CDA HW 7

Joshua Ingram

11/18/2019

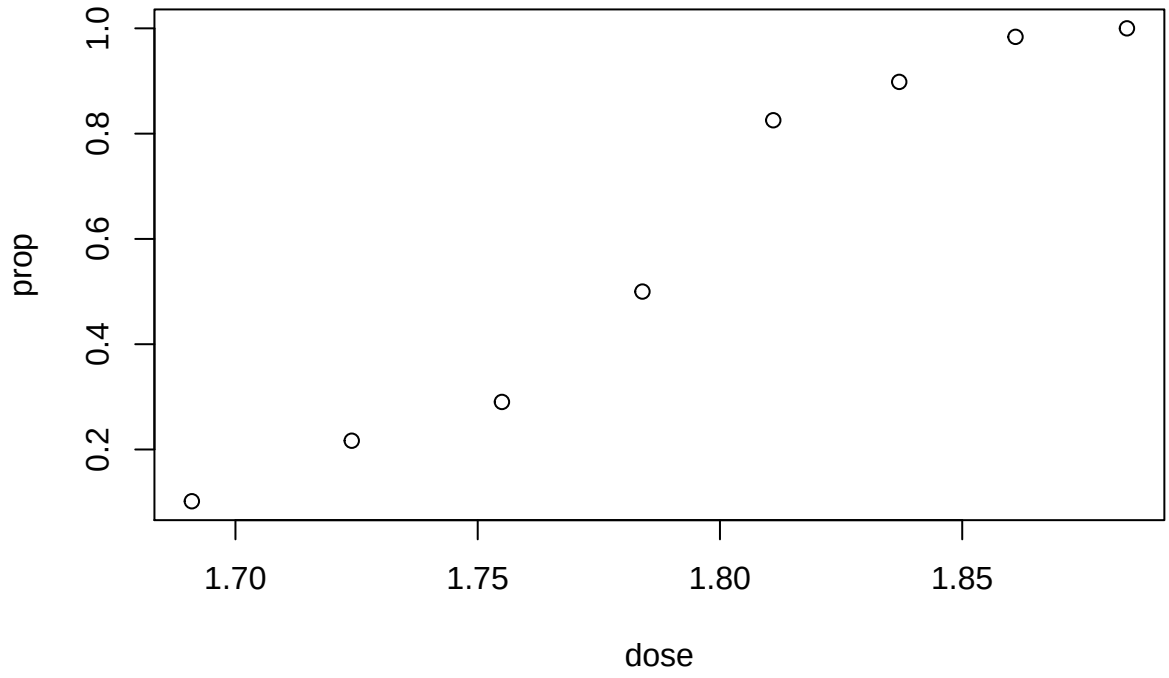
Problem 1: Beetles

```
dose <-c(1.691, 1.724, 1.755, 1.784, 1.811, 1.837, 1.861, 1.884)
exposed <-c(59, 60, 62, 56, 63, 59, 62, 60)
killed <-c(6, 13, 18, 28, 52, 53, 61, 60)
beetles <-data.frame(dose = dose, exposed = exposed, killed = killed)
beetles
```

```
##    dose exposed killed
## 1 1.691      59      6
## 2 1.724      60     13
## 3 1.755      62     18
## 4 1.784      56     28
## 5 1.811      63     52
## 6 1.837      59     53
## 7 1.861      62     61
## 8 1.884      60     60
```

a.

