

Logistic Regression Handout

1 Initialization

```
> library(NCStats)
```

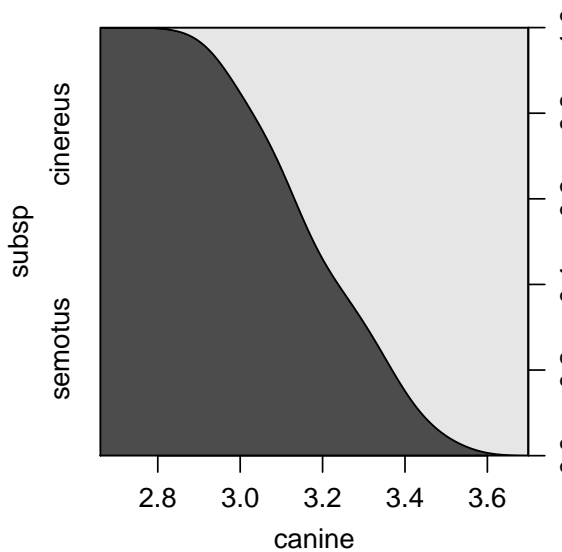
2 Bat Subspecies Example

You must change the directory to where the following file is located. In addition, I changed the canine measurements to mm (from cm) for ease of explanation later on.

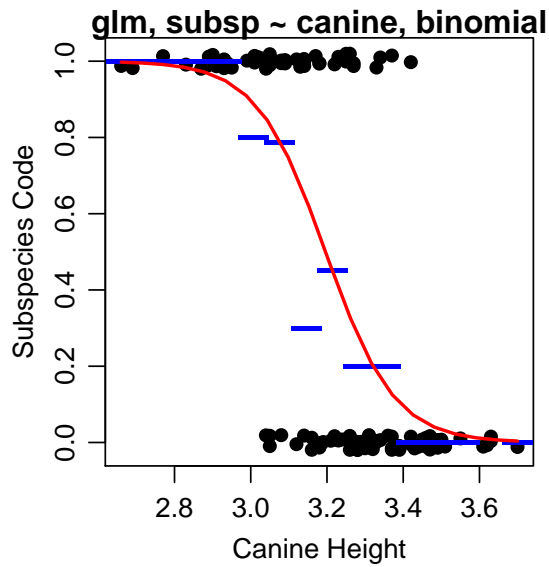
```
> bat <- read.table("BatMorph.txt",head=T)
> str(bat)
```

```
'data.frame':      118 obs. of  7 variables:
 $ subsp      : Factor w/ 2 levels "cinereus","semotus": 2 2 2 2 2 2 2 2 2 2 ...
 $ bodymass   : num  19.5 16.2 17.0 16.5 14.3 ...
 $ skulllength: num  1.60 1.55 1.56 1.56 1.53 ...
 $ canine     : num  0.326 0.308 0.291 0.287 0.301 0.305 0.277 0.313 0.289 0.293 ...
 $ coronoid   : num  0.303 0.282 0.292 0.303 0.279 0.284 0.286 0.281 0.278 0.28 ...
 $ wingspan   : num  0.358 0.358 0.359 0.353 0.351 0.361 0.351 0.363 0.34 0.365 ...
 $ hab       : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 2 2 ...
```

```
> bat$canine <- bat$canine*10
> attach(bat)
> cdplot(subsp~canine,ylevels=2:1)
```



```
> glm1 <- glm(subsp~canine,family=binomial)
> logreg.plot(glm1,p.ints=15,xlab="Canine Height",ylab="Subspecies Code")
```



```
> coef(glm1)
```

```
(Intercept)    canine
 35.51574    -11.11193
```

```
> confint(glm1)
```

```
                2.5 %    97.5 %
(Intercept) 24.21618 49.662954
canine      -15.52481 -7.589218
```

```
> predict(glm1, data.frame(canine=c(3,4)))
```

```
      1      2
2.179940 -8.931994
```

```
> -8.931994-2.179940
```

```
[1] -11.11193
```

```
> exp(coef(glm1))
```

```
(Intercept)    canine
2.656377e+15 1.493306e-05
```

```
> exp(predict(glm1, data.frame(canine=c(3,4))))
```

```
      1      2
8.8457728416 0.0001320944
```

```
> 0.0001320944/8.8457728416
```

```
[1] 1.493305e-05
```

```
> summary(glm1)
```

```

Call:
glm(formula = subsp ~ canine, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9483  -0.6384  -0.1377   0.5923   2.2658

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  35.516     6.428   5.525 3.29e-08 ***
canine      -11.112     2.005  -5.543 2.97e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 163.040  on 117  degrees of freedom
Residual deviance:  97.178  on 116  degrees of freedom
AIC: 101.18

Number of Fisher Scoring iterations: 5

> predict(glm1,data.frame(canine=c(3,3.4)))

      1      2
2.179940 -2.264834

> predict(glm1,data.frame(canine=c(3,3.4)),type="response")

      1      2
0.8984336 0.0940776

> detach(bat)

