Analysis on Pesticides data set

Suhas Shastry

November 7, 2018

Background

Data set is about an experiment performed with 4 different organophosphorus pesticides and their mixtures. These pesticides are acephate (A), diazinon (B), chlorpyrifos (C) and dimethoate (D). Furthermore, two mixtures E and F are considered where the mixing ratio for E is (0.045 : 0.002 : 0.035 : 0.918) for (A: B: C: D) respectively and the mixing ratio for F is (0.229 : 0.011 : 0.177 : 0.583) for (A: B: C: D) respectively. The experiment is performed with different amount of doses, and the number of dead pests among the total number of pests present are reported. Below are the top 6 rows of the data.

pesticide	amount of dose	# of dead pests	# of pests present
1	0	0	8
1	3	0	8
1	10	0	8
1	30	5	8
1	60	8	8
1	120	8	8

Goal of this project to fit a logistic regression model for kill rate and dose of the pesticide. Probit and c log-log regressions are also explored.

Analysis

Before fitting any model, graphical representation of relationship between the dose and the kill rate are shown below. These pictures will help us decide on the type of relation that exist between dose and kill rate.

From these pictures, relationship between dose and kill rate looks like curved 'S'. Hence logistic regression would be a good fit.

A linear logistic regression model was fit on *kill rate* against $\log(1 + dose)$ and pesticide A, B, C and D. If β_1, β_2, β_3 and β_4 are the effects of pesticides A, B, C and D respectively, then the effect of E is

$$0.045\beta_1 + 0.002\beta_2 + 0.035\beta_3 + 0.918\beta_4$$

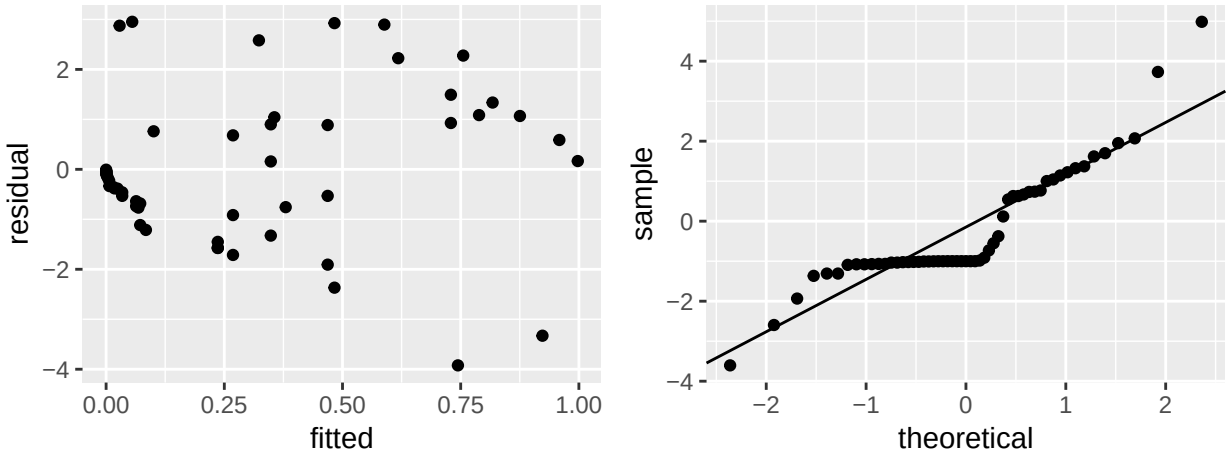
and the effect of F is

$$0.229\beta_1 + 0.011\beta_2 + 0.177\beta_3 + 0.583\beta_4$$

Below is the summary of the logistic model

```
##
## Call:
## glm(formula = rate ~ log(1 + dose) + A + B + C, family = binomial(link = "logit"),
##     data = df, weights = data1$m)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5795  -1.0246  -0.1662   0.8785   3.5668
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.44703    0.92354  -9.146   <2e-16 ***
## log(1 + dose)  2.40042    0.24876   9.650   <2e-16 ***
## A              0.07846    0.42178   0.186    0.852
## B              0.13612    0.36902   0.369    0.712
## C             -4.33559    0.51697  -8.386   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 493.06  on 56  degrees of freedom
## Residual deviance: 144.26  on 52  degrees of freedom
## AIC: 204.34
##
## Number of Fisher Scoring iterations: 6
```

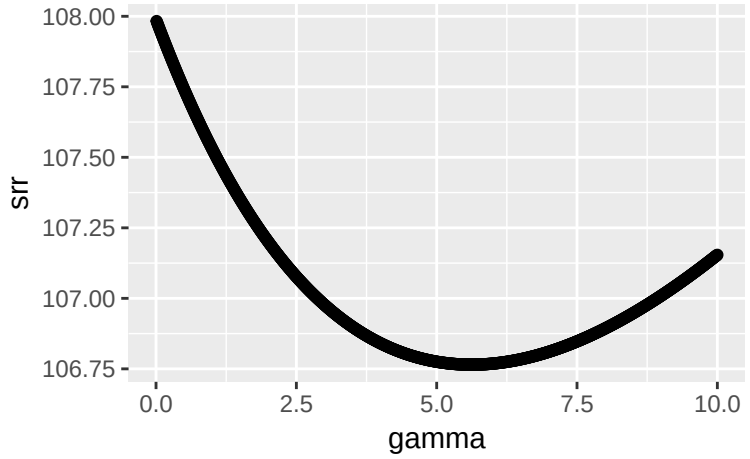
Once the regression is fit, model is evaluated using Pearson residuals plots.



