MATH 588

HW9

Md Ismail Hossain

4/23/2022

Exercise 20.11

```
library(Sleuth3)
head(ex2011)
```

```
##   Temperature Failure
## 1          53     Yes
## 2          56     Yes
## 3          57     Yes
## 4          63      No
## 5          66      No
## 6          67      No
```

a

```
lm1_1 <- glm (as.factor(Failure) ~ Temperature, data = ex2011, family = binomial)
summary (lm1_1)
```

```
##
## Call:
## glm(formula = as.factor(Failure) ~ Temperature, family = binomial,
##     data = ex2011)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2125  -0.8253  -0.4706   0.5907   2.0512
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.87535    5.70291   1.907   0.0565 .
## Temperature -0.17132    0.08344  -2.053   0.0400 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.975  on 23  degrees of freedom
## Residual deviance: 23.030  on 22  degrees of freedom
## AIC: 27.03
##
## Number of Fisher Scoring iterations: 4
```

The fit of a logistic Regression model to the space shuttle data, where pi represents survival probability, gives

$$logit(\hat{\pi}) = -10.87535 + 0.17132(Temperature)$$

with intercept standard error is 5.70291 and slope standard error is 0.08344.

b

H_0: coefficient of Temperature $(\beta_1) = 0$
H_a: coefficient of Temperature $(\beta_1) \neq 0$

The Wald statistic (z - value) for the Temperature is 2.053 and the correspondent two sided p-value is 0.04 which is smaller than the critical value. We can reject the H0 at 5% level of significance and conclude that there is strong evidence of an association between Temperature and the incidence of O-ring failure.

Consider the coefficient is negative then the hypothesis can be expressed as:

H_0: coefficient of Temperature (β_1) < 0
H_a: coefficient of Temperature (β_1) ≥ 0

The Wald statistic for Temperature is 2.053 and the one sided p-value is 0.02, which is less than 0.05. So, at 5% level of significance we can reject the H_0. So, we can conclude that there is association between Temperature and the incidence of O-ring failure.

c

```
lm1_2 <- update(lm1_1, ~ . -Temperature)
summary(lm1_2)
```

```
##
## Call:
## glm(formula = as.factor(Failure) ~ 1, family = binomial, data = ex2011)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8305  -0.8305  -0.8305   1.5698   1.5698
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8873     0.4491  -1.976   0.0482 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.975  on 23  degrees of freedom
## Residual deviance: 28.975  on 23  degrees of freedom
## AIC: 30.975
##
## Number of Fisher Scoring iterations: 4
```

Here, in the model with the Temperature the deviance is 23.030 and without Temperature the deviance is 28.975. So, drop in deviance is $(28.975 - 23.030) = 5.945$ with the drop in degrees of freedom $23 - 22 = 1$. The drop in deviance follows a χ^2 distribution with the drop in degrees of freedom so at 5% level of significance.

```
pchisq (5.945, 1, lower.tail = F)
```

```
## [1] 0.01475909
```

So, we can conclude that Temperature have a significant impact because the p-value of the test is very small.

d

```
exp(confint.lm(lm1_1))
```

```
##                 2.5 %       97.5 %
## (Intercept) 0.3860581 7.236951e+09
## Temperature 0.7086726 1.001722e+00
```

The 95% confidence interval for Temperature is $[0.07086726, 1.001722]$.

e

From the above output, the estimated logit of failure probability at 31^0 F (the launch temperature on January 28,1996 is given by,

$$logit(\pi_1) = -10.87535 + 0.17132 * (31) = -5.56443$$

The estimate failure probability is:

