ENV 710: Lecture 18

binary logistic regression

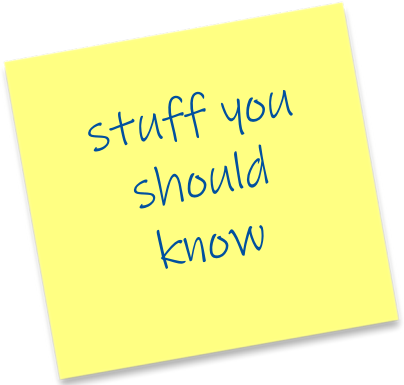
Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

generalized linear models

binary logistic regression

learning goals

- generalized linear models
- binary logistic regression
  - binary and dichotomous responses
  - odds, log-odds, probabilities
  - what is it? when to use it?
  - interpretation of coefficients
  - assumptions of logistic regression



review – GLM's

- link function g(•) links the response to the linear predictor

(transforms *Y* to be linear with predictors)

$$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

- identity link — $g(y) = y$
- log link (log of the mean) — $g(y) = log(y)$
- logit link (log of the odds) — $g(y) = log[y/(1-y)]$

linear regression: $Y \sim N(\mu, \sigma^2)$ — $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$

Poisson regression: $Y \sim Pois(\lambda)$ — $log(y) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

logistic regression: $Y \sim Binomial(p, n)$ — $log[y/(1-y)] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

logistic regression

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

- model as a conditional expectation

$$E[y|x] = P(y = 1|x)$$

$$E[y|x] = \beta_0 + \beta_1 x$$

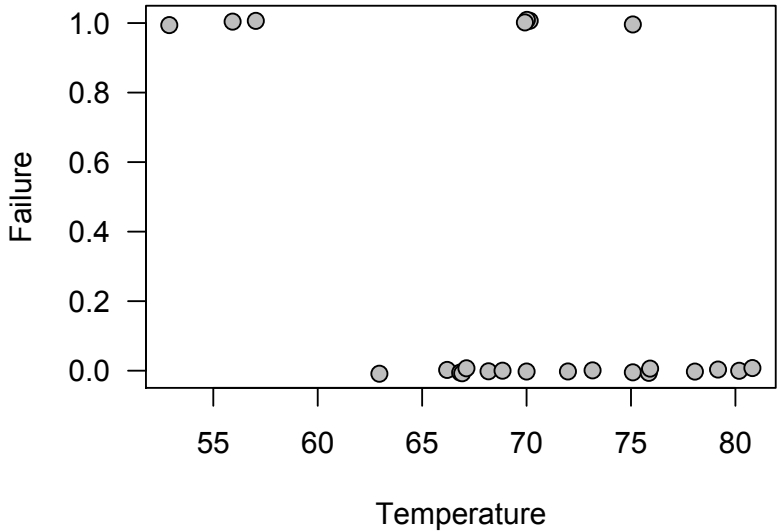
probability of success given *x*, when *y* is binary

example

space shuttle

How does outside temperature affect the odds of O-ring failure?

O-ring failure is the "success" that is being modeled.

