# Nonlinear Regression and Heteroscedasticity

Statistics 516, Homework 2

This homework assignment concerns specifying and the interpreting (via inference) nonlinear regression models, and methods for accounting for heteroscedasticty. You will likely need to install several packages to access the data. You will need to install the **bootstrap**, **drc**, and **alr4** packages, as well as the **trtools** and **ggplot2** packages which you should have already installed.

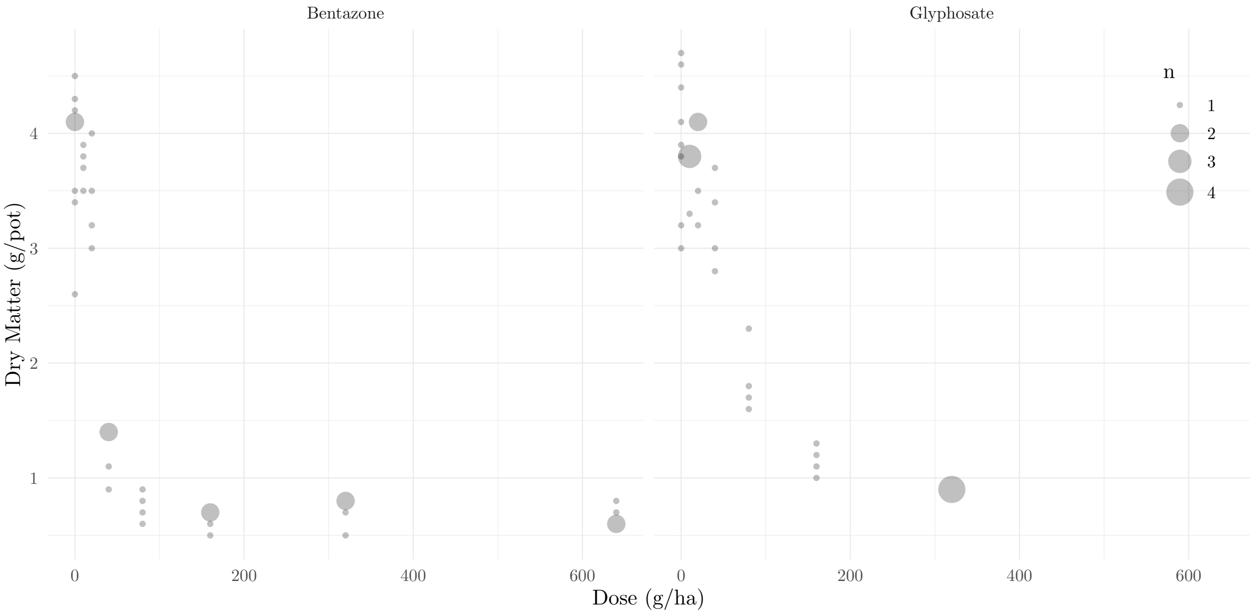
## Instructions

1. This assignment is due by 5:00 PM on Friday, March 11th. Email me your homework at trjohns@uidaho.edu. If possible, save/export your homework as a PDF file. Late assignments will be penalized by 10% if turned-in within 12 hours of the deadline, and 10% more for each additional 12 hour interval.
2. Your solutions must be **typed** and **very** neatly organized. I will not try to infer your solutions if they are not clearly presented. Mathematical expressions need not be typeset perfectly but they should be clear. You may substitute letters for symbols (e.g., b1 for $\beta_1$) and use other shortcuts for mathematical notation if no meaning is lost.
3. You must include with your solutions the relevant R output **and** R code that created them. Be sure that you provide sufficient code that I can replicate your results. Include both the code and the output within the text of your solutions (not in an appendix) using cut-and-paste. But edit your output so as to provide only that which is relevant to answering the questions. Use a monospace font (e.g., Courier or Monaco) for R code and output for clarity. Do not use a monospace font for text that is not R code or output.
4. Plots from R Studio can be exported in various formats or directly to the clipboard using the "export" menu in the top-left part of the plot panel.
5. It is permitted for you to discuss the homework with other students in the course. However your work including R code, output, and written answers must be your own.
6. You are very welcome to ask me questions. I will be happy to clarify what I am asking in any of the questions and will provide you some help with solving problems by showing you how to work through similar problems from class. I will also be open to helping with any R problems. If you email me with a R question, it will usually be helpful for you to include enough of your R script so that I can replicate your issue. But please avoid saving all your questions for just before the assignment is due. I can usually respond quickly to questions, but I will sometimes need time to respond.

## Modeling the Potency of Two Herbicides

The data frame `S.alba` in the **drc** package contains data from an experiment investigating the potency of two herbicides, bentazone and glyphosate, for use with white mustard.[1] The data are shown in the plot below.

```r
library(ggplot2)
library(drc)
p <- ggplot(S.alba, aes(x = Dose, y = DryMatter)) + theme_minimal() +
  geom_count(alpha = 0.25) + facet_wrap(~ Herbicide) +
  labs(x = "Dose (g/ha)", y = "Dry Matter (g/pot)") +
  theme(legend.position = c(0.95, 0.8))
plot(p)
```



Note the use of `geom_count` here. It can be used instead of `geom_point` to make the size of the points proportional to the number of points at a given location. Pots of plants were randomly assigned to receive a specified dose of one of the two herbicides. The amount of dry matter from each pot was later measured.