$$\pi_1 = \frac{e^-5.56443}{1 + e^-5.56443} = 0.003817$$

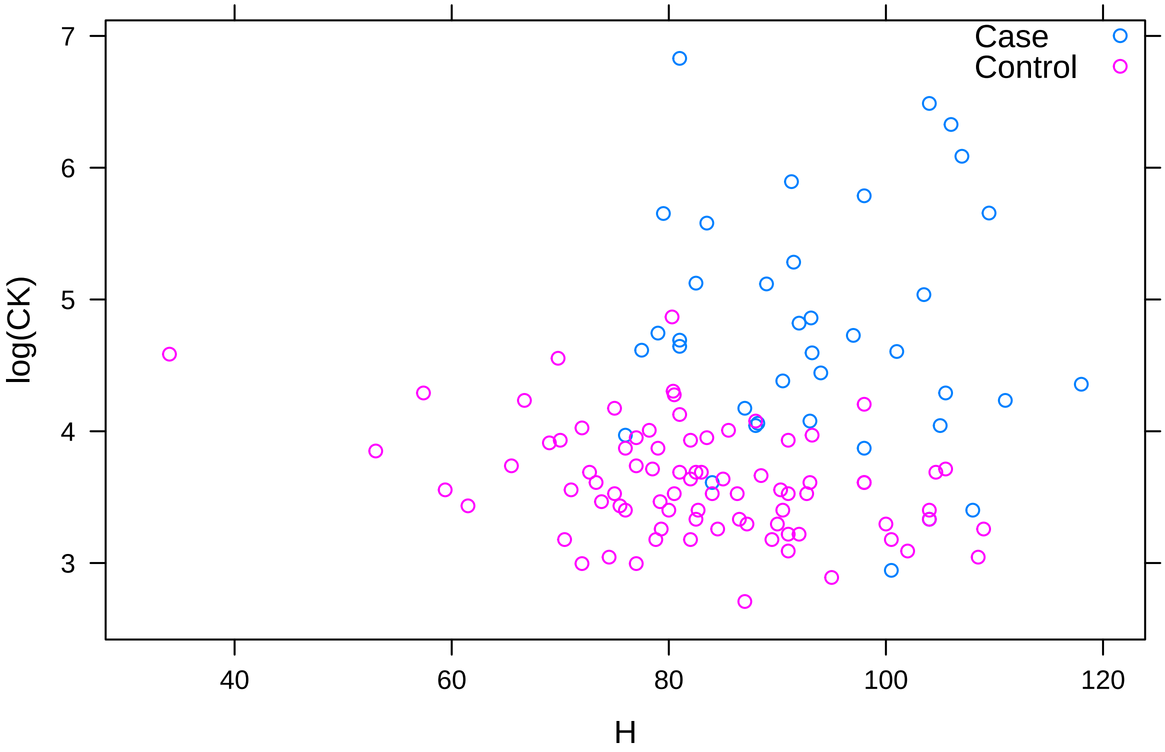
f

As the prediction outside the range of the independent variables values, so we have to be more cautious in estimating.

