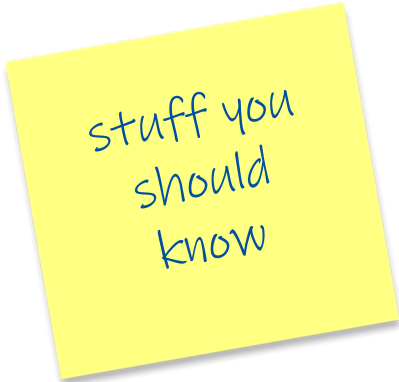# ENV 710: Lecture 19

binomial logistic regression

generalized linear models

binomial logistic regression

learning goals

- binomial logistic regression
  - grouped responses and modeling proportions
  - odds, log-odds, probabilities
  - what is it? when to use it?
  - how does it work?
  - interpretation of coefficients

logistic regression

- binary response variable
  - success or failure (1 or 0)
  - model the probability of success (π)

  $$\pi = \beta_0 + \beta_1 X_1$$

goal is to model the probability of 'success'

- model log odds

  $$logit(\pi) = log[\pi/(1 - \pi)]$$

  $$logit(\pi) = \beta_0 + \beta_1 X_k + ... + \beta_k X_k$$

transform response to
be linearly related to
the IV's through the
logit transform

- calculate probabilities

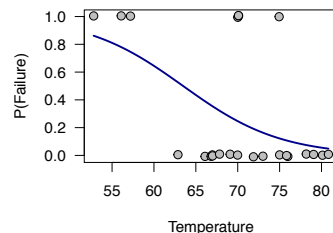 $$\pi = \frac{exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}$$
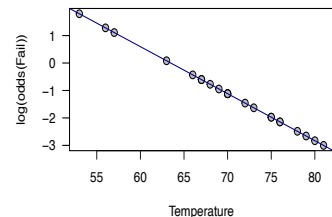
back transform to
probability using
the inverse logit

logistic regression for binomial counts

- binomial count = sum of independent binary responses
- $Y \sim$ binomial(n, π), with population proportion, π, and n trials
- observed proportion of 1's is Y/n – binomial proportion
- continuous proportions of amounts cannot be modeled
  - e.g., the proportion of fat that is saturated fat
  - numerator and denominator are not integers, and no n is involved
  - when response variable is a continuous proportion, use ordinary regression methods

- model how population proportion depends on the explanatory variables through a nonlinear link function

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- interpretation of model is same as for binary logistic regression, which is a special case of the binomial model in which all n_i's are 1

example

beetles

Beetles are exposed to different doses of carbon disulphide (mf/L) for 5 hours. Does dosage level affect mortality?

dose-response curve

Dose	# Exposed	# Killed	Proportion
49.1	59	6	0.102
53.0	60	13	0.217
56.9	62	18	0.290
60.8	56	28	0.500
64.8	63	52	0.825
68.7	59	53	0.898
72.6	62	61	0.985
76.5	60	60	1.000

- lowest dose, $n = 59$ beetles, $k = 6$
- for each beetle, outcome is death or survival
  - death = success
  - p = probability that an individual beetle will die after the dose
  - probability that y assumes a value k is given by:

