Assignment 5 — Due Nov 27, 2018

NAIQING CAI
ncai5@wisc.edu

1. Eight isolates of rose blackspot fungus

(a) Consider an additive model Y ~ Isolate + Temp. Complete the decomposition of the total sum of squares:

By running r function, we can get the table as followed:

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
isolate    7 0.34814 0.04973  4.2338  0.001297 **
temp       6 2.72335 0.45389 38.6397 1.467e-15 ***
Residuals 42 0.49336 0.01175
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.1 Analysis of Variance Table

SOURCE	SS	DF	MS	F	P-VALUE
ISOLATE	0.35	7	0.0497	4.2338	0.001
TEMP	2.72	6	0.4539	38.6397	1.467e-15
RESIDUAL	0.49	42	0.01175		

Table 1.2 Anova table for model Y ~ Isolate + Temp

(b) Now consider a linear model where the covariates are polynomials of temperature.

First, fit the linear regression model:

lm0=lm(Y~1)

lm1=lm(Y~1+x)

lm2=lm(Y~1+x+x2)

lm3=lm(Y~x+x2+x3)

lm4=lm(Y~x+x2+x3+x4)

lm6=lm(Y~x+x2+x3+x4+x5+x6)

Then by using anova on either two linear regression models, we can get the table.

(P1|1)=anova(lm0,lm1)

(P2|P1)=anova(lm1,lm2)

(P3|P2)=anova(lm2,lm3)

(P4|P3)=anova(lm3,lm4)

(TEMP|P4)=anova(lm4,lm6)

(TEMP|P3)=anova(lm3,lm6)

SOURCE	SS	DF	MS	F	P-VALUE
P1|1	0.0117	1	0.0117	0.1781	0.6747
P2|P1	2.56781	1	2.56781	138.12	< 2.2e-16
P3|P2	0.09992	1	0.09992	5.8683	0.01894
P4|P3	0.01345	1	0.01345	0.7864	0.3793
TEMP(P6)|P4	0.03045	2	0.01523	0.8867	0.4185
TEMP|P3	0.0439	3	0.01463	0.8521	0.4722

Table 1.3 Anova table for linear model

(c) What does the preceding table tell you about the effect of temperature on growth? Estimate the temperature at which the growth rate is a maximum.

Based the preceding table, we can find that the model: Y~x+x^2+x^3 is the best regression model since its p-value is small.

By optimize() function, we can calculate that temperature =71.74756, the growth rate is a maximum reached 1.05296

2. Flushot

(a) Fit a multiple logistic regression model with the three explanatory variables by maximum likelihood estimation. Report the summary table from the R output. Obtain and interpret the maximum likelihood estimates of β0, β1, β2, and β3. State the fitted logistic response function.

Multiple Logistic Regression Model

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

X1: age

X2: health awareness index, for which higher values indicate greater awareness

X3: gender, where males were coded X3 = 1 and females were coded X3 = 0

R summary table

```
Call:
glm(formula = y ~ x1 + x2 + x3, family = binomial("logit"), data = b)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4037  -0.5637  -0.3352  -0.1542   2.9394

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
x1           0.07279    0.03038   2.396  0.01658 *
x2          -0.09899    0.03348  -2.957  0.00311 **
x3           0.43397    0.52179   0.832  0.40558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.09  on 155  degrees of freedom
AIC: 113.09

Number of Fisher Scoring iterations: 6
```

Table 2.1 R summary table

Interpretation

β0=-1.17716, which means the average odds of a person having a flu shot is -1.17716..

β1=0.07279, representing the odds of a person having a flu shot increase by about 7.3% with each unit increase in age.

β2=-0.09899, representing the odds of a person having a flu shot decrease by about 9.9% with each unit increase in awareness index.

β3=0.43397, representing the difference between the log odds for male and the log odds for female is 0.43397.

Fitted Logistic Response Function

$$\log(\frac{p_i}{1-p_i}) = -1.177 + 0.07279 X_1 - 0.09899 X_2 + 0.43397 X_3$$

(b) Obtain exp(β1), exp(β2), and exp(β3) and the respective 95% CI. Interpret the results.

		2.5 %	97.5 %
(Intercept)	0.308	0.001	106.515
x1	1.076	1.013	1.141
x2	0.906	0.848	0.967
x3	1.543	0.555	4.292

Table 2.2 Result of Confidence Interval

Interpretation

exp(β1)=1.076, representing estimated odds ratio for X1
95% CI for exp(β1) is (1.013,1.141)

exp(β2)=0.906, representing estimated odds ratio for X2
95% CI for exp(β2) is (0.848,0.967)

exp(β3)=1.543, representing estimated odds ratio for X3
95% CI for exp(β3) is (0.555,4.292)

(c) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? Provide a 95% CI. Interpret the results.

Estimated Probability

$$X_h'\hat{\beta} = -2.679033$$

$$s\{X_h'\hat{\beta}\} = 0.5085794$$

$$\hat{\pi}_h = \frac{\exp(X_h'\hat{\beta})}{1+\exp(X_h'\hat{\beta})} 0.06422197$$

The result shows that male clients aged 55 with a health awareness index of 60 receive a flu shot at the probability 6.4%.

95% CI

$$L = X_h'\hat{\beta} - z_{1-\alpha/2}s\{X_h'\hat{\beta}\} = -3.67583$$

$$U = X_h'\hat{\beta} + z_{1-\alpha/2}s\{X_h'\hat{\beta}\} = -1.682236$$

$$L^* = \{1+\exp(-L)\}^{-1} = 0.0247027$$

$$U^* = \{1+\exp(-U)\}^{-1} = 0.1567997$$

$$\Rightarrow CI \ \ for \ \pi_h \ \ is \ (0.0247, 0.1568)$$

The result shows that male clients aged 55 with a health awareness index of 60 have 95% probability to receive a flu shot at the probability between 0.0247 and 0.1568.