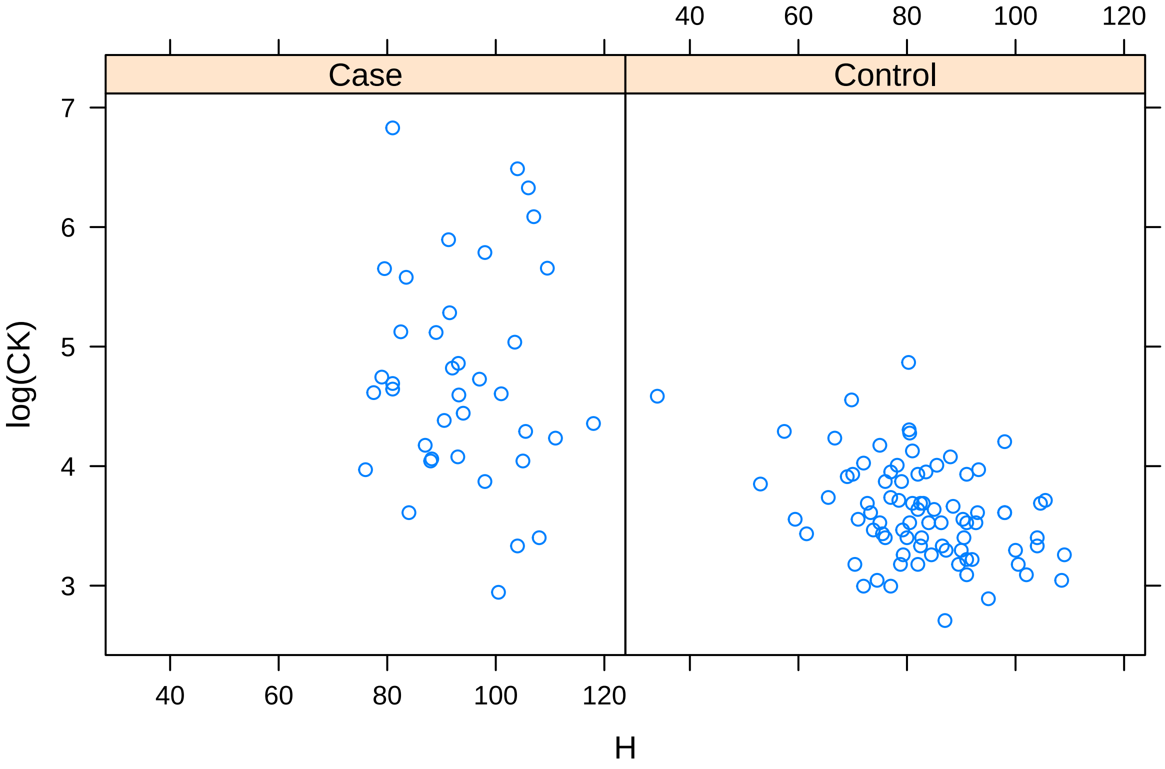
Exercise 20.12

a

```
library(lattice)
xyplot(log(CK)~H, data=ex2012, groups=Group, auto.key=list(corner=c(1,1)))
```



```
xyplot(log(CK)~H|Group, data=ex2012)
```



We couldn't see any relationship between log of creatine kinase and hemopexin for case and control group. But mascular dystropy carriers have hemopexin values more than around 75.

b

```
lm2_1 <- glm (as.factor(Group) ~ CK + I(CK^2) , data = ex2012, family = binomial)
summary(lm2_1)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ CK + I(CK^2), family = binomial,
##     data = ex2012)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50536  -0.03915   0.37969   0.51841   2.27337
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.177e+00  7.264e-01   5.751 8.87e-09 ***
## CK          -5.798e-02  1.299e-02  -4.463 8.10e-06 ***
## I(CK^2)      5.054e-05  3.268e-05   1.547    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.84  on 119  degrees of freedom
## Residual deviance:  85.47  on 117  degrees of freedom
## AIC: 91.47
##
## Number of Fisher Scoring iterations: 9
```

The quadratic term of CK of the model doesn't seems statistically significant at 5% level of significance.

```
lm2_2 <- glm (as.factor(Group) ~ log(CK) + I((log(CK))^2) , data = ex2012, family = binomial)
summary(lm2_2)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ log(CK) + I((log(CK))^2), family = binomial,
##     data = ex2012)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39368  -0.03111   0.38041   0.50222   2.28558
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -9.735     16.298  -0.597    0.550
## log(CK)           8.516      8.358   1.019    0.308
## I((log(CK))^2)   -1.446      1.063  -1.360    0.174
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  85.017  on 117  degrees of freedom
## AIC: 91.017
##
## Number of Fisher Scoring iterations: 7
```

In this model none of the CK variables are statitically significant at 5% level of significance.

c

```
lm2_3 <- glm (as.factor(Group) ~ log(CK) + H, data = ex2012, family = binomial)
summary (lm2_3)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ log(CK) + H, family = binomial,
##     data = ex2012)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.60371  -0.09903   0.16696   0.38782   1.89706
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 28.91340    5.80017   4.985 6.20e-07 ***
## log(CK)     -4.02043    0.82910  -4.849 1.24e-06 ***
## H           -0.13652    0.03654  -3.736 0.000187 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.840  on 119  degrees of freedom
## Residual deviance:  61.992  on 117  degrees of freedom
## AIC: 67.992
##
## Number of Fisher Scoring iterations: 7
```