$$P(y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

example

beetles

```
 b1 <- glm(cbind(killed, exposed - killed) ~ dose,
        family = binomial)
   summary(b1)
```

```
glm(formula = cbind(killed, alive) ~ dose,
family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.82300    1.28959  -11.49   <2e-16 ***
dose          0.24942    0.02139   11.66   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   7.3849  on 6  degrees of freedom
AIC: 37.583
```
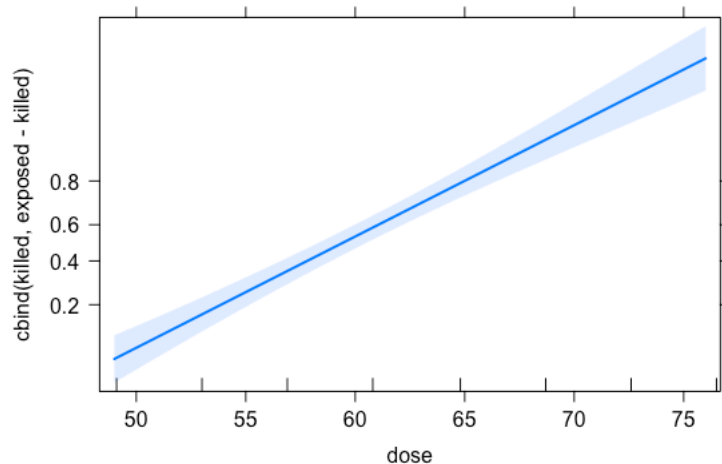
```
betas <- coef(b1)

exp(betas)
 (Intercept)         dose
3.651351e-07 1.283275e+00


exp(cbind(OR = coef(b1), confint(b1)))

                     OR        2.5 %        97.5 %
(Intercept) 3.651351e-07 2.530115e-08 4.02275e-06
dose        1.283275e+00 1.233305e+00 1.34146e+00
```


dose effect plot

example   beetles

```r
b1 <- glm(cbind(killed, exposed - killed) ~ dose,
     family = binomial)

plot(dose, killed/exposed, xlab="Dose, mf/L",
    ylab="Probabilty of beetle mortality",
     ylim=c(0,1),
      pch = 21, bg = "black")

x <- seq(min(dose), max(dose), length = 100)

lines(x, predict(b1, data.frame(dose=x),
     type="response"), lwd = 2, type = "l",
      col = "red")

# predict specific values
predict(b1, data.frame(dose = c(50, 65, 75)), type =
"response")

1          2          3
0.08688964 0.80043514 0.97982979
```
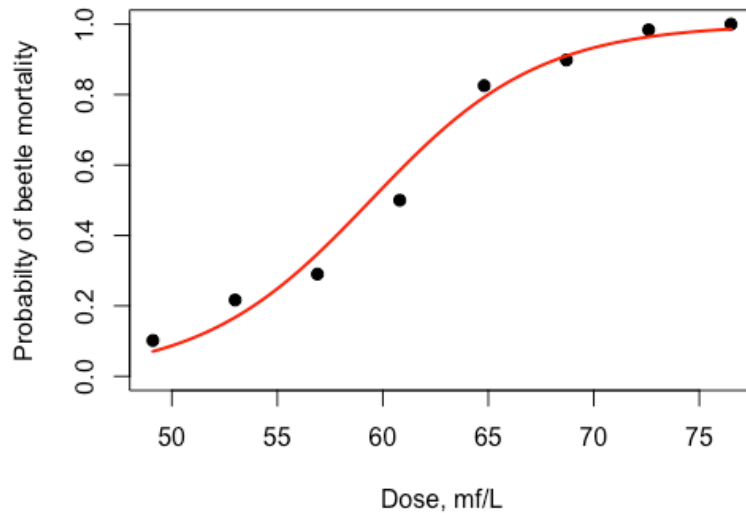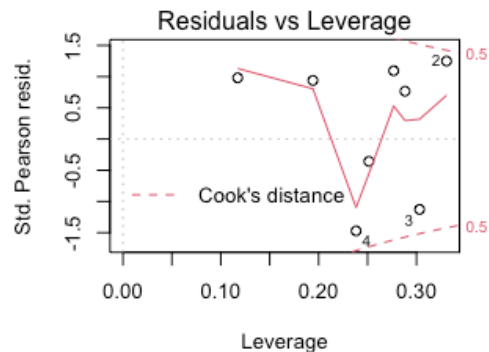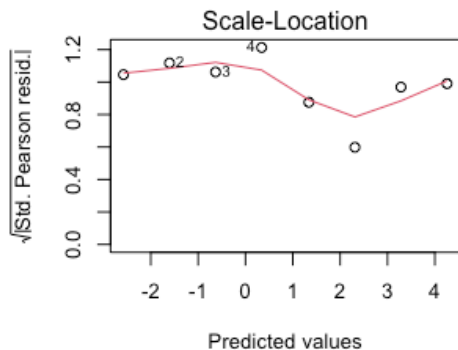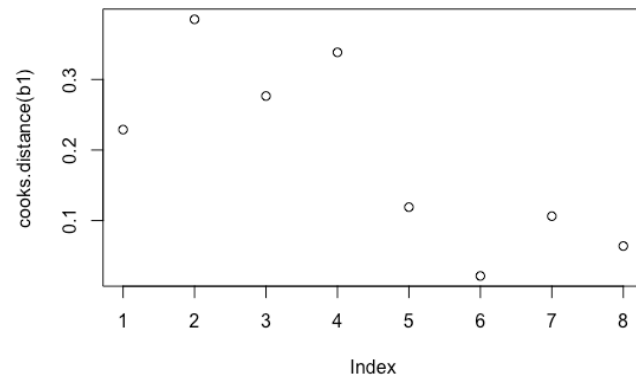
example   beetles
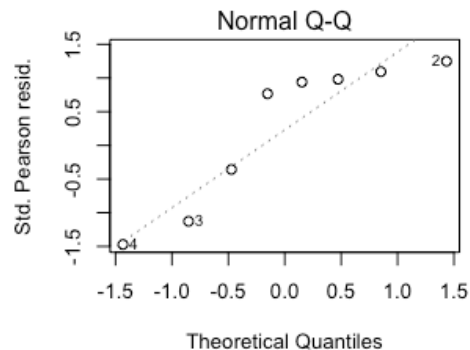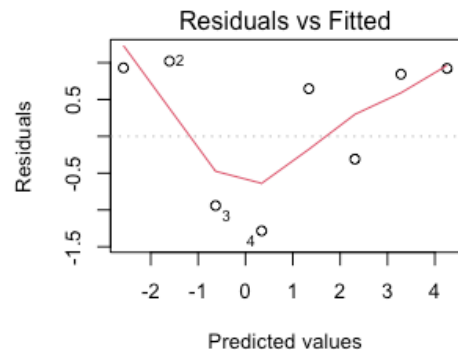

```
plot(cooks.distance(b1))
```

