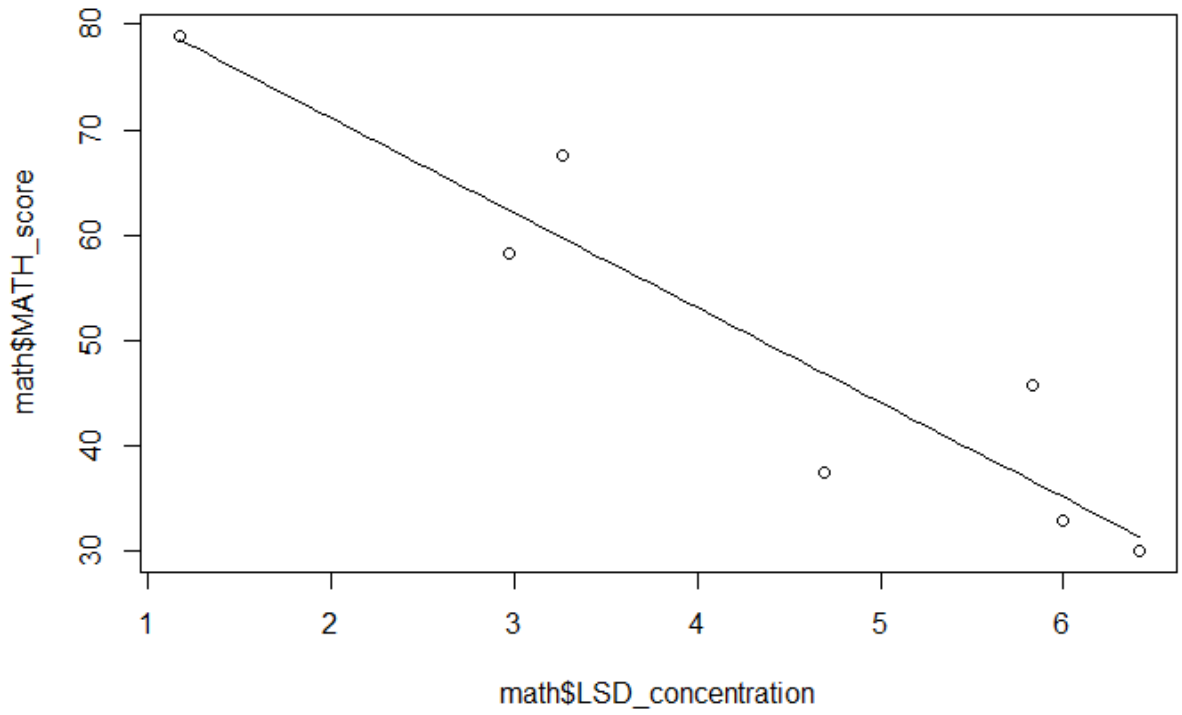
Louis Jochems

Professor Caughlin

22 January 2019 Homework 1

Question 1 a)



b) coef(modelq1)

	(Intercept)	math$LSD_concentration
	89.123874	-9.009466

confint(modelq1)

	2.5 %	97.5 %
(Intercept)	71.00758	107.240169
math$LSD_concentration	-12.87325	-5.145685

c) r2(y_hat,y)

[1] 0.877835

rmse(y_hat,y)