When we are using H in the model then both the coefficient of H and log(CK) becomes statistically significant at 5% level of significance.

d

```r
lm2_4 <- update(lm2_3, ~ . - log(CK) - H)
summary(lm2_4)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ 1, family = binomial, data = ex2012)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5165  -1.5165   0.8727   0.8727   0.8727
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7691     0.1962   3.919 8.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 149.84  on 119  degrees of freedom
## Residual deviance: 149.84  on 119  degrees of freedom
## AIC: 151.84
##
## Number of Fisher Scoring iterations: 4
```

Here in the model, with the log(CK) and H the deviance is 61.992 and deviance without these two predictors is 149.84. So, the drop in deviance is $(149.84 - 61.992) = 87.848$ with the drop in degrees of freedom $119 - 117 = 2$. The drop in deviance follows a χ^2 distribution with the drop in degrees of freedom so at 5% level of significance.

```r
pchisq (87.848, 2, lower.tail = F)
```

```
## [1] 8.39555e-20
```

log(CK) and H are useful predictors according to the p-value we go from the test.

e

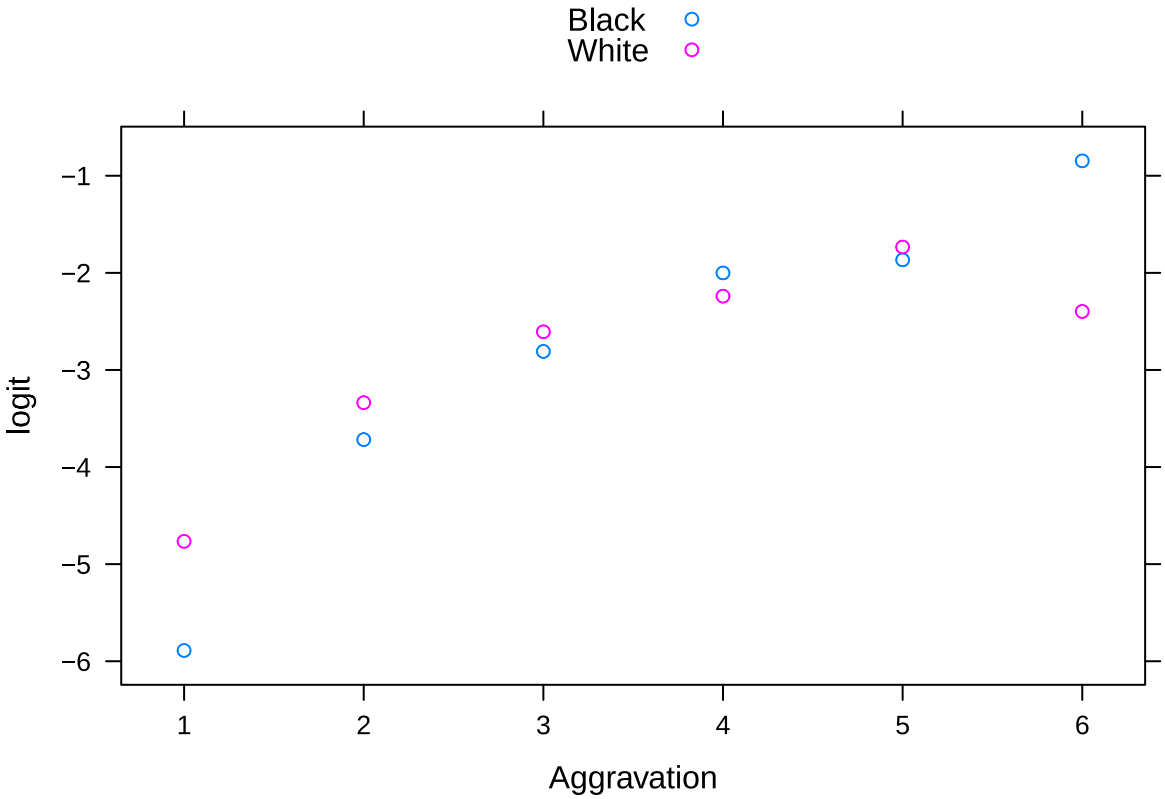
Odds of a suspected carrier with the values CK = 300 and H = 100 is -7.67 and the odds of suspected carrier with the values CK = 80 and H = 85 is -0.308. The odds ratio value is 25.1948 which indicates that the odds of a suspected carrier with CK and H values of 300 and 100 respectively is almost 25 times more than odds of a suspected carrier with the typical values (CK = 80 and H = 85).

