Megan Kelly-Slatten

Stats

1/22/19

Homework 1

**Question 1**

|  |
| --- |
| > math<-read.csv("R/math\_scores.csv")  > str(math)  'data.frame': 7 obs. of 2 variables:  $ LSD\_concentration: num 1.17 2.97 3.26 4.69 5.83 6 6.41  $ MATH\_score : num 78.9 58.2 67.5 37.5 45.6 ...  > summary(math)  LSD\_concentration MATH\_score  Min. :1.170 Min. :29.97  1st Qu.:3.115 1st Qu.:35.20  Median :4.690 Median :45.65  Mean :4.333 Mean :50.09  3rd Qu.:5.915 3rd Qu.:62.84  Max. :6.410 Max. :78.93  > plot(math$MATH\_score~math$LSD\_concentration)  > mathmod<-lm(math$MATH\_score~math$LSD\_concentration)  > coef(mathmod)  (Intercept) math$LSD\_concentration  89.123874 -9.009466  > curve(89.123874+-9.009466\*x, add = T, col="red")  C:\Users\megankelly-slatten\Documents\quest1.png  > confint(mathmod)  2.5 % 97.5 %  (Intercept) 71.00758 107.240169  math$LSD\_concentration -12.87325 -5.145685  > a<-89.123874  > b<--9.009466  > yhat<-a+b\*math$LSD\_concentration  > SS<-sum((math$MATH\_score-50.09)^2)  > RSS<-sum((math$MATH\_score-yhat)^2)  > 1-(RSS/SS)  [1] 0.877835 |

1. To ensure a test score of over 85% you need a dosage of LSD that is equal or less than 0.4576.

85=89.123+-9.009(X)…. X=0.4576

1. The LSD dosage predicts math scores fairly well with a R2=0.877.
2. The normal distribution may be inappropriate because the sample size is so small that it is difficult to determine if the data has a normal structure.

**Question 2**

> food<-read.csv("R/miracle\_food.csv")

> str(food)

'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(food)

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

> plot(food$Weight\_loss~food$pomegranate)

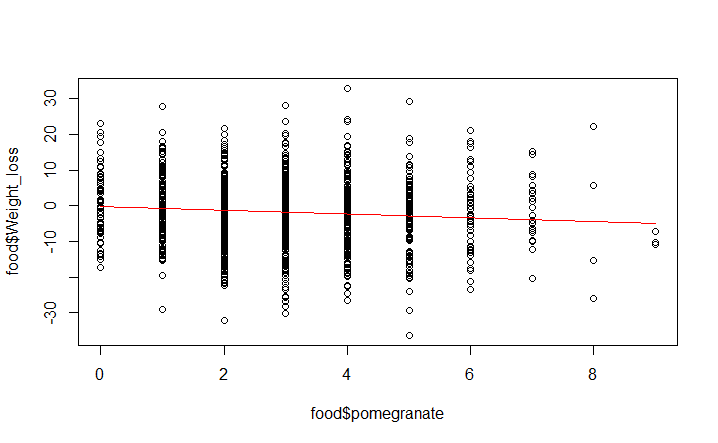
> foodmod<-lm(food$Weight\_loss~food$pomegranate)

> coef(foodmod)

(Intercept) food$pomegranate

-0.1789802 -0.5251053

> curve(-0.17898+-0.52510\*x, add = T, col="red")



> confint(foodmod)

2.5 % 97.5 %

(Intercept) -1.408937 1.0509767

food$pomegranate -0.886420 -0.1637906

> m<--0.17898

> n<--0.52510

> yhat<-m+n\*food$pomegranate

> SS<-sum((food$Weight\_loss--1.724)^2)

> RSS<-sum((food$Weight\_loss-yhat)^2)

> 1-(RSS/SS)

[1] 0.008083812

I do not agree with the claim that pomegranates helps with weight loss. The extremely low R2 value (0.008) tells me that the model does a horrible job of predicting the data. Therefore, the amount of pomegranates eaten does not accurately predict weight loss.

