Assignment 4

**Simple Linear Regression**

Part 1 of 4

**To predict weight gained using calories consumed**

Submitted

To

logo



Submitted

By

Munawwara Shakila

(21\_Nov-2019)

**Q1) Calories\_consumed: Predict weight gained using calories consumed**

**Introduction**

Linear regression is a supervised learning technique. Here target variable is known. The regression model is used to predict the outcome of target variable using one or more independent variables. Here target variable should be of continuous nature and the independent variables can be both continuous and discrete (categorical).

There are two types:

* Simple linear regression
* Multiple linear regression

Simple linear regression is used to predict the outcome of target variable using one independent variable. In Multi linear regression two or more independent variables are used. The main assumption is that the independent variables should be linearly related to the target variable.

The line formula is

Y = m \* X + c

where, c is the y intercept, m is the slope

and Y, X are values on y and x axes respectively

The regression line formula can be given by:

Y = B0 + B1\*X1 + B2\* X2 + ………………… + Bn\*Xn + ε

Where, B0 is intercept, X1 to Xn are variables and B1 to Bn are regression coefficients, ε is the error term

We need to find a line where the difference between actual values and predicted values is the least. The core idea is to find a line that best fits the data. This line is the best fit line. It can be obtained using the OLS (ordinary least squares method).

Since we are finding linear relationship, we need to find whether the predictor is related to the target variable. This can be done using correlation. Correlation coefficient gives the strength and direction of linear relationship between two variables. This relationship also be visualized using scatter plot.

Here we will be using the R package to perform simple linear regression on the given datasets.

--------------------------------------------------------------------------------------------------

**1. Business problem**

**W**e need to predict the weight gained (target variable) using calories consumed. Here we have to find out if a person’s weight gain is related to his food consumption. Food consumption is measured here in terms of calories that he/she consumes.

**2. Data acquisition**

As per the business problem, we need to predict weight gained using calories consumed. The dataset considered includes 2 variables. The target variable here is weight\_gained which is measured in grams. Food consumption here is given in terms of calories that have been consumed. So, the independent variable here is calories\_consumed, measured in calories.

The below table, shows first 6 records of the dataset. The dataset contains 2 variables weight\_gained and calories\_consumed. It has 14 observations.

|  |
| --- |
| weight\_gained calories\_consumed |
| 1 108 1500 |
| 2 200 2300 |
| 3 900 3400 |
| 4 200 2200 |
| 5 300 2500 |
| 6 110 1600 |

**3. Exploratory Data Analysis (EDA)**

Let us understand our data. Both are continuous variables. To know more about our data we need to find the 4 business moments.

|  |
| --- |
| > str(cals) |
| 'data.frame': 14 obs. of 2 variables: |
| $ weight\_gained : int 108 200 900 200 300 110 128 62 600 1100 ... |
| $ calories\_consumed: int 1500 2300 3400 2200 2500 1600 1400 1900 2800 3900 ... |

**Business Moments**

Moments are popularly used to describe the characteristics of a distribution. They summarize many of the descriptive statistical measures.

|  |  |  |
| --- | --- | --- |
|  | weight\_gained | calories\_consumed |
| Mean | 357.71 | 2340.71 |
| Median | 200 | 2250 |
| Mode | 200 | 1900 |
| Variance | 111350.7 | 565668.7 |
| Std deviation | 333.6925 | 752.1095 |
| Range | 62 to 1100  1038 | 1400 to 3900  2500 |
| Skewness | 1.116977 | 0.5825597 |
| Kurtosis | 2.891938 | 2.403367 |

* First business moment decisions: mean, median and mode

These are measures of central tendency: give measures around which most of the data points lie. For weight, mean is 358. Median is 200, it is the middlemost value. Mode is the most frequent value and is 200 with a frequency of 2. Since mean and median differ, it indicates that the data may have outliers.

For calories, mean is 2341, median is 2250 and mode is 1900 (count 2). Since mean and median differ, it indicates that the data may have outliers.

* Second business moment decisions: variance, standard deviation and range

These are measures of dispersion, show the spread of the data.

For weight, the variance is 111,351 which is square of the deviations from the mean. The standard deviation (square root of variance) of weight is 334. The range indicates minimum and maximum values of the dataset. Range value (max – min) is 1038 with a minimum of 62 and maximum of 1100.

For calories, the variance is 565,669 and standard deviation is 752. Range value (max – min) is 2500 with a minimum of 1400 and maximum of 3900.

* Third business moment decision: skewness

Skewness measures the symmetry of the data. Variation tells the amount of variation and skewness tells the direction of variation. The value of skewness for weight is 1.117 and for calories it is 0.583. Also, for both variables, the mean is greater than the median. For weight the distribution could be positively skewed. Since the value for calories is closer to zero, the distribution could be normal. To know more accurately about normality of the distribution, we will be using some more measures in the report.

* Fourth business moment decision: kurtosis

Kurtosis measures the flatness or peakedness of the data. It measures the heaviness of the distribution tails. The value of kurtosis for weight is 2.892 and for calories it is 2.403. Since both values are positive, the distributions could be more peaked than normal. More measures will follow to confirm these results.

