric(0)

>

> sqrt(mean(reg$Calories.Consumed^2))

[1] NaN

> confint(reg,level=0.95)

2.5 % 97.5 %

(Intercept) -845.4266546 -406.0780569

Calories.Consumed 0.3305064 0.5098069

> predict(reg,interval="predict")

fit lwr upr

1 4.482599 -258.20569 267.1709

2 340.607908 88.93791 592.2779

3 802.780209 533.81393 1071.7465

4 298.592245 46.63271 550.5518

5 424.639236 172.59086 676.6876

6 46.498263 -213.75953 306.7561

7 -37.533065 -302.93258 227.8664

8 172.545254 -82.18110 427.2716

9 550.686227 295.69632 805.6761

10 1012.858527 724.99432 1300.7227

11 75.909227 -182.81852 334.6370

12 172.545254 -82.18110 427.2716

13 508.670563 254.97398 762.3671

14 634.717554 376.22600 893.2091

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 = Calories\_consumed, aes(x = Calories.Consumed, y = Weight.gained..grams.)) +

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

+ geom\_line(color='red',data = Calories\_consumed, aes(x=Calories.Consumed, y=pred))

>

> ?ggplot2

>

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

> # A simple ggplot code for directly showing the line

>

> # ggplot(Calories\_consumed,aes(Calories.Consumed,Weight.gained..grams.))+stat\_summary(fun.data=mean\_cl\_normal) + geom\_smooth(method='lm')

>

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

>

> # Logarithmic Model

>

> # x = log(Calories.Consumed); y = Weight.gained..grams.

>

> plot(log(Calories.Consumed), Weight.gained..grams.)

> cor(log(Calories.Consumed), Weight.gained..grams.)

[1] 0.8987253

>

> reg\_log <- lm(Weight.gained..grams. ~ log(Calories.Consumed)) # lm(Y ~ X)

>

> summary(reg\_log)

Call:

lm(formula = Weight.gained..grams. ~ log(Calories.Consumed))

Residuals:

Min 1Q Median 3Q Max

-187.44 -142.96 23.13 113.20 213.82

Coefficients:

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

(Intercept) -6955.7 1030.9 -6.747 2.05e-05 \*\*\*

log(Calories.Consumed) 948.4 133.6 7.100 1.25e-05 \*\*\*

---

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

Residual standard error: 152.3 on 12 degrees of freedom

Multiple R-squared: 0.8077, Adjusted R-squared: 0.7917

F-statistic: 50.4 on 1 and 12 DF, p-value: 1.248e-05

> predict(reg\_log)

1 2 3 4 5 6 7 8

-19.99870 385.37711 756.06367 343.22032 464.45388 41.20781 -85.42959 204.18573

9 10 11 12 13 14

571.93160 886.18133 81.81708 204.18573 537.44155 637.36248

>

> reg\_log$Calories.Consumed

NULL

> sqrt(sum(reg\_log$Calories.Consumed^2)/nrow(Calories\_consumed)) #RMSE

[1] 0

>

> confint(reg\_log,level=0.95)

2.5 % 97.5 %

(Intercept) -9201.8063 -4709.494

log(Calories.Consumed) 657.3251 1239.418

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

fit lwr upr

1 -19.99870 -165.95156 125.95416

2 385.37711 296.28370 474.47053

3 756.06367 605.03197 907.09537

4 343.22032 254.42082 432.01982

5 464.45388 369.90963 558.99812

6 41.20781 -90.32308 172.73869

7 -85.42959 -247.78936 76.93018

8 204.18573 103.75898 304.61248

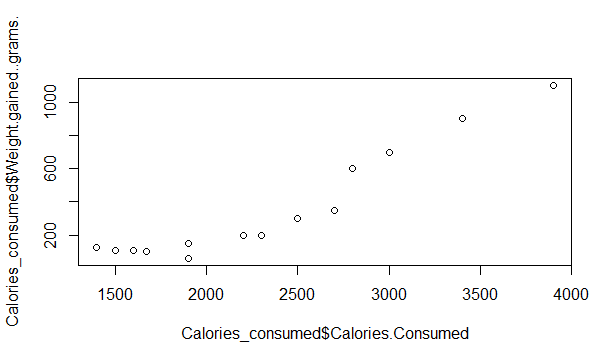
9 571.93160 461.53464 682.32856

10 886.18133 701.33418 1071.02849

11 81.81708 -40.79870 204.43286

12 204.18573 103.75898 304.61248

13 537.44155 433.00103 641.88207

14 637.36248 513.94899 760.77598

> attach(Calories\_consumed)