Assume that the goal of this study is to assess how the dose of each of the two herbicides affect dry matter. To do this a nonlinear regression model can be used with dry matter as the response variable, and dose and herbicide as the explanatory variables. This model will have the form

$$E(M) = \gamma + \frac{\delta - \gamma}{1 + e^{\beta(\log d - \log \alpha)}},$$

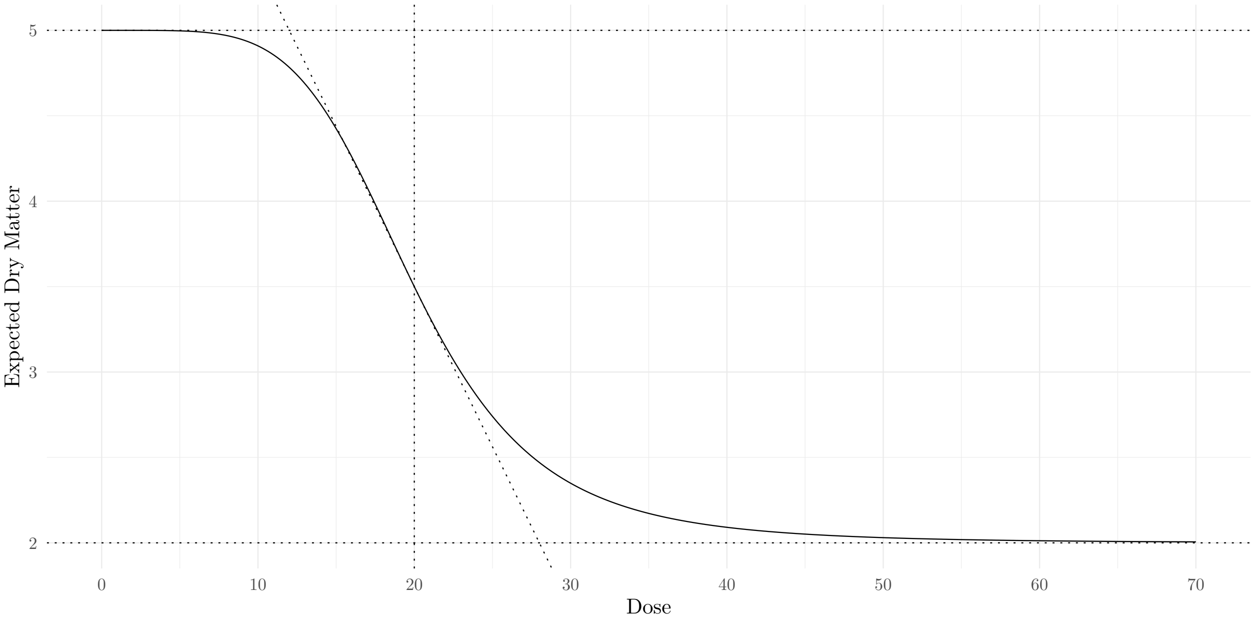
where $M$ and $d$ are dry matter and dose, respectively.[2] The four parameters of this model (i.e., $\alpha$, $\beta$, $\delta$, and $\gamma$) have useful interpretations in terms of how the expected dry matter is related to dose. The parameter $\delta$ is the expected dry matter at zero dose, and $\gamma$ is the asymptote of expected dry matter as dose increases. The parameter $\alpha$ is the dose value where the expected dry matter is half way between its maximum value of $\delta$ and its minimum value of $\gamma$ — i.e., when $E(M) = (\delta + \gamma)/2$.[3] The parameter $\beta$ is related to "how quickly" the expected dry matter decreases as dose increases when dose equals $\alpha$. Specifically, it can be shown that the

[1] Christensen, M. G., Teicher, H. B., & Streibig, J. C. (2003). Linking fluorescence induction curve and biomass in herbicide screening. *Pest Management Science*, *59*, 1303–1310.

[2] Note that $e^x$ is the exponential function where $e \approx 2.718$ is Euler's number. This function is also written as $\exp(x)$, and in R it is written as `exp(x)`.

[3] Because $\log(0)$ is not defined, $E(M)$ is not defined *mathematically* if the dose equals $\alpha$. But computers will typically evaluate $\log(0)$ as $-\infty$ because of the one-sided limit $\lim_{x \to 0+} \log(x) = -\infty$. And for a similar reason $e^{-\infty}$ is evaluated as 0 by computers.

slope of a tangent line when dose equals $\alpha$ is $-\beta(\delta - \gamma)/(4\alpha)$, so everything else being equal as $\beta$ increases the expected dry matter decreases "more quickly" as dose increases.[4] The plot below shows this model with $\alpha = 20$, $\beta = 5$, $\delta = 5$, and $\gamma = 2$.



The **drc** package provides functions the help automate the estimation of a variety of nonlinear regression models like this one for dose-response relationships. But here you will consider how to use the `nls` function to estimate this model. Being proficient at using a function like `nls` is very useful because then you are not limited to using only those models programmed by other authors. A feature of the **drc** package is that it provides "self-starter" features that find good starting values for you automatically. But when using `nls` it is up to you to find good starting values. Fortunately for this particular model this is not too difficult. You can relatively easily "eyeball" reasonable starting values for $\alpha$, $\delta$, and $\gamma$ by looking at a plot of the data. Finding a good starting value for $\beta$ can be a bit trickier, but here is one strategy that can be used. Suppose we compute the mean value of `DryMatter` for each combination of `Herbicide` and `Dose` as follows.

```
library(dplyr)
S.alba %>% group_by(Herbicide, Dose) %>%
  summarize(drymatter = mean(DryMatter))
```

```
# A tibble: 15 x 3
# Groups:   Herbicide [2]
   Herbicide   Dose drymatter
   <fct>      <int>     <dbl>
 1 Bentazone      0     3.84
 2 Bentazone     10     3.72
 3 Bentazone     20     3.42
 4 Bentazone     40     1.2
 5 Bentazone     80     0.75
 6 Bentazone    160     0.625
 7 Bentazone    320     0.7
 8 Bentazone    640     0.675
```