(d) Perform a **Wald test** to determine whether X3, client gender, can be dropped from the regression model.

Multiple Logistic Regression Model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Hypothesis:

$$H_0 : \beta_3 = 0 \ v.s \ H_1 : \beta_3 \neq 0$$

Wald test Statistic:

$$z^* = \frac{\hat{\beta}_3 - \beta_3}{s\{\hat{\beta}_3\}} \sim N(0,1)$$

$$z^* = 0.832 < z_{1-\alpha/2} = 1.96$$

Conclusion:

We accept null hypothesis, that is β3=0 and thus X3 can be dropped from the regression model.

(e) Perform a **likelihood ratio test** to determine whether X3, client gender, can be dropped from the regression model.

Multiple Logistic Regression Model:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Hypothesis:

$$H_0 : \beta_3 = 0 \ v.s \ H_1 : \beta_3 \neq 0$$

Likelihood ratio test Statistic:

$$G^2 = -2\log\{\frac{L(R)}{L(F)}\} = -2\{\log(L(R)) - \log(L(F))\} \sim \chi^2_{df_R - df_F}$$

```
Analysis of Deviance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + x3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       156     105.80
2       155     105.09  1  0.70221    0.402
```

Table 2.3 Likelihood Radio Test

$$G^2 = 0.70221 < \chi^2_{p-q,1-\alpha} = 3.84146$$

Thus, based on the rejection rule, we should reject null hypothesis and beta3=0 could be dropped from the model.

(f) Perform stepwise model selection based on AIC. Report the summary table for your final model.

(1) Forward stepwise+ AIC

Step 1

```
Start:  AIC=136.94
y ~ 1

       Df Deviance    AIC
+ x2    1   113.20 117.20
+ x1    1   116.27 120.27
+ x3    1   132.88 136.88
<none>      134.94 136.94
```

Model start with y~1. Based on the step one, AIC=136.94. These three variables AIC are all greater than 136.94, but we should first add x2 into the model since its AIC is the smallest and thus it has most effect on the model. Then the model becomes y~x2.

Step 2

```
Step:  AIC=117.2
y ~ x2

       Df Deviance    AIC
+ x1    1   105.80 111.80
+ x3    1   111.19 117.19
<none>      113.20 117.20
```

After add x2 as a variable into the model, based on the step two, AIC=117.2. We should then add x1 into the model since its AIC is smaller than 117.2 and it is the smallest and thus it has most effect on the model. Then the model becomes y~x2+x1.

Step 3

```
Step:  AIC=111.8
y ~ x2 + x1

       Df Deviance    AIC
<none>      105.80 111.80
+ x3    1   105.09 113.09

Call:  glm(formula = y ~ x2 + x1, family = binomial("logit"), data = b)

Coefficients:
(Intercept)           x2           x1
   -1.45778     -0.09547      0.07787

Degrees of Freedom: 158 Total (i.e. Null);  156 Residual
Null Deviance:      134.9
Residual Deviance: 105.8        AIC: 111.8
```

Now, the model becomes y~x2+x1, and then after step three, AIC=111.8 and we find that x3's AIC is greater than 111.8, so it does not have much effect on the model. We will not add it into the model.

Finally, we get the model

$$y = -1.458 - 0.0955 * x_2 + 0.078 * x_1$$

(2) Backward stepwise+ AIC

Step 1

```
Start:  AIC=113.09
y ~ x1 + x2 + x3

       Df Deviance    AIC
- x3    1   105.80 111.80
<none>      105.09 113.09
- x1    1   111.19 117.19
- x2    1   115.80 121.80
```

Model start with y~x1+x2+x3. Based on the step one, AIC=113.09. x3's AIC is 111.8 which is smaller than 113.09, so we should first remove x3 from the model since its AIC is the smallest and thus it has least effect on the model. Then the model becomes y~x1+x2.

Step 2

```
Step:  AIC=111.8
y ~ x1 + x2

       Df Deviance    AIC
<none>      105.80 111.80
- x1    1   113.20 117.20
- x2    1   116.27 120.27

Call:  glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = b)

Coefficients:
(Intercept)           x1           x2
   -1.45778      0.07787     -0.09547

Degrees of Freedom: 158 Total (i.e. Null);  156 Residual
Null Deviance:      134.9
Residual Deviance: 105.8        AIC: 111.8
```

Now, the model is y~x1+x2, and AIC=111.8. x1 and x2's AIC are both greater than 111.8, as they both don't need to be removed from the model. And the model is still y~x1+x2.

Finally, we get the model

$$y = -1.458 - 0.0955 * x_2 + 0.078 * x_1$$

(3) Both direction selection+ AIC

Step 1

```
Start:  AIC=136.94
y ~ 1

       Df Deviance    AIC
+ x2    1   113.20 117.20
+ x1    1   116.27 120.27
+ x3    1   132.88 136.88
<none>      134.94 136.94
```

Based on the step one, AIC=136.94. These three variables AIC are all greater than 136.94, but we should first add x2 into the model since its AIC is the smallest and thus it has most effect on the model. Then the model becomes y~x2.

Step 2

```
Step:  AIC=117.2
y ~ x2

       Df Deviance    AIC
+ x1    1   105.80 111.80
+ x3    1   111.19 117.19
<none>      113.20 117.20
- x2    1   134.94 136.94
```

After add x2 as a variable into the model, based on the step two, AIC=117.2. We should then add x1 into the model since its AIC is smaller than 117.2 and it is the smallest and thus it has most effect on the model. Then the model becomes y~x2+x1.