The summary() function in r gives the below results. It summarises some of the descriptive statistics values for all the variables. Below table gives the following values: minimum, first quartile (Q1), mean, median (Q2), third quartile (Q3), maximum

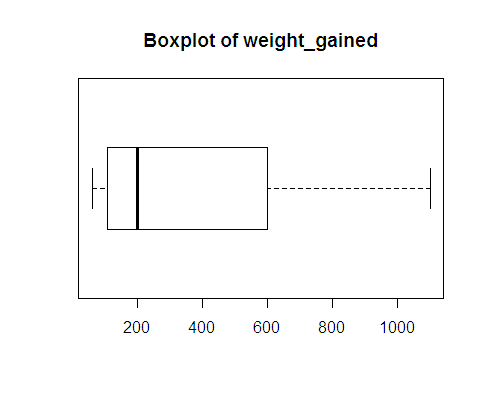
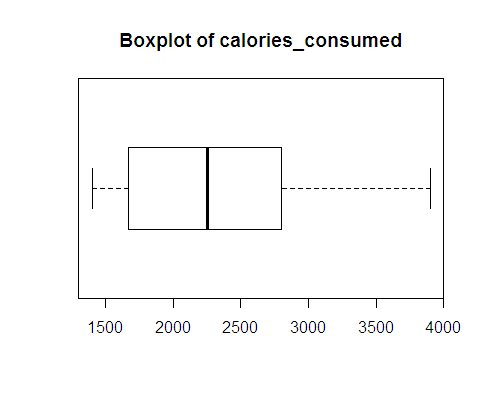
|  |
| --- |
| > summary(cals) |
| weight\_gained calories\_consumed |
| Min. : 62.0 Min. :1400 |
| 1st Qu.: 114.5 1st Qu.:1728 |
| Median : 200.0 Median :2250 |
| Mean : 357.7 Mean :2341 |
| 3rd Qu.: 537.5 3rd Qu.:2775 |
| Max. :1100.0 Max. :3900 |

**Visualizations**

Visualization by way of plots, charts or other images is an easy way to understand our data. They help us to gain useful insights. Depending on the variables plotted, there are different types. There are univariate plots like boxplot, histogram, barplot, dotplot. There are bivariate plots like scatterplot. Also multivariate plots like scatterplot matrix.

Boxplot

Box plot is a 5 point summary of the data. It has a middle box and on sides, whiskers, also called as box and whisper plot. The middle thick line is the median, for weight it is 200. From boxplot of weight, it is clearly visible that the data is positively skewed. Boxplot of calories\_consumedis also slightly skewed towards right. There are no outliers in the data which are generally represented by small circles.

Finding outliers:

* Using r

The below code can be used to find outliers. Since value is 0, indicates both variables do not have outliers.

|  |
| --- |
| > boxplot\_wt$out # no outliers in weight variable |
| numeric(0) |
| > boxplot\_cals$out # no outliers in calories variable |
| numeric(0) |

* Manual calculation of outliers

Also, we can manually calculate outliers using IQR. Since the minimum and maximum values of weight are within the outlier limit, there will be no outliers. Similarly for calories, there are no outliers.

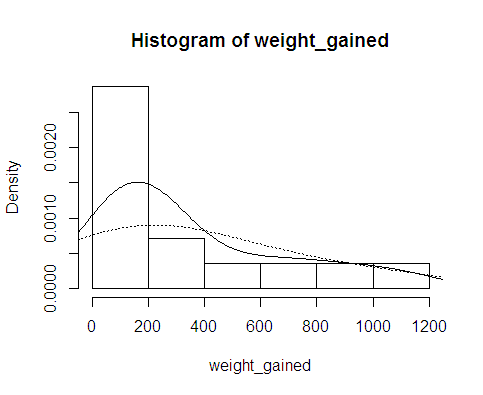
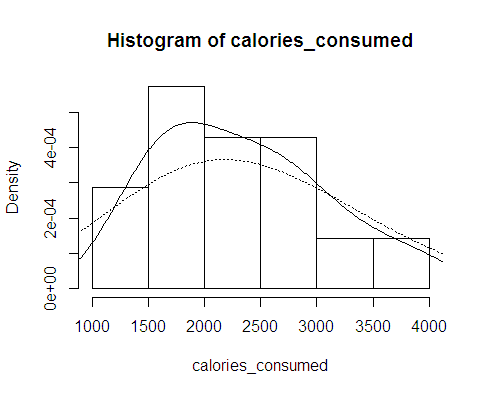
IQR = Q3 – Q1

Outlier lower limit = Q1 – 1.5 \*(IQR)

Outlier upper limit = Q3 + 1.5 \*(IQR)

|  |  |  |
| --- | --- | --- |
|  | weight\_gained | calories\_consumed |
| Q1 | 114.5 | 1728 |
| Q2 | 200 | 2250 |
| Q3 | 537.5 | 2775 |
| IQR = Q3 – Q1 | 423 | 1047.5 |
| outlier lower limit | 114.5 - (1.5 \* 423) = -520 | 1728 - (1.5 \* 1047.5) = 156.75 |
| outlier upper limit | 537.5 + (1.5 \* 423) = 1172 | 2775 + (1.5 \* 1047.5) = 4346.25 |
| minimum value | 62 | 1400 |
| maximum value | 1100 | 3900 |

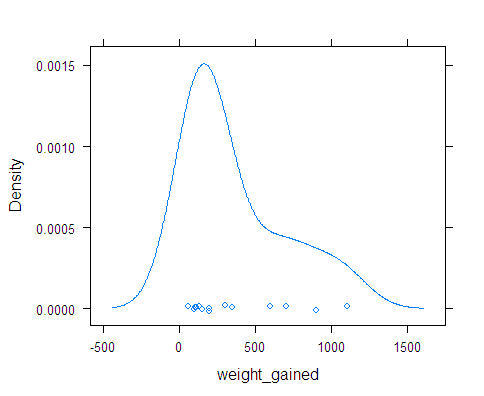
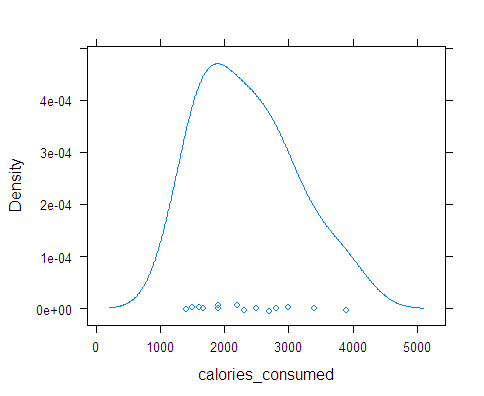
Histogram

In histogram the variable is divided into bins and plotted on X axis. Their frequency is plotted on Y axis. We can see that weight is right skewed with mode lying between 0 and 200 (mode is 200). Calories is almost normally distributed with mode lying between 1500 and 2000 (mode is 1900). The curve of density estimates also shows the same distribution of data.

