PATRICIA KAYE T. DUMANDAN **23 JANUARY 2019**
MS RAPTOR BIOLOGY **EEB 622 HW 1**

Q1: MATH SCORES

```
> str(math)
'data.frame':  7 obs. of  2 variables:
 $ LSD_concentration: int  1 2 3 4 5 6 6
 $ MATH_score       : num  78.9 58.2 67.5 37.5 45.6 ...
> summary(math)
 LSD_concentration   MATH_score
 Min.   :1.000     Min.   :29.97
 1st Qu.:2.500     1st Qu.:35.20
 Median :4.000     Median :45.65
 Mean   :3.857     Mean   :50.09
 3rd Qu.:5.500     3rd Qu.:62.84
 Max.   :6.000     Max.   :78.93
> #test LM assumptions:
> #normality
> shapiro.test(math$MATH_score)
        Shapiro-Wilk normality test
data:  math$MATH_score
W = 0.92745, p-value = 0.5294          #data normally distributed

> #Homoscedasticity
> math$LSD_concentration=as.integer(math$LSD_concentration)
> fligner.test(math$MATH_score, math$LSD_concentration)
Fligner-Killeen test of homogeneity of variances
data:  math$MATH_score and math$LSD_concentration
Fligner-Killeen:med chi-squared = 6, df = 5,
p-value = 0.3062                       #variances are equal

> #model
> math.lm=lm(math$MATH_score~math$LSD_concentration)
> summary(math.lm)
Call:
lm(formula = math$MATH_score ~ math$LSD_concentration)

Residuals:
      1       2       3       4       5       6       7
  3.950  -8.067   9.915 -11.372   5.520   1.503  -1.448
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             83.692      7.381  11.339 9.33e-05 ***
math$LSD_concentration  -8.712      1.733  -5.028  0.00401 **
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.284 on 5 degrees of freedom
Multiple R-squared:  0.8349,   Adjusted R-squared:  0.8019
F-statistic: 25.28 on 1 and 5 DF,  p-value: 0.004008

> #plot####
> #eyeball method plot
> plot(math$MATH_score~math$LSD_concentration, ylab="math score (%)", xlab="
LSD concentration(mg)")
> curve(65+-5*x, a=T, col="green")
```

```
> #parameter estimates
> coef(math.lm)
           (Intercept) math$LSD_concentration
               83.6925                 -8.7125
> curve(83.6925+-8.7125*x, a=T, col="red")
> confint(math.lm)
                            2.5 %     97.5 %
(Intercept)              64.71972 102.665284
math$LSD_concentration -13.16679  -4.258213
```

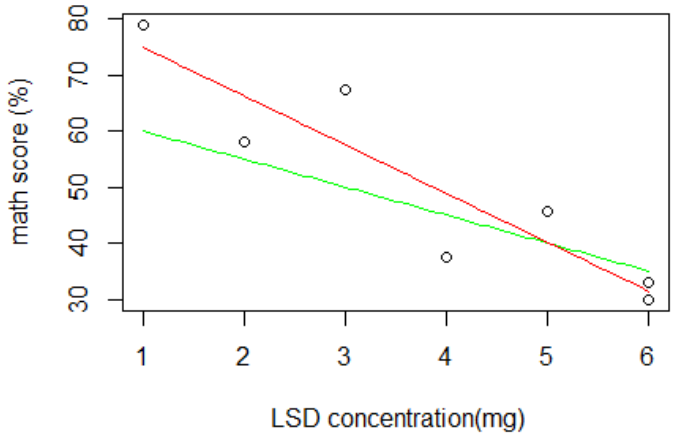

Fig.1. Effect of LSD concentration(mg) on student's math scores (%). Green line shows eyeball estimate of LSS and red line shows fit of calculated parameter estimates. At increasing LSD concentrations, math scores decline.

```
> #calculate R2####
> a=83.6925
> b=-8.7125
> y_hat=a+b*math$LSD_concentration
> RSS=sum((((y_hat))-(math$MATH_score))^2)
> TSS=sum(((math$MATH_score)-50.09)^2)
> r2=(1-RSS/TSS)
[1] 0.8348793
> #calculate Rmse####
> rmse=sqrt(mean((math$MATH_score-y_hat)^2))
 [1] 7.001543
> #calculate MAE####
> MATH.MAE=MAE(y= math$MATH_score, y_hat=83.6925+-8.7125*math$LSD_concentrati
on, n=7)
 [1] 5.967857
```

A. **LSD CONCENTRATION TO GET 85% AND ABOVE MATH SCORE**
   -Given the equation of the line: math score(%)= 83.692+ (-8.712) x, LSD concentration must be less than (-0.150) to get a math score 85% and higher.
B. **HOW WELL DOES OUR MODEL PREDICT EFFECT OF LSD CONC.ON MATH SCORES**
   -Model predicts approximately 83% of the variation in the data which is relatively high meaning it can effectively describe the patterns observed.