Step 3

```
Step:  AIC=111.8
y ~ x2 + x1

       Df Deviance    AIC
<none>      105.80 111.80
+ x3    1   105.09 113.09
- x1    1   113.20 117.20
- x2    1   116.27 120.27

Call:  glm(formula = y ~ x2 + x1, family = binomial("logit"), data = b)

Coefficients:
(Intercept)           x2           x1
   -1.45778     -0.09547      0.07787

Degrees of Freedom: 158 Total (i.e. Null);  156 Residual
Null Deviance:      134.9
Residual Deviance: 105.8        AIC: 111.8
```

Now, the model becomes y~x2+x1, and then after step three, AIC=111.8 and we find that x3's AIC is greater than 111.8, so it does not have much effect on the model. We will not add it into the model.

Finally, we get the model

```
Call:  glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = b)

Coefficients:
(Intercept)           x1           x2
   -1.45778      0.07787     -0.09547

Degrees of Freedom: 158 Total (i.e. Null);  156 Residual
Null Deviance:      134.9
Residual Deviance: 105.8        AIC: 111.8
```

Table 2.4 R summary table for final model

$$y = -1.458 - 0.0955 * x_2 + 0.078 * x_1$$

(g) Create an ROC plot for the final model. Interpret the results.

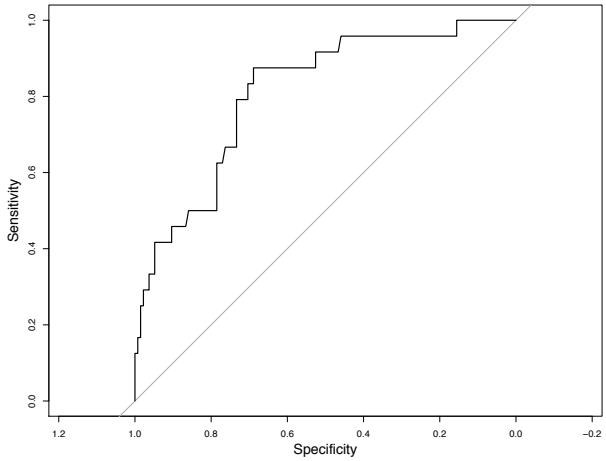

Figure 2.5 ROC Plot

Area under the curve: 0.8094

The result shows excellent discrimination, which means that the probability that predictions and the outcomes are concordant is 80.94%.

3. In a clinical trial, m subjects are assigned to each of the treatment group and the control group. In the treatment group, y1 of the m subjects have positive response, while in the control group, y2 subjects have positive response. We are interested in estimating the treatment effect and providing its confidence interval.

(a) Present the above problem as a logistic regression problem and give the interpretation of the regression coefficients β.

Y	control group(Xi=0)	treatment group(Xi=1)
# of positive response	y2	y1
# of negative response	m-y2	m-y1
total	m	m

$Y_i: \ reaction \ of \ ith \ subject$

$$Y_i = \begin{cases} 1, & ith \ subject \ has \ positive \ response \\ 0, & ith \ subject \ has \ negative \ response \end{cases}$$

$$Y_i \sim Bernoulli(p_i)$$

$y_1,...,y_m$ belongs to control group, $y_{m+1},...,y_{2m}$ belongs to treatment group

$$\sum_{i=1}^{2m} y_i = y_1 + y_2$$

$$\sum_{i=m+1}^{2m} y_i = y_1$$

$$\sum_{i=1}^{m} y_i = y_2$$

$$X_i = \begin{cases} 1, & ith \ subject \ belong \ to \ treatment \ group \\ 0, & ith \ subject \ belong \ to \ control \ group \end{cases}$$

Logistic Regression Model

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i}$$

Interpretation of the regression coefficients β

$$\beta = (\beta_0, \beta_1)'$$

β0: representing the average odds of of response variable corresponding to the random samples from control group.

β1: representing the difference between the log odds for subjects from treatment group and the log odds for subjects from control group.

(b) Express the likelihood function as a function of β, and derive the MLE β^.

Likelihood Function

$$\Pr(Y_i = y_i; p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = (\frac{p_i}{1-p_i})^{y_i}(1-p_i)$$

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 X_{1i}}$$

$$1-p_i = \frac{1}{1+e^{\beta_0+\beta_1 X_{1i}}}$$

$$\Pr(Y_i = y_i; \beta_0, \beta_1) = (e^{\beta_0+\beta_1 X_{1i}})^{y_i}\frac{1}{1+e^{\beta_0+\beta_1 X_{1i}}}$$

$$L(Y;\beta_0,\beta_1) = \prod_{i=1}^{2m}\Pr(Y_i = y_i;\beta_0,\beta_1) = \prod_{i=1}^{2m}(e^{\beta_0+\beta_1 X_{1i}})^{y_i}\frac{1}{1+e^{\beta_0+\beta_1 X_{1i}}}$$

MLE β^

$$X_i = 0, p_i = \frac{e^{\beta_0}}{1+e^{\beta_0}} \triangleq p_c$$

$$X_i = 1, p_i = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \triangleq p_t$$

$$l(Y;\beta_0,\beta_1) = \log L(Y;\beta_0,\beta_1) = \sum_{i=1}^{2m}(\beta_0+\beta_1 X_{1i})y_i - \log\prod_{i=1}^{2m}(1+e^{\beta_0+\beta_1 X_{1i}})$$

$$\frac{\partial l(Y;\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{2m} y_i - \sum_{i=1}^{2m}\frac{e^{\beta_0+\beta_1 X_{1i}}}{1+e^{\beta_0+\beta_1 X_{1i}}} = 0$$

$$\frac{\partial l(Y;\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{2m} X_{1i}y_i - \sum_{i=1}^{2m}\frac{e^{\beta_0+\beta_1 X_{1i}}}{1+e^{\beta_0+\beta_1 X_{1i}}}X_{1i} = 0$$

$$\Rightarrow \begin{cases} \hat{p}_t = \frac{y_1}{m} \\ \hat{p}_c = \frac{y_2}{m} \end{cases} \Rightarrow \begin{cases} \hat{\beta}_0 = \log(\frac{y_2}{m-y_2}) \\ \hat{\beta}_1 = \log(\frac{y_1(m-y_2)}{y_2(m-y_1)}) \end{cases}$$