>

>

> #Correlation Coefficient (r)

> cor(Calories.Consumed,Weight.gained..grams.) # cor(X,Y)

[1] 0.946991

>

> # Simple Linear Regression model

> reg <- lm(Weight.gained..grams. ~ Calories.Consumed) # lm(Y ~ X)

>

> summary(reg)

Call:

lm(formula = Weight.gained..grams. ~ Calories.Consumed)

Residuals:

Min 1Q Median 3Q Max

-158.67 -107.56 36.70 81.68 165.53

Coefficients:

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

(Intercept) -625.75236 100.82293 -6.206 4.54e-05 \*\*\*

Calories.Consumed 0.42016 0.04115 10.211 2.86e-07 \*\*\*

---

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

Residual standard error: 111.6 on 12 degrees of freedom

Multiple R-squared: 0.8968, Adjusted R-squared: 0.8882

F-statistic: 104.3 on 1 and 12 DF, p-value: 2.856e-07

>

> pred <- predict(reg)

>

> reg$Calories.Consumed

NULL

> sum(reg$Calories.Consumed)

[1] 0

Error: Unable to establish connection with R session

Error: Unable to establish connection with R session

> mean(reg$Calories.Consumed)

[1] NA

Warning message:

In mean.default(reg$Calories.Consumed) :

argument is not numeric or logical: returning NA

> sqrt(sum(reg$Calories.Consumed^2)/nrow(Weight.gained..grams.)) #RMSE

numeric(0)

>

> sqrt(mean(reg$Calories.Consumed^2))

[1] NaN

> confint(reg,level=0.95)

2.5 % 97.5 %

(Intercept) -845.4266546 -406.0780569

Calories.Consumed 0.3305064 0.5098069

> predict(reg,interval="predict")

fit lwr upr

1 4.482599 -258.20569 267.1709

2 340.607908 88.93791 592.2779

3 802.780209 533.81393 1071.7465

4 298.592245 46.63271 550.5518

5 424.639236 172.59086 676.6876

6 46.498263 -213.75953 306.7561

7 -37.533065 -302.93258 227.8664

8 172.545254 -82.18110 427.2716

9 550.686227 295.69632 805.6761

10 1012.858527 724.99432 1300.7227

11 75.909227 -182.81852 334.6370

12 172.545254 -82.18110 427.2716

13 508.670563 254.97398 762.3671

14 634.717554 376.22600 893.2091

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 = Calories\_consumed, aes(x = Calories.Consumed, y = Weight.gained..grams.)) +

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

+ geom\_line(color='red',data = Calories\_consumed, aes(x=Calories.Consumed, y=pred))

>

> ?ggplot2

>

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

> # A simple ggplot code for directly showing the line

>

> # ggplot(Calories\_consumed,aes(Calories.Consumed,Weight.gained..grams.))+stat\_summary(fun.data=mean\_cl\_normal) + geom\_smooth(method='lm')

>

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

>

> # Logarithmic Model

>

> # x = log(Calories.Consumed); y = Weight.gained..grams.

>

> plot(log(Calories.Consumed), Weight.gained..grams.)

> cor(log(Calories.Consumed), Weight.gained..grams.)