[1] 6.022355

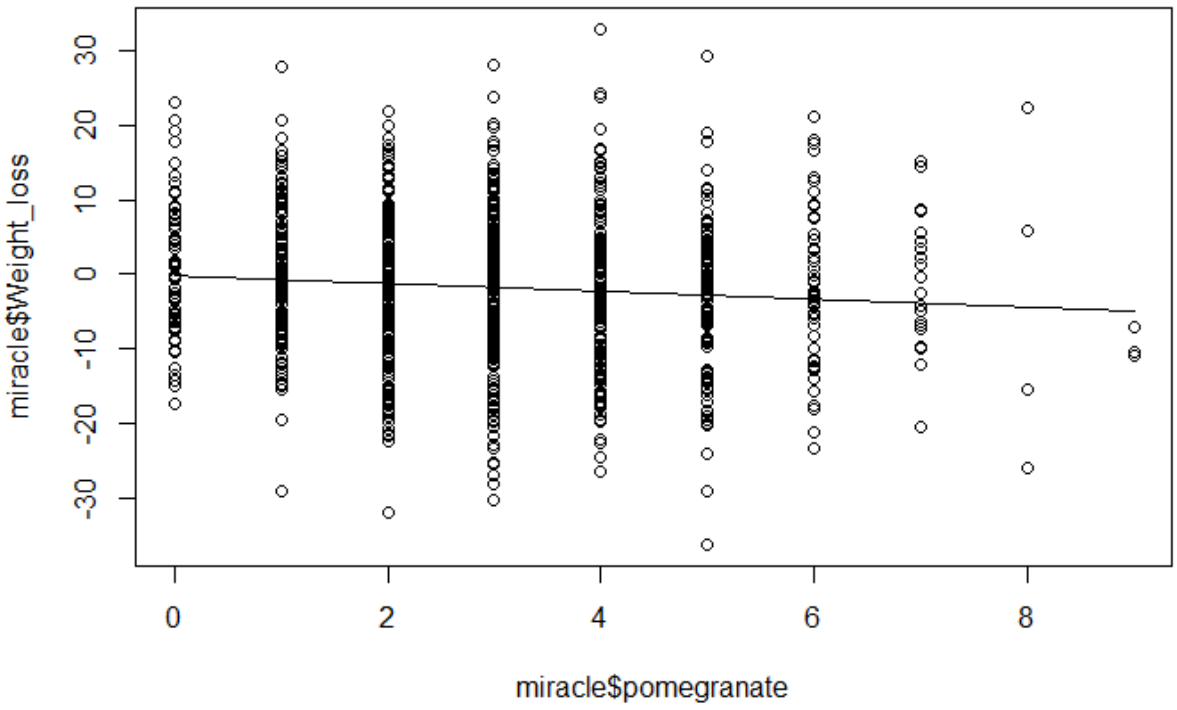
A) 85 = 89.123874+ -9.009466*x
-4.123874 = -9.009466*x
x= 0.458 ug/kg

Based on the parameter estimates of this regression model, there needs to be a max level of 0.458 ug/kg tissue of LSD to ensure a score of >85%.

B) Based on these data, his model seems show that LSD concentration predicts Math Score fairly well. There is a high R^2 and then sig p value, and the range of the confidence inte rvals is large and far from zero. However, this indicates a low sample size (n=7).

C) However, a normal distribution might be inappropriate to model these data because of a small sample size and considerably high variance.

Question 2 a)



b)
```
coef(modelq2)
        (Intercept) miracle$pomegranate
         -0.1789802          -0.5251053
 confint(modelq2)
                          2.5 %     97.5 %
(Intercept)          -1.408937  1.0509767
miracle$pomegranate -0.886420 -0.1637906
```

c) R
```
r2(y_hat,y)
[1] 0.008083812

rmse(y_hat,y)
[1] 9.961044
```

Disagree. R squared is low (and plus a poor predictor of fit anyways), there is high variance (both explained and unexplained) in the data, residual standard error is high, and CI intervals are only slightly negative.

Question 3 A.

```
> n_math=nrow(math)
> mae=function(y_hat,y) {
   ABS<-(sum(abs((y)-(y_hat))))
   n <- n_math
   return(ABS/n)
   }


> n_miracle=nrow(miracle)
> mae=function(y_hat,y) {
   ABS<-(sum(abs((y)-(y_hat))))
   n <- n_miracle
   return(ABS/n)
   }
```

B.