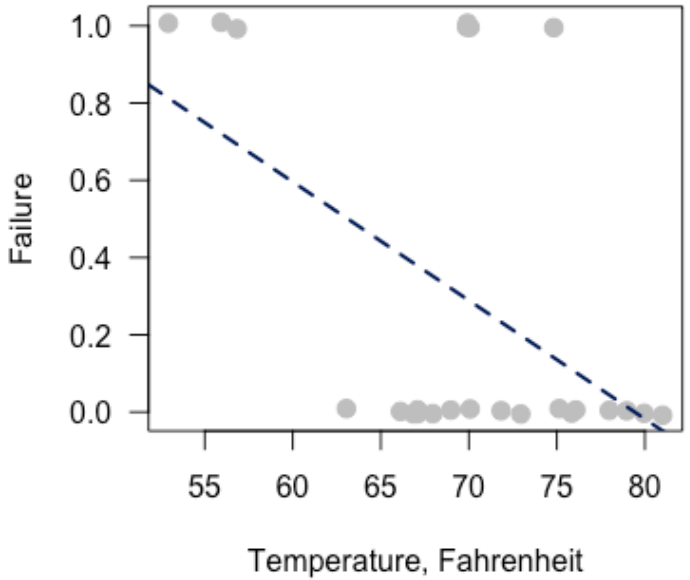


	Temp	Failure	Fail
1	53	Yes	1
2	56	Yes	1
3	57	Yes	1
4	63	No	0
5	66	No	0
6	67	No	0
7	67	No	0
8	67	No	0
9	68	No	0
10	69	No	0
11	70	No	0
12	70	Yes	1
13	70	Yes	1
14	70	Yes	1
15	72	No	0
16	73	No	0
17	75	No	0
18	75	Yes	1
19	76	No	0
20	76	No	0
21	78	No	0
22	79	No	0
23	80	No	0
24	81	No	0

example space shuttle

linear regression fit

```
lm(formula = Fail ~ Temperature, data = ss)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.43729    0.82340    2.96  0.00723 **
Temperature -0.03069    0.01171   -2.62  0.01565 *

Residual standard error: 0.4145 on 22 degrees of freedom
Multiple R-squared:  0.2378, Adjusted R-squared:  0.2031
F-statistic: 6.863 on 1 and 22 DF,  p-value: 0.01565
```



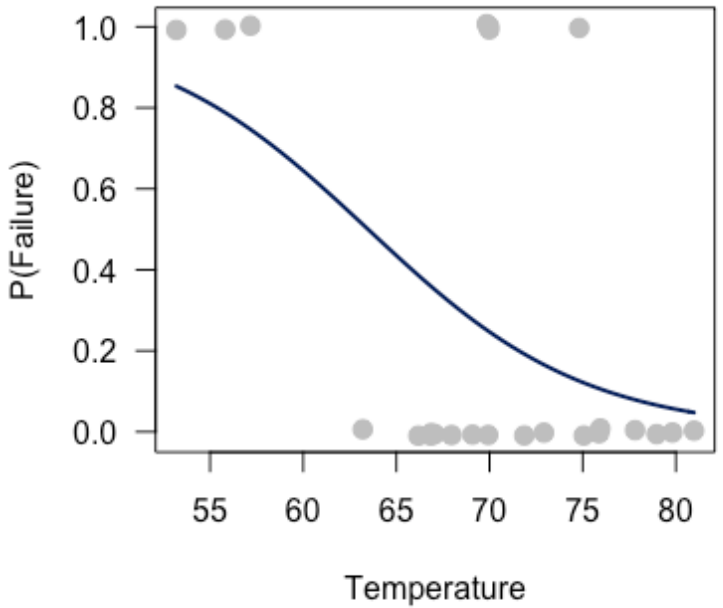
- error does not satisfy: $\varepsilon \sim N(0, \sigma^2)$
- predictions can be outside of [0, 1]

  for $x = 81, \hat{y} = -0.05$

better way to model a 0/1 response?

- at each value of *X* there is a certain chance of a 0 and a chance of a 1
- 1 is more likely at low temps, 0 is more like at high temps
- *P*(*Y*=1) changes with *X*
- consider the probability of getting a 1 given the *X*-value(s) in our modeling

$$\pi_i = P(Y_i = 1|X_i)$$



odds of an event

in gambling, odds represent the ratio between the amounts staked by the bookmaker and the gambler

"6 to 5 odds" – $6 in profit for every $5 wagered

6/5 +1 = 2.2 times original $5 wager = $11.00

"20 to 1 odds" – $20 in profit for every $1 wagered

10/1 + 1 = 21 times original $1 wager = $21

odds of an event

in statistics odds express relative probabilities, generally articulated as odds in favor

odds are the ratio of the probability of an event occurring to the probability of it not occurring

$$\text{odds} = \frac{P(\text{success})}{P(\text{failure})} = \frac{p}{q} = \frac{\pi}{1-\pi}$$