Exercise 21.10

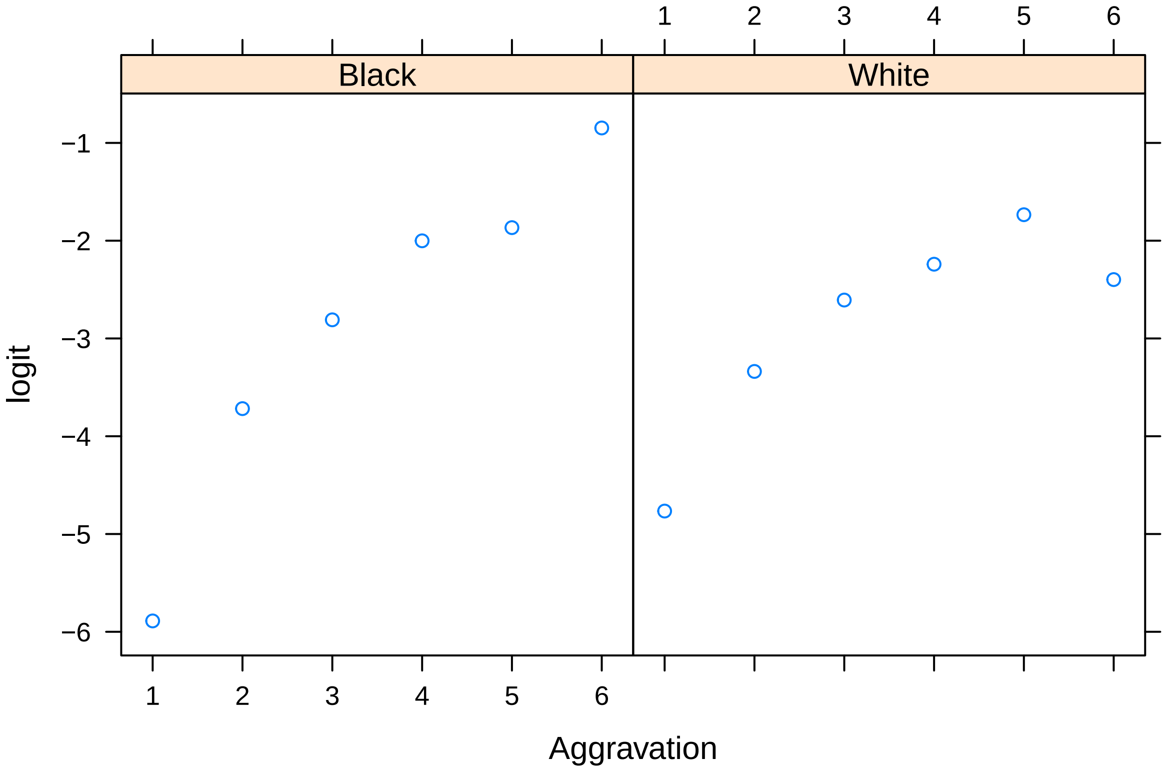
```
ex2110 = case1902
```

a

```
ex2110$y <- with(ex2110, Death/(Death + NoDeath))
ex2110$logit <- with (ex2110, log((y+0.5)/(Death+NoDeath-y+0.5)))
xyplot(logit ~ Aggravation, data=ex2110, groups=Victim, auto.key=list(bottom=c(1,1)))
```



```
xyplot (logit ~ Aggravation | Victim, data = ex2110)
```