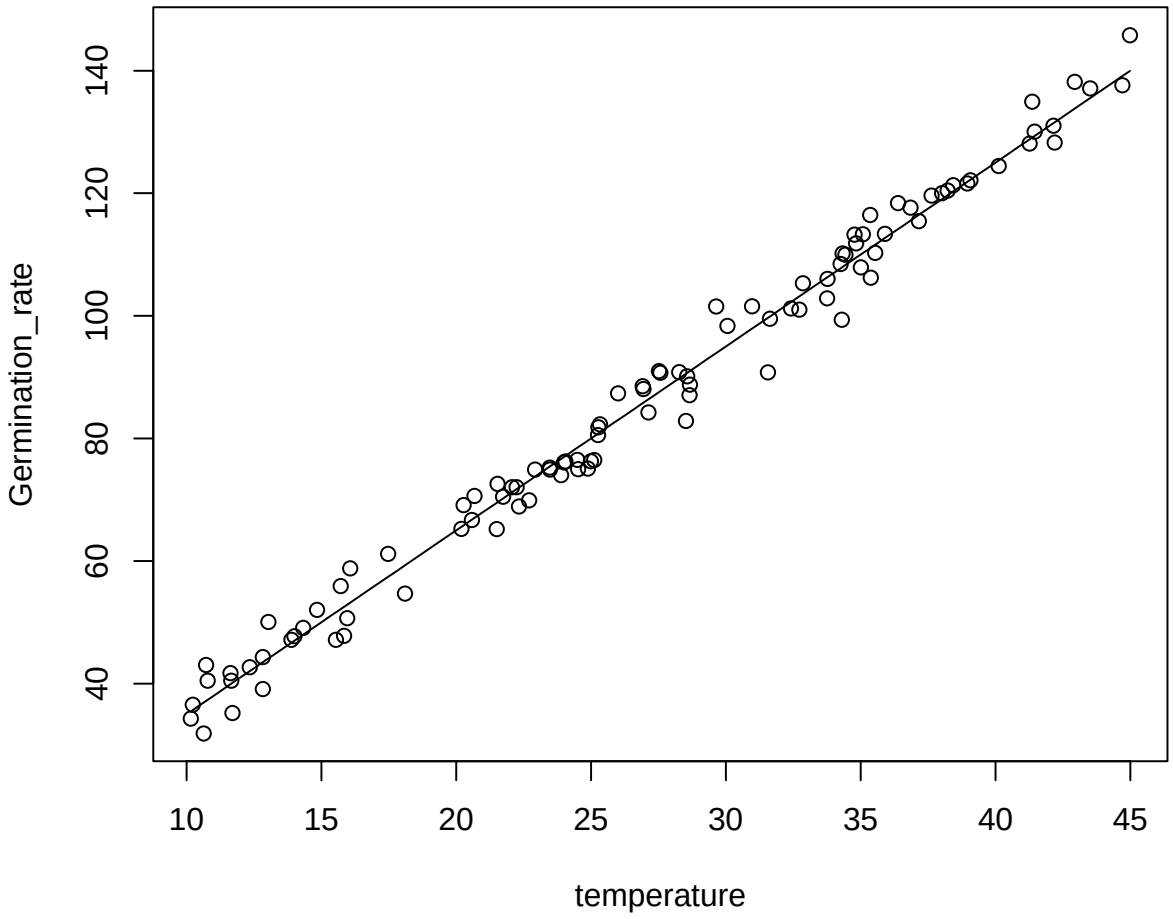
Metric	Math	Miracle
RMSE	6.022355	9.961044
R2	0.877835	0.008083812
MAE	4.980145	7.981461

Mean absolute error for both models shows that it gives about equal weight to all errors compared to RMSE that is higher for both models because of given weight to larger errors, especially with the case of the Miracle model. R2 is a lot smaller because its range is only between 0 and 1 because it represents only explained variance but not prediction error.

