Homework 3

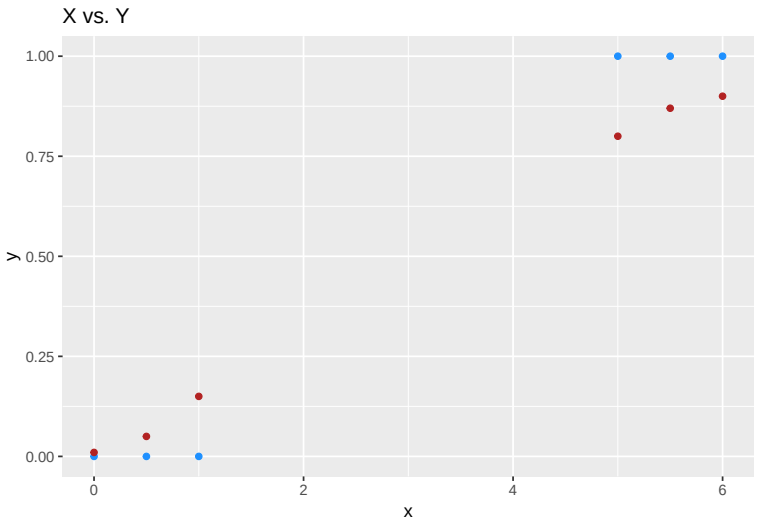
ST623

Nick Sun

October 14, 2019

Question 1

In the scatterplot, I encode the actual observed data as *blue* and the possible fits as *red*.



Let's try fitting a model to this data.

```
model <- glm(y ~ x,
             data = data,
             family=binomial(link = "logit"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Note that we get a warning message here because the data is *perfectly separated*, meaning that there is a clear boundary between the data points that have an observed value of 0 and an observed value of 1. This is not a good sign since this tells us that the MLEs for our coefficients will not exist. Proceeding with the analysis anyways, we get the following summary:

```
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"), data = data)
##
## Deviance Residuals:
```

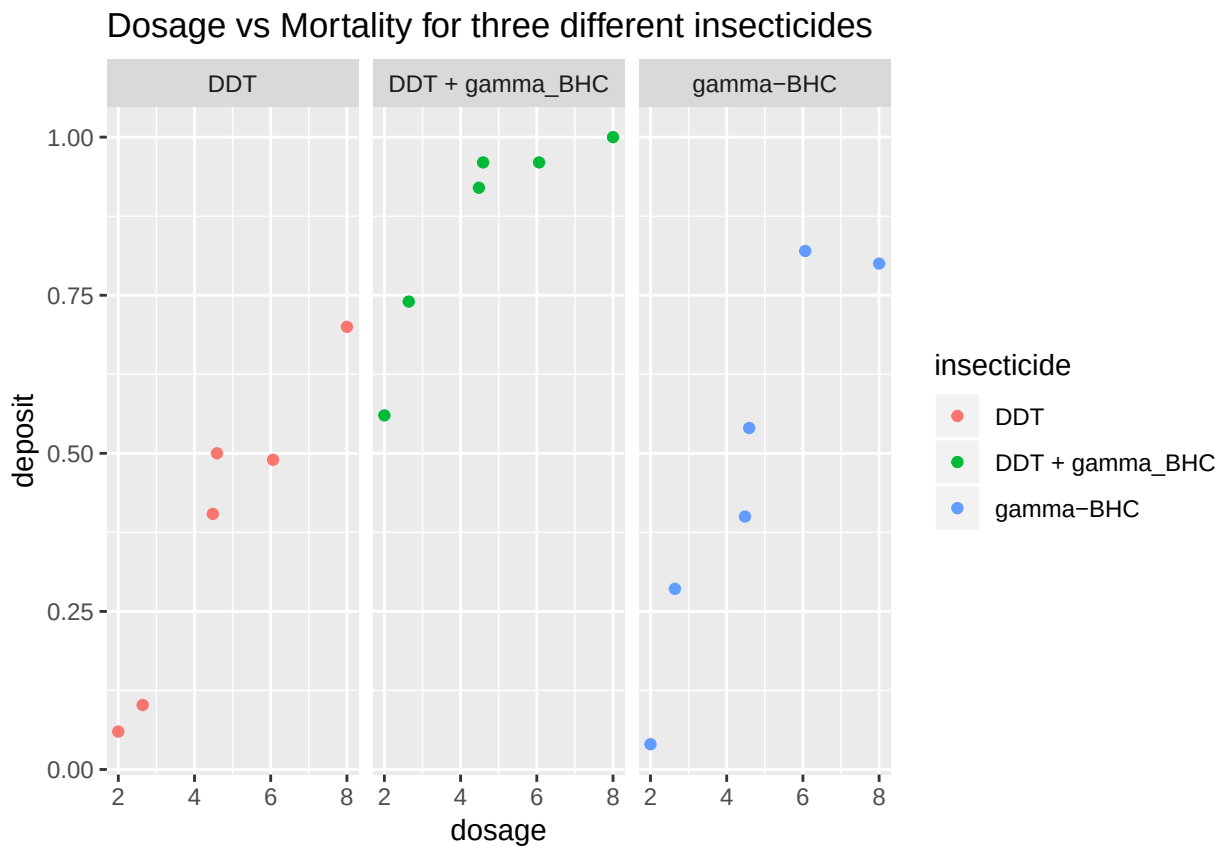
```
##          1           2           3           4           5           6
## -2.110e-08  -7.595e-07  -1.363e-05   1.363e-05   7.595e-07   2.110e-08
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -34.65   80040.88   0.000        1
## x              11.55   22191.72   0.001        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.3178e+00  on 5  degrees of freedom
## Residual deviance: 3.7280e-10  on 4  degrees of freedom
## AIC: 4
##
## Number of Fisher Scoring iterations: 23
```

In the summary of the model, we can see that the estimated coefficients all have enormous standard errors. The estimates themselves also appear somewhat large. The proper remedy for a scenario like this would be to use some form of penalized regression or to identify the cutoff value which separates and the 0 and 1 values and simply use that as a predictor.

Question 2

Part a.

