**Homework 1.**

(Please also see my r script file for the full code associated with this document. For the future assignments I’ll be sure to learn R markdown!)

As output, include:

a)  a scatterplot of the dependent and independent variable with a line added (using curve) representing best fit least squares model

b)  Parameter estimates and 95% CI for slope and intercept parameters

c)  A metric of model fit (calculate either R2 or RMSE using the functions above) reminder: y\_hat is calculated using the equation for a line applied to the predictor variable

-----Question 1------

a)



b) Parameter estimates and 95% CI:

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| Intercept (score w/no  LSD) | 71.00758 | 107.240169 |
| Slope (change in score for every unit increase in LSD) | 12.87325 | -5.145685 |

c) RMSE = 6.022356

R2= 0.877835

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

LSDconForScore=function(score)

{

return((score-interceptMS)/slopeMS)

}

LSDconForScore(85)

Answer= 0.4577645

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

There seems to be a pretty strong negative correlation between how well a student does on the test, and how much LSD is concentrated in their tissue. The R2 value of .878 indicates that the model can explain about 87.8% of the variance in the data.. The RMSE value is 6.022, which considering the scale of the dependent variable in this case is 0 to 100, corresponds to about half a letter grade and is a reasonably good resolution, and also indicates this model fits the actual data quite well. As far as the predictive power of this model, however, I’m cautious to say at this point given the small number of data points. It’s pretty easy to accidentally fit a good line through a small number of points, and with a little outside knowledge of how test scores vary even without LSD present, I’d be curious to explore this further before I got too excited about this model’s predictive power.

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

The scores on the test presumably only go up to 100 and down to 0. Also, presumably you can’t have more than 100% concentration of LSD in your body. At a certain point most likely way below that, its effects will also likely be lethal. This is also a super small dataset, and with only seven values it doesn’t (at least visually) look exceptionally normal:



-----Question 2-------

a)



b) Parameter estimates and 95% CI:

|  |  |  |
| --- | --- | --- |
|  | 2.5 % | 97.5 % |
| Intercept (weight loss w/no  Pomegranate) | -1.408937 | 1.0509767 |
| Slope (change in weight for every unit increase in pomegranate) | -0.886420 | -0.1637906 |

c) RMSE= 9.961044

R2=0.008083812

**With the least squares output in mind, do you agree or disagree with the miracle claim?**

I really don’t think there’s much of a claim to be made about this data given the stats outcomes here. With an R2 of .008, precious little (less than 1%) of the variation is explained by this model—one gains very little to nothing from fitting a line through the data as opposed to just using the mean. As can be seen from the graph, the ranges of weight *gains* look pretty similar to the ranges of weight losses, so anyone privy to this would rightly be skeptical about their weight loss prospects regrading eating pomegranates at all, let alone the promise that with more pomegranate consumption one will achieve more weight loss.

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

MeanAbs<- function(y, yhat) {

return((1/length(y))\*sum((abs(y-yhat))))

}

MAE for the math scores:

MeanAbs(math$MATH\_score, yhat\_MS)

4.890816

MAE for the Miracle pomegranate data

MeanAbs(miracle$Weight\_loss, yhat\_WL)

7.981461

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

Math data:

R2= 0.877835

RMSE = 6.022356

MAE: 4.890816

Miracle pomegranate data:

R2=0.008083812

RMSE= 9.961044

MAE: 7.981461

The R2 values show us the proportion of variance explained by the model, or in other words, how much better the model is to explain the data than the mean alone model. It’s a relative measure of fit. RMSE, in contrast, shows us the square root of the variance of the residuals, so it gives us an idea of the spread and magnitude of how far off the model is from the actual values. It’s an absolute measure of fit, so a more direct connection to the actual data, although it can be a little harder to interpret because how “good” it is isn’t on a standard scale, but is in the units of the response variable, and how good it is depends on how it measures up on the scale of the response variable. MAE is similar to RMSE, but differs in that it gives equal weight to all errors, while RMSE gives extra weight to large errors.

In the case of the two datasets here, we can see from all three metrics that the math score model does a decent job of fitting the data, and the pomegranate model does a poor job. In both cases, the MAE values were quite a bit better than the RMSEs, suggesting that some of more extreme or outlier data points were playing an arguably outsized role in reducing the fit of the models. The difference in the R2 values is the most dramatic. I think this is probably because the R2 isn’t measuring simply how well the model fits the data, it’s telling us how much *better* the model is than the mean model. In the case of math scores, this is a lot better. In the case of the pomegranate, the mean model is almost the same as the least squares model so we have little improvement. As far as how well the actual model fits the data itself—the math model is still a lot better than the pomegranate models, but we don’t see as dramatic of a difference between the two.

One thing I’m curious about, is that I’m not sure any of these three metrics do a good enough job of describing just how bad the pomegranate model is. All three are designed to tell us how well the model *fits* the data, but don’t account for the outcomes that actually break down the claims. The claim is that pomegranate consumption can help you lose weight, and the kind of implied null hypothesis is that there is no relationship between pomegranate consumption and weight loss—i.e. you have no or only a background change in weight without eating pomegranates, and you lose weight if you eat them. But what we actually see is a good portion of the people actually *gained* weight. It would be interesting to try to re-create the model to account for this.

----Question 4----

Simulate linear data.

**Step 1: Create a predictor variable using runif or seq**

#Predictor variable:

#NumKittens, number of cute kitten pics an account has on Instagram

NumKittens<-seq(300)

**Step 2: Decide on a value for the intercept and slope**

slopeK<-20

interceptK<-15

sigmaK<-50

**Step 3: Use rnorm to simulate draws from a normal distribution for your dataset**

NumLikes <- rnorm(300, mean=interceptK + slopeK\*NumKittens, sd=50)

hist(NumLikes)

**A. Plot your data**

plot(NumLikes~NumKittens)



**B. Estimate the slope and intercept parameters from the data using linear regression.**

kittenMod<-lm(NumLikes~NumKittens)

kittenMod

Coefficients:

Intercept: 9.378

Slope: 20.028

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

They're pretty close, but they aren't the same. They also differed quite a bit between different runs of my simulation, and unsurprisingly varied a lot less when I played around with lower sigma values.

----Question 5---------

**Step 1: Create a predictor variable using runif or seq**

Scenario: number of cats Clara will have as she transforms into an extreme cat lady and lives to an unlikely old age.

#Predictor variable: Age

ClaraAge<-seq(150)

**Step 2: Decide on a value for the intercept and slope**

slopeC<-.3

interceptC<-0

sigmaC<-.2\*ClaraAge

**Step 3: Use rnorm to simulate draws from a normal distribution for your dataset**

NumCats <- rnorm(150, mean=interceptC + slopeC\*ClaraAge, sd=sigmaC)

hist(NumCats)

**A. Plot your data**

plot(NumCats~ClaraAge)



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

What we might be seeing here is that the further into the future we look, the less certain we can be about Clara’s capacity to take in more cats and the potential effects of *Toxoplasma gondii-*induced erratic behavior. It is one possible scenario that Clara will continue accumulating lots of cats into very old age. It is also possible that after her first few cats she will only slowly accumulate more, and or that she will accumulate them at a rate only slightly exceeding replacement. In a few troubling outlier scenarios where the number of cats Clara has actually decreases and goes *negative*, we can only guess that Clara not only gets rid of her own cats, but goes on to start getting rid of *other people’s* cats too.