- odds of rolling a 3 with a fair die are 1:5
- odds against rolling a 3 are 5:1
- probability of rolling a 3 is 1/6 of 0.17

odds of an event

- when two coins are flipped
  - P(2 heads) = ½ x ½ = 1/4
  - P(not 2 heads) = 3/4

- odds in favor of getting two heads is:

$$odds = \frac{P(\text{success})}{P(\text{failure})} = \frac{p}{q} = \frac{\pi}{1-\pi}$$

$$odds = \frac{P(2 \text{ heads})}{P(\text{not } 2 \text{ heads})} = \frac{1/4}{3/4} = 1/3$$

- odds for 2 heads are 1 to 3
- odds are 3 to 1 against getting 2 heads

- for a binary variable *Y*, odds in favor of *Y*=1:

$$odds = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$$

- if *P*(pollution) = 0.33, then the odds of pollution is:

$$\frac{0.333}{1-0.333} = \frac{0.333}{0.667} = 0.50$$

logistic regression

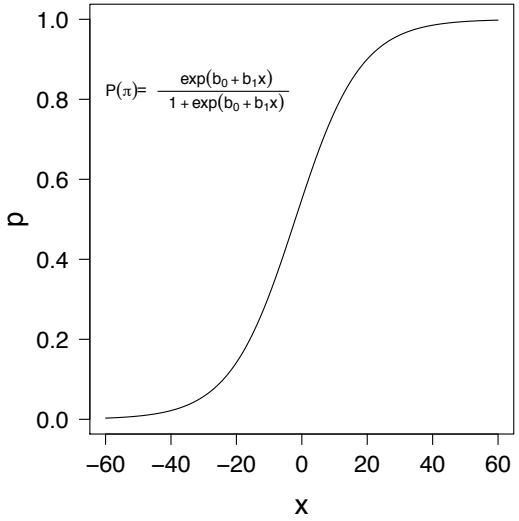
- model response variable as a transformation of P(Y_i=1) or P(π_i)
- link is the logit transformation: the $\log_e$ of the odds that $Y_i = 1$ or the $\log_e$ odds of "success"

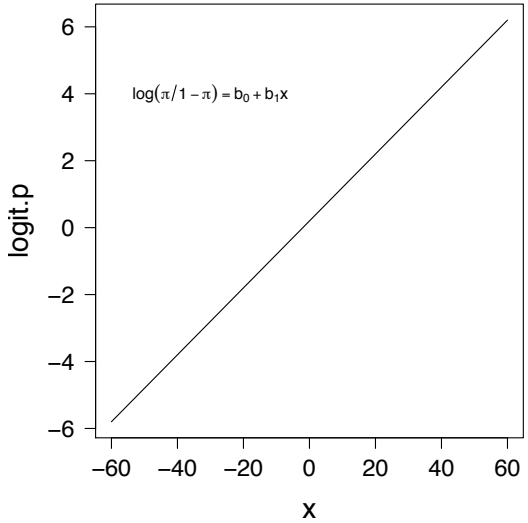
$$logit(\pi) = log[\pi/(1-\pi)]$$

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

- inverse of the logit: $\pi = \frac{exp(x)}{1+exp(x)}$

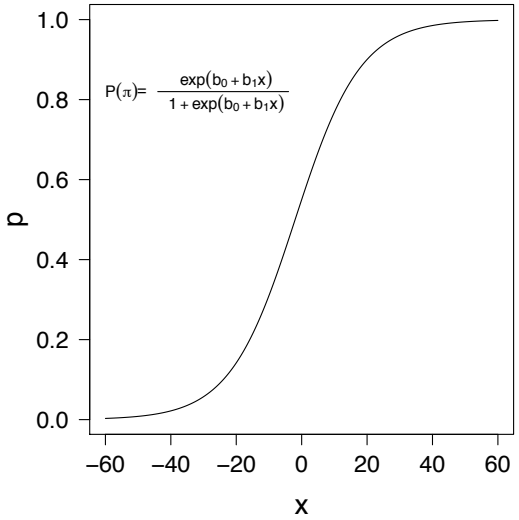
$$\pi = \frac{exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}{1+exp(\beta_0 + \beta_1 X_k + ... + \beta_k X_k)}$$