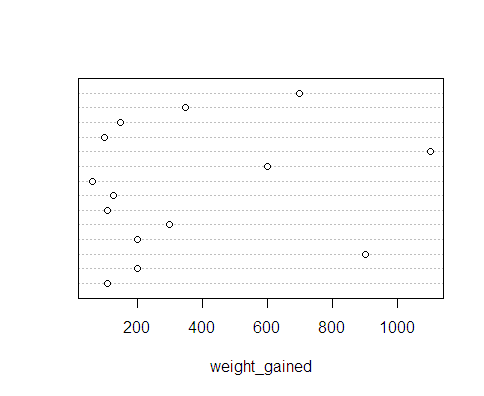
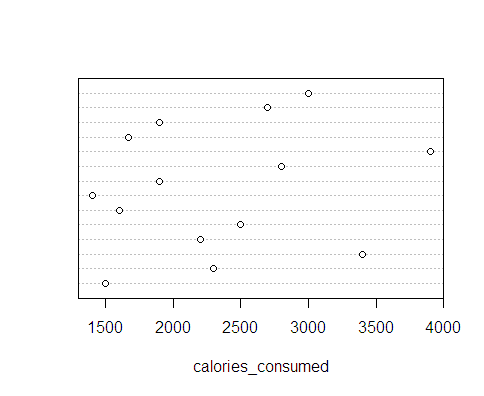
Densityplot

It is a smoothed version of histogram. It shows the probability density function of the variable using kernel density estimate. Weight is right skewed and calories is normal.

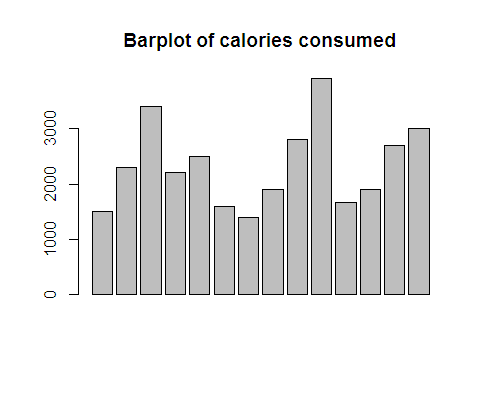
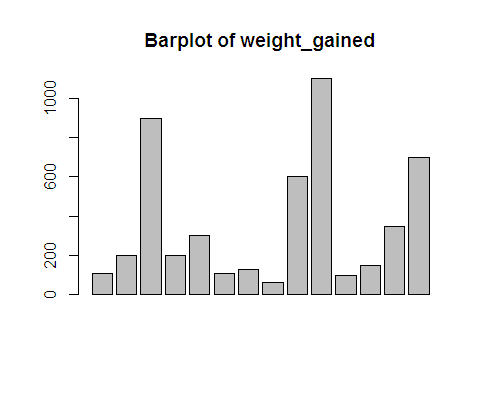
Dot chart

Dot chart or dot plot is a graphical distribution of data using dots. It is a simple histogram like chart used for small datasets.

**Barplot**

Barplot can be used for small datasets. The length of the bar gives value of each data point.

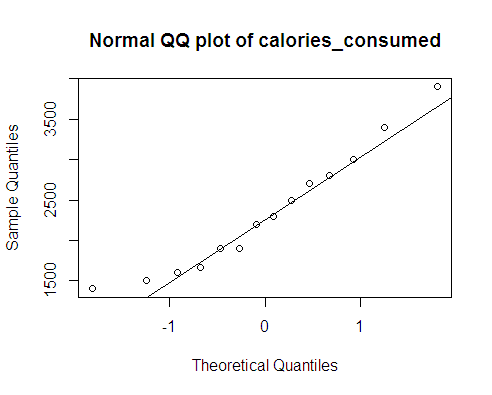
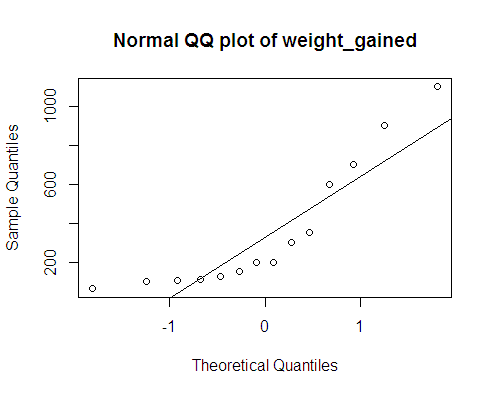


**Finding Normality of data**

Most of the analytical models developed work with underlying assumption that the variables are normally distributed. There are different ways to know whether variables are normally distributed or not. We have already seen boxplot, histogram and density plot.

Normal Quantile-Quantile plot

QQ plot compares two distributions. It plots the theoretical quantiles here of normal distribution against the variable quantiles.