Question 4

1. temperature <- runif(100, min=10,max=45)
2. slope <- 3
   intercept <- 5
   sd <- 3
3. Germination_rate <- rnorm(mean=intercept+slope*temperature, n=100, sd=3)

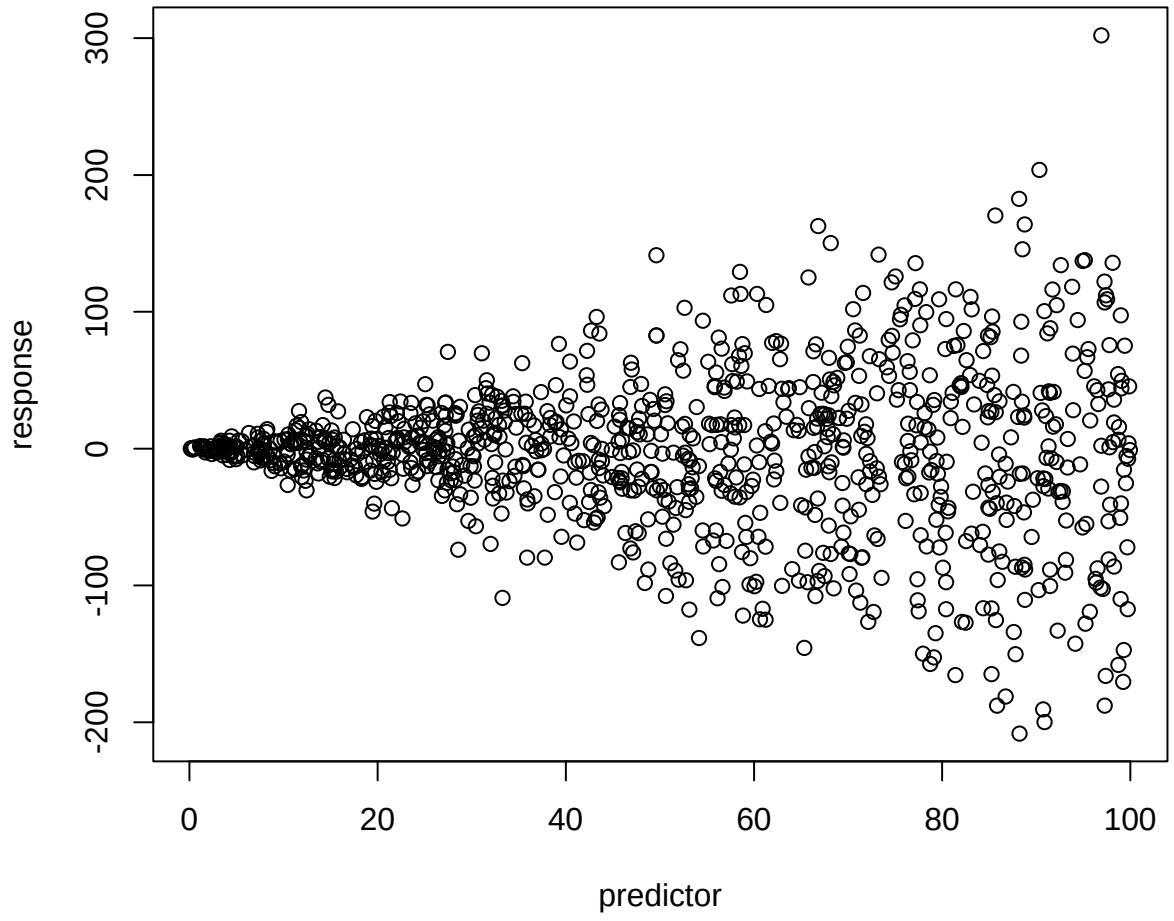
A.



B. modelq4=lm(Germination_rate~temperature)
coef(modelq4)
(Intercept) temperature
4.528822 3.021530

C. The slope is similar to the slope I set and the intercept is about 0.5 off of my parameter estimate, likely due to the considerable Residual Std. Error of about 3.3 (I set a sigma of 3).

Question 5
A.



B. A particular biological phenomena that could explain these data is the increasing range extent of a plant species results in the increase in variance of a particular trait, such as sap content of a tree species. Among a population within a given area, the variation of sap content might have considerably less variation as a selective pressure may particular narrow range of sap content phenotypes among those individuals. When you increase the range of the species in question (especially if it's a generalist species) then you may observe more variation in sap content due to more likely variation in microhabitats, or selective pressures, that could result in a high variation of sap content for this species as a whole. There tend to be many spatial and temporal (autocorrelation) examples wherein heteroscedasticity increases as a function of increase of the predictor variable.