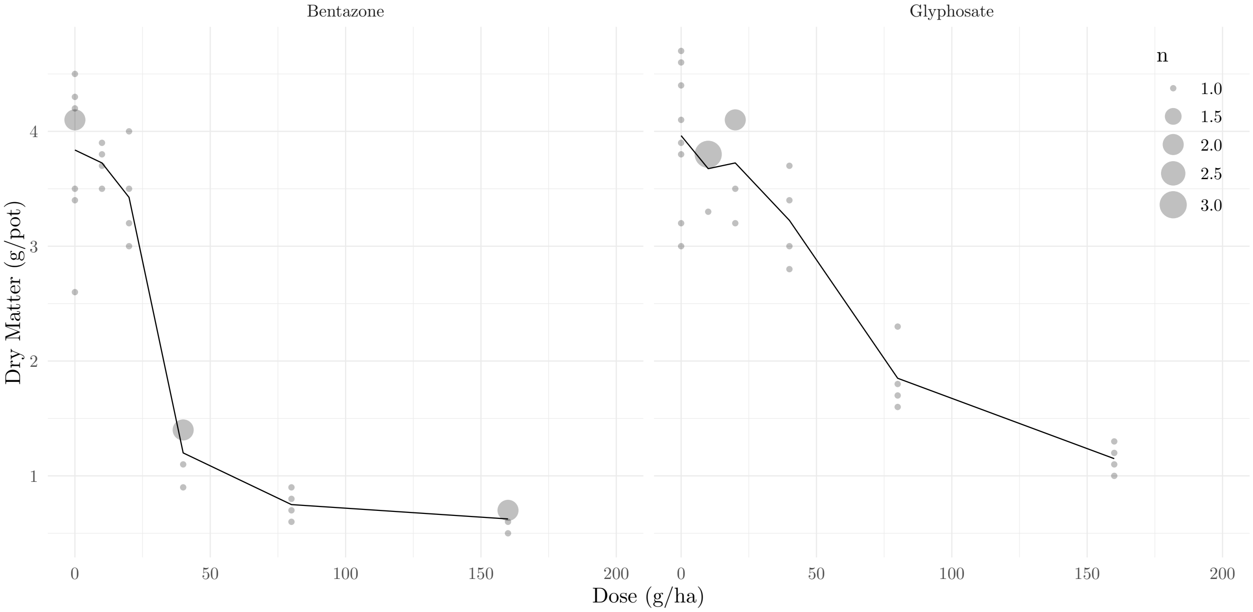
[4] This can be seen by differentiation to show that

$$\left.\frac{\partial E(M)}{\partial d}\right|_{d=\alpha} = \frac{-\beta(\delta - \gamma)}{4\alpha}.$$

```
 9 Glyphosate     0      3.96
10 Glyphosate    10      3.68
11 Glyphosate    20      3.72
12 Glyphosate    40      3.22
13 Glyphosate    80      1.85
14 Glyphosate   160      1.15
15 Glyphosate   320      0.9
```

We can actually plot these means and connect them with line segments by "adding" `stat_summary` to the earlier plot to provide a very crude approximation to the model as shown below. Note that the following also "zooms-in" on dose values between 0 and 200.

```
p <- p + stat_summary(fun = "mean", geom = "line") +
  scale_x_continuous(limits = c(0, 200))
plot(p)
```



Recall that the slope of the tangent line when dose equals $\alpha$ is $-\beta(\delta-\gamma)/(4\alpha)$. We can approximate this slope by computing the slope of the line segment that "contains" what we guess is the value of $\alpha$. For example, for the bentazone herbicide if we guessed that $\alpha$ was between 20 and 40 g/ha, then the slope of that line segment (using the means computed above) equals $(1.2 - 3.42)/(40 - 20)$. Thus we might find an approximate value of $\beta$ to use as a starting value if we solve for $\beta$ in the equation

$$\frac{1.2 - 3.42}{40 - 20} = \frac{-\beta(\delta-\gamma)}{4\alpha},$$

where $\alpha$, $\gamma$, and $\delta$ are replaced the values that you "eyeballed" from the plot of the data to use as starting values. You may find it useful to use this strategy to find a good starting value for $\beta$ in your models.

1. Estimate the nonlinear model described above using `nls`. In this model assume that the type of herbicide does not matter so your model will simply be

$$E(M_i) = \gamma + \frac{\delta - \gamma}{1 + e^{\beta(\log d_i - \log \alpha)}},$$

   where $M_i$ and $d_i$ are the $i$-th observations of dry matter and dose, respectively. To find your starting values you can make a plot of the data for both herbicides combined by omitting `facet_wrap(~ Herbicide)` from the code given earlier to produce a plot of the raw data without accounting for the type of herbicide. And to compute the sample means for each dose but not for each combination of dose use `group_by(Dose)` instead of `group_by(Herbicide, Dose)` in the code given earlier for computing these means. Give the parameter estimates and their standard errors using the `summary` function, and plot the model by adding a smooth curve to the plot to show the estimated expected response as a function of dose. Note that if you add this curve to the original plot then the data frame of predicted values must include the type of herbicide even though it is not part of your model (see the first problem from the in-class exercise with the Michaelis-Menten model).

2. Estimate a nonlinear model where the $\alpha$, $\beta$, and $\gamma$ parameters vary by herbicide, but $\delta$ does not, using the `nls` function. This model can be written case-wise as

$$E(M_i) = \begin{cases} \gamma_b + \frac{\delta-\gamma_b}{1+e^{\beta_b(\log d_i - \log \alpha_b)}}, & \text{if the herbacide used was bentazone,} \\ \gamma_g + \frac{\delta-\gamma_g}{1+e^{\beta_g(\log d_i - \log \alpha_g)}}, & \text{if the herbacide used was glyphosate.} \end{cases}$$

   The rationale for this model is that when the dose is zero there should be no difference in the expected response as a function of the type of herbicide, so $\delta$ should not depend on the type of herbicide used. Report the estimates and standard errors of the seven parameters using the `summary` function. Also plot this model with the raw data by adding a smooth curve to the first plot shown above to show the estimated expected response as a function of dose and type of herbicide.

3. A researcher might like to make inferences about the difference in the $\alpha$, $\beta$, and $\gamma$ parameters between the two herbicides. Use the `lincon` function to produce estimates, standard errors, confidence intervals, and tests concerning $\alpha_b - \alpha_g$, $\beta_b - \beta_g$, and $\gamma_b - \gamma_g$.

## Jevon's Gold Sovereigns

The data frame `jevons` in the **alr4** package contains summary statistics on the weights of sovereigns (i.e., British gold coins) that were collected from circulation in Manchester, England. These data are from a paper by the 19th century economist and philosopher William Stanley Jevons.[5] This data frame (shown below) gives the mean and standard deviation of the weights of five samples of sovereigns that vary by age (in decades).

```
library(alr4)
jevons
```

```
  Age   n Weight      SD   Min   Max
1   1 123  7.973 0.01409 7.900 7.999
2   2  78  7.950 0.02272 7.892 7.993
3   3  32  7.928 0.03426 7.848 7.984
4   4  17  7.896 0.04057 7.827 7.965
5   5  24  7.873 0.05353 7.757 7.961
```

We do not have the original data, so for the purpose of this exercise you will create an artificial data set that produces data with the same sample sizes, means, and standard deviations.[6]

```
library(dplyr)
library(tidyr)

set.seed(123)
```

[5] Jevons, W. S. (1868). On the condition of the metallic currency of the United Kingdom, with reference to the question of international coinage. *Journal of the Statistical Society of London*, *31*, 426–464.

