Homework 8

Question 1

Using the diamonds data set, make a data frame that shows the slope of the relationship of price as a function of weight (carat) for each unique color, in decreasing order of slope. Your answer should look exactly like the data frame below.

Hints:

* remember the difference between map() and map\_lgl()
* you are likely to find the arrange() function useful. Remember using ? to see what arguments it takes.

Question 2: nonlinear least squares models

Fitting nonlinear models

The DNase dataset in the datasets package (which automatically loads when you load R) has examples of ELISA band optical density as a function of protein concentration, for 12 different runs. These relationships are clearly nonlinear (please plot density as a function of conc to see for yourself!).

Let’s imagine that optical density is proportional to the square root of concentration, as follows:

density=β1conc−−−−√+β0density=β1conc+β0

We could model that with a nonlinear least squares model, fit using the nls2() function in the nls2 package[1](file:///C:\Users\taraw\Downloads\hmk_08.html#fn1), and the following formula:

sqrt\_form <- formula(density ~ beta1\*sqrt() + beta0)

Note that when we specify formulas for nonlinear least squares models, we need to explcitly name each parameter (beta1 and beta0). Furthermore we need to specify each of the variables (density and conc) using the same name that they have in the data frame.

Note also that we’ll need to give nls2() a list of starting “guesses” for each parameter. R will take the each guess and change it slightly until it has found the combination of parameters that fits the data the best (more precisely: that minimizes the sum of the squares of the distances between each value and the model’s prediction of that value, aka the residuals).

For instance, to fit the model above to the full data set, we could do as follows:

**library**(nls2)

nls\_mod <- formula(density ~ beta\_1 \* sqrt(conc) + beta\_0)

single\_sqrt\_model <- nls2(nls\_mod,

data = DNase,

start = list(beta\_1 = 0.5, beta\_0 = 0.1))

We can examine the model using summary, as with lm objects:

summary(single\_sqrt\_model)

##

## Formula: density ~ beta\_1 \* sqrt(conc) + beta\_0

##

## Parameters:

## Estimate Std. Error t value Pr(>|t|)

## beta\_1 0.550521 0.006287 87.57 < 2e-16 \*\*\*

## beta\_0 -0.053297 0.011081 -4.81 3.26e-06 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.08897 on 174 degrees of freedom

##

## Number of iterations to convergence: 1

## Achieved convergence tolerance: 6.129e-08

Comparing models

I want you to consider two possible models for these data: the square-root model mentioned above, and the monod model:

density=conc×dmaxconc+kdensity=conc×dmaxconc+k

Here, dmaxdmax is the maximum possible density at infinitely high concentration, and kk is a constant that tells you how fast density approaches the maximum[2](file:///C:\Users\taraw\Downloads\hmk_08.html#fn2).

Your assignment is to use list-columns to create a separate square root model and monod model for density as a function of concentration, **for each Run**.

Next, extract the AIC value for each model. My approach to do this was to write a function that extracts the AIC value from the output of broom::glance().

AIC is a statistic used to evaluate the quality of a model. If you use different models on the same data, the model with the lowest AIC is the “best”[3](file:///C:\Users\taraw\Downloads\hmk_08.html#fn3).

Make a plot of the AIC values for each model type, for each Run, and use that plot to determine whether, in general, it is better to use the square root model or a monod-type model for this sort of experiment.

Extra credit

Make a plot of the raw DNase data, with density on the y axis and conc on the x axis. Then use the base R predict() function, or modelr::add\_prediction(), to plot each kind of model as a smooth line on the data.

1. Base R has a nls() function. nls2() works basically identically to nls(), but a bit better.[↩︎](file:///C:\Users\taraw\Downloads\hmk_08.html#fnref1)
2. More specifically: kk is the concentration at which density=dmax/2density=dmax/2[↩︎](file:///C:\Users\taraw\Downloads\hmk_08.html#fnref2)
3. Model selection is a difficult task. Properly speaking, AIC, or Akaike Information Criterion, estimates the out-of-sample error of a statistical model. Often people will select a model on the basis of having the lowest AIC, but there are many other bases on which to select models.[↩︎](file:///C:\Users\taraw\Downloads\hmk_08.html#fnref3)