1) Estimation of a Simple Regression Model, Significance, and Confidence Interval

FOSSEE (R Team)

28 January, 2020

What is a Simple Regression Model, and what do we mean by its Significance and Confidence Interval?

1. **Simple Regression Model** refers to a predictive modelling technique using which the relationship between an independent and a dependent variable can be determined. Let us represent the independent variable by 'X' and the dependent variable by 'Y'. We can represent the Simple Regression Model by the equation "$Y = a + bX + e$". Here, 'a' represents intercept, 'b' represents slope and 'e' represents error term.
2. To determine the **significance** of the generated model we can analyze the following parameters associated with the model -

- *Residual standard error* - A measure of the deviation of predicted values from the actual values. If residual standard error is zero, then there is probably an overfitting situation.
- *R-squared* - It is a measure of the proportion of variability in predicted values that can be explained using model's inputs. The value should be as high or close to one as possible for best fit.
  - *Multiple R-squared* - It represents the percentage of the variance of the generated model intact, after subtracting error of the model.
  - *Adjusted R-squared* - It gives an adjusted value of R-squared by increasing R-squared only when a newly added term in the model improves the model's fit; otherwise R-squared is reduced.
- *p-value* - Summary function performs an F-test on the generated model and compares it to a model that has fewer parameters. Theoretically, a model with more parameters should fit better, and hence the p-value should be as low as possible and very near to absolute zero.

3. **Confidence Interval** of a regression model represents the difference between the upper and lower limit of the predicted values for each data point.

Null Hypothesis (H_o)

There is a relationship between feed supplement type and chicken weight.

Introduction

The purpose of this experiment is to create a univariate linear regression model fitted over the dataset "chickwts". This dataset contains the measure of the growth rate of chickens depending upon the type of feed supplements provided. After successfully creating the model, we calculate its significance and confidence interval.

Procedure

Step by step procedure to conduct the required experiment -

1. Splitting the dataset for training and testing
2. Normalizing data
3. Labelling feed supplements
4. Creating a linear regression model and predicting feed
5. Plotting predictions against actual values
6. Calculating the significance of the generated model
7. Calculating the confidence interval of the generated model

Code and Results

Data used for analysis

```
## R has a predefined dataset with the name "chickwts"
  # To know more about the dataset type "?chickwts" in the console
```

Splitting the dataset for training and testing

```
  # 1) Creating training and testing data from "chickwts" dataset
    # 1.1) Generating Random Numbers
    set.seed(100)
    # 1.2) Creating sample for splitting the dataset
    Sample <- sample(nrow(chickwts),0.7*nrow(chickwts))
    # 1.3) Training dataset contains 70% of data
    Training <- chickwts[Sample,]
    # 1.4) Testing dataset contains 30% of data
    Testing <- chickwts[-Sample,]
```

Normalizing data

```
  # 2) Normalizing data
    # 2.1) Normalizing training data
    Training[,1] <- scale(Training[,1])
    # 2.2) Normalizing testing data
    Testing[,1] <- scale(Testing[,1])
```