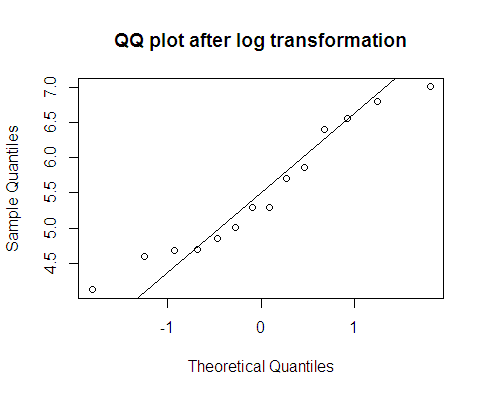
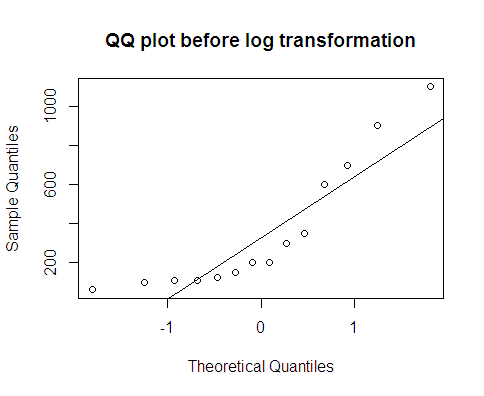


Shapiro and Anderson-Darling tests for normality

The values of both test are shown below. The null hypothesis for the tests is that data is normal. For weight variable, p-value is less than 0.05. Hence the data is not normal. For calories variable, p-value is more than 0.05. It follows normal distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Hypothesis | weight\_gained | calories\_consumed | Weight\_log (log transformation) |
|  |  | p value | p value | p value |
| Shapiro test |  | 0.006646 | 0.4887 | 0.4051 |
| Anderson-darling test |  | 0.004069 | 0.6086 | 0.3827 |
| Inference | H0: data normal  Ha: data not normal | pvalue < 0.05,  reject null hyp  data is not normal | pvalue > 0.05,  cannot reject null hyp  data is normal | pvalue > 0.05,  cannot reject null hyp  data is normal |

For large datasets, non-normality is not a problem. But since out dataset has just 14 records, we can try transformations to normalize our weight variable



The above qq plots show the weight variable. Before applying any transformation, the data is not normal. After log transformation, the data is normal. The same is confirmed by shapiro test and AD test. The p-value for both is > 0.05. So the data is now normally distributed.

Both variables are now normally distributed.

**Missing Values**

The missing patterns in the dataset can be found using the following r codes. There are no missing values in the dataset.

|  |
| --- |
| > library(mice) |
| > md.pattern(cals) |
|  |
| No need for mice. This data set is completely observed. |
|  |
| weight\_gained calories\_consumed |
| 14 1 1 0 |
| 0 0 0 |
|  |
| > cals[!complete.cases(cals),] |
| [1] weight\_gained calories\_consumed |
| <0 rows> (or 0-length row.names) |

**4. Model building and Interpretation**

We have pre-processed our dataset. The dataset has no missing values, there are no outliers. The target variable weight\_gained is not normal, predictor variable calories\_consumed is normally distributed.

The business problem is to predict weight gained using calories consumed. We need to build a prediction model. As seen, both are continuous variables. Here target variable is weight\_gained. We need to predict the outcome using only one independent variable calories\_consumed. So we will be using simple linear regression technique.

Linear relationship

First, linear regression needs the predictor to be linearly related to the target variable. We can check this visually using scatter plot. Also, we can find magnitude using pearson correlation coefficient.



The scatter plot shows a positive correaltion between both the variables. The correlation coefficient shows a strong positive linear relationship between the variables. This means that calories\_consumed is a good input to predict weight\_gained.

|  |
| --- |
| > cor(weight\_gained,calories\_consumed) |
| [1] 0.946991 |

**I Standard Linear Regression Model using weight\_gained and calories\_consumed**

Let us now build the Simple linear regression model in r. Below table gives summary statistics of regression.

|  |
| --- |
| > # Regression Model |
| > reg\_simple <- lm(weight\_gained~ calories\_consumed) |
| > summary((reg\_simple)) |
|  |
| Call: |
| lm(formula = weight\_gained ~ 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 |

Formula call: Shows the syntax used to build the linear model

Coefficients: These are the estimates of regression equation.

The Regression equation is :

ŷ = B0 + B1\*X1 + ε

Where, ŷ is the predicted value of dependent variable,

B0 is Y intercept,

X1 is independent variable and B1 is regression coefficient of X1 variable,

ε is the error term

Weight\_gained = B0 + B1\* (calories\_consumed) + ε

Using above coefficients we can write this equation as:

Weight\_gained = -625.75236 + 0.42016 \* calories\_consumed

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intercept | Coeff of calories\_consumed | calories\_consumed | Predicted(y) | If X1 increases by 100 units, increase in Y is |
| 1 | 2 | 3 | 4 | 5 |
| -625.75236 | 0.42016 | 1600 | 46.50364 | 88.51964  -- 46.50364  = **42.016** |
| -625.75236 | 0.42016 | 1700 | 88.51964 |  |

If there are no variables, then the loss in weight\_gained is 625.752 gms. We are now adding a new variable calories\_consumed.

If calories\_consumed is increased by 1 unit, then predicted weight\_gained increases by 0.42016 gms i.e., if there is 100 units increase in calories\_consumed, the weight\_gained increases by 42.016 gms (as shown in 5th column of above table).

If we have a new value of 1700 calories\_consumed, then we can predict the weight\_gained as 88.51964 gms.

t-statistic: The test statistic used is t-value. The null hyothesis here is that the coefficients are zero. The pvalue is = 0.000 for both intercept and calories\_consumed. This indicates that both coefficients are statistically different from zero. They are highly significant in estimating weight\_gained.

F-statistic: It is a measure to know if the model is good or not.The p-value is equal to 0.00, indicating that the model is good and is significant in predicting weight\_gained.