[1] 0.8987253

>

> reg\_log <- lm(Weight.gained..grams. ~ log(Calories.Consumed)) # lm(Y ~ X)

>

> summary(reg\_log)

Call:

lm(formula = Weight.gained..grams. ~ log(Calories.Consumed))

Residuals:

Min 1Q Median 3Q Max

-187.44 -142.96 23.13 113.20 213.82

Coefficients:

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

(Intercept) -6955.7 1030.9 -6.747 2.05e-05 \*\*\*

log(Calories.Consumed) 948.4 133.6 7.100 1.25e-05 \*\*\*

---

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

Residual standard error: 152.3 on 12 degrees of freedom

Multiple R-squared: 0.8077, Adjusted R-squared: 0.7917

F-statistic: 50.4 on 1 and 12 DF, p-value: 1.248e-05

> predict(reg\_log)

1 2 3 4 5 6 7 8

-19.99870 385.37711 756.06367 343.22032 464.45388 41.20781 -85.42959 204.18573

9 10 11 12 13 14

571.93160 886.18133 81.81708 204.18573 537.44155 637.36248

>

> reg\_log$Calories.Consumed

NULL

> sqrt(sum(reg\_log$Calories.Consumed^2)/nrow(Calories\_consumed)) #RMSE

[1] 0

>

> confint(reg\_log,level=0.95)

2.5 % 97.5 %

(Intercept) -9201.8063 -4709.494

log(Calories.Consumed) 657.3251 1239.418

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

fit lwr upr

1 -19.99870 -165.95156 125.95416

2 385.37711 296.28370 474.47053

3 756.06367 605.03197 907.09537

4 343.22032 254.42082 432.01982

5 464.45388 369.90963 558.99812

6 41.20781 -90.32308 172.73869

7 -85.42959 -247.78936 76.93018

8 204.18573 103.75898 304.61248

9 571.93160 461.53464 682.32856

10 886.18133 701.33418 1071.02849

11 81.81708 -40.79870 204.43286

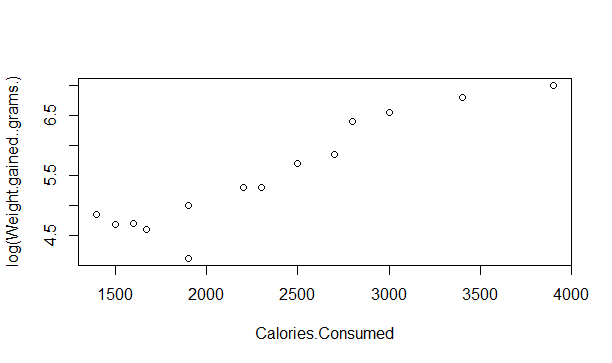
12 204.18573 103.75898 304.61248

13 537.44155 433.00103 641.88207

14 637.36248 513.94899 760.77598

# Exponential Model

# x = Calories.Consumed and y = log(Weight.gained..grams.)



> cor(Calories.Consumed, log(Weight.gained..grams.))

[1] 0.9368037

>

> reg\_exp <- lm(log(Weight.gained..grams.) ~ Calories.Consumed) #lm(log(Y) ~ X)

>

> summary(reg\_exp)

Call:

lm(formula = log(Weight.gained..grams.) ~ Calories.Consumed)

Residuals:

Min 1Q Median 3Q Max

-0.86537 -0.10532 0.02462 0.13467 0.42632

Coefficients:

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

(Intercept) 2.8386724 0.2994581 9.479 6.36e-07 \*\*\*

Calories.Consumed 0.0011336 0.0001222 9.276 8.02e-07 \*\*\*

---

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

Residual standard error: 0.3314 on 12 degrees of freedom

Multiple R-squared: 0.8776, Adjusted R-squared: 0.8674

F-statistic: 86.04 on 1 and 12 DF, p-value: 8.018e-07

>

> reg\_exp$residuals

1 2 3 4 5 6 7

0.14306263 -0.14762919 0.10949103 -0.03426944 0.03111643 0.04805202 0.42632141

8 9 10 11 12 13 14

-0.86537319 0.38418438 -0.25663700 -0.12660998 0.01812772 -0.04145238 0.31161557

>

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

[1] 0.3068228

>

> logat <- predict(reg\_exp)

> at <- exp(logat)

>

> error =Calories\_consumed $Weight.gained..grams. - at

> error

1 2 3 4 5 6 7

14.396423 -31.816603 93.338812 -6.972681 9.191190 5.160737 44.427973

8 9 10 11 12 13 14

-85.305340 191.396489 -321.833419 -13.497427 2.694660 -14.813232 187.415917

>

> sqrt(sum(error^2)/nrow(Calories\_consumed)) #RMSE

[1] 118.0452

>

> confint(reg\_exp,level=0.95)