There are 2 outliers in the residual. The above plots are excluding those residuals. There is a visible curved pattern in residuals. Constant variance assumption seems to be violated.

Hypothesis of Parallelism

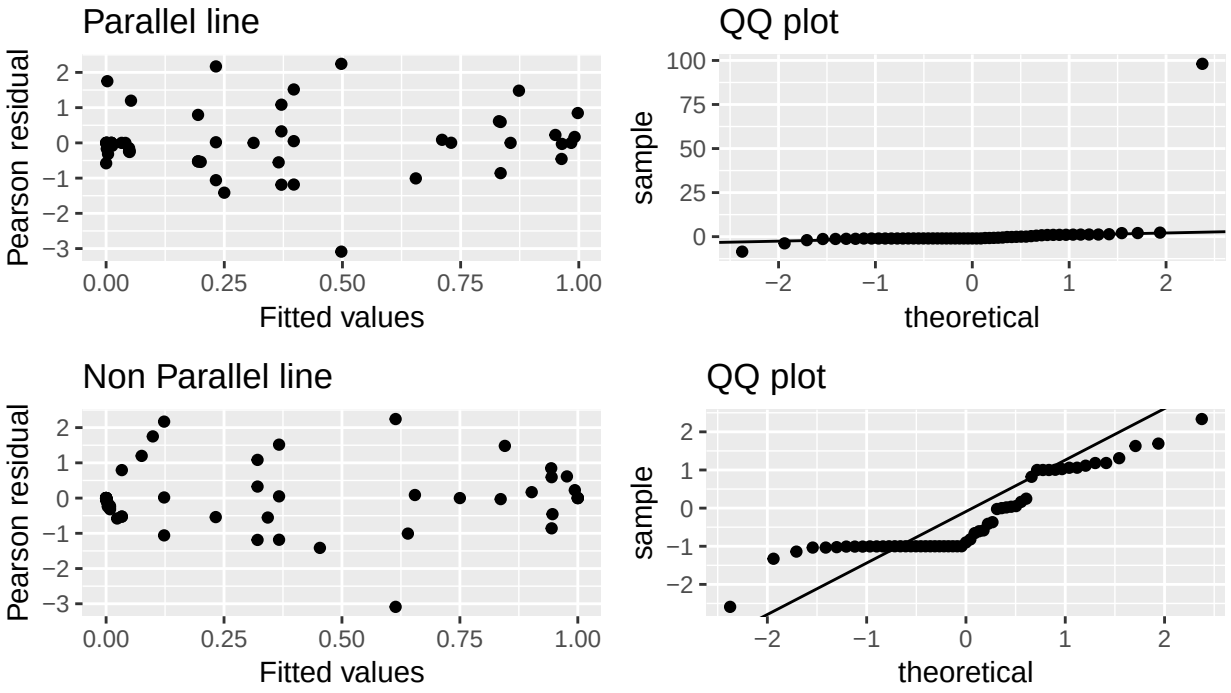
Two models, one in which the relationship is described by six parallel straight lines in the log dose model and one in which the six lines are straight but not parallel model are tested for best fit under the assumption that substances combine additively. Before comparing these models, instead of $\log(1 + dose)$, $\log(\gamma + dose)$ which minimizes residual deviance is obtained by plotting residual deviance vs gamma.



Residual deviance is minimized at $\gamma = 5.62$. Hence $\log(5.62 + dose)$ will be our new explanatory variable.

$H_0 : Model\ is\ parallel$ vs $H_A : Model\ is\ not\ parallel$

Two different models were fit for above hypothesis one with interaction and the other without interaction term. Residual plot and QQ plot are drawn for both the models.