R2 and Adjusted R2: R2 is the coefficient of determination. It indicates that about 89.68% of the variance of weight\_gained is explained by the model. R2  increases with increase in number of predictors even if they are not significant. Adj. R2 penalises if more variables are used in the model. Here we have just one predictor and adj. R2  value (0.8882) is close to R2 value, indicating that the model is very good.

Now let us plot the regression line on the scatter plot of actual weight\_gained vs calories\_consumed. The straight line gives the predicted Y values. Not many actual values of Y are lying on the regression line. We can say that model is not very good in predicting Y.



Residuals: Residuals are the difference between actual values of weight\_gained and the predicted values. This section of the model breaks down residuals into 5 summary points.

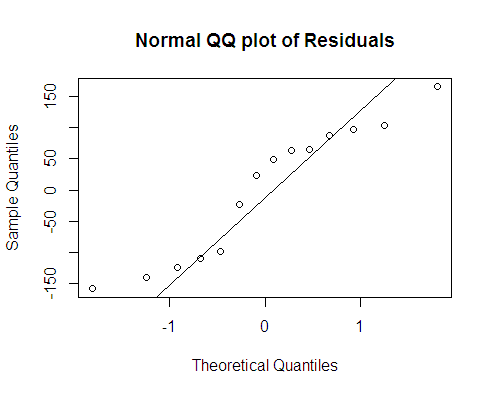
Residual = observed value – predicted value

e = y – ŷ

Both the sum and mean of the residuals should be equal to zero, which is true in this case.

RMSE: Root mean squared error should be lower for a good model. Here it is 103. We can compare this value with different models and select the model with least RMSE.

|  |
| --- |
| > sum(reg\_simple$residuals) # =0, assumption of error |
| [1] 3.552714e-15 |
| > mean(reg\_simple$residuals) # =0 |
| [1] 2.542609e-16 |
| > sqrt(mean(reg\_simple$residuals^2)) # RMSE = 103.3025 |
| [1] 103.3025 |



Residual plot: The plot shows residuals on Y axis and calories\_consumed (predictor) on X axis. If the pattern is random, they we can say that the model fits the data. Since here the pattern is not random (U-shaped curve), it is not a good model. We can try transformations to get a good linear model.

Normality of residuals: The residuals should be normally distributed across the mean. From below plot we find that the distribution of errors is not normal. But when we use Shapiro test and Anderson-Darling tests, the p-value is >0.05, indicating that the residuals follow normal distribution.

Confidence and Prediction intervals:

These are the types of confidence intervals used for predictions in regression and other linear models. Both these quantify for uncertainty in statistical analysis.

Confidence interval of prediction: It represents a range that the mean resonse is likely to fall.

When we substitute these values in the regression equation, we get:

Lower range for mean of weight\_gained = -845.4267 + 0.33059 \* (calories\_consumed)

Upper range for mean of weight\_gained = -406.0781 + 0.5098 \* (calories\_consumed)

Prediction interval: it represents a range that a single new observation is likely to fall. The prediction interval table is given below for all predicted values.The fit value is the predicted value got by susbtituting the actual values in regression equation. For example the first value 4.4826 (in below prediction interval table), is the predicted weight\_gained value for 1500 calories\_consumed. Weight\_gained predicted value can range from -258 to 267, when 1500 calories is consumed. For the entire dataset, the prediction interval ranges from -302.93 to 1300.72. So, we see that prediction inteval is wider than confidence interval.

|  |
| --- |
| > # confidence and prediction intervals |
| > confint(reg\_simple,level = 0.95) |
| 2.5 % 97.5 % |
| (Intercept) -845.4266546 -406.0780569 |
| calories\_consumed 0.3305064 0.5098069 |

|  |
| --- |
| > predict(reg\_simple,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 |

We have now built a standard regression model. We have alos looked at various criteria or parameters that make a good model. From above results, we can conclude that the regression model obtained is not a good model. So let us try transformations and try to get a good linear model.

**II Exponential Model: taking log(y)**

We have seen that the target variable, weight\_gained is not normally distributed. But the predictor calories\_consumed is normal. Also in the residual plot, the errors were not random. The model obtained was not very good.

To overcome these shortcomings, let us try transforming our Y variable.

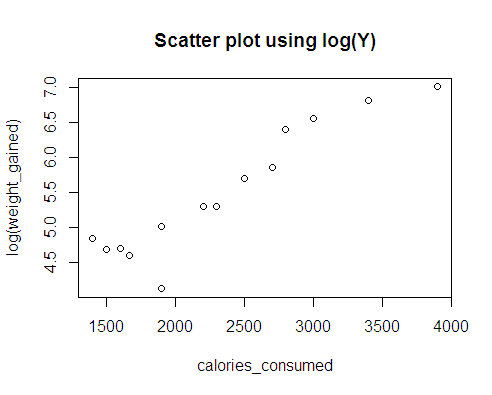
* First, let us try log transformation of Y – exponential model
  + log(Y) = B0 + B1 \* X1+ ε
* Second, let us try taking square root of Y – quadratic model
  + sqrt(Y) = B0 + B1 \* X1 + ε
* Our standard regression equation without transformation was
  + Y = B0 + B1 \* X1  + ε

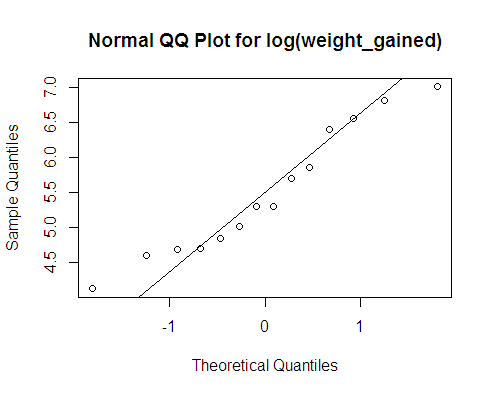
We are now building the model using log(Y).