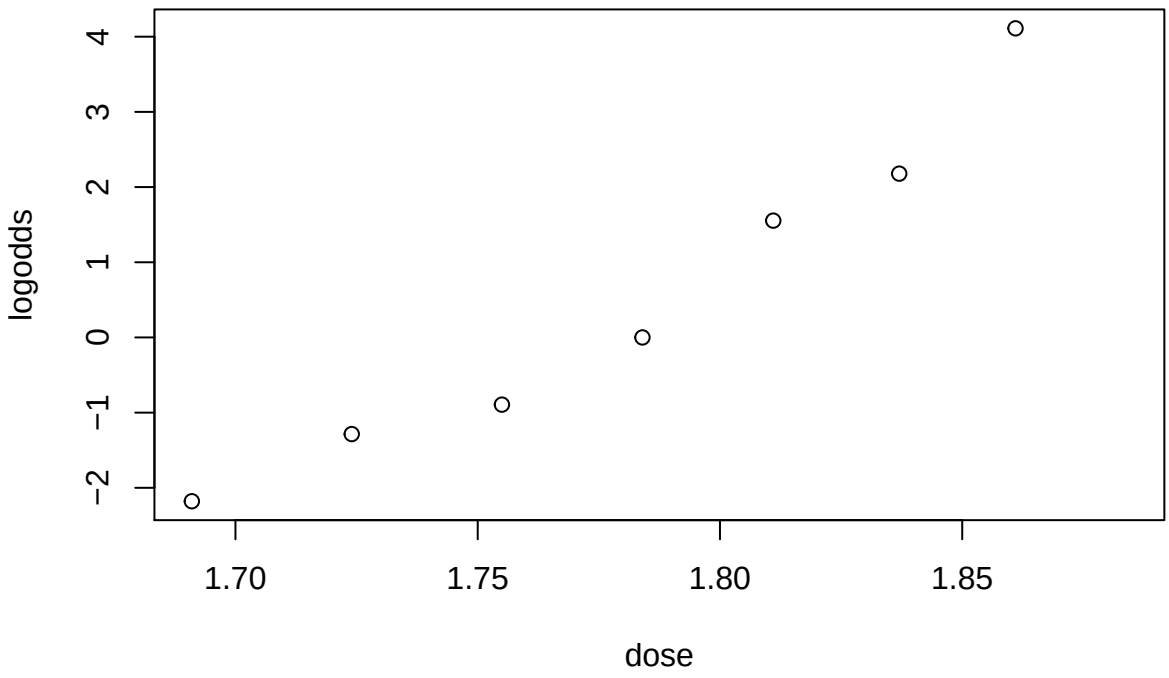
```
beetles$prop <- beetles$killed / beetles$exposed
plot(prop~dose, data=beetles)
```



There seems to be a non-linear relationship between dose and the proportion of beetles killed, it is s-shaped.

b.

```
beetles$odds <- (beetles$killed/beetles$exposed)/((beetles$exposed - beetles$killed) / beetles$exposed)
beetles$logodds <- log(beetles$odds)
plot(logodds ~ dose, data = beetles)
```



If logistic regression is appropriate, the log-odds plot should be roughly linear. It seems the plot of the log-odds is more linear than the proportion plot and follows a roughly linear trend.

c.

```
fit.logit <-glm(cbind(killed, exposed-killed)~dose, family = binomial,data = beetles)
summary(fit.logit)
```

```
##
## Call:
## glm(formula = cbind(killed, exposed - killed) ~ dose, family = binomial,
##     data = beetles)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5878  -0.4085   0.8442   1.2455   1.5860
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.740      5.182  -11.72   <2e-16 ***
## dose          34.286      2.913   11.77   <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: 284.202  on 7  degrees of freedom
## Residual deviance:  11.116  on 6  degrees of freedom
## AIC: 41.314
##
## Number of Fisher Scoring iterations: 4
```

d.

```
anova(fit.logit, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(killed, exposed - killed)
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     7    284.202
## dose  1   273.09         6     11.116 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We receive a p-value of 2.2e-16, which yeilds signicant evidence to suggest that dose has an effect on the probability of a beetle being killed.

e.

Based on our model fit, for every .1 unit increase in dose, we predict the odds of beetle being killed to increase by a factor of $e^{3.4286}$

f.

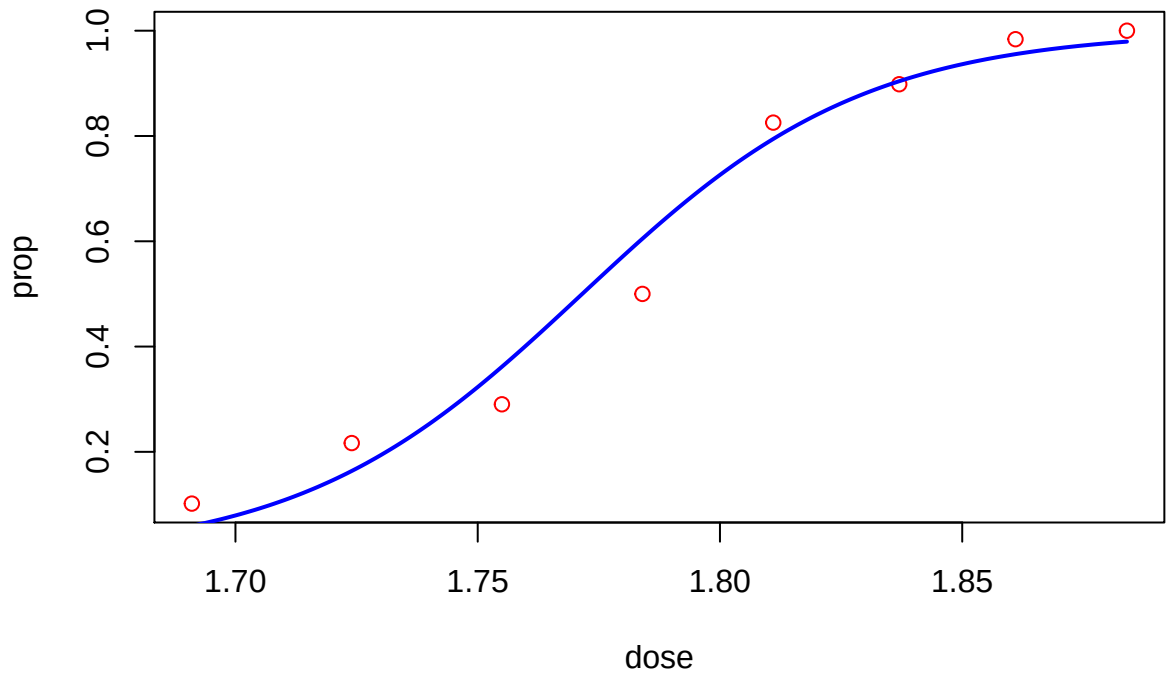
```
predict(fit.logit, type = "response", newdata = (dose = 1.8))
```