(c) Find the asymptotic variance-covariance matrix of β^.

$$H_{2m}(\beta_0,\beta_1)=\begin{bmatrix} \frac{\partial^2 l}{\partial\beta_0^2} & \frac{\partial^2 l}{\partial\beta_0\partial\beta_1} \\ \frac{\partial^2 l}{\partial\beta_1\partial\beta_0} & \frac{\partial^2 l}{\partial\beta_1^2} \end{bmatrix}=\begin{bmatrix} -m\left(\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2}+\frac{e^{\beta_0}}{(1+e^{\beta_0})^2}\right) & -m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} \\ -m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} & -m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} \end{bmatrix}$$

$$\sqrt{n}\left[\begin{pmatrix}\hat\beta_0\\ \hat\beta_1\end{pmatrix}-\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}\right]\xrightarrow{d}N(0,\mathrm{I}_1(\beta_0,\beta_1)^{-1}),\ where\ \mathrm{I}_1=\frac{1}{n}\mathrm{I}_n\ \ and\ \ n=2m$$

$$\mathrm{I}_n(\beta_0,\beta_1)=-E(H_n(\beta_0,\beta_1))$$

$$Asymptotic\ variance-covariance\ matrix\ of\ \hat\beta\ \text{is}\ \text{-}H_{2m}^{-1}(\hat\beta)$$

$$-H^{-1}{}_{2m}(\beta_0,\beta_1)=\frac{(1+e^{\beta_0+\beta_1})^2(1+e^{\beta_0})^2}{m^2e^{2\beta_0+\beta_1}}\begin{bmatrix} m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} & -m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} \\ -m\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2} & m\left(\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2}+\frac{e^{\beta_0}}{(1+e^{\beta_0})^2}\right) \end{bmatrix}$$

(d) Suppose that m = 20, y1 = 12, and y2 = 9. Give the point estimate and the corresponding 95% confidence interval for the odds ratio of having positive response between the treatment group and the control group. Can you conclude that the treatment is effective?

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1X_{1i}$$

$$\hat\beta_0=\log\left(\frac{y_2}{m-y_2}\right)=\log\frac{9}{11}=-0.2007\qquad\hat\beta_1=\log\left(\frac{y_1(m-y_2)}{y_2(m-y_1)}\right)=\log\frac{11}{6}=0.6061$$

$$\hat{OR}=\exp(\hat\beta_1)=\frac{11}{6}\ \ is\ the\ point\ estimate\ for\ the\ odds\ ratio\ of\ having\ positive\ response\ between\ the\ treatment\ group\ and\ the\ control\ group$$

$$Var(\ln(\hat{OR}))=Var(\hat\beta_1)=\frac{(1+e^{\beta_0+\beta_1})^2(1+e^{\beta_0})^2}{m^2e^{2\beta_0+\beta_1}}m\left(\frac{e^{\beta_0+\beta_1}}{(1+e^{\beta_0+\beta_1})^2}+\frac{e^{\beta_0}}{(1+e^{\beta_0})^2}\right)=0.4104$$

$$\hat{se}(\ln(\hat{OR}))=\hat{se}(\hat\beta_1)=\sqrt{Var(\hat\beta_1)}=0.6406$$

$$95\%\ \text{CI for}\ \hat\beta_1\ \ is\ \ \hat\beta_1\pm z_{1-\alpha/2}\hat{se}(\hat\beta_1)=(-0.6495,1.8617)$$

$$95\%\ \text{CI for}\ \hat{OR}\ \ is\ \ e^{\hat\beta_1\pm z_{1-\alpha/2}\hat{se}(\hat\beta_1)}=(0.5223,6.4347)$$

From the result we can conclude that the treatment is not effective.

4. Flour beetles Tribolium castaneum were sprayed with one of three insecticides in solution at different doses. The number of insects killed after a six-day period is recorded below

(a) Investigate graphically the relationship between the dose, either in original units or in log units, and the kill rate.

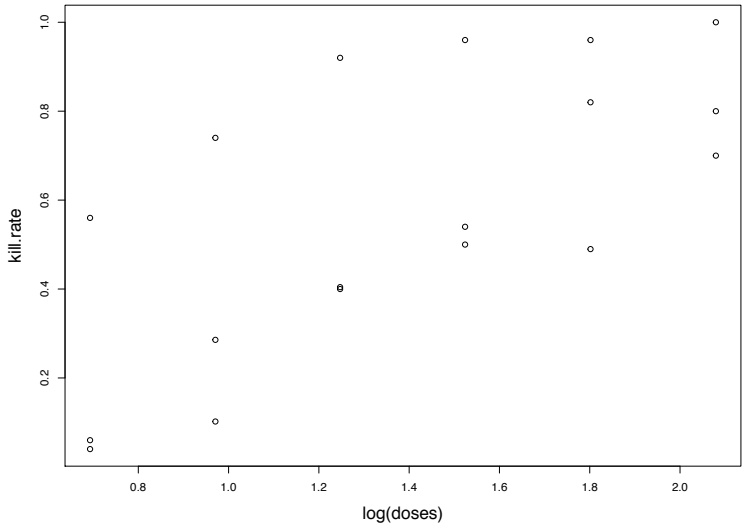

Figure 4.1 Relationship between the doses (log units) and the kill rate

Based on the graph above, the dose and kill rate are positively related.

(b) On the graph for part (a), plot the linear logistic fitted curve for each of the insecticides plus the combination.