Let us plot the scatter plot using log(Y). When we compare with sccatter plot without transfomation of Y, it looks more linear but for one data point. The correlation coefficient with log(Y) is 0.937, a strong positive value but less than 0.947 obtained using only Y.

The QQ plot plots the quantiles of log(weight\_gained) against normal distribution quantiles. Shapiro test for normality gives value of 0.4051 > 0.05. The log(weight\_gained) is normally distributed against only weight\_gained, which was not normal

|  |
| --- |
| > cor(log(weight\_gained),calories\_consumed) # 0.9368037, strong positive |
| [1] 0.9368037 |





Summary of regression using log(Y)

The coefficients for intercept and calories\_consumed are 2.8387 and 0.0011 respectively: both are very significant. From F-statistic also it is clear that calories consumed is a significant varaible in predicting weight gained.

Here the regresssion equation is:

log(Y) = B0 + B1 \* X1 + ε

When we substitute the coefficents, we get:

log(weight\_gained) = 2.8386724 + 0.0011336 \* calories\_consumed

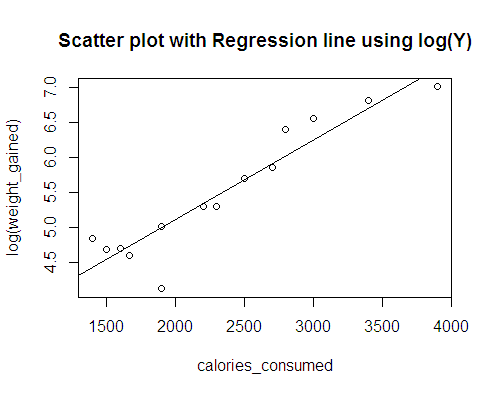
For interpreting above equation, we need to exponentiate the coefficients. The exp(0.0011336) is 1.001134243. if we increase calories by 1 unit, weight will increase by 0.1%. To find for 100 unit increase then we have to find exp(0.0011336 \* 100) which is 1.120035073. For a 100 unit increase in calories, we can see about 12% increase in weight.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Calories\_consumed | pred\_log | exp(pred\_log) |  |  |
| 1400 | 4.425709 | 83.57203 |  |  |
| 1500 | 4.539069 | 93.60358 | 93.60358 / 83.57203 | 1.120035 |
| 1600 | 4.652428 | 104.8393 | 104.8393 / 93.60358 | 1.120035 |

The R2 value is 0.8776 which is less than R2 value of standard regression of 0.8968. The adjusted R2 of 0.8674 is also close to R2 value. If we only consider R2 values, it indicates that standard regression is a better model than the transformed log(Y) model. But we should always look at more than one measure to say if a model is better or not.

|  |
| --- |
| > #linear eqn using log(y) |
| > reg\_log <- lm(log(weight\_gained)~ calories\_consumed) |
| > summary((reg\_log)) |
|  |
| Call: |
| lm(formula = log(weight\_gained) ~ 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 |

Now let us plot the regression line on the scatter plot of log transformed weight\_gained vs calories\_consumed. The straight line gives the predicted Y values. When we compare with previous regression line, we can see that most of the values lie on the line itself. This is a better model than the previous one.

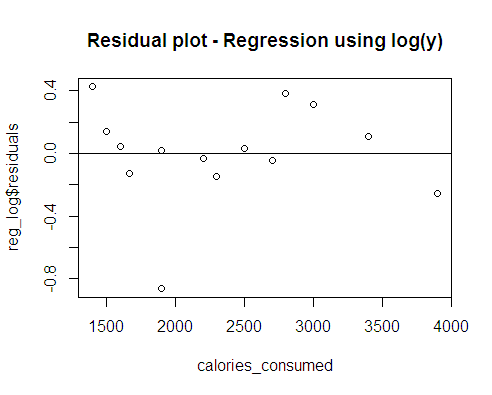


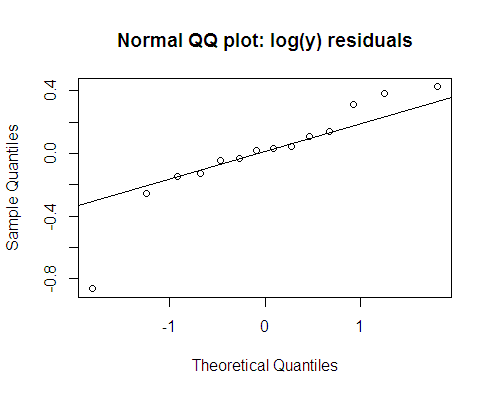
Residuals: Now let us look at the residuals. The sum and mean of the residuals is zero. RMSE is 0.3068 which is very less compared to RMSE of 103.3025 of previous model. When we look at the residual plot, we find that the errors are random in nature, following homoscedasticity. When we compare residual plots of this model and standard regression model, we can say that this is a better model.

The residuals are normally distributed as shown in the above QQ plot and also the Anderson-Darling test gives a p-value of 0.1096, which is > 0.05.

