HW 1 Edward Trout

### Question 1: LSD and Math Score Data

# read in data for math scores and LSD concentration  
lsd <- read.file("lsd.txt")

## Data from the .txt file lsd.txt has been loaded.

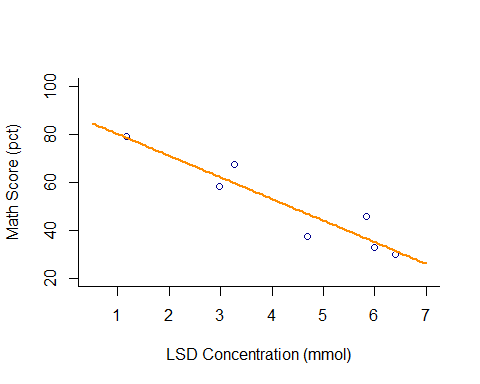
# build model  
lsd.mod <- lm(lsd$MATH\_score~lsd$LSD\_concentration)  
  
#return parameters and confidence intervals  
coef(lsd.mod)

## (Intercept) lsd$LSD\_concentration   
## 89.123874 -9.009466

confint(lsd.mod, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 71.00758 107.240169  
## lsd$LSD\_concentration -12.87325 -5.145685

#### a) Scatterplot



#### b) Parameter Coefficients and Confidence Intervals

Intercept: 89.124 [71.008, 107.240]  
Slope: -9.009 [-12.873,-5.146]  
Variance: 7.126 (residual standard error)

# make objects of parameters  
a <- 89.124  
b <- -9.009  
  
y.hat <- a + b\*(lsd$LSD\_concentration)  
  
#R-squared function  
  
r2<-function(y\_hat,y) {   
 RSS<-sum((((y\_hat))-(y))^2)   
 TSS<-sum(((y)-(mean(y)))^2)   
 return((TSS-RSS)/TSS)}   
  
#RMSE function  
  
rmse=function(y\_hat,y) {   
 return(sqrt(mean((y-y\_hat)^2)))  
}  
  
# return R squared and RMSE  
  
r2(y.hat, lsd$MATH\_score)

## [1] 0.877835

rmse(y.hat, lsd$MATH\_score)

## [1] 6.022356

#### c) Metrics of Model Fit

R2: 0.877  
RMSE: 6.022

### A. What level of LSD tissue concentration do you need to ensure a test score of >85%?

*If it is to be assumed that LSD concentration in the bloodstream is completely explanatory to test scores, these data would predict a level of .458 or lower to score above 85% on the math test.*

### B. How well does LSD tissue concentration predict test performance?

*The model for LSD has a .877 R2 value, which is a good predictive strength, within the bounds of the given LSD concentrations. This value would not be useful for ascertaining the correctness of predictions outside of the experimental LSD values*

### C. Why might the normal distribution be inappropriate to model these data?

*The normal distribution is best used on data that are unlimited and continuous. Math scores, while mostly continuous, are strictly bounded between 0 and 100.*

## Question 2: Miracle Weight Loss Pomegranates

# read in pomegranate weight loss data  
pom <- read.csv("miracle\_food.csv")  
head(pom)

## Weight\_loss pomegranate  
## 1 -0.89 2  
## 2 6.31 2  
## 3 -30.21 3  
## 4 -6.28 7  
## 5 11.38 4  
## 6 1.67 2

#build and inspect model  
pom.mod <- lm(pom$Weight\_loss ~ pom$pomegranate)  
coef(pom.mod)

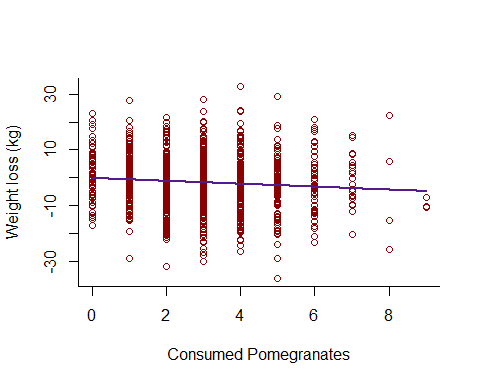
## (Intercept) pom$pomegranate   
## -0.1789802 -0.5251053

confint(pom.mod, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -1.408937 1.0509767  
## pom$pomegranate -0.886420 -0.1637906

#### a) Scatterplot

# scatterplot  
plot(pom$Weight\_loss~pom$pomegranate, col = "darkred", bty = "l",  
 xlab = "Consumed Pomegranates",  
 ylab = "Weight loss (kg)")  
curve(-.1790 -.5251\*x, add = T, col = "purple4", lwd = 2)



#### b) Parameter Coefficients and Confidence Intervals

Intercept: -0.179 [-1.409, 1.051]  
Slope: -0.525 [-0.886,-0.164]  
Variance: 9.971 (residual standard error)

# make objects of parameters  
c <- -.1790  
d <- -.5251  
  
y.hat2 <- c + d\*(pom$pomegranate)  
  
# return r2 and rmse  
r2(y.hat2, pom$Weight\_loss)

## [1] 0.008083812

rmse(y.hat2, pom$Weight\_loss)

## [1] 9.961044

#### c) Metrics of Model Fit

R2: 0.008  
RMSE: 9.961

### Given the significant p-value for the effect of pomegranates on weight loss, do you agree or disagree with the miracle claim?