```
ggplot(data) +
  geom_point(mapping = aes(x = dosage, y = deposit, color = insecticide)) +
  facet_grid(~insecticide) +
  labs(
    title = "Dosage vs Mortality for three different insecticides",
    xlab = "Dosage",
    ylab = "Deposit"
  )
```

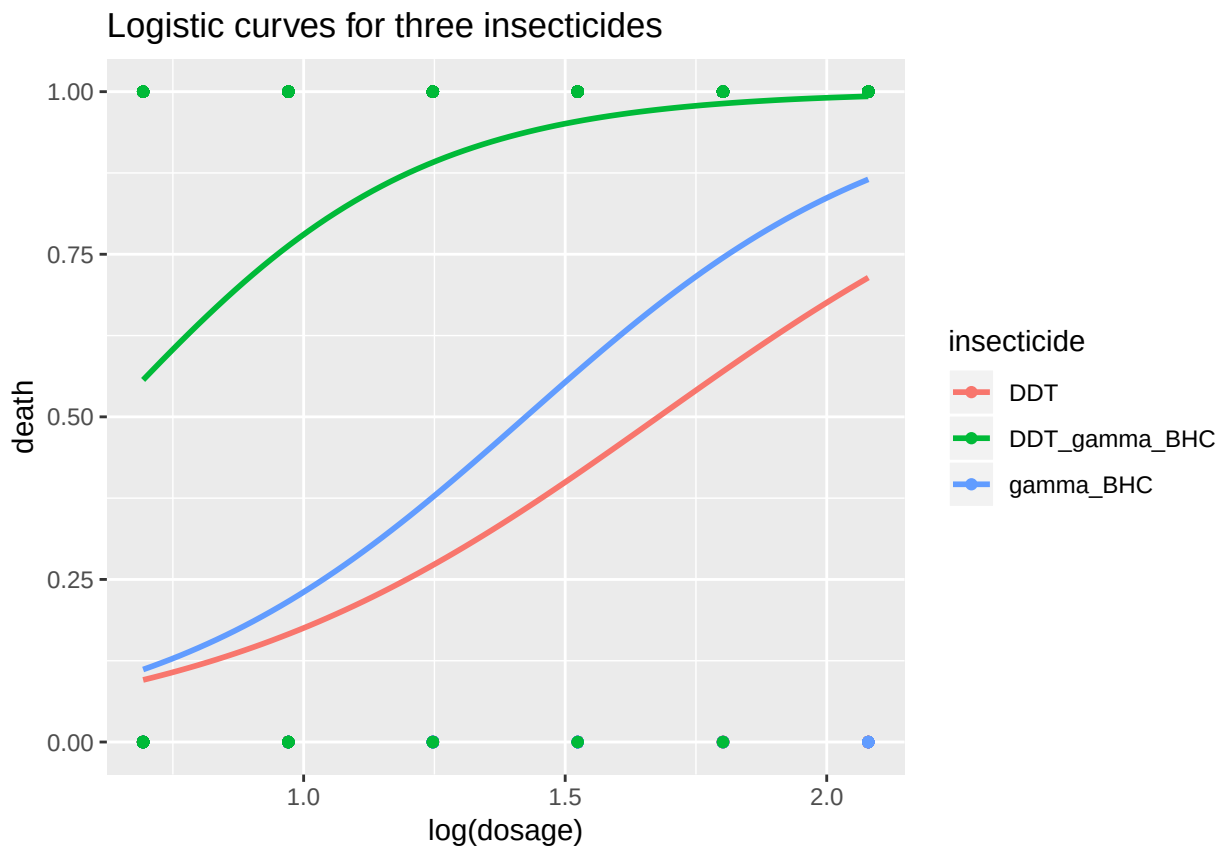


For all insecticides, as the dosage increases the mortality also increases. However, the combination of DDT and γ-BHC appears to have higher mortality than its component insecticides after accounting for dosage.

Part b.

Creating a logistic curve for each insecticide can actually be done within `ggplot()`.