goodness-of-fit

deviance test for lack-of-fit

- significant p-value suggests that a residual deviance as large or larger than that observed under the model is highly unlikely
  - model is not adequate or lack of fit
- deviance: -2 log(likelihood) evaluated at the maximum likelihood of the parameters
- change in deviance has a Chi-squared distribution with df's equal to the change in number of parameters in the model

```
1-pchisq(b1$dev, b1$df.resid)
[1] 0.2867122

pchisq(b1$dev, b1$df.resid, lower.tail = F)
[1] 0.2867122
```

pseudo-R^2

```
require(pscl)
pR2(b1)
      llh         llhNull          G2
-16.7914741  -155.2002438  276.8175395

  McFadden          r2ML            r2CU
 0.8918077      1.0000000      1.0000000
```
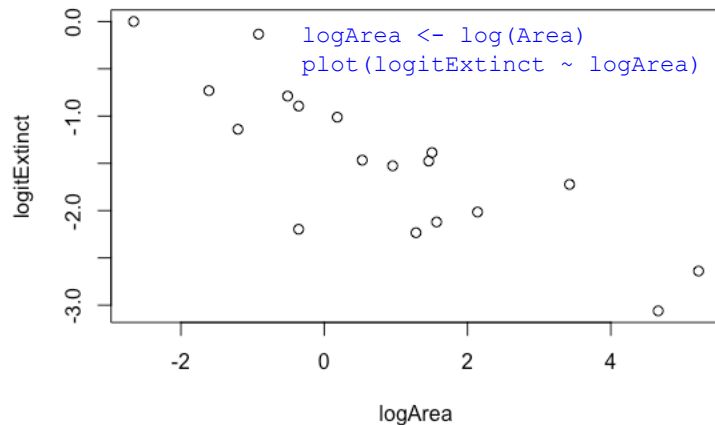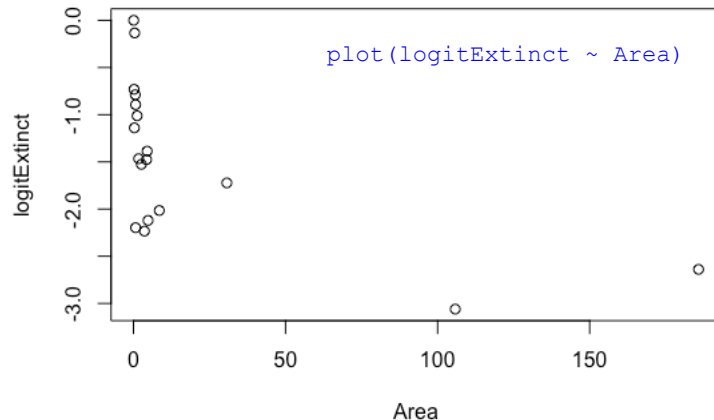
example — Krunnit Island Archipelago