**Question 3**

**Function:**

MAE<-function(y, yhat, n) {(1/n)\*sum(abs(y-yhat))}

**Math data:**

RMSE:

> sqrt(mean((math$MATH\_score-yhat)^2))

[1] 6.022355

R2: 0.877835

MAE:

>MAE(y=math$MATH\_score, yhat=89.123874+-9.009466\*math$LSD\_concentration, n=7)

[1] 4.890145

**Food data:**

RMSE:

> sqrt(mean((food$Weight\_loss-yhat)^2))

[1] 9.961044

R2: 0.008083812

MAE:

>MAE(y=food$Weight\_loss, yhat=-0.17898+-0.52510\*food$pomegranate, n=1000)

[1] 7.981461

The food data has a very high RMSE especially when you look at how small the slope is. Taking the MAE decreases the error a little bit, but still shows high variation from predicted points to actually observed data.

**Question 4**

How does the level of oxygen in the water effect number of salmon offspring.

> oxygen<-runif(50, min=0, max=100)

> slope<-1.5

> intercept<-27

> babynumba<-rnorm(n=50, mean=intercept+slope\*oxygen, sd=11)

> oxygen

[1] 91.513178 82.921415 14.025232 6.186944 60.339769 30.766070 56.600732 94.622946 88.660579 61.714800 64.949382

[12] 77.794945 28.458825 54.485391 12.010367 68.340633 58.992279 32.002447 63.850225 28.016477 47.635212 54.331951

[23] 79.004674 87.721825 76.860253 71.680344 3.544301 80.895438 90.602979 37.831771 19.394107 19.524388 70.366727

[34] 76.530286 31.317946 91.244335 79.575077 42.296468 45.402274 42.749988 32.955032 72.535210 60.617658 59.452078

[45] 49.085682 24.537979 13.175931 96.371900 75.117814 61.982155

> babynumba

[1] 164.01754 161.88521 44.67567 42.00623 138.08213 69.72432 123.43693 168.28017 139.75281 128.46999 118.87943

[12] 142.63865 57.61675 128.74246 42.01822 123.67976 113.54334 75.67099 112.91671 90.27277 102.92433 109.38802

[23] 136.93224 160.13431 158.11499 145.49042 29.45521 138.03383 171.28849 62.59388 54.83637 46.22218 121.86613

[34] 134.95248 72.50642 153.94686 147.73836 79.71988 102.19336 97.03057 89.45860 143.55504 135.03310 107.40421

[45] 88.69211 69.46224 40.34806 156.28155 137.38401 112.44494

> plot(babynumba~oxygen)

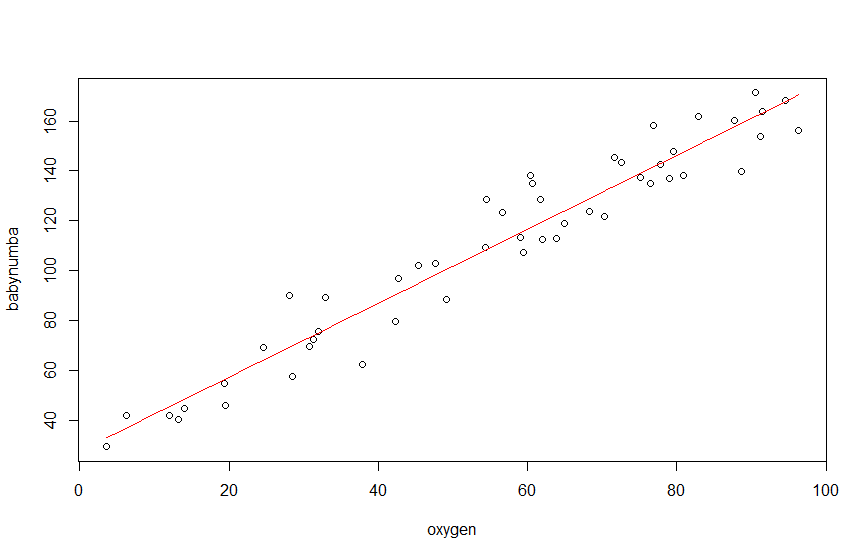
> model<-lm(babynumba~oxygen)

> coef(model)

(Intercept) oxygen

27.829322 1.479927

> curve(27.829322+1.4799\*x, add=T, col="red")



My estimates are extremely close to my set parameters.