From above results we can conclude that this is a better model as compared to previous standard regression model.

|  |
| --- |
| > sum(reg\_simple\_log$residuals) # =0, assumption of error |
| [1] 4.510281e-17 |
| > mean(reg\_simple\_log$residuals) # =0 |
| [1] 3.221629e-18 |
| > sqrt(mean(reg\_simple\_log$residuals^2)) # RMSE = 0.3068228 |
| [1] 0.3068228 |



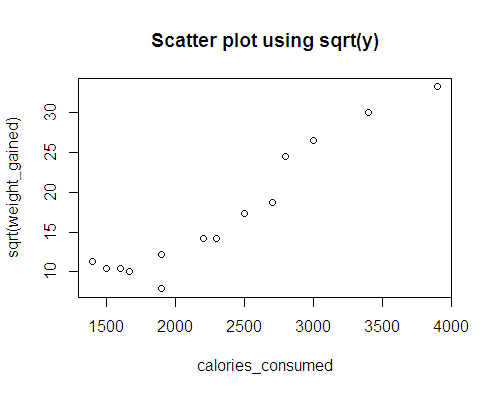
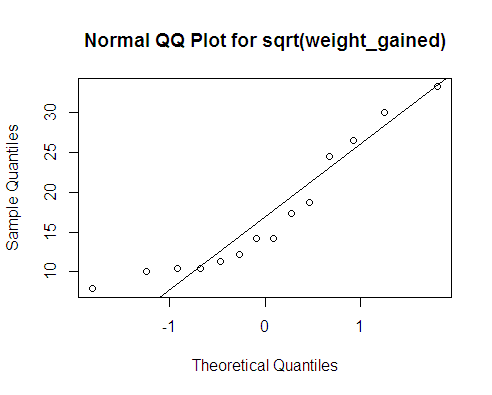


**III Quadratic Model: taking squareroot(Y)**

We are now building the model using squareroot(Y).

First we have to see if sqrt(Y) is normally distributed or not. Since the dataset is small with only 14 records, linear regression with non-normal data does not give good results. From QQ plot below the data looks normal. The Shapiro test value is 0.06, indicating that now target value is normal. As seen earlier, weight\_gained without any transformation was not normal.

Next we have to check if both targte and predictor variables share a linear relationship or not. To find this let us plot the scatter plot using sqrt(Y). The correlation coefficient with sqrt(Y) is 0.956, a strong positive value and more than 0.947 obtained using only Y.

Summary of regression using sqrt(Y)

The coefficients for intercept and calories\_consumed are -7.1154342 and 0.0103864 respectively: both are significant. From F-statistic also it is clear that calories consumed is a significant varaible in predicting weight gained.

The R2 value is 0.8776 which is less than R2 value of standard regression of 0.8968. The adjusted R2 of 0.8674 is also close to R2 value. If we only consider R2 values, it indicates that standard regression is a better model than the transformed log(Y) model. But we should always look at more than one measure to say if a model is better or not.

|  |
| --- |
| > # regression model |
| > reg\_sqrt <- lm(sqrt(weight\_gained) ~ calories\_consumed) |
| > summary(reg\_sqrt) |
|  |
| Call: |
| lm(formula = sqrt(weight\_gained) ~ calories\_consumed) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -4.7448 -1.5770 -0.2277 1.8965 3.8881 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -7.1154342 2.2552542 -3.155 0.0083 \*\* |
| calories\_consumed 0.0103864 0.0009204 11.285 9.56e-08 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 2.496 on 12 degrees of freedom |
| Multiple R-squared: 0.9139, Adjusted R-squared: 0.9067 |
| F-statistic: 127.3 on 1 and 12 DF, p-value: 9.56e-08 |

Regression equation and its interpretation

Here the regresssion equation is:

sqrt(Y) = B0 + B1 \* X1 + ε

When we substitute the coefficents, we get:

sqrt(weight\_gained) = -7.1154342 + 0.0103864 \* calories\_consumed

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | Coeff of calories\_consumed | calories\_consumed | Predicted(y) in terms of sqrt | If X1 increases by 100 units | Sqaure of pred(y) |
| 1 | 2 | 3 | 4 | 5 | 6 |
| -7.1154342 | 0.0103864 | 1600 | 9.5028058 | 10.5414458  -- 9.5028058  = **1.03864** | 90.3033181 |
| -7.1154342 | 0.0103864 | 1700 | 10.5414458 |  | 111.12208 |

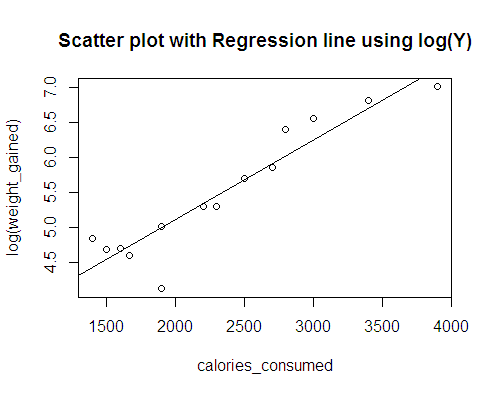
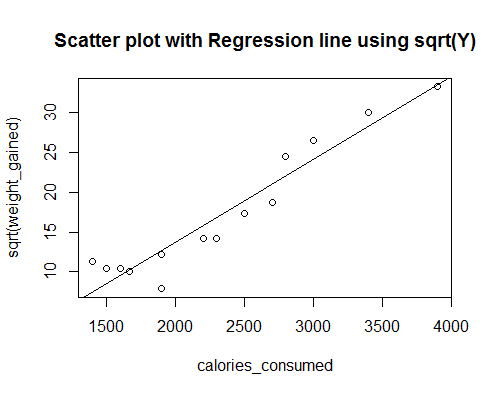
If there are no variables, then the squareroot of weight\_gained decreases by 7.1154. We are now adding a new variable calories\_consumed.

If calories\_consumed is increased by 1 unit, then sqrt of predicted weight\_gained increases by 0.0103864 i.e., if there is 100 units increase in calories\_consumed, the sqrt of weight\_gained increases by 1.03864 units (as shown in 5th column of above table).

If we have a new value of 1700 calories\_consumed, then we get predicted weight\_gained as 10.5414. But this is in terms of square root. So if we square it, the weght\_gained is 111.122 gms.

Scatter plot with Regression line:

Now let us plot the regression line on the scatter plot. Below 3 plots show the regression lines obtained by 3 different models. The regression line obtained using log(Y) appears to be more fitting to the data points than other 2 models.

