Name: Parth Pareek
UNI: PP2547
Date: 02/04/2016
Assignment: HW1

1. Summary of Egg Production 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
```

2. Regression of eggs on feed

```
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 ***
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
```

Interpretation
The p-value of the F-Statistic is <0.05 and hence, the model is relevant. The adjusted R-squared value is pretty low, and we can think of better models which is more effective. The p-value for the coefficient is also <0.05 indicating that the coefficient is relevant to the model

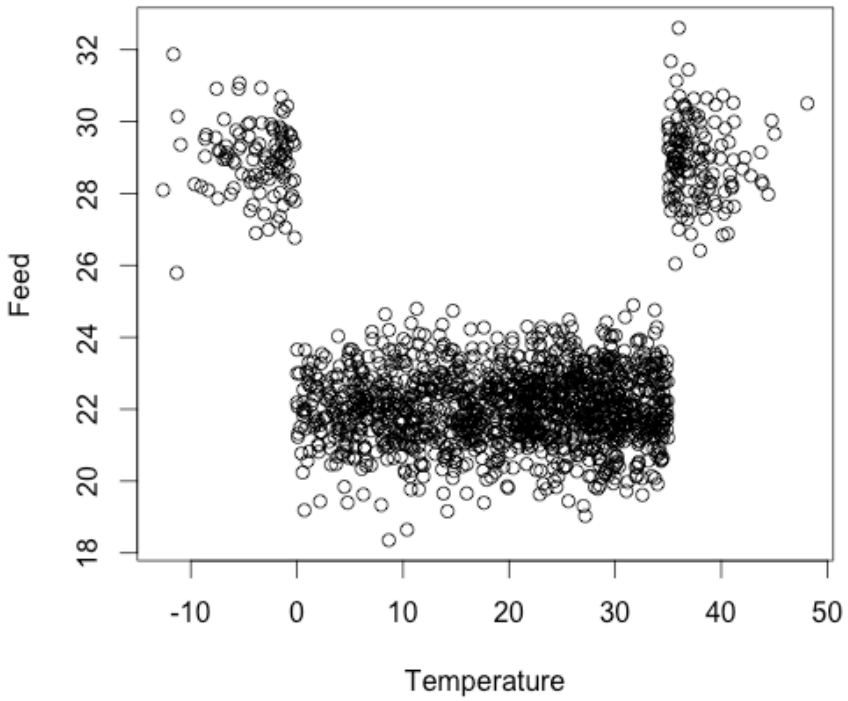
3. Regression of eggs on feed and temperature

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

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
```

Interpretation

The p-value of the F-Statistic is still <0.05 and hence, the model is relevant. The adjusted R-squared value is lower than in previous model. The p-value of temperature variable is >0.05 indicating that the variable is irrelevant for the model. The coefficients for feed is not very different from the first model and the coefficient for temperature is extremely low. RSE has also increased by very little suggesting we may revert to the old model.

4. Plot of Feed vs. Temperature



Summary of binary variable

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1701  0.0000  1.0000
```

5. Regression of eggs on feed, temperature and binary variable

```
Call:
lm(formula = eggs ~ feed + temperature + binary, data = EggProd)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55182 -0.33482 -0.00793  0.33607  1.75906

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.087861   0.219811   9.498   <2e-16 ***
feed        -0.006949   0.010024  -0.693   0.4883
temperature -0.001967   0.001078  -1.825   0.0682 .
binary      -0.673228   0.072255  -9.317   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5077 on 1548 degrees of freedom
Multiple R-squared:  0.2199,    Adjusted R-squared:  0.2184
F-statistic: 145.5 on 3 and 1548 DF,  p-value: < 2.2e-16
```

Interpretation

The p-value of the F-Statistic is continues to be <0.05 and hence, the model is relevant. The adjusted R-squared value is higher than in all previous models, which mean this model will be more effective than the previous ones. The addition of the binary variable renders the previous variables irrelevant; p-values of feed previous models.

6. We will check all the models on the training data and choose the one with highest R-squared value

Model 1: 0.148
Model 2: 0.148
Model 3: 0.192

Model 3 gives the highest R-squared value on the raining data set and is chosen to be the best model

7. 99% confidence interval for coefficients

```
                   0.5 %        99.5 %
(Intercept)  1.520966184  2.6547560512
feed        -0.032801391  0.0189033845
temperature -0.004745748  0.0008124497
binary      -0.859574751 -0.4868808202
```

Model 3 places little weightage on feed and temperature variables (evident in the summary in Qs. 5). The extremely small confidence intervals of feed and temperature confirms the interpretation of the summary of model 3. However, the coefficients of binary variable are fairly high indicating that number of eggs is *more dependent* on it.

8. Prediction at 90% confidence level

```
       fit       lwr      upr
1 1.242875 0.4034307 2.082319
```

Using model 3, the number of eggs is predicted to be 1.24 (best fit). With a 90% confidence, we can say that the range of eggs will be between 0.4 and 2.08

9. The plot of temperature vs. feed suggests that as the temperature increases, the feed reduces when the temperature is between 0 and 35 degrees. However, beyond 35 degrees, the feed increases again. Model 1 and Model 2 fail to capture this effect in number of eggs. When we create a binary variable, we are implicitly capturing the feed vs. temperature effect. This is expressed by a better R-squared and lower RSE value in the model summary.