Business Analytics for OR : IEOR 4574

Shravan Kumar Chandrasekaran | sc3940

February 4th, 2016

Assignment #1

In this assignment we needed to analyze the Egg Production dataset. The data includes daily information about egg production (number of eggs laid per chicken), feed (amount of feed per chicken) and daily temperature (in degrees Celsius). We then interpreted the results based on questions asked.

1. Download EggProduction.csv to your computer. You can use setwd() to set the current working directory. Load the data and print a summary of the variables.
   a. **Code Block:**
   ```
   getwd()
   dir = "/Users/Shravan/BAOR/Assignment1"

   #Setting working directory
   setwd(dir)

   #Reading EggProduction Data
   Data<-read.csv("EggProduction.csv")
   #Summaziring the Data
   summary(Data)
   ```
   b. **Summary:**
   ```
   > summary(Data)
        eggs           feed        temperature
    Min.   :0.000   Min.   :18.36   Min.   :-12.61
    1st Qu.:1.418   1st Qu.:21.50   1st Qu.: 10.71
    Median :1.782   Median :22.27   Median : 21.76
    Mean   :1.773   Mean   :23.11   Mean   : 19.96
    3rd Qu.:2.174   3rd Qu.:23.30   3rd Qu.: 29.63
    Max.   :3.652   Max.   :32.60   Max.   : 48.12
   >
   ```
   Run a regression of eggs on feed and interpret the result. Does the result make sense to you?

2. Run a regression of eggs on feed and interpret the result. Does the result make sense to you?
   a. **Code Block:**
   ```
   #Running a regression of eggs on feed
   ModelEggs<-lm(Data$eggs ~ Data$feed, data = Data)
   summary(ModelEggs)
   ```

b. Regression Thesis:

Residuals:
```
    Min      1Q   Median      3Q     Max
-1.54185 -0.34831 -0.02782  0.36793  1.81521
```

Residuals:
```
    Min      1Q   Median      3Q     Max
-1.54185 -0.34831 -0.02782  0.36793  1.81521
```


The 5-point residual information gives us a quick snapshot of the data. Ideally, the median should be zero and the values should represent a normal distribution.
We can see that there aren't big outliers. Although the median is close to 0, the data is not normally distributed. It is roughly balanced equally between the median.

Model Summary

R-Squared	Adjusted R-Squared	Residual Std. Error	P-value
0.1755	**0.1755**	**0.5215**	**<2.2e-16**

Model Output
```
Call:
lm(formula = Data$eggs ~ Data$feed, data = Data)

Residuals:
    Min      1Q   Median      3Q     Max
-1.54185 -0.34831 -0.02782  0.36793  1.81521

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.832768   0.113951   33.63   <2e-16 ***
Data$feed   -0.089108   0.004897  -18.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5215 on 1550 degrees of freedom
Multiple R-squared:  0.176,	Adjusted R-squared:  0.1755
F-statistic: 331.1 on 1 and 1550 DF,  p-value: < 2.2e-16
```

R-Squared is the total variance in the data that is accounted for by the regression equation.
We can see that the p-value of feed is close to zero, hence the input affects the output.
Also, the Estimate for feed is negative, hence it is inversely correlated.
We can see that the adjusted R-Squared, that should be closer to 1 is 0.1755, although the P-value is very close to zero.
Hence the model doesn't make much sense.

3. Now include temperature in the model. Run a regression of eggs on feed and temperature and interpret the result. Does the result make sense to you?
   a. Code Block:

#Running a regression of eggs on feed and temperature
ModelEggsFeedTemp<-lm(Data$eggs ~ feed + temperature, data = Data)
summary(ModelEggsFeedTemp)

b. Regression Thesis:

Residuals:

```
     Min       1Q   Median       3Q      Max
-1.55172 -0.34901 -0.02884  0.36528  1.81519
```

The 5-point residual information gives us a quick snapshot of the data. Ideally, the median should be zero and the values should represent a normal distribution.
We can see that there aren't big outliers. Although the median is close to 0, the data is not normally distributed. It is roughly balanced equally between the median.