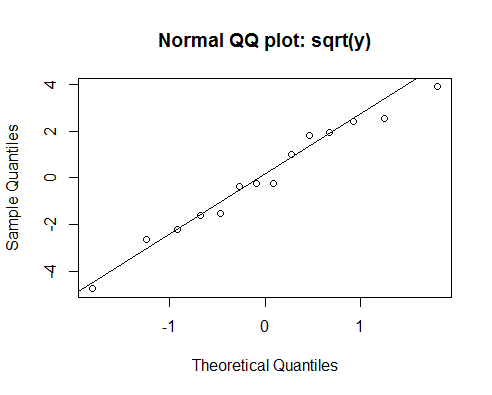
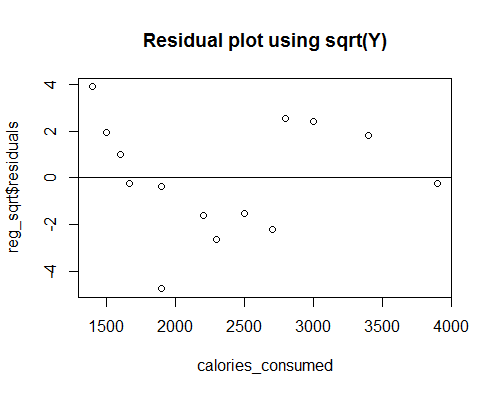




Residuals: Now let us look at the residuals. The sum and mean of the residuals is zero. RMSE is 2.3107 which is very less compared to RMSE of 103.3025 of standard regression model but more than the RMSE obtained using log(Y). When we look at the residual plot, we find that the errors are random in nature. When we compare residual plots of this model and standard regression model and with model using log(Y) we can say that this is a better model.

The residuals are normally distributed as shown in the below QQ plot and also the Shapiro-Wilk test gives a p-value of 0.9201, which is > 0.05.

|  |
| --- |
| > # RESIDUALS |
| > sum(reg\_sqrt$residuals) |
| [1] 1.498801e-15 |
| > mean(reg\_sqrt$residuals) |
| [1] 1.070417e-16 |
| > sqrt(mean(reg\_sqrt$residuals^2)) # RMSE |
| [1] 2.310718 |



**Comparison of all three models**

We have used linear regression technique and arrived at three models. Let us now compare them and find out which is the best model. Below table gives important criteria to compare them.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Standard model** | **Exponential model** | **Quadratic model** |
| Regression equation | Y = B0 + B1 \* X1 + ε | log(Y) = B0 + B1 \* X1+ ε | sqrt(y) = B0 + B1 \* X1+ ε |
| Dependent variable | weight\_gained | log(weight\_gained) | sqrt(weight\_gained) |
| Normally distributed  (Shapiro test p-value) | Y not normal  p-value = 0.006646 | log(Y) is normal  p-value = 0.4051 | sqrt(Y) is normal  p-value = 0.0631 |
| Independent variable calories\_consumed | X is normal  p-value= 0.4887 | X is normal | X is normal |
| Correlation | 0.946991 | 0.9368037 | **0.9559736** |
| Intercept value: B0  (p-value) | -625.75236  (0.00) | 2.8386724  (0.00) | 7.1154342  (0.0083) |
| Coefficient B1  (p-value) | 0.42016  (0.00) | 0.0011336  (0.00) | 0.0103864  (0.00) |
|  |  |  |  |
| R2 | 0.8968 | 0.8776 | **0.9139** |
| Adjusted R2 | 0.8882 | 0.8674 | **0.9067** |
| F statistic p-value | 0.00  Calories\_consumed is significant in predicting Y | 0.00  Calories\_consumed is significant in predicting Y | 0.00  Calories\_consumed is significant in predicting Y |
| Sum of residuals | 0.00 | 0.00 | 0.00 |
| Mean of residuals | 0.00 | 0.00 | 0.00 |
| RMSE | 103.3025 | **0.3068228** | 2.310718 |
| Normality of residuals | Shapiro test p-value 0.155 >0.05, hence normal | Anderson-Darling test p-value is 0.1096 >0.05 ,  Hence residuals normal | Shapiro test p-value 0.9201 >0.05, hence normal |
| Residual plot | Not random | Random errors | Random errors |
| Good linear model | Not a good model | Better than standard model | Better than both models |

Standard model: The dependent variable weight\_gained (Y) is not normal, since this is a small dataset, non-normality creates problem with regression model. The errors are not random

Exponential Model: Y is normal. Errors generated are random. R2 is the least as compared to all 3 models, which is not good. But RMSE is least compared to all 3 models, which is good.

Quadratic Model: Y is normal and errors are random. R2 is the highest comapred to all 3 models, which is a good feature. RMSE is less than standard model but more than exponential model.

Limitations:

The dataset is very small, only 14 records which is a major limitation for generating a good mode. We can increase the sample size to get a better model. More transormations can be done for both target and predictor variables but not possible due to time limit. More input varaibles can be added, to predict weight\_gained.

**Conclusions**

The dataset has 2 variables and both are continuous. As per the business problem we have predicted weight\_gained using calories\_consumed. The predictor is normally distributed. Target variable is not normal and so we have used transformations. We have used log(Y) and squareroot(Y) to normalize the data.

We are using simple linear regreession technique to predict weight\_gained. We have built three models: standard model using only Y, exponential model using log(Y) and quadratic model using squareroot(Y).

The exponential and quadratic models are better than standard regression model. The exponential model has least RMSE and quadratic model has the highest R2. Of these 2 models, it is difficult to conclude which is the better one. Also if a large dataset is given, we can be more confident of finding the best model.

----------------------------------X-------------------X-----------------X--------------------------------------