```
ggplot(fulldata,
       mapping = aes(color = insecticide,
                     x = log(dosage),
                     y = death)) +
  geom_point() +
  stat_smooth(
    method = "glm",
    method.args = list(family = "binomial"),
    se = FALSE
  ) +
  labs(
    title = "Logistic curves for three insecticides"
  )
```



Part c.

We can enforce parallel lines by specifying in our model formula that there is only one slope associated with `log(dosage)`. We enable nonparallel lines by specifying an additional slope interaction terms for each level of insecticide.

```
parallel <- glm(death ~ log(dosage) + factor(insecticide),
                family = binomial(link = "logit"),
                data = fulldata)
summary(parallel)
```

```
##
## Call:
## glm(formula = death ~ log(dosage) + factor(insecticide), family = binomial(link = "logit"),
##     data = fulldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6316  -0.7178   0.2524   0.7426   2.3063
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -4.4541     0.3575 -12.460  < 2e-16 ***
## log(dosage)                      2.6938     0.2146  12.551  < 2e-16 ***
```

```
## factor(insecticide)DDT_gamma_BHC   3.0314     0.2521  12.022  < 2e-16 ***
## factor(insecticide)gamma_BHC       0.6144     0.1999   3.074  0.00211 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1203.48  on 881  degrees of freedom
## Residual deviance:  819.04  on 878  degrees of freedom
## AIC: 827.04
##
## Number of Fisher Scoring iterations: 5
```

```
# Not parallel
notparallel <- glm(death ~ factor(insecticide)*log(dosage),
                   family = binomial(link = "logit"),
                   data = fulldata)
summary(notparallel)
```

```
##
## Call:
## glm(formula = death ~ factor(insecticide) * log(dosage), family = binomial(link = "logit"),
##     data = fulldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8293  -0.6981   0.1920   0.7678   2.1675
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                                -3.8308     0.5002  -7.659
## factor(insecticide)DDT_gamma_BHC            1.7101     0.7741   2.209
## factor(insecticide)gamma_BHC               -0.2120     0.7053  -0.301
## log(dosage)                                 2.2824     0.3190   7.156
## factor(insecticide)DDT_gamma_BHC:log(dosage)   1.1060     0.6561   1.686
## factor(insecticide)gamma_BHC:log(dosage)    0.5557     0.4656   1.193
##                                           Pr(>|z|)
## (Intercept)                               1.87e-14 ***
## factor(insecticide)DDT_gamma_BHC            0.0272 *
## factor(insecticide)gamma_BHC                0.7637
## log(dosage)                               8.33e-13 ***
## factor(insecticide)DDT_gamma_BHC:log(dosage)   0.0919 .
## factor(insecticide)gamma_BHC:log(dosage)    0.2327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1203.48  on 881  degrees of freedom
## Residual deviance:  815.64  on 876  degrees of freedom
## AIC: 827.64
##
## Number of Fisher Scoring iterations: 6
```