C. WHY NORMAL DIST.IS INAPPROPRIATE

-Normal distribution is inappropriate because of the relatively small sample size and given that this is based on real data which rarely follows the normal distribution, doing parametric tests based on normal distribution would be inappropriate.

Q2: MIRACLE FOOD

```
> str(mir)
'data.frame':  1000 obs. of  2 variables:
 $ Weight_loss: num  -0.89 6.31 -30.21 -6.28 11.38 ...
 $ pomegranate: int  2 2 3 7 4 2 3 3 5 5 ...
> summary(mir)
  Weight_loss       pomegranate
 Min.   :-36.240   Min.   :0.000
 1st Qu.: -8.570   1st Qu.:2.000
 Median : -1.650   Median :3.000
 Mean   : -1.724   Mean   :2.942
 3rd Qu.:  5.037   3rd Qu.:4.000
 Max.   : 32.890   Max.   :9.000
> #test assumptions
> #normality
> shapiro.test(mir$Weight_loss)
        Shapiro-Wilk normality test

data:  mir$Weight_loss
W = 0.99941, p-value = 0.9924          # data normally distributed

> #homscedasticity
> fligner.test(mir$Weight_loss, mir$pomegranate)
        Fligner-Killeen test of homogeneity of        variances

data:  mir$Weight_loss and mir$pomegranate
Fligner-Killeen:med chi-squared = 17.832, df
= 9, p-value = 0.03717                 #variances not equal

> #model
> mir.lm=lm(mir$Weight_loss~mir$pomegranate)
> summary(mir.lm)
Call:
lm(formula = mir$Weight_loss ~ mir$pomegranate)
Residuals:
    Min      1Q  Median      3Q     Max
-33.435  -6.780  -0.041   6.807  35.169
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.1790     0.6268  -0.286  0.77528
mir$pomegranate  -0.5251     0.1841  -2.852  0.00444 **
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.971 on 998 degrees of freedom
Multiple R-squared:  0.008084, Adjusted R-squared:  0.00709
F-statistic: 8.133 on 1 and 998 DF,  p-value: 0.004435
> #parameter estimates
> coef(mir.lm)
    (Intercept) mir$pomegranate
     -0.1789802      -0.5251053
```

```
> #confinterval
> confint(mir.lm)
                    2.5 %     97.5 %
(Intercept)     -1.408937  1.0509767
mir$pomegranate -0.886420 -0.1637906
> #plot
> plot(mir$weight_loss~mir$pomegranate, ylab=" Weight loss (kg)", xlab= "No.
of Pomegranates")
> a2=-0.1789802
> b2=-0.5251053
> curve(a2+b2*x, add=T, col= "blue", lwd=2)
```

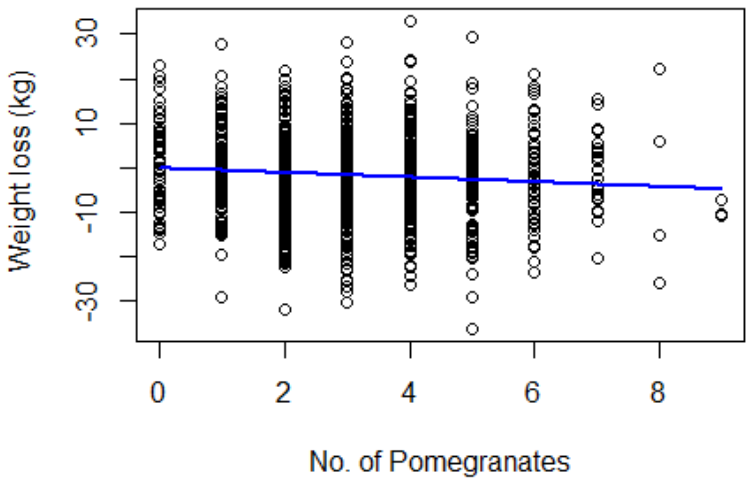

Fig.2. Effect of pomegranate consumption on weight loss. Blue line shows best fit of LSS suggesting weight loss is highest at 3 pomegranates consumed.

```
> y_hat1=a2+b2*mir$pomegranate
> RSS2=sum((((y_hat1))-(mir$weight_loss))^2)
> TSS2=sum(((mir$weight_loss)-(-1.724))^2)
> r2.2=(1-RSS2/TSS2)
 [1] 0.008083812

> #CALCULATE RMSE####
> rmse2=sqrt(mean((mir$weight_loss-y_hat1)^2))
 [1] 9.961044
> MIR.MAE=MAE(y= mir$weight_loss, y_hat=-0.1789802 +-0.5251053*mir$pomegranat
e, n=1000)
> MIR.MAE
[1] 7.981461
```

A. Despite a low p value that may suggest significant effect of pomegranate consumption on weight loss, I will not support/ believe the claim that it works because of the low R2 value which suggests that only 0.08% of the noise/ variation is explained by the model. The significance in p value might have just been brought about by the huge sample size.