[6] There are a couple of things to note here. One is the use of `set.seed`. The simulated data are generated using the random number generators in R. This initializes the state of the random number generator so that anyone using this code would produce the *same* random numbers. The other thing to note is the use of `dplyr::select`. The function `select` from the **dplyr** package is used to select certain variables from a data frame (and thus deselect others). But there is a function of the same name in the **MASS** package that does something very different. The **MASS** package is frequently loaded with other packages, so to avoid potential conflicts I will often use `dplyr::select` out of habit to avoid problems.

```
coins <- jevons %>% uncount(n) %>%
  group_by(Age) %>% mutate(y = rnorm(n(), Weight, SD)) %>%
  mutate(y = SD * (y - mean(y))/sd(y) + Weight) %>%
  dplyr::select(Age, y) %>% rename(Weight = y)
```

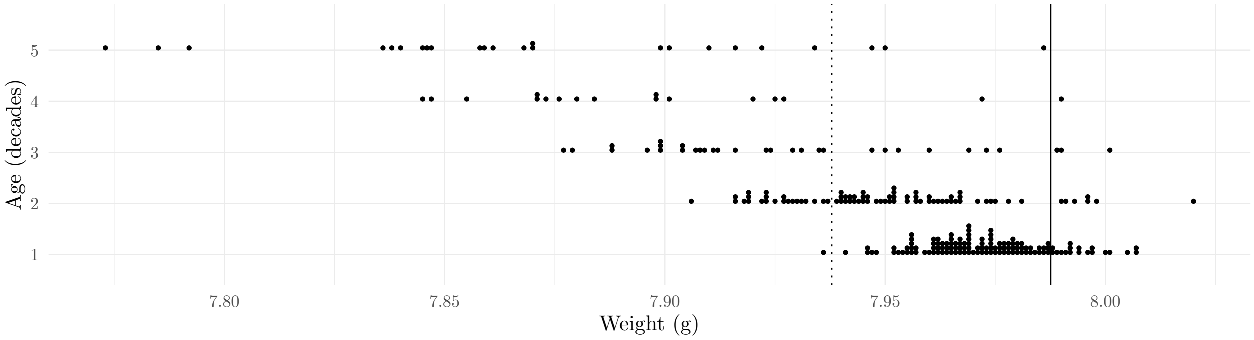
We can confirm that these artificial data give the same means and standard deviations as the original data.

```
coins %>% group_by(Age) %>%
  summarize(n = n(), meanweight = mean(Weight), sdweight = sd(Weight))
```

```
# A tibble: 5 x 4
    Age     n meanweight sdweight
  <int> <int>      <dbl>    <dbl>
1     1   123       7.97   0.0141
2     2    78       7.95   0.0227
3     3    32       7.93   0.0343
4     4    17       7.90   0.0406
5     5    24       7.87   0.0535
```

The figure below shows a plot of the simulated data.

```
p <- ggplot(coins, aes(x = factor(Age), y = Weight)) + theme_minimal() +
  geom_hline(yintercept = 7.9876) +
  geom_hline(yintercept = 7.9379, linetype = 3) +
  geom_dotplot(binaxis = "y", binwidth = 0.001, method = "histodot") +
  coord_flip() + labs(x = "Age (decades)", y = "Weight (g)")
plot(p)
```



The solid line shows the intended standard weight of newly minted sovereigns (7.9876 g), and the dotted line shows the minimum legal weight (7.9379 g). Perhaps not unsurprisingly it can be seen that, on average, older sovereigns have less weight, presumably due to wear while in circulation. But note also that the *variability* of weight appears to increase with the age of the coins. This may be due to differences in how much the coins are in circulation. Some coins are frequently being exchanged thus losing more material, whereas others may being exchanged less and thus not losing as much material. In this problem you will consider various ways of dealing with the heteroscedasticity as well as the consequences of failing to account for heteroscedasticity.

1. Estimate a linear model using the `lm` function with `Weight` as the response variable and `Age` as the explanatory variable. In your model treat age as a *factor* (i.e., a categorical variable) and not a quantitative variable. You can do this by either using `factor(Age)` instead of `Age` in the model formula, or by creating a new variable such as `coins$Agef <- factor(coins$Age)` which will coerce the variable into a factor. Use either `contrast` *or* functions from the **emmeans** package to produce estimates, standard errors, and confidence intervals for (a) the expected weight of coins from each age group and (b) the difference in the expected weight between the the newest coins (i.e., age of one

decade) and the other four groups of coins.[7] For this model you should find that the estimated expected weights are equal the corresponding sample means, and that the estimated differences in the expected weights are equal to the differences in the corresponding sample means.

2. Assume that the variances vary by decade so that

$$Y_i = \begin{cases} \sigma_1^2, & \text{if the } i\text{-th observation is of a coin one decade old,} \\ \sigma_2^2, & \text{if the } i\text{-th observation is of a coin two decades old,} \\ \sigma_3^2, & \text{if the } i\text{-th observation is of a coin three decades old,} \\ \sigma_4^2, & \text{if the } i\text{-th observation is of a coin four decades old,} \\ \sigma_5^2, & \text{if the } i\text{-th observation is of a coin five decades old.} \end{cases}$$

   There are a couple of different ways to account for this kind of variance structure. One is to use *weighted* least squares where the weights are estimated as the reciprocals of the sample variances. We discussed how to compute these weights using functions from the **dplyr** package. Another approach is to use a parametric model where the five variances are effectively estimated from the data. We discussed how to do this with using the `gls` function from the **nlme** package. Use both of these approaches and for each show the parameter estimates and their standard errors as given by `summary`, and use either the `contrast` function or functions from the **emmeans** package to produce estimates, standard errors, and confidence intervals for (a) the expected weight of coins from each age group and (b) the difference in the expected weight between the the newest coins, just as you did in the previous problem.[8]

3. Compare the estimates and standard errors for estimating the model parameters as well as the expected weight and differences in expected weight when accounting for heteroscedasticity as you did in the last problem, and when not accounting for heteroscedasticity as you did in the first problem. Discuss briefly how failing to account heteroscedasticity (i.e., incorrectly assuming homoscedasticity) may affect your inferences.