b

```
ex2110$prop <- with (ex2110, Death/(Death+NoDeath))
ind_vic <- with (ex2110, ifelse(Victim=="White",1,0))
binResponse <- with (ex2110, cbind(Death, NoDeath))
lm3_1 <- glm (binResponse ~ Aggravation + ind_vic, data = ex2110, family = binomial)
summary (lm3_1)
```

```
##
## Call:
## glm(formula = binResponse ~ Aggravation + ind_vic, family = binomial,
##     data = ex2110)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.93570  -0.22548   0.05142   0.65620   1.01444
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6760     0.7574  -8.814  < 2e-16 ***
## Aggravation   1.5397     0.1867   8.246  < 2e-16 ***
## ind_vic       1.8106     0.5361   3.377 0.000732 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 212.2838  on 11  degrees of freedom
## Residual deviance:   3.8816  on  9  degrees of freedom
## AIC: 31.747
##
## Number of Fisher Scoring iterations: 4
```

c

```
pchisq(lm3_1$deviance, df=lm3_1$df.residual, lower.tail=FALSE)
```

```
## [1] 0.9190319
```

The null hypothesis is that our model is correctly specified, and we have strong evidence to support the null hypothesis since the p value is larger than the critical value at 5% level of significance. So we have strong evidence that our model fitted well.

d

For β_2: From the above output we can say that, the Wald statistic (z - value) for the Dose is 3.377 and the correspondent two sided p-value is 0.000732 which is smaller than the critical value. We can reject the H0 at 5% level of significance and conclude that β_2 not equal to 0.

e

```
confint.lm(lm3_1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -8.3894220 -4.962528
## Aggravation  1.1172590  1.962064
## ind_vic      0.5978728  3.023420
```

```
exp(confint.lm(lm3_1)[3,1:2])
```

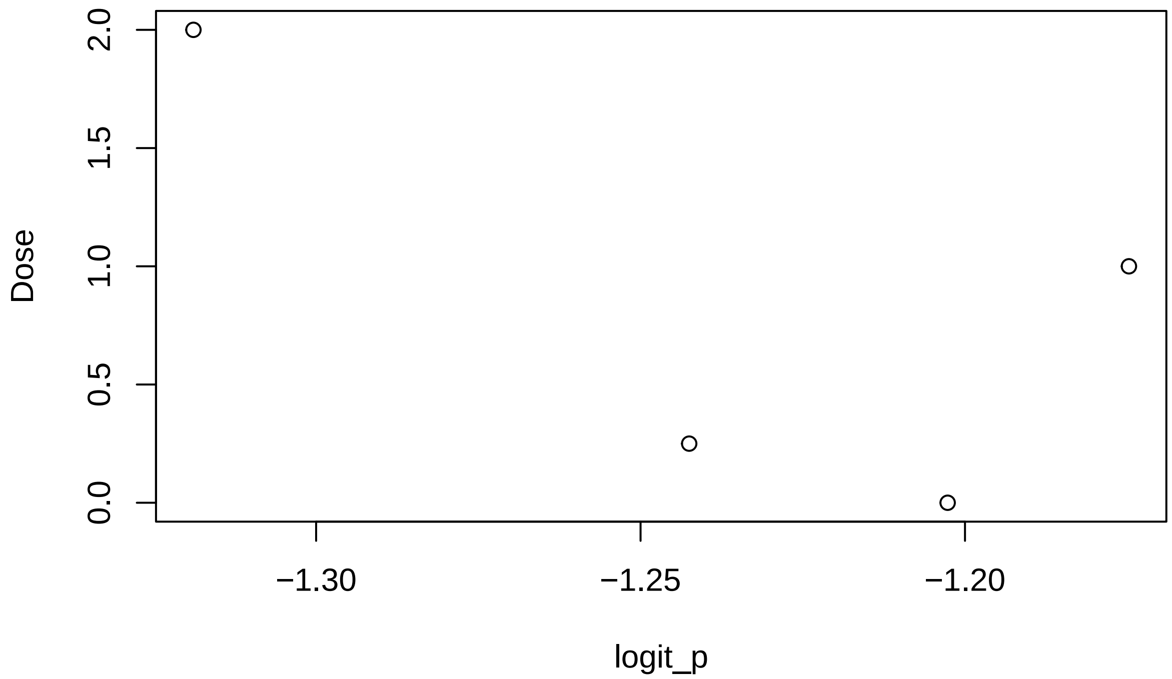
```
##     2.5 %    97.5 %
##  1.818247 20.561501
```

95% confidence interval for the odds of death sentence for white-victim murderers relative to the black-victim murderers, accounting for aggravation level of the crime is 1.818 to 20.652.