```r
anova(parallel, notparallel, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: death ~ log(dosage) + factor(insecticide)
## Model 2: death ~ factor(insecticide) * log(dosage)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       878     819.04
## 2       876     815.64  2   3.3925   0.1834
```

We get a high p-value of around .18 for this test, telling us that the "reduced model" performs adequately and we don't require the extra terms of the "full model". In other words, the model with non-parallel lines does not perform significantly better than the model with parallelism enforced. We should use the model with parallel lines.

Part d.

The model formulas specify one model with an intercept term included and another model with no intercept. We can surmise that not having a single model intercept would cause the three separate insecticide groups to have three individual and different intercepts, but the effect of log(dosage) should be the same between the two models.

If we denote β as the coefficients from the model with the intercept included and β^* as the coefficients from the model without the intercept, then we can see that we can rewrite the β^* in terms of β:

- $\beta_0 = \beta^*_{Mix}$
- $\beta_0 + \beta_{DDT} = \beta^*_{DDT}$
- $\beta_0 + \beta_{\gamma BHC} = \beta^*_{\gamma BHC}$

This allows us to create a matrix A

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

such that the covariance matrix for β^* is $A\Sigma A^T$ where Σ is the covariance matrix for β.

This can be checked by doing the algebra on the output of the `vcov()` function.

Part e.

As discussed in class, Fieller's method uses the following formula to calculate the relative potency of insecticde 1 vs insecticide 2: $\text{Potency}_{01} = e^{\frac{\beta_0 - \beta_1}{\beta_3}}$ where β_3 was the calculated coefficeint related to log(dosage).

The 90% confidence interval for that relative potency is the roots of the equation:

$$\frac{(\hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_3\theta)^2}{V_1(1,1) - 2\theta V_1(1,2) + \theta^2 V_1(2,2)} = 1.645^2$$

where V is the covariance matrix of the estimated parameters.

Now we have to make a choice for what model to extract our estimated β parameters from. I will use the no-intercept model that was used in part d. since it is easier to calculate the covariance matrix and it appears to be sufficient in modeling the data given the model diagnostics.

```r
fieller_model <- glm(death ~ as.factor(insecticide) + log(dosage) -1,
                     data = fulldata,
                     family = binomial(link = "logit"))
betas <- summary(fieller_model)$coe[,1]
betas
```

```
##           as.factor(insecticide)DDT as.factor(insecticide)DDT_gamma_BHC
##                           -4.454087                           -1.422692
##     as.factor(insecticide)gamma_BHC                        log(dosage)
##                           -3.839654                            2.693770
```

```r
potency_mix_gamma <- exp((betas[2] - betas[3])/betas[4])
potency_mix_ddt <- exp((betas[2] - betas[1])/betas[4])
potency_mix_gamma
```

```
## as.factor(insecticide)DDT_gamma_BHC
##                            2.452828
```

```r
potency_mix_ddt
```

```
## as.factor(insecticide)DDT_gamma_BHC
##                             3.08125
```

The potency of the mixture of the insecticides relative to DDT is 3.081 and the potency of the mixture relative to γ-BHC is 2.453.

The confidence intervals here can be found by calculating the variance-covariance matrices for each of the potencies. From `vcov(fieller_model)`, we get the variance covariance of each of the parameters. We can then plug these values into the following to get the variance-covariance for the potencies:

$$V_{\text{mix vs. ddt}} = \begin{pmatrix} var(\beta_2) + var(\beta_1) - 2cov(\beta_2, \beta_1) & cov(\beta_2 - \beta_1, \beta_4) \\ cov(\beta_2 - \beta_1, \beta_4) & var(\beta_4) \end{pmatrix}$$

Plugging in the values we get rom `vcov()`, we get

$$V_{\text{mix vs. ddt}} = \begin{pmatrix} .0636 & .0226 \\ .0226 & .0461 \end{pmatrix}$$

$$V_{\text{mix vs. gamma}} = \begin{pmatrix} .0566 & .0172 \\ .0172 & .0461 \end{pmatrix}$$

Then plugging in to the formula above for the confidence interval, we get the following intevals:

- The relative potency of the mix to γ-BHC is between (2.12, 2.88) with 90% confidence
- The relative potency of the mix to DDT is between (2.64, 3.67) with 90% confidence.

Part f.

Since we are using different link functions in each of the models, the only tool we have at our disposal to compare the models is AIC.

```
logistic <- glm(death ~ log(dosage)*as.factor(insecticide),
              family = binomial(link = "logit"),
              data = fulldata)
summary(logistic)$aic
```

```
## [1] 827.6443
```

```
probit <- glm(death ~ log(dosage)*as.factor(insecticide),
              family = binomial(link = "probit"),
              data = fulldata)
summary(probit)$aic
```

```
## [1] 827.0653
```

```
# Can't get loglog model to work directly, so calculate it by taking the complement of the death data
fulldata$c_death <- ifelse(fulldata$death == 0, 1, 0)
loglog <- glm(c_death ~ log(dosage)*as.factor(insecticide),
              family = binomial(link = "cloglog"),
              data = fulldata)
summary(loglog)$aic
```

```
## [1] 822.0261
```

```
cloglog <- glm(death ~ log(dosage) + as.factor(insecticide),
              family = binomial(link = "cloglog"),
              data = fulldata)
summary(cloglog)$aic
```