```
##         1
## 0.7260234
```

The predicted probability of a beetle being killed at dosage 1.8 is 72.6%

g.

```r
pred.plot.data <- seq(min(beetles$dose), max(beetles$dose), by = .001)
prediction.prob <- predict(fit.logit, newdata = data.frame(dose = pred.plot.data), type = "response")
plot(prop~dose, data=beetles, col = "red")
lines(prediction.prob~pred.plot.data, data = beetles, col = "blue", lwd = 2)
```



Problem 2: Crabs

```r
head(crab)
```

```
##   color spine width satellite weight sat
## 1     2     3  28.3         8   3.05   1
## 2     3     3  22.5         0   1.55   0
## 3     1     1  26.0         9   2.30   1
## 4     3     3  24.8         0   2.10   0
## 5     3     3  26.0         4   2.60   1
## 6     2     3  23.8         0   2.10   0
```

a.

```
crab.glm.fit <- glm(sat ~ width, family="binomial", data = crab)
summary(crab.glm.fit)
```

```
##
## Call:
## glm(formula = sat ~ width, family = "binomial", data = crab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0281  -1.0458   0.5480   0.9066   1.6942
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
## width         0.4972     0.1017   4.887 1.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 194.45  on 171  degrees of freedom
## AIC: 198.45
##
## Number of Fisher Scoring iterations: 4
```

```
confint(crab.glm.fit, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -17.8100090 -7.4572470
## width         0.3083806  0.7090167
```

Based on our model, for every 1 unit increase in the width of the shell the odds of a female crab having a satellite increases by a factor of $e^{0.4972}$ We are 95% confident that the true effect of a 1 unit increase in the width of a shell on the odds of a female crab having at least one satellite is an increase by factor between $e^{0.3083806}$ and $e^{0.7090167}$

b.

```
crab$col <-as.numeric(crab$color>2)
crab.glm.fit2 <- glm(sat~width+col, family = "binomial", data = crab)
summary(crab.glm.fit2)
```

```
##
## Call:
## glm(formula = sat ~ width + col, family = "binomial", data = crab)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1080  -0.9708   0.5346   0.8958   1.8188
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.2970     2.7011  -4.182 2.88e-05 ***
## width         0.4670     0.1037   4.506 6.61e-06 ***
## col          -0.6531     0.3571  -1.829   0.0675 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 191.12  on 170  degrees of freedom
## AIC: 197.12
##
## Number of Fisher Scoring iterations: 4
```

```
anova(crab.glm.fit2, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sat
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    172     225.76
## width  1  31.3059       171     194.45 2.204e-08 ***
## col    1   3.3344       170     191.12   0.06785 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

we expect the odds of a female crab having a satellite to be $e^{-0.6531}$ times less for a dark colored crab than a light colored crab.

After performing a likelihood ratio test using the anova() command, the effects of width are very signifcant but col is borderline with a p-value of 0.06785. Being so close to .05, it seems that it may be worth keeping in the model.