2.5 % 97.5 %

(Intercept) 2.1862091856 3.491135698

Calories.Consumed 0.0008673238 0.001399871

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

fit lwr upr

1 4.539069 4.243509 4.834628

2 5.445947 5.252660 5.639233

3 6.692904 6.351144 7.034664

4 5.332587 5.136001 5.529173

5 5.672666 5.475078 5.870254

6 4.652428 4.376488 4.928368

7 4.425709 4.109503 4.741915

8 4.992508 4.766646 5.218369

9 6.012745 5.784276 6.241215

10 7.259702 6.801848 7.717557

11 4.731780 4.468840 4.994721

12 4.992508 4.766646 5.218369

13 5.899386 5.683992 6.114779# Polynomial model with 2 degree (quadratic model)

>

> plot(Calories.Consumed,Weight.gained..grams.)

> plot(Calories.Consumed\*Calories.Consumed, Weight.gained..grams.)

>

> cor(Calories.Consumed\*Calories.Consumed, Weight.gained..grams.)

[1] 0.9710636

>

> plot(Calories.Consumed\*Calories.Consumed, log(Weight.gained..grams.))

>

> cor(Calories.Consumed, log(Weight.gained..grams.))

[1] 0.9368037

> cor(Calories.Consumed\*Calories.Consumed, log(Weight.gained..grams.))

[1] 0.9267624

>

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

>

> reg2degree <- lm(log(Weight.gained..grams.) ~ Calories.Consumed + I(Calories.Consumed\*Calories.Consumed))

>

> summary(reg2degree)

Call:

lm(formula = log(Weight.gained..grams.) ~ Calories.Consumed +

I(Calories.Consumed \* Calories.Consumed))

Residuals:

Min 1Q Median 3Q Max

-0.86562 -0.10529 0.02403 0.13541 0.42759

Coefficients:

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

(Intercept) 2.829e+00 1.062e+00 2.664 0.022 \*

Calories.Consumed 1.142e-03 8.807e-04 1.297 0.221

I(Calories.Consumed \* Calories.Consumed) -1.675e-09 1.707e-07 -0.010 0.992

---

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

Residual standard error: 0.3461 on 11 degrees of freedom

Multiple R-squared: 0.8776, Adjusted R-squared: 0.8553

F-statistic: 39.44 on 2 and 11 DF, p-value: 9.611e-06

>

> logpol <- predict(reg2degree)

> expy <- exp(logpol)

>

> err = Calories\_consumed$Weight.gained..grams. - expy

>

> sqrt(sum(err^2)/nrow(Calories\_consumed)) #RMSE

[1] 117.4145

>

> confint(reg2degree,level=0.95)

2.5 % 97.5 %

(Intercept) 4.920101e-01 5.165429e+00

Calories.Consumed -7.962832e-04 3.080576e-03

I(Calories.Consumed \* Calories.Consumed) -3.774780e-07 3.741279e-07

> predict(reg2degree,interval="confidence")

fit lwr upr

1 4.538170 4.166828 4.909512

2 5.446795 5.167854 5.725736

3 6.692653 6.327710 7.057596

4 5.333334 5.066650 5.600018

5 5.673616 5.375476 5.971756

6 4.651865 4.334501 4.969229

7 4.424441 3.986028 4.862855

8 4.992750 4.748299 5.237202

9 6.013597 5.706046 6.321148

10 7.257612 6.584380 7.930845

11 4.731432 4.443227 5.019637

12 4.992750 4.748299 5.237202

13 5.900303 5.593646 6.206960

14 6.240083 5.931876 6.548290

>

> # visualization

> ggplot(data = Calories\_consumed, aes(x = Calories.Consumed + I(Calories.Consumed^2), y = log(Weight.gained..grams.))) +

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

+ geom\_line(color='red',data = Calories\_consumed, aes(x=Calories.Consumed+I(Calories.Consumed^2), y=logpol))

>

>

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

>

14 6.239465 5.978582 6.500348

plot(Calories.Consumed, log(Weight.gained..grams.))