I fit the logistic regression and I will have the model:

$$\log\left(\frac{\Pr(Y_i=1\mid X)}{1-\Pr(Y_i=1\mid X)}\right)\triangleq \log it(p_i)=X^T\beta \quad \Leftrightarrow p_i=\frac{e^{X^T\beta}}{1+e^{X^T\beta}}$$

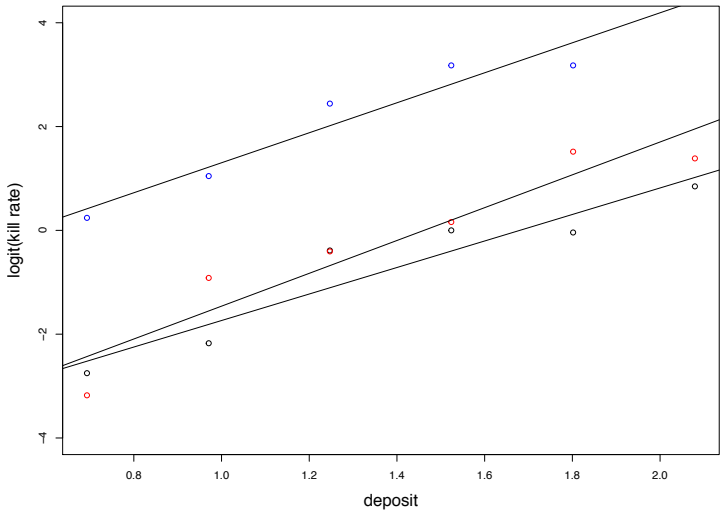

Figure 4.2 Linear Logistic Fitted Curve

(c) Consider the two models, one in which the relationship is described by three parallel straight lines in the log dose and and one in which the three lines are straight but not parallel. Assess the evidence against the hypothesis of parallelism.

Model one:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

$$X_{1i} = \begin{cases} 1, \ if \ \ DDT \ \ insecticide \\ 0, \qquad\quad otherwise \end{cases}$$

$$X_{2i} = \begin{cases} 1, if \ \ \gamma - BHC \ \ insecticide \\ 0, \qquad\quad otherwise \end{cases}$$

$$X_{3i} : \log \ dose$$

Model two:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{3i} + \beta_5 X_{2i} X_{3i}$$

Assess the evidence against the hypothesis of parallelism：

$H_0 : \beta_4 = 0 \ \ v.s \ \ H_1 : \beta_4 \neq 0$ and $H_0 : \beta_5 = 0 \ \ v.s \ \ H_1 : \beta_5 \neq 0$

```
Call:
glm(formula = y ~ log(XX$deposit) + chem$ddt + chem$bhc + log(XX$deposit) *
    chem$ddt + log(XX$deposit) * chem$bhc, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8293  -0.6981   0.1920   0.7678   2.1675

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.1207     0.5908  -3.590 0.000331 ***
log(XX$deposit)            3.3884     0.5733   5.910 3.42e-09 ***
chem$ddt                  -1.7101     0.7741  -2.209 0.027163 *
chem$bhc                  -1.9221     0.7722  -2.489 0.012803 *
log(XX$deposit):chem$ddt  -1.1060     0.6561  -1.686 0.091861 .
log(XX$deposit):chem$bhc  -0.5503     0.6662  -0.826 0.408769
---
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
```

Figure 4.3 Anova summary table for parallelism

Based on the result, we find that p-value is 0.09186 and 0.408769 respectively, they are both >0.025. So we can conclude that we should accept the null hypothesis that beta4=0 and beta5=0 which means three straight lines are parallel.

(d) Let chem denote a 3-level categorical factor, and let ldose be the log dose. Explain the relationship between the regression coefficients in the model formulae chem + ldose and chem + ldose - 1. Explain the relationship between the two covariance matrices.

Model 1: lm.reg1=lm(killrate~chem+ldose)

```
Call:  glm(formula = y ~ log(XX$deposit) + XX$type, family = binomial(link = "logit"))

Coefficients:
    (Intercept)  log(XX$deposit)       XX$typeDDT     XX$typeMIXED
        -3.8397           2.6938          -0.6144           2.4170

Degrees of Freedom: 881 Total (i.e. Null);  878 Residual
Null Deviance:      1203
Residual Deviance: 819  AIC: 827
```

Figure 4.4 model 1 regression

Based on the table, we can find that
regression coefficient for intercept is -3.8397
regression coefficient for chemDDT is -0.6144
regression coefficient for chemMIXED is 2.4170
regression coefficient for deposit is 2.6938

Model 2: lm.reg2=lm(killrate~chem+ldose-1)

```
Call:  glm(formula = y ~ log(XX$deposit) + XX$type - 1, family = binomial(link = "logit"))

Coefficients:
log(XX$deposit)       XX$typeBHC       XX$typeDDT     XX$typeMIXED
          2.694           -3.840           -4.454           -1.423

Degrees of Freedom: 882 Total (i.e. Null);  878 Residual
Null Deviance:      1223
Residual Deviance: 819  AIC: 827
```

Figure 4.5 model 2 regression

Based on the table, we can find that
regression coefficient for chemDDT is -4.454
regression coefficient for chemBHC is -3.840
regression coefficient for chemMIXED is -1.423
regression coefficient for deposit is 2.694

Comparison

Regression coefficient for deposit is the same for two models, and the coefficients for others are different.

ChemDDT in model 2 equal to coefficient for intercept added regression coefficient for chemDDT in model 1.

chemMIXED in model 2 equal to coefficient for intercept added regression coefficient for chemMIXED in model 1.