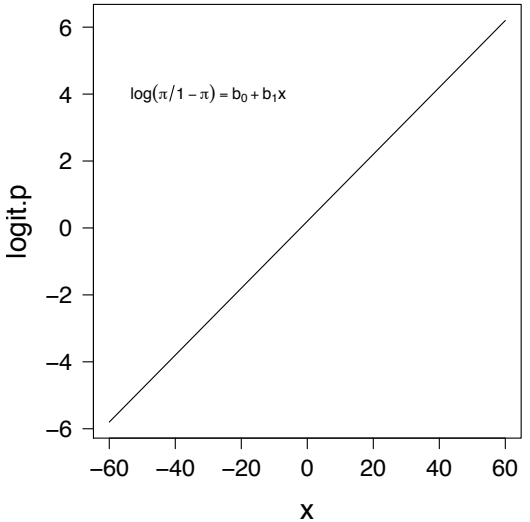




assumptions

- data Y_1, Y_2, ..., Y_n are independently distributed
- response, Y_i, is binary or dichotomous
- linear relationship between the logit of the response and the explanatory variables
- little or no multicollinearity between predictor variables
- large sample sizes, the larger the sample size the more reliable the results
- uses maximum likelihood estimation (MLE) to estimate the parameters





example

space shuttle

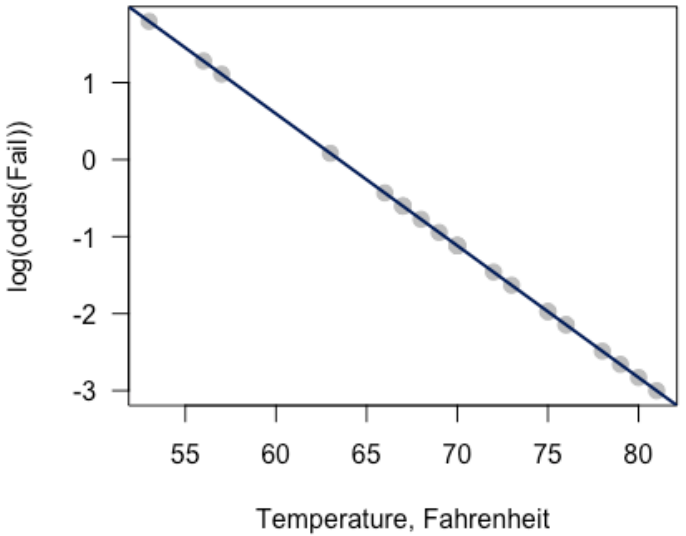
```
glm(Fail ~ Temperature, family = binomial,
      data = ss)
```

```
Call:
glm(formula = Fail ~ Temperature, family =
binomial, data = ss)

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 *

(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
```

interpretation of parameters

- intercept β_0: when X_i or temperature is 0
  - β_0 are the log-odds that an O-ring will fail: 10.8535
  - $\exp(\beta_0)$ are the odds that an O-ring will fail: 52857 or 52857:1!
  - probability of failure is: $\pi_i = \frac{e^{\beta_0}}{1+e^{\beta_0}} = 0.999$