## Mortality of Confused Flour Beetles from Carbon Disulphide

The data frame `bliss` in the **trtools** package are from an experiment investigating the effect of gaseous carbon disulphide ($CS_2$) on the mortality of confused flour beetles (*Tribolium confusum*).[9] Cloth cages of batches of approximately thirty beetles were suspended in a flask above a fixed volume of liquid carbon disulphide. The number of dead beetles after five hours of exposure was recorded. The figure below shows the proportion of dead beetles by concentration of carbon disulphide. Note that there are two observations for each dose.

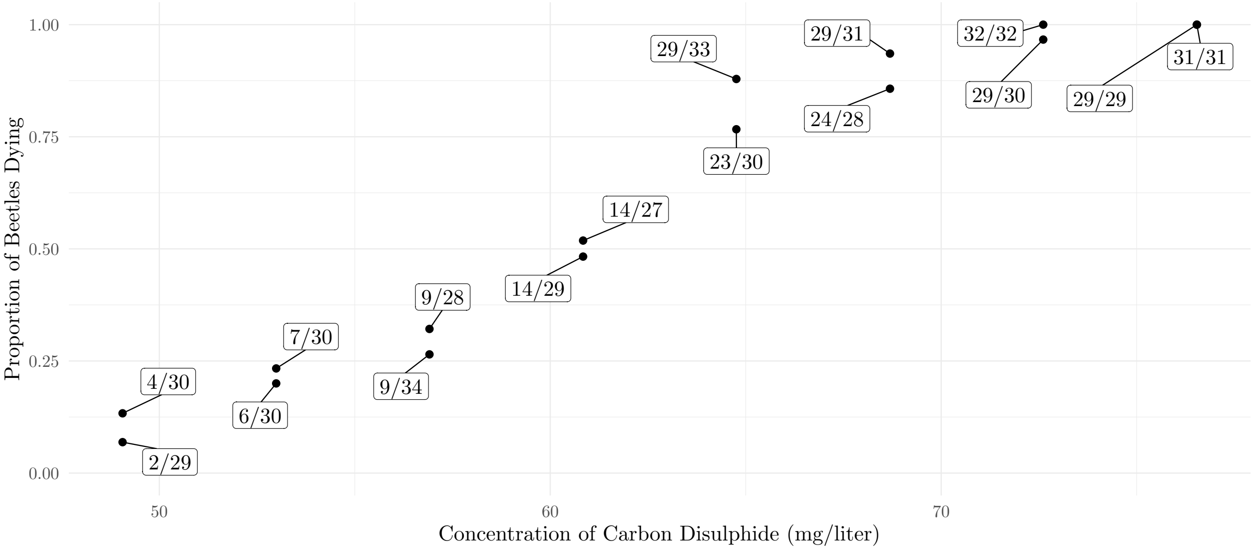
```
library(trtools)
library(ggplot2)
library(ggrepel)

bliss$proportion <- paste(bliss$dead, "/", bliss$exposed, sep = "")

p <- ggplot(bliss, aes(x = concentration, y = dead/exposed)) +
  geom_point() + ylim(0, 1) + theme_minimal() +
  geom_label_repel(aes(label = proportion), box.padding = 0.75) +
  labs(x = "Concentration of Carbon Disulphide (mg/liter)",
```

[7] For an example of how to specify comparisons between one level and the other levels with functions in the **emmeans** package, see the example from the lecture on February 18 where I use the `trt.vs.ctrl` contrast method with the `contrast` function from the **emmeans** package (not the **trtools** package).

[8] Both the `contrast` function from the **trtools** package and functions from the **emmeans** package should give the same estimates and standard errors. They will give somewhat different confidence intervals and p-values, however, because of how the two functions compute the degrees of freedom by default. They can be brought into agreement by using some extra options, but for the purpose of this problem that is not necessary.

[9] The data are featured in Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, *22*, 134–167. But the original source is Strand, A. L. (1930). Measuring the toxicity of insect fumigants. *Industrial and Engineering Chemistry: Analytical Edition*, *2*, 4–8.

```
    y = "Proportion of Beetles Dying")
plot(p)
```