So, in this sense -1 remove the intercept:

killrate~chem+ldose doesn't have chemBHC but have intercept

killrate~chem+ldose-1 doesn't have intercept but have chemBHC

Covariance Matrix

	(Intercept)	log(XX$deposit)	XX$typeDDT	XX$typeMIXED
(Intercept)	0.10974859	-0.064832766	-0.010956108	-0.04278991
log(XX$deposit)	-0.06483277	0.046061045	-0.005355279	0.01726135
XX$typeDDT	-0.01095611	-0.005355279	0.039941979	0.01648699
XX$typeMIXED	-0.04278991	0.017261346	0.016486990	0.05660964

Figure 4.6 Covariance Matrix for model 1

	log(XX$deposit)	XX$typeBHC	XX$typeDDT	XX$typeMIXED
log(XX$deposit)	0.04606105	-0.06483277	-0.07018804	-0.04757142
XX$typeBHC	-0.06483277	0.10974859	0.09879248	0.06695868
XX$typeDDT	-0.07018804	0.09879248	0.12777835	0.07248956
XX$typeMIXED	-0.04757142	0.06695868	0.07248956	0.08077841

Figure 4.7 Covariance Matrix for model 2

Relationship

Actually, the relationship between the two covariance matrices is a kind of linear transformation.

$$Model1: y \sim \beta_0 + \beta_1 \log(X) + \beta_1 DDT + \beta_3 MIXED$$
$$Model2: y \sim \alpha_0 \log(X) + \alpha_1 BHC + \alpha_2 DDT + \alpha_3 MIXED$$

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = A * (\alpha_0, \alpha_1, \alpha_2, \alpha_3)', where\ A\ is\ a\ transformation\ matrix$$

$$\mathrm{cov}(\beta) = A * \mathrm{cov}(\alpha), where\ A\ is\ a\ transformation\ matrix)$$

(e) On the assumption that three parallel straight lines suffice, estimate the potency of the combination relative to each of the components. Obtain a 90% confidence interval for each of these relative potencies.

Model

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

$$X_{1i} = \begin{cases} 1, if\ DDT\ insecticide \\ 0, \qquad otherwise \end{cases}$$

$$X_{2i} = \begin{cases} 1, if\ \gamma - BHC\ insecticide \\ 0, \qquad otherwise \end{cases}$$

$$X_{3i}: \log\ dose$$

Combination

$$\log(\frac{\Pr(Y_i = 1 \mid X_{1i} = 0, X_{2i} = 0)}{1 - \Pr(Y_i = 1 \mid X_{1i} = 0, X_{2i} = 0)}) = \beta_0 + \beta_3 X_{3i}$$

Components(ddt&bhc)

$$\log(\frac{\Pr(Y_i = 1 \mid X_{1i} = 1, X_{2i} = 0)}{1 - \Pr(Y_i = 1 \mid X_i = 1, X_{2i} = 0)}) = \beta_0 + \beta_1 + \beta_3 X_{3i}$$

$$\log\left(\frac{\Pr(Y_i=1\mid X_{1i}=0,X_{2i}=1)}{1-\Pr(Y_i=1\mid X_i=0,X_{2i}=1)}\right)=\beta_0+\beta_2+\beta_3X_{3i}$$

Potency

$$\log\left(\frac{P_{00}/(1-P_{00})}{P_{10}/(1-P_{10})}\right)=\log(OR1)=-\beta_1$$

$$\log\left(\frac{P_{00}/(1-P_{00})}{P_{01}/(1-P_{01})}\right)=\log(OR2)=-\beta_2$$

So, we need to estimate the coefficient beta1 and beta2.

```
Call:  glm(formula = y ~ log(XX$deposit) + XX$type - 1, family = binomial(link = "logit"))

Coefficients:
log(XX$deposit)       XX$typeBHC       XX$typeDDT     XX$typeMIXED
          2.694           -3.840           -4.454           -1.423

Degrees of Freedom: 882 Total (i.e. Null);  878 Residual
Null Deviance:      1223
Residual Deviance: 819  AIC: 827
```

Figure 4.8 R summary table for general linear logistic regression

By r function, beta1=-4.454, beta2=-3.840.

Potency of the combination relative to DDT is exp(4.454)= 85.97.

Potency of the combination relative to BHC is exp(3.840)= 46.5255.

90% CI

90% CI for beta1 (-5.0582, -3.8812)

90% CI for beta2 (-4.3977, -3.3070)

90% CI for potency of the combination relative to DDT is (48.4825, 157.3024)

90% CI for potency of the combination relative to BHC is (27.302, 81.264)

(f) Check to see if one of the alternative link functions probit, c-log log or log log, gives an appreciably better fit. Give the answer to part (e) for the c-log log model.

1) link =probit

```
> glm(y~log(XX$deposit)+XX$type,family=binomial(link="probit"))

Call:  glm(formula = y ~ log(XX$deposit) + XX$type, family = binomial(link = "probit"))

Coefficients:
    (Intercept)  log(XX$deposit)       XX$typeDDT     XX$typeMIXED
        -2.2734           1.5902           -0.3535           1.4304

Degrees of Freedom: 881 Total (i.e. Null);  878 Residual
Null Deviance:      1203
Residual Deviance: 817  AIC: 825
> glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="probit"))

Call:  glm(formula = y ~ log(XX$deposit) + XX$type - 1, family = binomial(link = "probit"))

Coefficients:
log(XX$deposit)       XX$typeBHC       XX$typeDDT     XX$typeMIXED
         1.5902           -2.2734           -2.6269           -0.8431

Degrees of Freedom: 882 Total (i.e. Null);  878 Residual
Null Deviance:      1223
Residual Deviance: 817  AIC: 825
```

Figure 4.9 R summary table for link=probit

When link =probit, AIC=825 which is smaller than link=logit AIC=827.

So link function probit, gives an appreciably better fit.