**Question 5**

> oxygen<-runif(50, min=0, max=100)

>

> slope<-1.5

> intercept<-27

**> sd<-(oxygen\*2)**

>

> babynumba<-rnorm(n=50, mean=intercept+slope\*oxygen, sd)

>

> oxygen

[1] 68.5905026 18.2767907 79.7402110 53.7840533 36.9795841 93.9449254 76.2368807 42.2443310 43.0241601 85.8370586

[11] 37.1110472 52.1868965 45.3243309 14.7441561 21.2900886 99.0807190 42.4547376 80.6264146 10.3413749 84.1171689

[21] 49.5465235 47.5817754 47.0207197 97.9896036 27.5865719 78.3505681 36.1298476 78.8850430 32.0332619 0.5884682

[31] 63.9368161 90.3575582 37.5027739 5.7579422 24.0502437 36.5825688 72.1248978 2.5668381 29.8082433 84.9902869

[41] 6.6890839 86.8893877 77.5981656 48.6108971 49.6832243 32.9304707 53.1817149 28.4625526 39.7312838 37.3026041

> babynumba

[1] 242.609474 -69.033102 50.335957 142.248821 52.385574 -15.586052 38.541274 153.372738 80.695223 -35.286654

[11] 26.458508 76.367237 166.549167 40.932470 19.655087 166.061261 429.383611 77.243707 39.714623 155.213390

[21] 99.581220 141.639133 -53.788295 126.914196 -1.053097 -30.844851 47.241508 93.007863 18.522905 28.855519

[31] 282.844938 59.484262 2.816173 33.980686 91.993027 -44.874966 297.532317 35.268772 58.287985 397.720713

[41] 32.099104 8.766778 99.406381 12.941841 117.059179 -14.333822 152.516832 26.776455 68.007161 69.205766

>

> plot(babynumba~oxygen)

>

> model<-lm(babynumba~oxygen)

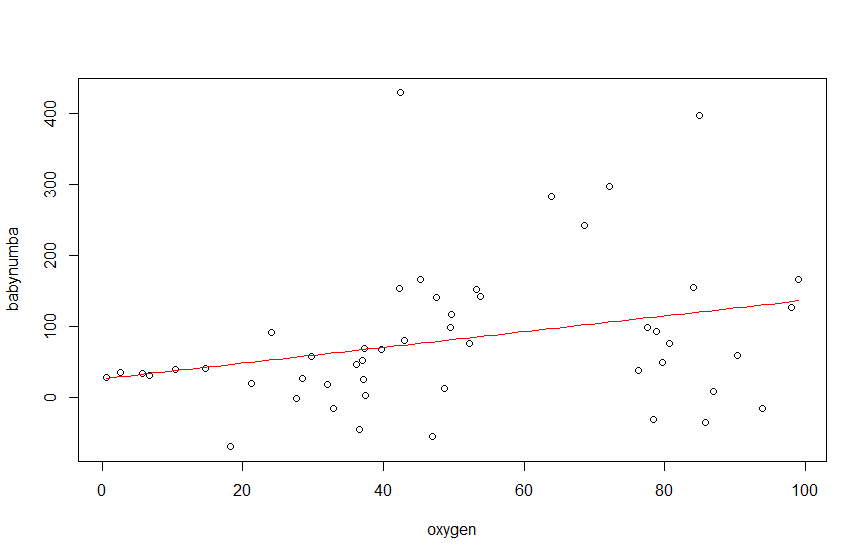
>

> coef(model)

(Intercept) oxygen

26.776073 1.106902

> curve(26.776073+1.106902\*x, add=T, col="red")



Lower oxygen levels seem to predict lower levels of salmon offspring. However, as oxygen levels increase this can either greatly reduce or greatly increase the number of salmon offspring. There appears to be something else interacting with the high level of oxygen to create this large variation in data.