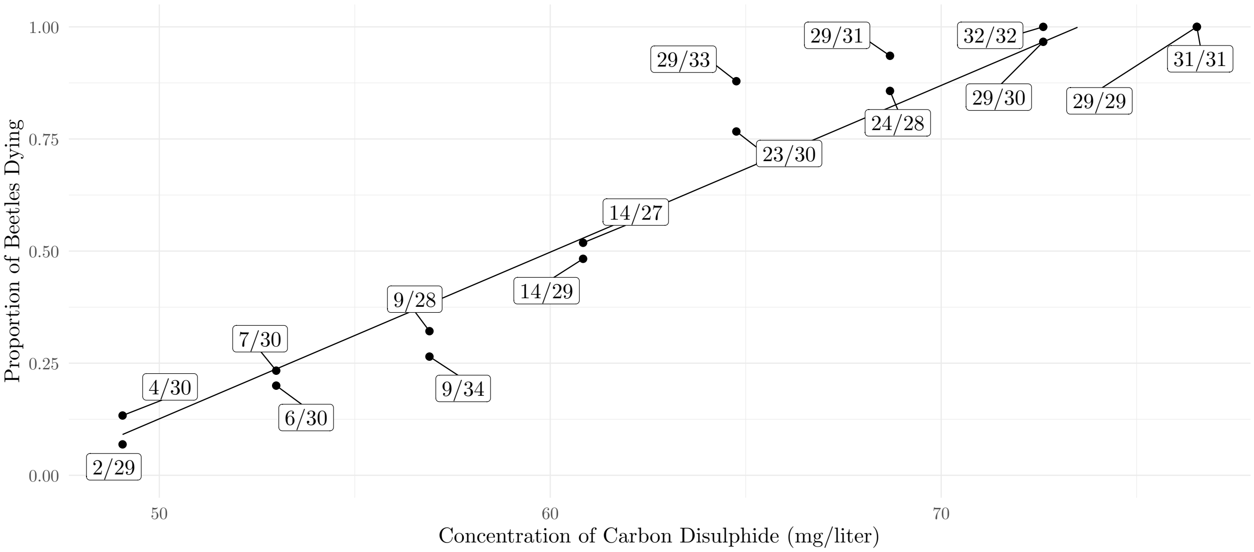


A naive approach to modeling these data would be to use linear regression where the proportion is the response variable.

```
m <- lm(dead/exposed ~ concentration, data = bliss)
cbind(summary(m)$coefficients, confint(m))
```

```
              Estimate Std. Error t value  Pr(>|t|)    2.5 %   97.5 %
(Intercept)   -1.73277   0.159636  -10.85 3.352e-08 -2.07515 -1.39038
concentration  0.03717   0.002516   14.77 6.225e-10  0.03178  0.04257
```

```
d <- data.frame(concentration = seq(49.06, 76.54, length = 100))
d$yhat <- predict(m, newdata = d)

p <- ggplot(bliss, aes(x = concentration, y = dead/exposed)) +
  geom_line(aes(y = yhat), data = d) +
  geom_point() + ylim(0, 1) + theme_minimal() +
  geom_label_repel(aes(label = proportion), box.padding = 0.75) +
  labs(x = "Concentration of Carbon Disulphide (mg/liter)",
    y = "Proportion of Beetles Dying")
plot(p)
```



This model is probably not adequate for two reasons. One is that the relationship between the expected of dead beetles and concentration is probably not linear. Secondly, proportions tend to exhibit heteroscedasticity where the variance of a proportion tends to decrease as its expected value gets farther from 0.5. As we will discuss in lecture, if the number of dead beetles has a *binomial distribution*, then it can be shown that

$$\text{Var}(P) = E(P)[1 - E(P)]/m,$$

where $P$ is the proportion (i.e., `dead/exposed`) and $m$ is the denominator of the proportion (i.e., `exposed`). This implies that the variance of $P$ decreases as $E(P)$ gets farther from 0.5, and also decreases as $m$ increases.

These data will be used in lecture to demonstrate logistic regression, but for this problem you will consider modeling the data using nonlinear regression. Later we will also discuss the relationship between logistic and nonlinear regression.

1. Consider the nonlinear regression model

$$E(P_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 d_i}},$$

where $P_i$ and $d_i$ are the $i$-th observations of the proportion of dead beetles (i.e., `dead/exposed`) and the concentration, respectively.[10] Use the `nls` function to estimate this nonlinear regression model. For starting values you can cheat and use the parameter estimates from a logistic regression model estimated as followed.

```
m <- glm(cbind(dead, exposed - dead) ~ concentration,
  family = binomial, data = bliss)
summary(m)$coefficients
```

```
              Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -14.8084    1.28976  -11.48 1.633e-30
concentration   0.2492    0.02138   11.65 2.250e-31
```

The two estimates reported above are estimates of $\beta_0$ and $\beta_1$ from the logistic regression model. You can use these as your starting values for `nls`. The estimates you obtain using nonlinear regression should be similar but not necessarily equal to those shown above. Report the parameter estimates and their standard errors by showing the output from `summary`. Also plot the estimated model by adding a curve to the plot shown above.[11]

[10] Note that $e^x$ is the exponential function where $e \approx 2.718$ is Euler's number. This function is also written as $\exp(x)$, and in R it is written as `exp(x)`.

[11] Note that when using `ggplot` the *order* that you specify the various geometric objects matters. For example, if `geom_line` appears before `geom_label_repel` then the point labels will be shown in front of rather than behind the curve.

2. As described above, the number of dead beetles has a binomial distribution then the variance of $P_i$ is

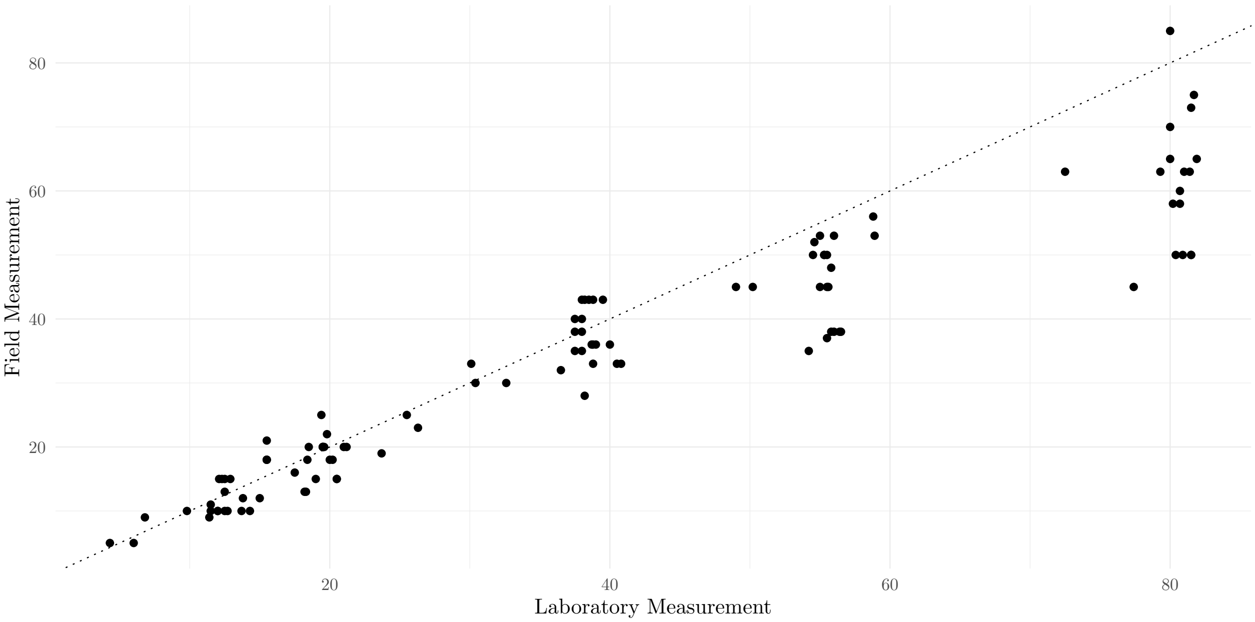
$$\text{Var}(P_i) = E(P_i)[1 - E(P_i)]/m_i,$$

where $m_i$ is the number of exposed beetles for the $i$-th observation. Since $E(P_i)$ is unknown it can be estimated as the predicted value $\hat{Y}_i$. Use an iteratively weighted least squares algorithm with weights implied by the variance above to estimate the model shown above and show the parameter estimates and standard errors using `summary`. You should find that the parameter estimates will be equal to or very close to those obtained using the `glm` function above, but the standard errors will be somewhat different.