Below is the comparison of two models using Analysis of Deviance.

```
## Analysis of Deviance Table
##
## Model 1: rate ~ log(gamm + dose) + as.factor(pest)
## Model 2: rate ~ log(gamm + dose) * as.factor(pest)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        50    106.765
## 2        45     48.908  5   57.857 3.367e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p value is very small, we reject null hypothesis and go with non-parallel model.

Effect of E and F

We test the hypothesis that the effects of E and F are greater than the effects when substances combine additively.

$$H_{01} : \beta_5 = 0.045\beta_0 + 0.002\beta_2 + 0.035\beta_3 + 0.918\beta_4$$

$$H_{A1} : \beta_5 > 0.045\beta_0 + 0.002\beta_2 + 0.035\beta_3 + 0.918\beta_4$$

and

$$H_{02} : \beta_6 = 0.229\beta_0 + 0.011\beta_2 + 0.177\beta_3 + 0.583\beta_4$$

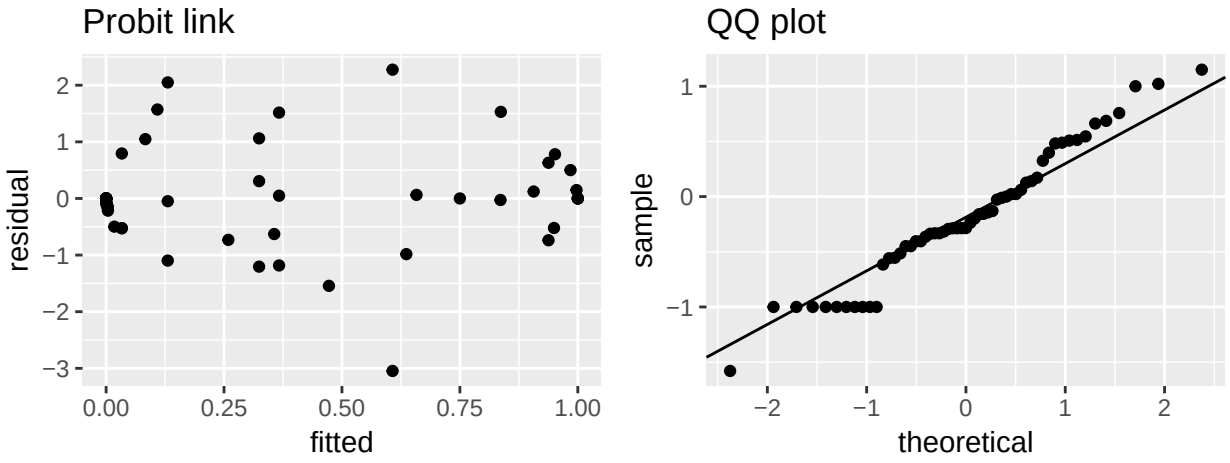
$$H_{A2} : \beta_6 > 0.229\beta_0 + 0.011\beta_2 + 0.177\beta_3 + 0.583\beta_4$$

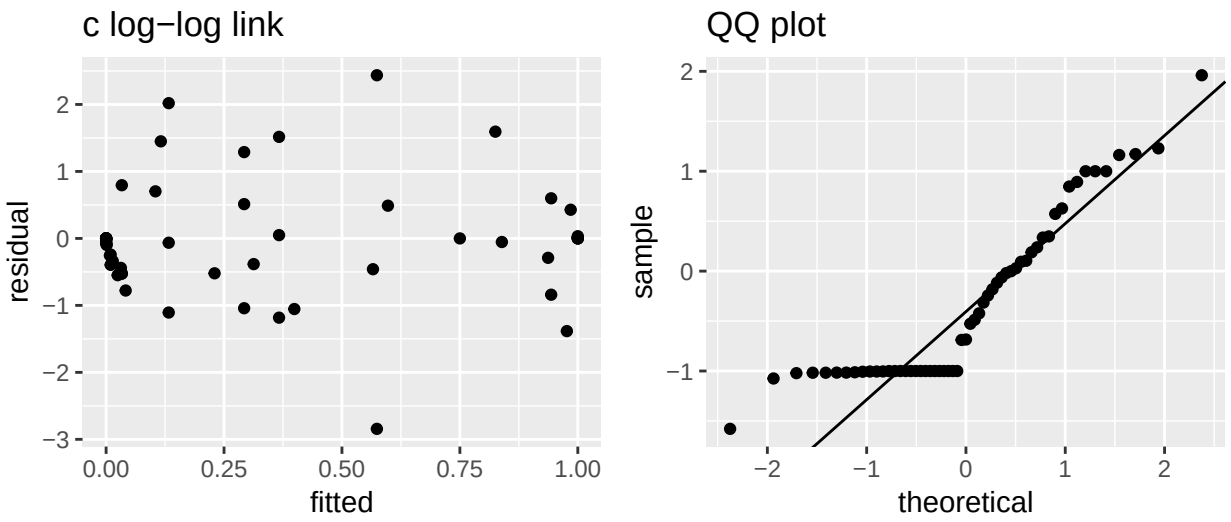
test	test_statistic	critical_value	p_value
1	28.5127	1.795885	1
2	80.8006	1.795885	1

We reject null in both the cases and conclude that effects of E and F are greater than the effects when substances combine additively,

Alternative models

Other link functions like probit link function and complementary log-log function are fitted for pesticides data.





AIC for logit model is 122.988 and for Probit model is 122.053. For c-loglog model AIC is 121.845. Of all the models, c-loglog model fares better.

Conclusion

Pest kill rate is regressed on $\log(\gamma + dose)$ and its interaction 6 different types of pesticide. Logistic regression with 6 lines model is a good fit for pesticides data. Among all the pesitcides, E is more effective than the rest.