Labelling feed supplements

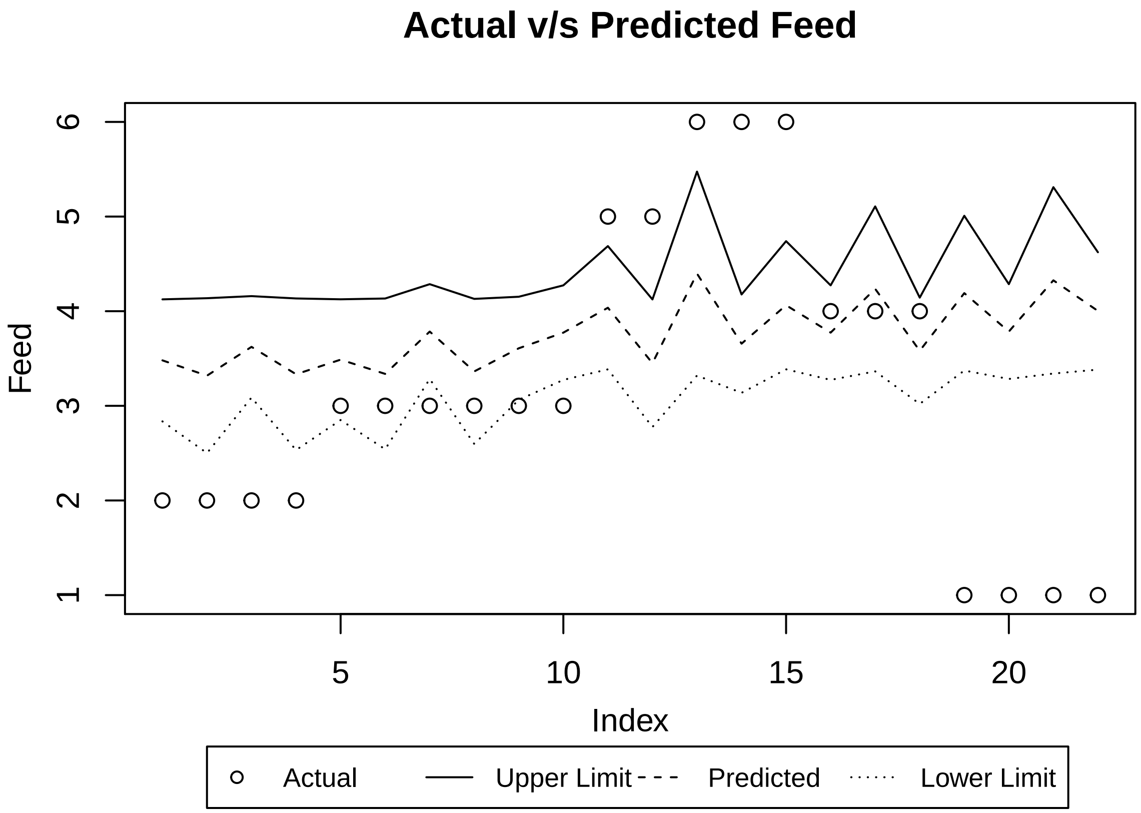
```
  # 3) Labelling feed
    # 3.1) Labelling feed for training data
    Training$feed <- as.integer(factor(Training$feed))
    # 3.2) Labelling feed for testing data
    Testing$feed <- as.integer(factor(Testing$feed))
```

Creating a linear regression model and predicting feed

```
# 4) Creating a Linear Regression model and predicting values
  # 4.1) Applying Linear Regression
  Model <- lm(feed~weight,data = Training)
  # 4.2) Predicting feed
  Prediction <- as.data.frame(predict(Model,Testing,interval='confidence'))
```

Plotting predictions against actual values

```
# 5) Plotting Predictions
  plot(Testing$feed,main = "Actual v/s Predicted Feed",ylab = "",xlab = "")
  # 5.1) Adding a line to represent Upper Limit of prediction
  lines(Prediction$upr,lty = 1)
  # 5.2) Adding a line to represent Best Fit of prediction
  lines(Prediction$fit,lty = 2)
  # 5.3) Adding a line to represent Lower Limit of prediction
  lines(Prediction$lwr,lty = 3)
  # 5.4) Clipping legend to the figure region instead of the plot region
  par(xpd=TRUE)
  # 5.5) Adding a legend
  legend(2,-0.60,legend=c("Actual","Upper Limit","Predicted","Lower Limit"),
  pch = c(1,NA,NA,NA) ,lty = c(NA,1,2,3) , cex=0.8, horiz=T, inset = c(0,0),
  xpd = TRUE)
  # 5.6) Setting and adjusting labels of x and y axes
  title(ylab="Feed",xlab="Index", line=2.25, cex.lab=1)
```



Calculating the significance of the generated model

```r
# 6) The Significance of resultant Linear Model
  summary.lm(Model)
```