## Estimating Bias in Field Measurements in Defects in the Alaska Pipeline

The data frame `pipeline` in the **alr4** package is from a study of the bias of field measurements of defects in the Alaska pipeline. That data includes field and laboratory measurements of the number of defects in observational units from the pipeline.[12]

```
p <- ggplot(pipeline, aes(x = Lab, y = Field)) + theme_minimal() +
  geom_abline(intercept = 0, slope = 1, linetype = 3) + geom_point() +
  labs(x = "Laboratory Measurement", y = "Field Measurement")
plot(p)
```



Assume that the laboratory measurements are very accurate and can be treated as the “true” number of defects. Field measurements are faster and cheaper than laboratory measurements, but are more prone to measurement error (both systematic error or *bias*, and random measurement error). The figure above suggests that the field measurements tend to underestimate the number of defects, particularly as the actual number of defects (as shown by the laboratory measurement) increases. A regression model can be used to estimate the bias of the field measurements so that they can be adjusted appropriately (a process sometimes called *calibration*). In this problem you will use linear and nonlinear regression to estimate a calibration model.

[12] The nature of the observational units and the measurement of the number of defects is not clear. The observational units may be select portions of the pipeline, but it is not clear if these units were removed from the pipeline and brought to a laboratory, or if only the data from the field was brought back to the laboratory for more thorough analysis. Also since the laboratory values are not all integers these measurements might be the number of defects per unit area or volume. Finally, note that in the help file (see `?pipeline`) the `Lab` variable is incorrectly labeled as “Number of defects measured in the field.”

1. Let $F_i$ and $L_i$ denote the field and laboratory measurements, respectively, for the $i$-th observation. If we assume that $L_i$ is the true number of defects in the $i$-th observational unit, then the bias is the expected difference between $F_i$ and $L_i$ which is $E(F_i - L_i)$. The figures suggests that the bias tends to increase as $L_i$ increases, so one possible model might be that the bias is proportional to $L_i$. This can be written as
$$E(F_i - L_i) = \theta L_i,$$
where $\theta$ is the constant of proportionality. If $\theta < 0$ then the field measurements tend to *underestimate* the number of defects by $(1-\theta)100\%$ (assuming that $\theta > 0$), and if $\theta > 0$ then the field measurements tend to *overestimate* the number of defects by $(\theta - 1)100\%$. The model shown above is linear so it can be estimated using `lm`. There are several ways to do this. One is to use $F_i - L_i$ as the response variable and estimate a model with $L_i$ as the explanatory variable but *without* a constant term (i.e., "intercept").[13] A second approach is to write the model as
$$E(F_i) = L_i + \theta L_i$$
if we regard $L_i$ as a fixed and not random variable so that $E(F_i - L_i) = E(F_i) - L_i$.[14] This model is a special case of the linear model
$$E(F_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$
where $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = \theta$, and $x_{i1} = x_{i2} = L_i$. This model is a little strange in that we have two explanatory variables that are the same variable, but $\beta_1$ is not estimated but instead is fixed at one. This can be done by specifying an *offset* in your model (see footnote).[15] One last approach is to write the model as
$$E(F_i) = \gamma L_i$$
where $\gamma = 1 + \theta$ since
$$E(F_i) = L_i + \theta L_i = (1+\theta)L_i = \gamma L_i.$$
Estimate the three models described above in the way described and show the estimate and standard error of the model parameter (i.e., $\theta$ or $\gamma$) using `summary`. The first two models should give you the same estimate and standard error of $\theta$, and the last model should give you an estimate of $\gamma$ that equals the estimate of $1 + \theta$ from the previous models.

2. Plot the model you estimated in the previous problem with the raw data to show a plot like that given earlier but with a line showing the estimated expected field measurement as a linear function of laboratory measurement. You will want to use either the second or third model you estimated in

[13] I have given several examples in lecture of how to estimate a model with `lm` that does not include an constant/intercept term.

[14] From a design perspective, $L_i$ may not be fixed since the values are not necessarily selected by the researchers. But in regression we frequently regard all variables except for the response variable as fixed. Technically what we are doing is *conditioning* on the values of the explanatory variables, so even if they are random we are only considering the distribution of the response variable *given* those values of the explanatory variables.

[15] An offset is an explanatory variable that has a $\beta_j$ fixed at one. This can be done by using `offset(variable)` in your model formula. For example, consider the model
$$E(V_i) = \beta_0 + \beta_1 g_i + \beta_2 h_i,$$
where $V_i$ is tree volume, $g_i$ is girth, and $h_i$ is height. We can estimate this model as follows.

```
m <- lm(Volume ~ Girth + Height, data = trees)
summary(m)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.750e-07
Girth         4.7082     0.2643  17.816 8.223e-17
Height        0.3393     0.1302   2.607 1.449e-02
```

But if I wanted $\beta_1 = 1$ then I could do the following.

```
m <- lm(Volume ~ offset(Girth) + Height, data = trees)
summary(m)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  -80.935    23.6206  -3.426 0.0018484
Height         1.288     0.3097   4.157 0.0002608
```

Note that no inferences for $\beta_1$ are given by `summary` because we are assuming we know it is one so there is nothing to infer.

the previous problem to do this since the first model uses the difference in the field and laboratory measurements as the response variable and so the predicted values from that model are not what you want for plotting purposes.