	Island	Area	AtRisk	Extinct
1	Ulkokrunni	185.8	75	5
2	Maakrunni	105.8	67	3
3	Ristikari	30.7	66	10
4	Isonkivenletto	8.5	51	6
5	Hietakraasukka	4.8	28	3
6	Kraasukka	4.5	20	4


```
plot(logitExtinct ~ Area)
```

- logit of the probability of extinctions is assumed to be linear in X = log(Area)
- logistic regression on binary responses is a special case of the binomial model in which all n_i's are 1

```
proportionExtinct <- Extinct/AtRisk
oddsExtinct   <- proportionExtinct/(1 - proportionExtinct)
logitExtinct <- log(oddsExtinct)    # Logit = Log Odds
```



```
logArea <- log(Area)
plot(logitExtinct ~ logArea)
```

example

```
k2 <- glm(cbind(Extinct, AtRisk - Extinct) ~
          log(Area), family = binomial)
 summary(k2)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
log(Area)   -0.29710    0.05485  -5.416 6.08e-08 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.338  on 17  degrees of freedom
Residual deviance: 12.062  on 16  degrees of freedom
AIC: 75.394
```
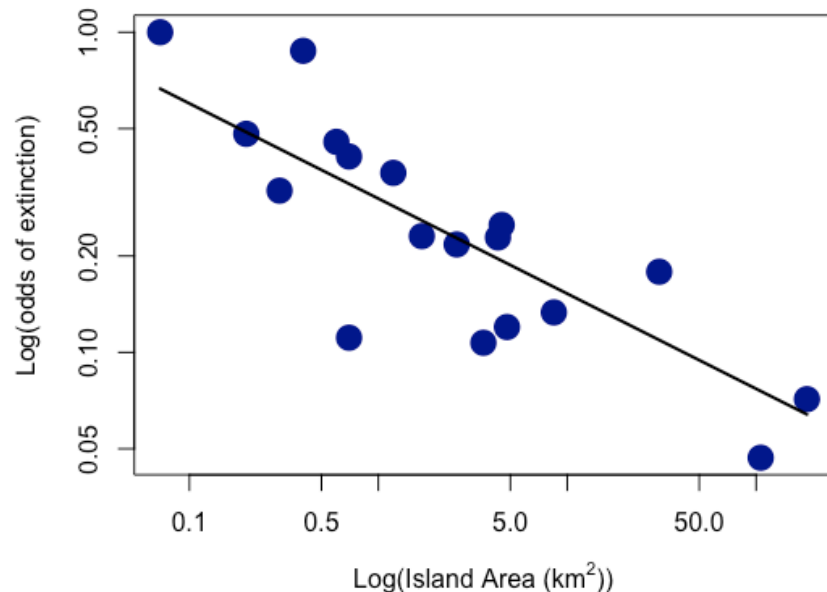


```
betas <- coef(k2)
 odds <- exp(betas)
  odds
(Intercept)    log(Area)
  0.3023423    0.7429670
```

for every 1 unit increase in log(Area), odds of extinction is 74.2% that of the previous area, or a reduction of 25.8% in odds of extinction