Model Summary

R-Squared	Adjusted R-Squared	Residual Std. Error	P-value
0.1762	**0.1751**	**0.5216**	**<2.2e-16**

Model Output

```
Call:
lm(formula = Data$eggs ~ feed + temperature, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55172 -0.34901 -0.02884  0.36528  1.81519

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.8448807  0.1160307  33.137   <2e-16 ***
feed        -0.0891043  0.0048985 -18.190   <2e-16 ***
temperature -0.0006112  0.0010969  -0.557    0.577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5216 on 1549 degrees of freedom
Multiple R-squared:  0.1762,	Adjusted R-squared:  0.1751
F-statistic: 165.6 on 2 and 1549 DF,  p-value: < 2.2e-16
```

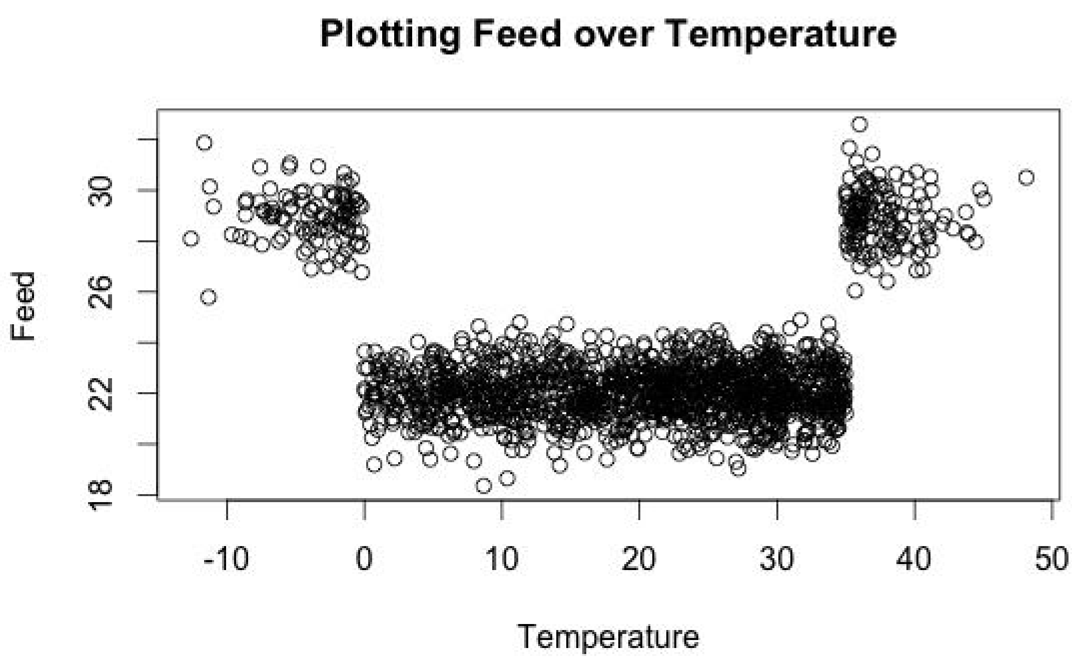
Adjusted R-Squared is the total variance in the data that is accounted for by the regression equation discounting additional variables.
The coefficients of both feed and temperature are negative and hence are inversely correlated with the egg production.
The p-value of temperature is significant, showing that the input is not affecting the output.
We can see that the adjusted R-Squared, that should be closer to 1 is still 0.1751, although the P-value is very close to zero.
Hence the model doesn't make much sense, although the model is slightly better than without including temperature variable.

4. Plot feed against temperature to understand how the temperature affects the amount of feed the chickens receive. From the graph you should be able to see that the amount of feed changes when temperature takes certain values. Create a new binary/discrete/categorical variable from the temperature column that captures this phenomenon. Print a summary of the new variable.

   a. **Code Block:**

```
#Plotting feed against temperature
plot(x = Data$temperature, y = Data$feed, xlab = "Temperature",
    ylab = "Feed",
    main = "Plotting Feed over Temperature")
#Adding Dummy Switch variable for Temperature
Dummyvar <- rep(0, length(Data$feed))
for (k in 1:length(Data$temperature))
{
  Dummyvar[k] <- as.numeric(Data$temperature[k] < 0 | Data$temperature[k] > 35)
}
col_names = c("eggs", "feed", "temperature", "Switch")
Data2 <- data.frame(Data$eggs, Data$feed, Data$temperature, Dummyvar)
colnames(Data2) <- col_names
summary(Dummyvar)
```

   b. **Plot:**



   c. **Summary:**

```
 summary(Dummyvar)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

0.0000 0.0000 0.0000 0.1559 0.0000 1.0000

5. Regress eggs on feed, temperature, and the new variable you created. Interpret the results.
   a. **Code Block:**
   #Running regression with new switch
   ModelEggswithSwitch <- lm(Data2$eggs ~ Data2$feed +
   Data2$temperature + factor(Data2$Switch), data = Data2)
   #Summary
   summary(ModelEggswithSwitch)

   b. Regression Thesis:

   Residuals:
   Min 1Q Median 3Q Max
   -1.56444 -0.34099 -0.00796 0.33876 1.74590

   The 5-point residual information gives us a quick snapshot of the data. Ideally, the median should be zero and the values should represent a normal distribution.
   We can see that there aren't big outliers. Although the median is close to 0, the data is not normally distributed. It is roughly balanced equally between the median.

   Model Summary

R-Squared	**Adjusted R-Squared**	**Residual Std. Error**	**P-value**
0.2355	**0.2341**	**0.5026**	**<2.2e-16**

   Model Output