Exercise 21.13

a

```
ex2113$logit_p <- log(ex2113$ProportionWithout/(1-ex2113$ProportionWithout))
with (ex2113, plot(logit_p, Dose), main = "Scatterplot of logit (p) with Dose")
```



b

```
binResponse <- with (ex2113, cbind(WithoutIllness, Number - WithoutIllness))
lm4_1 <- glm (binResponse ~ Dose, data = ex2113, family = binomial)
summary (lm4_1)
```

```
##
## Call:
## glm(formula = binResponse ~ Dose, family = binomial, data = ex2113)
##
## Deviance Residuals:
##        1         2         3         4
## -0.06857  -0.27405   0.57021  -0.35303
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.20031    0.06167 -19.464   <2e-16 ***
## Dose        -0.03465    0.07113  -0.487    0.626
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.76803  on 3  degrees of freedom
## Residual deviance: 0.52957  on 2  degrees of freedom
## AIC: 29.801
```

```
##
## Number of Fisher Scoring iterations: 3
```

The fit of a logistic regression model to the Vitamin C data, where π represents the proportion of the without illness gives,

$$logit(\hat{\pi}) = -1.20031 - 0.03465 * (Dose)$$

with intercept standard error is 0.06167 and slope (Dose) standard error is 0.07113.

H_0 : The model fits the data well
H_a : The model does not fit the data well

To deviance here is labelled as the 'residual deviance' by the glm function, and here is 0.52957. There are 4 observations, and our model has two parameters, so the degrees of freedom is 2, given by R as the residual df. To calculate the p-value for the deviance goodness of fit test we simply calculate the probability to the right of the deviance value for the chi-squared distribution on 2 degrees of freedom:

```
pchisq(lm4_1$deviance, df=lm4_1$df.residual, lower.tail=FALSE)
```

```
## [1] 0.7673694
```

The null hypothesis is that our model is correctly specified, and we have strong evidence to support the null hypothesis since the p-value is larger than the critical value at 5% level of significance. So we have strong evidence that our model fitted well.

c

Here is the hypothesis to test that the β_1 is 0, H_0: coefficient of Dose (β_1) $= 0$
H_a: coefficient of Dose (β_1) $\neq 0$

The calculated p-value for Wald statistic is 0.487 which is greater than 0.05. As a result we can not reject the null hypothesis.

```
lm4_2 <- update (lm4_1, ~ . - Dose)
summary (lm4_2)
```

```
##
## Call:
## glm(formula = binResponse ~ 1, family = binomial, data = ex2113)
##
## Deviance Residuals:
##       1        2        3        4
##  0.1934  -0.2006   0.4086  -0.7235
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.21860    0.04917  -24.79   <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: 0.76803  on 3  degrees of freedom
## Residual deviance: 0.76803  on 3  degrees of freedom
## AIC: 28.039
##
```

```
## Number of Fisher Scoring iterations: 3
```

H_0: coefficient of Dose $(\beta_1) = 0$
H_a: coefficient of Dose $(\beta_1) \neq 0$

Drop in the deviance is $(0.76803 - 0.52957) = 0.23846$ and drop in df is 1.

```
pchisq(lm4_1$deviance, df=lm4_1$df.residual, lower.tail=FALSE)
```

```
## [1] 0.7673694
```

The p-value is greater than 0.05. So, at 5% level of significance and conclude that there no evidence of association between Dose and the log of the proportion of without illness which contradicts the previous result got from the Wald test.