3. Plot the studentized residuals against the predicted values using either the second or third model you estimated in the first problem. Do you think that the expected field measurement is a linear function of the laboratory measurement? Why or why not? Is there any evidence of heteroscedasticity? Why or why not?

4. Consider an alternative nonlinear model where

   $$E(F_i) = L_i + \theta_1 L_i^{\theta_2}.$$

   This implies that the bias of the field measurements is proportional to some power $\theta_2$ of the laboratory measurements. Note that the second model you estimated in the first problem is a special case of this model where $\theta_2 = 1$. Estimate the model above using `nls`. For your starting values you can use the estimate of $\theta_1$ you obtained in the first problem and $\theta_2 = 1$, since that was the value you implicitly used for those models which can be viewed as an approximation to the model above. Show the parameter estimates and their standard errors using `summary`, and plot the model as a curve with the raw data like you did in the second problem.

5. Plot the standardized residuals against the predicted values based on the nonlinear model you estimated in the previous problem. You cannot use `rstandard` or `rstudent` with a `nls` object, but you can use the `nlsint` function from the **trtools** package to produce standardized residuals. The syntax for a basic plot would something like the following where `m` is your model object created using `nls`.

   ```
   d <- nlsint(m, residuals = TRUE)
   plot(d$fit, d$res)
   ```

   Now consider accounting for any heteroscedasticty in the data by assuming that the variance of the field measurements is proportional to some power $p$ of the expected field measurement so that

   $$\text{Var}(F_i) \propto E(Y_i)^p.$$

   Use an iteratively weighted least squares algorithm to estimate the nonlinear model described in the previous problem for several values of $p$, starting with $p = 1$ and trying increasingly larger values of increments of 0.5 up to $p = 3$. Using residual plots, decide on what you think is a good value of $p$ and then show the parameter estimates with standard errors for that model using `summary` and give another plot of the residuals against the predicted values for that model. Also discuss briefly why you selected that particular value of $p$.[16]

## Anti-Inflammatory Hormone Devices — Revisited

**Note**: This problem is *extra credit* for students in Stat 436, but is *required* for students in Stat 516.

The `nls` function computationally works very similarly to the `lm` function, but the interface is different. The `lm` function allows us to specify a model *symbolically* via the model formula (i.e., the first argument to `lm`), whereas `nls` requires us to specify the model *mathematically*. The `nls` function can be used to estimate a linear model. In practice, this is rarely necessary except maybe in cases where you are using a fairly unusual parameterization of a linear model that is difficult to express using the model formula argument to `lm`. But I think it can be a useful exercise for the student to use `nls` to specify a linear model. In this problem you will use the `nls` function to replicate several models for the `hormone` data from the **bootstrap** package that were featured in the last homework assignment.

In the following you are to estimate the model specified with `lm` by using `nls`. To do this I would recommend that you "decipher" the model from the output of `summary` given below, and then write the model case-wise. I would also suggest you use either the `case_when` function from the **dplyr** package or specify indicator

[16] There is not necessarily a correct value of $p$ here, although some values may be clearly better than others.

variables within the model itself using the `==` operator (see the second problem from the in-class exercise on February 25th for an example of doing this with another model — for example, an indicator variable for Lot A would be specified as `Lot == "A"`). I would recommend against using the `ifelse` function. If you do this correctly then the output from `summary` when applied to the model object created using `nls` should match that created by `lm`. When you do this make sure that the parameter estimates are in the same order (this can be controlled by the order you specify the parameter starting values). Note that since these models are all linear you do not need to specify good starting values. It is fine to specify them all as zero or some other number (as long as they are not very large in absolute value).

1. Estimate the following model using `nls` and show the estimates using `summary`.

```
m <- lm(amount ~ Lot, data = bootstrap::hormone)
summary(m)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   23.078      1.962 11.7630 1.887e-11
LotB          -1.011      2.775 -0.3644 7.187e-01
LotC           5.844      2.775  2.1065 4.581e-02
```

2. Estimate the following model using `nls` and show the estimates using `summary`.

```
m <- lm(amount ~ -1 + Lot, data = bootstrap::hormone)
summary(m)$coefficients
```

```
     Estimate Std. Error t value  Pr(>|t|)
LotA    23.08      1.962   11.76 1.887e-11
LotB    22.07      1.962   11.25 4.717e-11
LotC    28.92      1.962   14.74 1.583e-13
```

3. Estimate the following model using `nls` and show the estimates using `summary`.

```
m <- lm(amount ~ Lot:hrs, data = bootstrap::hormone)
summary(m)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 35.01572   0.736247  47.560 1.783e-24
LotA:hrs    -0.07728   0.005146 -15.016 2.238e-13
LotB:hrs    -0.05566   0.003142 -17.714 6.696e-15
LotC:hrs    -0.05722   0.007423  -7.709 8.045e-08
```

4. Estimate the following model using `nls` and show the estimates using `summary`.

```
m <- lm(amount ~ Lot + hrs + Lot:hrs, data = bootstrap::hormone)
summary(m)$coefficients
```

```
              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 33.360055   1.211583 27.5343 5.787e-18
LotB         1.846061   1.612797  1.1446 2.652e-01
LotC         3.833616   1.933112  1.9831 6.058e-02
hrs         -0.068296   0.007272 -9.3911 5.753e-09
LotB:hrs     0.012010   0.008291  1.4486 1.622e-01
LotC:hrs    -0.006222   0.014670 -0.4241 6.758e-01
```

5. Estimate the following model using `nls` and show the estimates using `summary`.

```
m <- lm(amount ~ -1 + Lot + Lot:hrs, data = bootstrap::hormone)
summary(m)$coefficients
```

```
         Estimate Std. Error t value  Pr(>|t|)
LotA     33.36006   1.211583  27.534 5.787e-18
```

```
LotB     35.20612   1.064509  33.073 1.340e-19
LotC     37.19367   1.506316  24.692 5.341e-17
LotA:hrs -0.06830   0.007272  -9.391 5.753e-09
LotB:hrs -0.05629   0.003982 -14.136 3.361e-12
LotC:hrs -0.07452   0.012740  -5.849 8.330e-06
```