2) link= c-log log

```
> glm(y~log(XX$deposit)+XX$type,family=binomial(link="cloglog"))

Call:  glm(formula = y ~ log(XX$deposit) + XX$type, family = binomial(link = "cloglog"))

Coefficients:
    (Intercept)  log(XX$deposit)       XX$typeDDT     XX$typeMIXED
        -2.8115           1.6655          -0.4242           1.4376

Degrees of Freedom: 881 Total (i.e. Null);  878 Residual
Null Deviance:      1203
Residual Deviance: 826.2        AIC: 834.2
> glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="cloglog"))

Call:  glm(formula = y ~ log(XX$deposit) + XX$type - 1, family = binomial(link = "cloglog"))

Coefficients:
log(XX$deposit)       XX$typeBHC       XX$typeDDT     XX$typeMIXED
          1.665           -2.811           -3.236           -1.374

Degrees of Freedom: 882 Total (i.e. Null);  878 Residual
Null Deviance:      1216
Residual Deviance: 826.2        AIC: 834.2
```

Figure 4.10 R summary table for link=cloglog

When link =cloglog, AIC=834.2 which is larger than link=logit AIC=827.

So link function cloglog, gives an appreciably better fit.

Answer to part (e) for the c-log log model

By r function, beta1=-3.236, beta2=-2.811.

Potency of the combination relative to DDT is exp(3.236)= 25.4318.

Potency of the combination relative to BHC is exp(2.811)= 16.6265.

90% CI

90% CI for beta1 (-3.630462, -2.856884)

90% CI for beta2 (-3.185522, -2.452735)

90% CI for potency of the combination relative to DDT is (17.4072, 37.73024)

90% CI for potency of the combination relative to BHC is (11.62008, 24.1799)

(g) Under the linear logistic model, estimate the combination dose required to give a 99% kill rate, and obtain a 90% confidence interval for this dose.

```
Call:  glm(formula = y ~ log(XX$deposit) + XX$type, family = binomial(link = "logit"))

Coefficients:
    (Intercept)  log(XX$deposit)       XX$typeDDT     XX$typeMIXED
        -3.8397           2.6938          -0.6144           2.4170

Degrees of Freedom: 881 Total (i.e. Null);  878 Residual
Null Deviance:      1203
Residual Deviance: 819  AIC: 827
```

Figure 4.11 Linear Logistic Model

$$\log it(p_i) = -8397+2.6938 \text{ *ldose } -0.6144*\text{DDT}+2.4170*\text{MIXED}$$

When give a 99% kill rate, the combination ldose=2.2339, dose=9.3367

90% confidence interval for this dose is (5.2834, 19.3627)

(h) Give a brief summary of your conclusions regarding the effectiveness of these three insecticides.

```
Call:  glm(formula = y ~ log(XX$deposit) + XX$type - 1, family = binomial(link = "logit"))

Coefficients:
log(XX$deposit)       XX$typeBHC       XX$typeDDT     XX$typeMIXED
          2.694           -3.840           -4.454           -1.423

Degrees of Freedom: 882 Total (i.e. Null);  878 Residual
Null Deviance:      1223
Residual Deviance: 819  AIC: 827
```

Figure 4.11 R summary table for effectiveness of insecticide

Based on the r summary table, we can find that the mixed insecticide can have a better effect on the kill rate. BHC ranks second and DDT has the worst effect on kill rate.