```
Call:
lm(formula = Data2$eggs ~ Data2$feed + Data2$temperature + factor(Data2$Switch),
    data = Data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56444 -0.34099 -0.00796  0.33876  1.74590

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.0528838  0.2780797   3.786 0.000159 ***
Data2$feed            0.0387500  0.0125787   3.081 0.002102 **
Data2$temperature    -0.0007344  0.0010570  -0.695 0.487319
factor(Data2$Switch)1 -1.0276280  0.0937132 -10.966  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5026 on 1548 degrees of freedom
Multiple R-squared:  0.2355,   Adjusted R-squared:  0.2341
F-statistic:   159 on 3 and 1548 DF,  p-value: < 2.2e-16
```

   Adjusted R-Squared is the total variance in the data that is accounted for by the regression equation discounting additional variables.
   We can see that the adjusted R-Squared, that should be closer to 1 is now 0.2341, and the P-value is very close to zero.

This model is better than the previous two models as

- Adjusted R-Square is higher
- Residual Std. Error is lower
- P-Value is almost zero

6. Randomly and evenly divide the data into a training and test dataset. What is the best model you would use to predict egg production based on this dataset? Why?
    a. **Code Block:**

```
#Seperate dataset to training set and test set
trainRows = runif(nrow(Data2))>0.50
#Choose 75% of data as our training data
train = Data2[trainRows,]
#Make the training dataset
test = Data2[!trainRows,]
#Put rest of data into test set

#fit three linear regression models
fit1 = lm(eggs~feed+temperature+factor(Switch),data=train)
summary(fit1)

fit2 = lm(eggs~feed+temperature,data=train)
summary(fit2)

fit3 = lm(eggs~feed,data=train)
summary(fit3

#test our models on test dataset
test1.pred = predict(fit1,newdata=test)
test.eggs = test$eggs

tss = sum((test.eggs-mean(test.eggs))^2)
rss1 = sum((test.eggs-test1.pred)^2)
rsq1 = 1 - rss1/tss

test2.pred = predict(fit2,newdata=test)
rss2 = sum((test.eggs-test2.pred)^2)
rsq2 = 1 - rss2/tss
test3.pred = predict(fit3,newdata=test)
rss3 = sum((test.eggs-test3.pred)^2)
rsq3 = 1 - rss3/tss
```

    b. **Regression Thesis:**

Model Summary

MODEL 1 -> *eggs~feed*
MODEL 2 -> *eggs~feed+temperature*

MODEL 3 -> *eggs~feed+temperature+factor(Switch)*

	MODEL 1	MODEL 2	MODEL 3
Adjusted R-Squared	**0.166**	**0.166**	**0.243**

Adjusted R-Squared is the total variance in the data that is accounted for by the regression equation discounting additional variables.
We can see that the adjusted R-Squared is highest for MODEL 3. Hence we would choose the model where the independent parameters are

- Feed
- Temperature
- Switch for the temperature

7. For your best model, what is a 99% confidence interval for the regression coefficients. Interpret the results.
   a. **Code Block:**

```
#Confidence interval of 99% for regression coefficients
confint(ModelEggswithSwitch, level = 0.99)
```

   b. **Regression Thesis:**

Model Summary

```
                          confint(ModelEggswithSwitch, level = 0.99)
                          0.5 %       99.5 %
(Intercept)               0.335713842  1.770053737
Data2$feed                0.006309352  0.071190737
Data2$temperature        -0.003460443  0.001991715
factor(Data2$Switch)1 -1.269315197 -0.785940877
```

99% means that If the procedure is repeated a large no. of times, the coefficients and egg production would lie in the interval given above.

8. For your best model, what is a 90% prediction interval if the feed was 25 and the temperature was -1. Interpret the results.
   Code Block:

```
#Finding 90% prediction interval
predict(fit1, Mynewdata,interval = "predict", level = 0.9)
```

   a. **Regression Thesis:**

Model Summary

	fit	*lwr*	*upr*
1	*1.13001*	*0.2770754*	*1.982945*

When we run the model a large number of times, the model would predict the egg production between the interval 90% of the times.

9. Provide an intuitive explanation for what changed between the first two regressions and the last regression.

Between the two models and the third model, we are using a switch variable to predict the egg production. The temperature is not linearly related to the egg production
The switch variable is a binary variable that very aptly captures the extremities displayed by the temperature variable against the egg production.
Hence adding a switch variable makes the model better.