goodness of fit


```
pchisq(k2$deviance, k2$df.residual, lower.tail = F)
[1] 0.7397351

pR2(k2)[4]
```

```
fitting null model for pseudo-r2
 McFadden
0.3179176
```
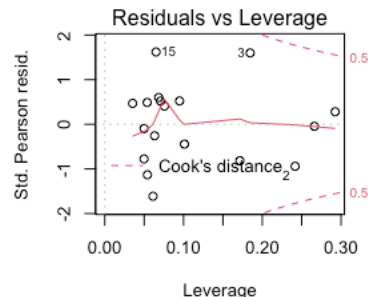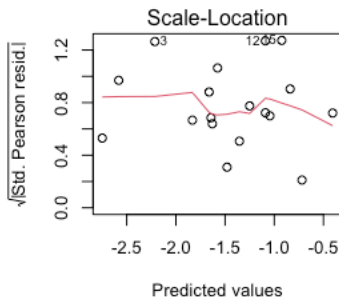
```
knull <- update(k2, .~.-log(Area))
 lrtest(knull, k2)
```
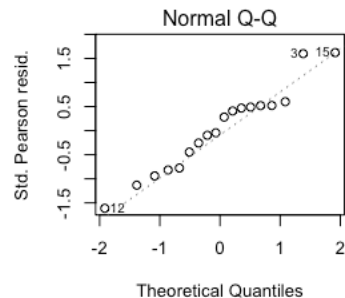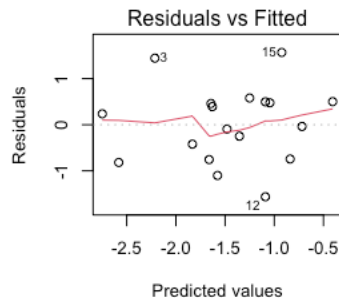
```
Likelihood ratio test

Model 1: cbind(Extinct, AtRisk - Extinct) ~ 1
Model 2: cbind(Extinct, AtRisk - Extinct) ~ log(Area)
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -52.335
2   2 -35.697  1 33.276  7.994e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(proportionExtinct ~ Area, log="xy",
  ylab="Proportions of species",
  xlab=expression(paste("Log(Island Area (km"^"2",")")")),
  pch=21, lwd=2, bg="darkblue", col="darkblue", cex=2)

dummyArea  <- seq(min(Area),max(Area),length=50)
lp <- betas[1] + betas[2]*log(dummyArea)
myProbability <- exp(lp)/(1 + exp(lp))

lines(myProbability ~ dummyArea,lwd=2)
```
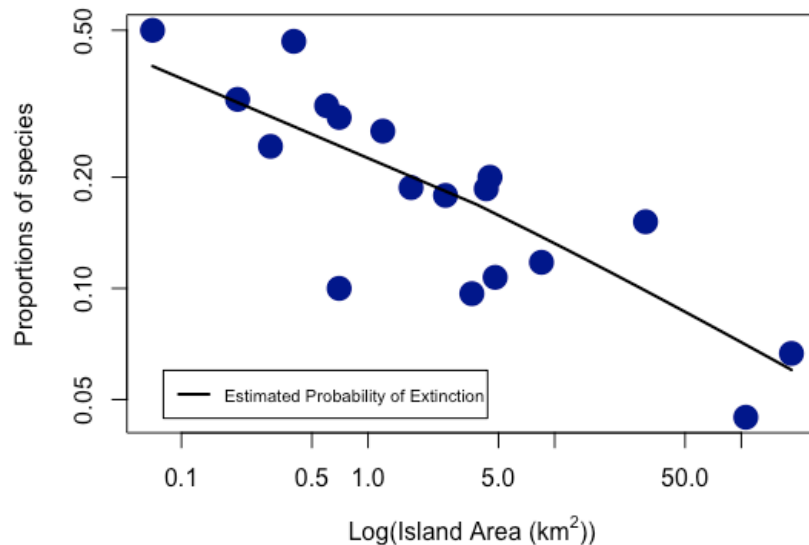
```
1-exp(betas[2]*log(2))
log(Area)
0.1861153

1-2^confint(k2, 2)
   2.5 %      97.5 %
0.2462041  0.1247743
```

$$ln(x^y) = y \cdot ln(x)$$

- with each doubling of island area, there is a ~ 19% reduction in the odds of extinction
- each halving of area is associated with a 23% increase in extinction

Questions?