Appendix

```
#1(a)
a=read.table("fungus.txt",header = TRUE)
Y=c(a$X1,a$X2,a$X3,a$X4,a$X5,a$X6,a$X7,a$X8)
temp=c("55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",
       "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",
       "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",
       "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85")
isolate=c("X1","X1","X1","X1","X1","X1","X1","X2","X2","X2","X2","X2","X2","X2","X3","X3","X
3","X3","X3","X3","X3",

"X4","X4","X4","X4","X4","X4","X4","X5","X5","X5","X5","X5","X5","X5","X6","X6","X6","X6","X
6","X6","X6",
          "X7","X7","X7","X7","X7","X7","X7","X8","X8","X8","X8","X8","X8","X8")
lm.reg=lm(Y~isolate+temp)
summary(lm.reg)
anova(lm.reg)

#1(b)
fug = read.table("fungus.txt",header = T)
fug = unlist(fug)[-(1:7)]
fung = data.frame("fungus" = fug,"temp" = factor(rep(seq(55,85,by=5),8)),"iso" =
factor(rep(1:8,each = 7)))
fung
temp

a=read.table("fungus.txt",header = TRUE)
a
Y=c(a$X1,a$X2,a$X3,a$X4,a$X5,a$X6,a$X7,a$X8)
x=rep(seq(55,85,by=5),8)
x2=x^2;x3=x^3;x4=x^4;x5=x^5;x6=x^6
lm6=lm(Y~x+x2+x3+x4+x5+x6)
anova(lm6)
lm0=lm(Y~1)
lm1=lm(Y~1+x)
lm2=lm(Y~1+x+x2)
lm3=lm(Y~x+x2+x3)
lm4=lm(Y~x+x2+x3+x4)
lm5=lm(Y~x+x2+x3+x4+x5)
anova(lm0,lm1)
anova(lm1,lm2)
anova(lm2,lm3)
anova(lm3,lm4)
```

```
anova(lm4,lm6)
anova(lm3,lm6)

#1(c)
lm3=lm(Y~x+x2+x3)
fun=function(x){
   return(-sum(c(1,x,x^2,x^3)*lm3$coefficients))
}
optimize(fun,c(55,85))
```

```
#2(a)
b=read.table("flushot.txt",header = TRUE)
b
glm.reg=glm(data=b,y~x1+x2+x3,family=binomial(link="logit"))
summary(glm.reg)
ci95=confint.default(glm.reg)

#2(b)
round(cbind(exp(glm.reg$coefficients),exp(confint.default(glm.reg))),3)

#2(c)
pnew = predict.glm(glm.reg, newdata=data.frame(x1=55,x2=60,x3=1), se.fit=T, type="link")
hat = pnew$fit
se.hat = pnew$se.fit
phat = exp(hat)/(1+exp(hat))
data.frame("lowerbound"=hat-qnorm(0.975)*se.hat,"upperbound"=hat+qnorm(0.975)*se.hat)
uphat = exp(hat-qnorm(0.975)*se.hat)/(1+exp(hat-qnorm(0.975)*se.hat))
lohat = exp(hat+qnorm(0.975)*se.hat)/(1+exp(hat+qnorm(0.975)*se.hat))
data.frame("lower bound"=uphat ,"upper bound"=lohat) # c.i for estimated prob

#2(e)
anova(glm(data=b,y~x1+x2,family = binomial("logit")),glm.reg,test = "Chisq")

#2(f)
b=read.table("flushot.txt",header = TRUE)
glm.reg=glm(data=b,y~x1+x2+x3,family=binomial(link="logit"))
glm0=glm(data=b,y~1,family = binomial("logit"))
#aic
step(glm0,scope=list(upper=glm.reg),direction = "both")
step(glm0,scope=list(upper=glm.reg),direction = "forward")
step(glm.reg)
#bic
step(glm0,scope=list(upper=glm.reg),direction = "both",k=log(length(y)))
step(glm0,scope=list(upper=glm.reg),direction = "forward",k=log(length(y)))
step(glm.reg,k=log(length(y)))
```

```
#final model
glm.reg=glm(data=b,y~x1+x2,family=binomial(link="logit"))
glm.reg

#2(g)
b=read.table("flushot.txt",header = TRUE)
glm.reg.final=glm(y~x1+x2,family=binomial(link="logit"))
install.packages("pROC")
library(pROC)
y=b$y
flushot.roc=roc(y~fitted(glm.reg.final))
plot(flushot.roc,cex.lab=1.5)
auc(flushot.roc)
```

```
#4(a)
par(mfrow=c(1,1))
kill.rate=c(3/50,5/49,19/47,19/38,24/49,35/50,2/50,14/49,20/50,27/50,41/50,40/50,28/50,37
/50,46/50,48/50,48/50,50/50)
doses=c(2.00,2.64,3.48,4.59,6.06,8.00,2.00,2.64,3.48,4.59,6.06,8.00,2.00,2.64,3.48,4.59,6.06,8
.00)
log(doses)
doses.killrate=lm(kill.rate~doses)
plot(log(doses),kill.rate,cex.lab=1.5)

#4(b)
data=data.frame("deposit"=c(2.00,2.64,3.48,4.59,6.06,8.00),
                "ddt"=c(3/50,5/49,19/47,19/38,24/49,35/50),
                "bhc"=c(2/50,14/49,20/50,27/50,41/50,40/50),
                "both"=c(28/50,37/50,46/50,48/50,48/50,50/50))
logit = function(x){
   log(x/(1-x))
}
data$deposit = log(data$deposit)
plot(data$deposit,logit(data$ddt),ylim=c(-4,4),xlab="deposit",ylab="logit(kill
rate)",cex.lab=1.5)
points(data$deposit,logit(data$bhc),col="red")
points(data$deposit,logit(data$both),col="blue")

d = lm(logit(ddt)~deposit,data=data)
abline(reg=d)
b = lm(logit(bhc)~deposit,data=data)
abline(reg=b)
combine = lm(logit(data$both[1:5])~deposit[1:5],data=data)
abline(reg=combine)
```

```
#4(c)
insect=c(rep(0,6),rep(1,6),rep(2,6))

insc_type = factor(insect, levels = c(0, 1, 2),
                           labels = c("DDT", "BHC", "MIXED"))

dose=rep(c(2.00, 2.64, 3.48, 4.59, 6.06, 8.00),3)
killed=c(3,5,19,19,24,35,2,14,20,27,41,40,28,37,46,48,48,50)
unkilled=c(50,49,47,38,49,50,50,49,rep(50,10))-killed
data=data.frame("killed"=killed,"unkilled"=unkilled,"deposit"=dose,"type"=insc_type)

# making binary var
y=c()
for(i in 1:length(killed)){
   yi=c(rep(1,killed[i]),rep(0,unkilled[i]))
   y=c(y,yi)
}

total=killed+unkilled
cumsum(total)
total=c(0,total)

XX = matrix(0,ncol = 2,nrow =sum(total))
X=as.matrix(data[,c(3,4)])
cum=cumsum(total)
for(i in 1:(length(cum)-1)){
   k=((cum[i]+1):(cum[i+1]))
   v=X[i,]
   for(j in 1:length(k)){
      XX[k[j],]=v
   }
}
XX=data.frame(XX)
XX$X1=as.numeric(as.character(XX$X1))
colnames(XX)=c("deposit","type")
head(XX)

ddt=c(rep(1,283),rep(0,299),rep(0,300))
bhc=c(rep(0,283),rep(1,299),rep(0,300))
chem=data.frame("ddt"=ddt,"bhc"=bhc)
parallel=
glm(y~log(XX$deposit)+chem$ddt+chem$bhc+log(XX$deposit)*chem$ddt+log(XX$deposit)*ch
em$bhc,family=binomial(link="logit"))
summary(parallel)
```

```
#4(d)
#regression
glm1=glm(y~log(XX$deposit)+XX$type,family=binomial(link="logit"))
glm2=glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="logit"))
#var-cor
summary(glm1)$cov.unscaled
summary(glm2)$cov.unscaled

#4(e)
confint(glm2,level = 0.9)

#4(f)
glm(y~log(XX$deposit)+XX$type,family=binomial(link="probit"))
glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="probit"))

glm(y~log(XX$deposit)+XX$type,family=binomial(link="cloglog"))
glmc=glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="cloglog"))
confint(glmc,level = 0.9)
```