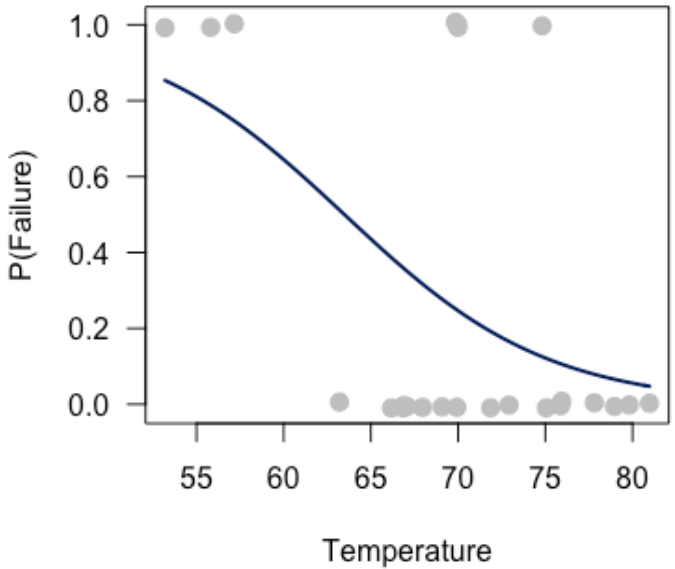
- slope β_1
  - β_1= -0.171: every 1-unit change in temperature, X, the log odds of failure, decreases by 0.171
  - for every 1-unit change in temperature, the odds of failure, $\exp(\beta_1)$, decrease by a factor of 0.84 (or 16%)



example space shuttle

- what is the predicted probability of failure at 70°?

$$P(\pi) = \frac{exp(10.87 - 0.171X_i)}{1 + exp(10.87 - 0.171X_i)} = \frac{exp(10.87 - 0.171 \cdot 70)}{1 + exp(10.87 - 0.171 \cdot 70)} = 0.247$$




```
require(boot)
inv.logit(10.87535 - 0.17132 * 70)
[1] 0.2465589

inv.logit(predict(lr1,
data.frame(Temperature = 78)))
[1] 0.07672851
```

likelihood ratio test

- compares two nested models: simpler model is a special case of the complex model

$$LRT = -2ln\frac{\ell_s}{\ell_c} = \frac{deviance_s}{deviance_c}$$

```
lr0 <- glm(Fail~1, family=binomial,
      data=ss)

lr1 <- glm(Fail~Temperature,
      family=binomial, data=ss)
```

```
require(lmtest)
lrtest(lr0, lr1)

Likelihood ratio test

Model 1: Fail ~ 1
Model 2: Fail ~ Temperature
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -14.487
2   2 -11.515   1  5.9441 0.01477 *
```

```
anova(lr0, lr1, test="Chisq")

Analysis of Deviance Table
Model 1: Fail ~ 1
Model 2: Fail ~ Temperature
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        23     28.975
2        22     23.030  1   5.9441  0.01477 *
```

pseudo R^2

```
require(pscl)
pR2(lr1)
llh       llhNull      G2         McFadden   r2ML       r2CU
-32.8613  -58.6387     51.5549    0.4395     0.3997     0.5820
```

example

African plots

Model the African tree plots as a binary response variable, with biomass ≤ 226 Mg ha^{-1} coded as 0 and biomass > 226 Mg ha^{-1} coded as 1.

What determines the probability of having a plot with biomass > 226 Mg ha^{-1}?

1. determine the best-fitting logistic regression model
2. interpret the coefficients
3. plot the effect of wood density on the probability of having a biomass greater than 225 Mg ha^{-1}.

```
glm(Bin ~ WD.avg + Ht.avg + DistVillage, family = binomial, data)
```