```

Homework Problems

1. The General Sociological Survey (GSS) is a very large survey that has been administered 25 times since 1972. The basic purposes of the GSS are to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop crossnational models of human society; and to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting. One question that was asked in the most recent GSS was “Have you watched an x-rated movie in the last year?” The respondent’s answer to this question (Yes or No) and age are recorded in **XMovAge.txt**. Read these data into R, remove all individuals older than age 95, fit a logistic regression model, and answer the following questions.
 - (a) Construct a “fitted-line plot” for the logistic regression model (I suggest you use as many intervals as ages in the data – i.e., 72). Comment on the adequacy of fit of this logistic regression model.
 - (b) Interpret the meaning of the slope (β_1) from the fit of the logistic regression model.
 - (c) Interpret the meaning of the “back-transformed” slope (e^{β_1}) from the fit of the logistic regression model.
 - (d) Show (“by hand”) how to predict the log odds of having seen an x-rated movie in the last year for a 50-year-old respondent.
 - (e) Confirm your hand-calculations to the previous question with R output.
 - (f) Show (“by hand”) how to predict the odds that a 50-year-old respondent has seen an x-rated movie in the last year.
 - (g) Show (“by hand”) how to predict the probability that a 50-year-old respondent has seen an x-rated movie in the last year.
 - (h) Confirm your hand-calculations to the previous question with R output.
 - (i) Use R to predict the probability that a 30-year-old respondent has seen an x-rated movie in the last year. Then show (“by hand”) how to compute the odds that a 30-year-old respondent will have seen an x-rated movie in the last year.
 - (j) Repeat the previous question but for a 31-year-old respondent. Use these results and the results from the previous question to show a computational example of the meaning of (e^{β_1}).
2. Ogle *et al.* (2004) examined the diet of larval ruffe (*Gymnocephalus cernuus*) in the St. Louis River Harbor, Lake Superior. In one part of their study, they recorded the occurrence (i.e., presence or absence) of six different prey items, plus an “other” category, for ruffe captured at two locations (Allouez Bay and Whaleback Bay) over six dates. In addition, they recorded the length of the larval ruffe (in mm). The results of this study can be found in **RuffeLarvalDiet.txt**. For the questions below, restrict the data set to **just Allouez Bay** and focus on the occurrence of *Daphnia* relative to the length of larval ruffe.
 - (a) Construct a “fitted-line plot” for the logistic regression model (I suggest you use as many intervals as lengths in the data – i.e., 13). Comment on the adequacy of fit of this logistic regression model.
 - (b) Is there a significant relationship between the logit probability of consuming a *Daphnia* and the length of the larval ruffe? Explain.
 - (c) Describe, in as much detail as possible, the relationship between the probability of having consumed a *Daphnia* and the length of the larval ruffe.
 - (d) Predict the odds that a 6-mm larval ruffe consumed a *Daphnia*. Describe what this number means.
 - (e) Predict the probability that a 6-mm larval ruffe consumed a *Daphnia*. Describe what this number means.
3. **Extra Credit** As a continuation of the previous question, fit a logistic regression model that will allow you to compare the relationship between the probability that a larval ruffe consumed *Daphnia* and the length of larval ruffe between larval ruffe captured in Allouez and Whaleback Bays. Note: (1) you should treat Allouez Bay as the reference group and (2) you will have to return to the original data set that included information from Whaleback Bay.

- (a) Construct (i.e., write) the ultimate full model on the transformed scale (i.e., the logit scale).
- (b) Explicitly define all variables and describe the meaning of each parameter in your ultimate full model.
- (c) Fit your full model in R. Is there a difference between the two locations in the relationship between the probability that a larval ruffe consumed *Daphnia* and the length of larval ruffe. Explain.
- (d) Predict, for fish from both locations, the probability that a 6-mm larval ruffe consumed a *Daphnia*.
- (e) Fit a similar model in R but focused on the occurrence of *Bosmina*. Is there a difference between the two locations in the relationship between the probability that a larval ruffe consumed *Bosmina* and the length of larval ruffe. Explain.
- (f) Predict, for fish from both locations, the probability that a 6-mm larval ruffe consumed a *Bosmina*.