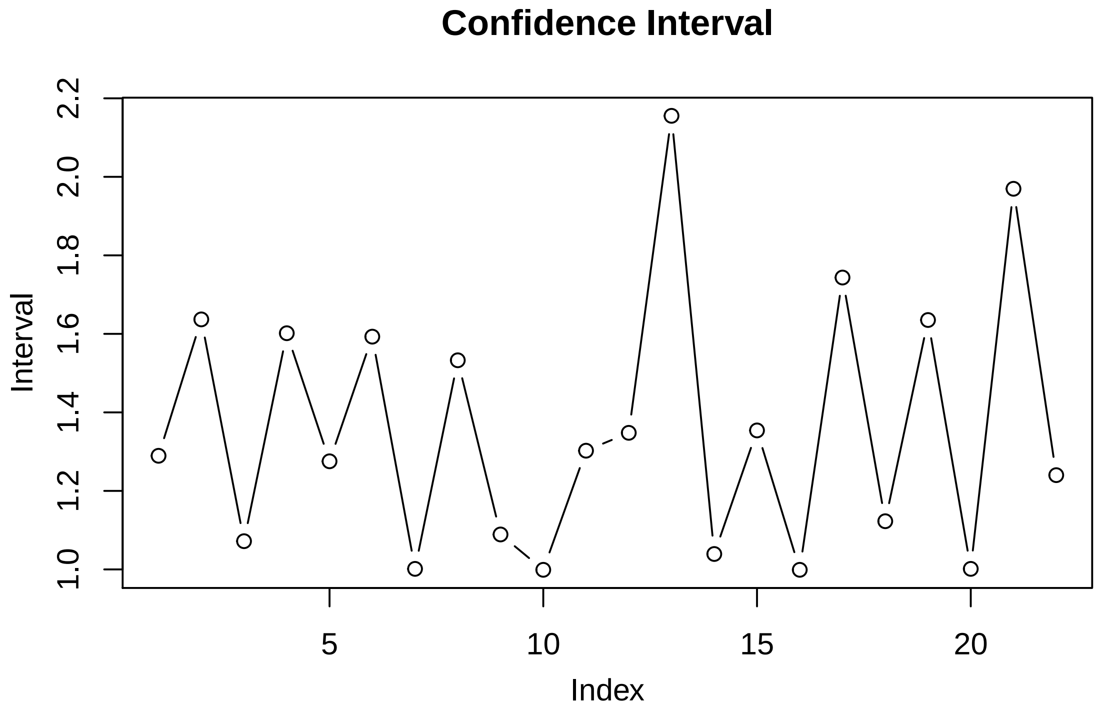
```
##
## Call:
## lm(formula = feed ~ weight, data = Training)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3360 -1.0954  0.2563  1.4102  2.1071
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.7551     0.2479  15.146   <2e-16 ***
## weight        0.3388     0.2505   1.353    0.183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.735 on 47 degrees of freedom
## Multiple R-squared:  0.03747,    Adjusted R-squared:  0.01699
## F-statistic: 1.829 on 1 and 47 DF,  p-value: 0.1827
```

Calculating the confidence interval of the generated model

```
# 7) Confidence Interval
  # 7.1) Printing Confidence Interval values
  print(Prediction$upr-Prediction$lwr)
```

```
##  [1] 1.2894438 1.6368668 1.0718653 1.6016437 1.2753927 1.5929124 1.0014088
##  [8] 1.5326840 1.0890268 0.9990063 1.3024192 1.3478944 2.1553602 1.0392681
## [15] 1.3540630 0.9990063 1.7434798 1.1227199 1.6352557 1.0014088 1.9695238
## [22] 1.2401226
```

```
  # 7.2) Plotting Confidence Interval
  plot(Prediction$upr-Prediction$lwr,main = "Confidence Interval",
  type = "b",ylab = "",xlab = "")
  # 7.3) Setting and adjusting labels of x and y axes
  title(ylab = "Interval",xlab="Index", line=2.25, cex.lab=1)
```



Conclusion

The *generated model isn't the best model for the prediction of feed supplements* as the **multiple R-squared and adjusted R-squared values are close to zero** instead of one. Also, the **p-value is significantly high,** whereas it should be as close to zero as possible. However, the **residual standard error is non-zero;** hence there is no overfitting.

We **reject our Null Hypothesis** as we can't observe any clear relationship between feed supplement type and chicken weight.