example

African plots

```
lr1 <- glm(Bin ~ WD.avg + Ht.avg +
           DistVillage, family =
           binomial, data = afr.dat1)
lr2 <- update(lr1, .~. -DistVillage)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -16.44943    3.85409  -4.268 1.97e-05 ***
WD.avg       18.01538    4.72513   3.813 0.000137 ***
Ht.avg        0.33660    0.09882   3.406 0.000659 ***
---
    Null deviance: 117.278  on 100  degrees of freedom
Residual deviance:  66.001  on  98  degrees of freedom
AIC: 72.001
```

```
log.odds <- coef(lr2)
odds <- exp(coef(lr2))

odds
 (Intercept)       WD.avg       Ht.avg
7.179667e-08 6.667789e+07 1.400176e+00


inv.logit(log.odds)
 (Intercept)       WD.avg       Ht.avg
7.179667e-08 1.000000e+00 5.833639e-01
```

example

African plots

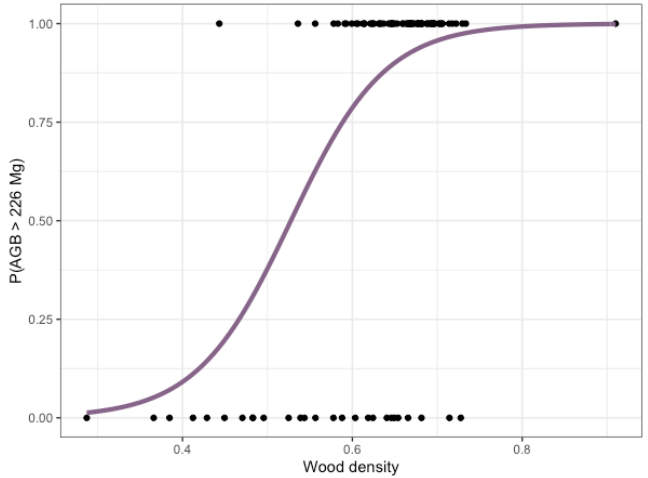
What determines the probability of having a plot with biomass > 226 Mg ha^{-1}?

To plot logistic regression models

```
# Extract coefficients from best model (here named lr2)
lrcoef <- coef(lr2)

# Create range of wood density values to graph over
x <- with(afr.dat1, seq(min(WD.avg), max(WD.avg), length = 100))

# Write the function for the final model, note use of inv.logit()
#   from boot package

lrfit <- function(x)inv.logit(lrcoef[1] + lrcoef[2]*x +
             lrcoef[3]*mean(afr.dat1$Ht.avg))
```

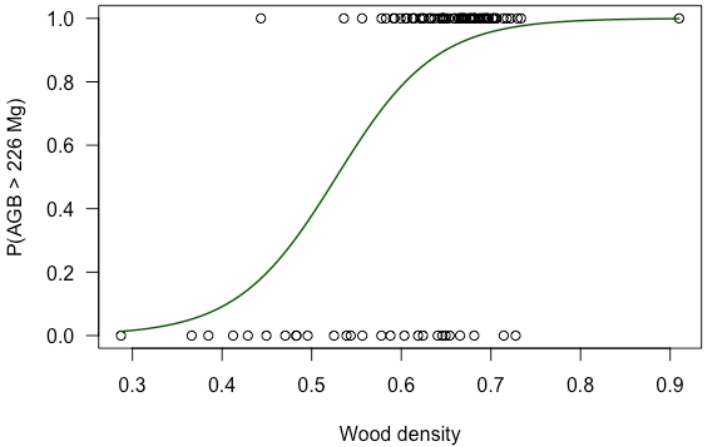
Plot logistic regression model with `ggplot()`

```
ggplot(afr.dat1, aes(x = WD.avg, y = Bin)) +
       geom_point() +
       stat_function(fun = lrfit, geom = "line",
                         col = "plum4", lwd = 1.5) +
       ylab("P(AGB > 226 mg)") +
       xlab("Wood density") +
       theme_bw()
```



Plot logistic regression models with `plot()`

```
with(afr.dat1, plot(WD.avg, Bin,
     ylab = c("P(AGB > 226 Mg)"),
     xlab = c("Wood density"), las = 1))

curve(lrfit,  add = T, col = "darkgreen",
      lwd = 1.5)
```



Questions?