```
## [1] 834.2001
```

Using AIC to compare betweent the link functions, we see that the loglog link function has the lowest AIC and therefore performs the "best" out of all these suggested models.

Reusing this model for Fieller's method, we get the following confidence intervals and estimates:

```
fieller_model <- glm(c_death ~ as.factor(insecticide) + log(dosage) - 1,
                     data = fulldata,
                     family = binomial(link = "cloglog"))
betas <- summary(fieller_model)$coe[,1]
vcov(fieller_model)
```

```
##                                      as.factor(insecticide)DDT
## as.factor(insecticide)DDT                           0.03937306
## as.factor(insecticide)DDT_gamma_BHC                 0.02583802
## as.factor(insecticide)gamma_BHC                     0.02979461
## log(dosage)                                        -0.02447640
##                                      as.factor(insecticide)DDT_gamma_BHC
## as.factor(insecticide)DDT                                     0.02583802
## as.factor(insecticide)DDT_gamma_BHC                           0.04407854
## as.factor(insecticide)gamma_BHC                               0.02439780
```

```
## log(dosage)                                                  -0.02004290
##                                     as.factor(insecticide)gamma_BHC
## as.factor(insecticide)DDT                                    0.02979461
## as.factor(insecticide)DDT_gamma_BHC                          0.02439780
## as.factor(insecticide)gamma_BHC                              0.03580940
## log(dosage)                                                 -0.02311208
##                                     log(dosage)
## as.factor(insecticide)DDT           -0.02447640
## as.factor(insecticide)DDT_gamma_BHC -0.02004290
## as.factor(insecticide)gamma_BHC     -0.02311208
## log(dosage)                          0.01898667
```

$$V_1 = Cov(\hat{\beta}_0 - \hat{\beta}_1, \beta_4) = \begin{pmatrix} .0228 & .0089 \\ .0089 & .0181 \end{pmatrix}$$

$$V_2 = Cov(\hat{\beta}_3 - \hat{\beta}_1, \beta_4) = \begin{pmatrix} .0195 & .0078 \\ .0172 & .0181 \end{pmatrix}$$

This gives us the relative potency and confidence interval between DDT and the mixed insecticide by plugging into the formula:

$$\frac{(\hat{\beta}_0 - \hat{\beta}_1 - \beta_4\theta)^2}{V_1(1,1) - 2\theta V_1(1,2) + \theta^2 V_1(2,2)} = 1.645^2$$

The relative potency of DDT vs the mixed insecticide is 3.31 and the 90% confidence interval is 2.77 to 3.84. In interpretable terms, this means that the mixed insecticide is between 2.77 and 3.84 more potent than DDT after accounting for log(dosage).

Similarly, we can compute the potency between γ-BHC and the mixed insecticde. We get that the relative potencyis 2.41 with a 90% confidence interval between 2.08 and 2.744. Again, in interpretable terms, this means that the mixed insecticide is between 2.08 and 2.744 more potent than γ-BHC after accounting for log(dosage).

Part g.

To estimate the dosage required for a 99% kill rate in the mixed insecticde, we have to solve

$$\beta_2 + \beta_4 log(dosage) = log\left(\frac{.99}{.01}\right)$$

Plugging in our estimates for β and solving, we get that $\hat{x} = 7.26$ mg/10cm3.

Using the fact that $-log(dosage) = -(\beta_2 - logit(.99))/\beta_4$, we use Fieller's method to get a 90% confidence interval for this dosage.

$$V = Cov(\hat{\beta}_2, \beta_4) = \begin{pmatrix} .349 & -.322 \\ -.322 & .329 \end{pmatrix}$$

Plugging this into

$$\frac{(logit(.99) - \hat{\beta}_2 - \hat{\beta}_4\theta)^2}{V(1,1) - 2\theta V(1,2) + \theta^2 V(2,2)} = 1.645^2$$

We get a 90% confidence interval of 3.79 to 22.87. We can interpret this as saying that the dosage of mixed insecticide that will give the 99% kill rate is between 3.79 and 22.87 mg/10cm3.

Part h.

The significant terms in the model are just the main factors effects of the insecticide type and the log(dosage). There is no evidence of nonparallelism in the model, meaning we can ignore the interaction terms between log(dosage) and the insecticdes.

Using a loglog link function provides a slightly better fit than the more standard probit or logit link functions.

The mixed insecticde is more potent than either of its components with appoximately 2.45 times more potency than γ-BHC and 3.08 times more than DDT.