*I would disagree, while as there is a significant effect of pomegranates on weight loss, the effect size itself is very small. The slope of the model line is around a half pound of weight loss per pomegranate over three months. Realistic weight loss goals (say 10–20 lbs) in that time span would require someone to eat 20-40 pomegranates per day (at which point the caloric intake itself might rule out any weight-loss power the pomegranates* do\* have). Additionally, the R2 value is incredibly low, this model does a terrible job at explaining the data. p-value isn’t everything!\*

## Question 3: Mean Absolute Error

### A. Translate the mathematical equation for MAE into a function in R.

# MAE function  
mae <- function(y\_hat,y) {  
 return(mean(sqrt((y-y\_hat)^2)))  
}

### B. Compare RMSE, R2, and MAE for the linear models in questions 1 and 2. How do these metrics of model fit differ?

##### **LSD data**

|  |  |
| --- | --- |
| Metric | Value |
| R2 | .878 |
| RMSE | 6.02 |
| MAE | 4.89 |

##### **Pomegranate data**

|  |  |
| --- | --- |
| Metric | Value |
| R2 | .008 |
| RMSE | 9.96 |
| MAE | 7.98 |

*R2 is different from both RMSE and MAE in that it is on a scale of 0-1 measuring how well the data are explained by the model. It is unitless. RMSE and MAE are both measured in the unit of the response or dependent variable, and give the average error in any of the response values given the explanatory power of the model. RMSE and MAE differ from each other in that MAE will always be lower in value and not give extra weight to more extreme distances, making it more even when trying to predict response variable values from experimental input variables.*

## Question 4: Simulating Datasets

#### Step 1- Create a Predictor Variable using runif or seq

# build predictor variable and simulated parameters  
predictor <- runif(1000)

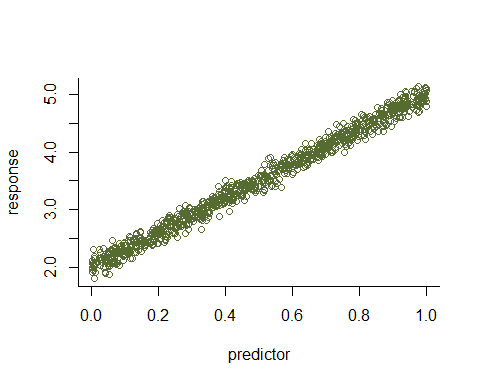
#### Step 2- Decide on a Value for the Intercept and Slope

intercept <- 2.0  
slope <- 3.0  
sigma <- .1

#### Step 3- Use rnorm to simulate draws from a normal distribution for your dataset.

# simulate response variable  
response <- rnorm(1000, mean = intercept + slope\*predictor, sd = sigma)

### A. Plot your Data



### B. Estimate Slope and Intercept Parameters from the data using Linear Regression

# check parameters for return data  
rand.mod <- lm(response~predictor)  
coef(rand.mod)

## (Intercept) predictor   
## 2.001063 2.999974

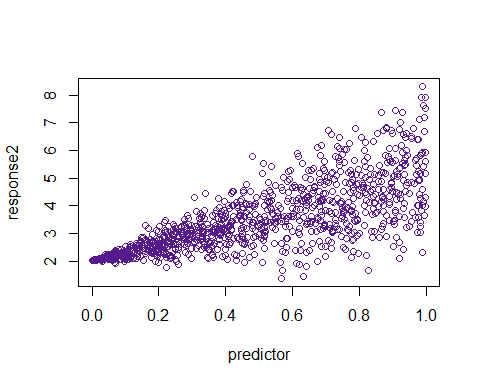
### C. How do your estimates compare to the true values you came up with in step 2?

*The parameters do indeed closely resemble the true values given, mostly due to the fact that the included variance was low (.1).*

## Question 5: Simulated Heteroscedasticity

# create increasing variance function  
var.obj <- 1.5\*predictor  
  
# generate new response to increasing variance model  
response2 <- rnorm(1000, mean = intercept + slope\*predictor, sd = var.obj)

### A. Construct a plot of simulated data



### B. What is a potential biological explanation for the data you have simulated?

*One example of seeing this kind of trend might be in animal limb length and home range size across all species. As a species limbs become longer, they may vary their home range sizes more, some may have large due to capability of movement, some may stay small due to geographic constraints, whereas species with smaller limb sizes will have a constant home range size regardless of geographic constraints.*