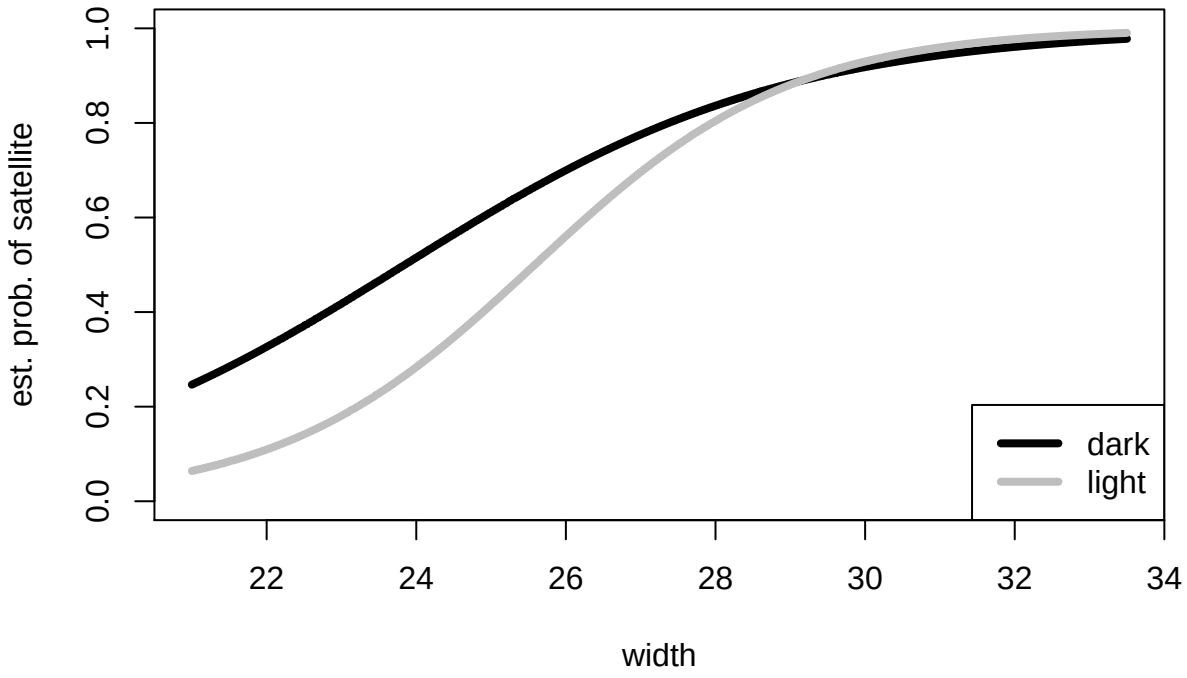
c.

```
crab.glm.fit3 <- glm(sat~width+col+width:col, family="binomial", data = crab)
summary(crab.glm.fit3)
```

```
##
```

```
## Call:
## glm(formula = sat ~ width + col + width:col, family = "binomial",
##     data = crab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0224  -0.9898   0.5662   0.8512   1.9783
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.3641     3.3597  -2.787  0.00532 **
## width         0.3927     0.1287   3.052  0.00227 **
## col          -5.6067     5.6084  -1.000  0.31745
## width:col     0.1925     0.2175   0.885  0.37613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 190.31  on 169  degrees of freedom
## AIC: 198.31
##
## Number of Fisher Scoring iterations: 4
```

```
width.new <- seq(min(crab$width), max(crab$width), .001)
fit.light <- predict(crab.glm.fit3, type = "response", newdata=data.frame(width = width.new, col = 0))
fit.dark <- predict(crab.glm.fit3, type = "response", newdata=data.frame(width=width.new, col = 1))
plot(fit.light~width.new, col = "black", xlab= "width", ylab="est. prob. of satellite", ylim=c(0,1), ty
lines(fit.dark~width.new, col = "gray", lwd = 4)
legend("bottomright", legend=c("dark", "light"), lwd=c(4,4), col= c("black", "gray"))
```



d.

```
anova(crab.glm.fit3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: sat
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        172     225.76
## width      1  31.3059       171     194.45 2.204e-08 ***
## col        1   3.3344       170     191.12   0.06785 .
## width:col  1   0.8061       169     190.31   0.36929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction effect is not needed, as we receive a p-value of 0.36929 which is not very significant.