Q3: RMSE, R2, MAE

A. Code for MAE

MAE= function (y, y_hat, n) {(sum(abs(y-y_hat))/n)}

B. Comparison among RMSE, R2, MAE

Metrics for model evaluation (REGRESSION ERROR)	Q1 (Math scores)	Q2 (Miracle food)
RMSE	7.001543	9.961044
R2	0.8348793	0.008083812
MAE	5.967857	7.981461

In question 1, R2 was higher suggesting good predictive capability of the model and for question 2, the opposite was shown (low R2). An inverse relationship between RMSE and MAE to R2 was observed such that higher R2 value showed lower RMSE and MAE values. Because RMSE gives extra weight to large errors, then maybe that is why the values are larger than the MAE which gives equal weight to all errors. Thus, low R2 and high RMSE and MAE may mean low quality of the model. Between the 2 questions, the model in question 1 seemed more plausible given the values of the different metrics for model fit.

Q4: DATA SIMULATION

How does air temperature influence raptor flights?

HO: proportion of raptor (birds/hour) flights increases with air temperature.

```
> airtemp=runif(100, min=0, max=50)
> slope= 7.5
> intercept= 10
> bph=rnorm(n=100, mean= intercept+slope*airtemp, sd=15)
> plot(bph~airtemp, ylab=" proportion of flights (birds per hour)", xlab=" ai
r temperature (Celsius)")
> a=13.104707
> b= 7.465501
> curve(a+b*x, a=T, col= "violet", lwd=2)
> #model
> q4.lm=lm(bph~airtemp)
> coef(q4.lm)
(Intercept)     airtemp
  11.925282    7.390855
```

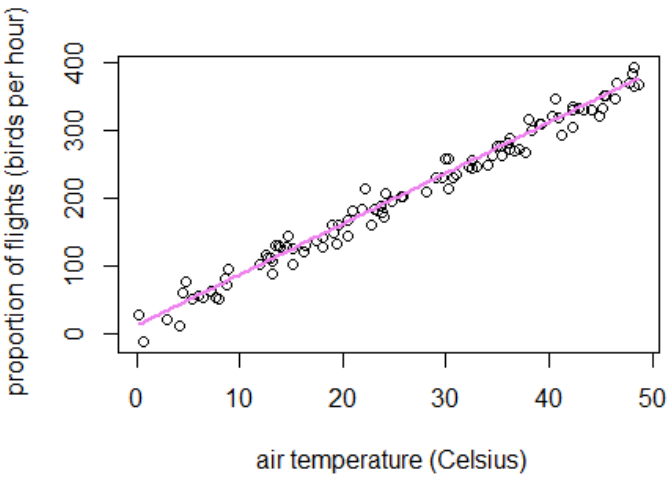

Fig.3. Effect of air temperature on proportion of raptor flights. Violet line shows line of best fit. Increasing air temperature results to higher proportion of flights.

A. Parameter estimates (based on regression) were close/ similar to the values I have set beforehand.

Q5: HETEROSCEDASTICITY

```
> airtemp=runif(100, min=0, max=50)
> slope= 7.5
> intercept= 10
> bph=rnorm(n=100, mean= intercept+slope*airtemp, sd)
> sd=airtemp*2
> plot(bph~airtemp, ylab=" proportion of flights (birds per hour)", xlab=" ai
r temperature (Celsius)")
> q4.lm=lm(bph~airtemp)
> coef(q4.lm)
(Intercept)     airtemp
  13.582728    7.312803
> c=18.814190
> d=7.056229
> curve(c+d*x, a=T, col= "pink", lwd=2)
```

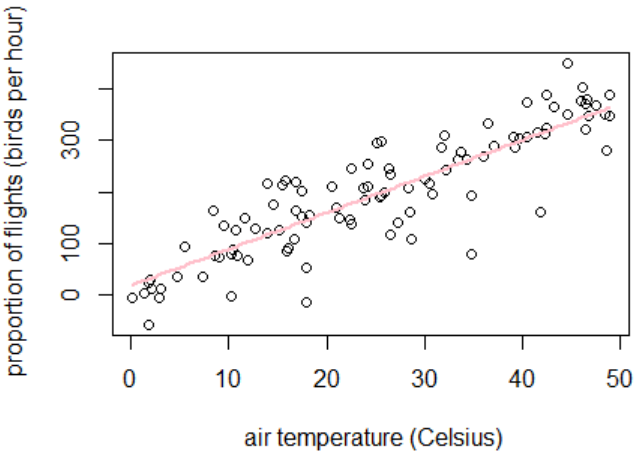

Fig.4. Effect of air temperature on proportion of raptor flights. Violet line shows line of best fit. Increasing air temperature results to higher proportion of flights. At increasing air temperature, variation in flight proportions also increased.

B. At increasing air temperature, variation also increased which may be due to increased human error. It is possible that proportion of flights also increased but the counts were more variable. This may be attributed to the reduced visibility of raptor flights as thermals become taller (which is directly correlated to increasing air temperature). It may also be due to species-specific responses to increasing air temperature.