# Chapter 8

Logistic Regression

STARTED: Monday 3/28/22 (Week 10_Lecture 26) @ 45 min mark.

# Introduction and Setup

START Wed 3/30/22 (Week 10, Lecture 27)

# Linear Models?

- Recall: a linear model is one that can be written in matrix form as $y = X\beta + e$. That is, we can express y as a linear combination of the *parameters* and the error term.

- Models with transformed response variables are not linear models. For example, $\log(y) = \beta_0 + \beta_1 x + e$ can be rewritten as $y = e^{\beta_0 + \beta_1 x} e^{e}$

  - (The e in the exponent is the error term; all others are the exponential function.)
  - The relationship between y and the parameters is nonlinear; note the error term is also *multiplicative.*

- Logistic regression models (this chapter) are also not linear models.

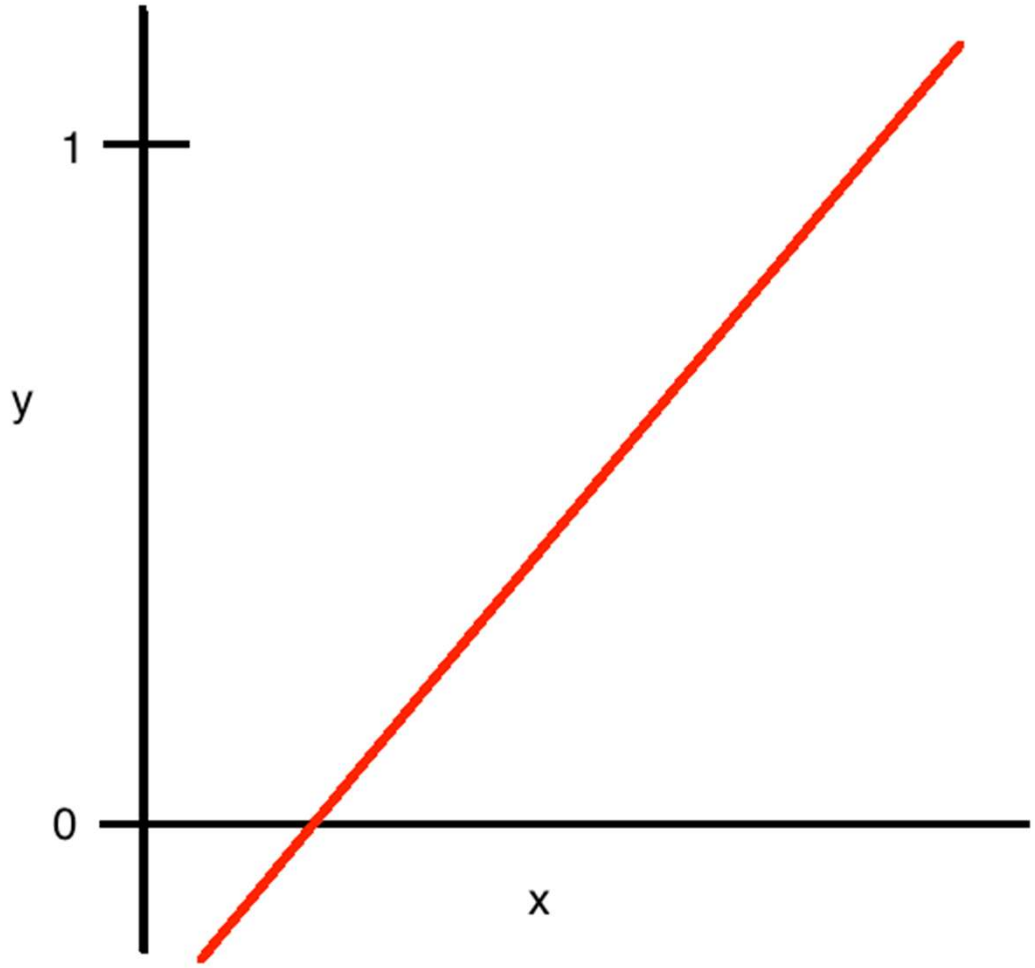
# Logistic Regression

- So far: response variable – quantitative
- Chapter 8: response variable – categorical
  - X = HSGPA, Y = Accepted to TAMU
  - X = Amount of credit card transaction, Y = Fraudulent (Y/N)
- Ideally such responses follow a binomial distribution in which case the appropriate model is a logistic regression model.

# Logistic Regression

| | Response Variable Quantitative | Response Variable Categorical |
|---|---|---|
| Explanatory Variable Quantitative | Regression | Logistic Regression |
| Explanatory Variable Categorical | ANOVA | $\chi^2$ Tests / Logistic Regression |
| Explanatory Variables Categorical & Quantitative | ANCOVA | Logistic Regression |

# Logistic Regression




Why not use our usual linear regression methods, creating a dummy variable for y, and fitting a least squares regression line between x and y?

# Logistic Regression

- Why not use the usual line?
  1. Possible predictions out of bounds
  2. Nonconstant variance: If the response Y has a Bernoulli (binomial with n = 1) distribution:
     - E[Y | X=x] = p(x) (This means the probability of success p is a function of the explanatory variable x: p changes with x.)
     - Var(Y | X=x) = p(x) (1 – p(x)) That is, the variance of Y changes with x when the mean changes.
  3. When the response Y has a Bernoulli distribution, the logistic regression model correctly models the mean.

# Logistic Regression

- How do we turn a binary variable into something continuous?
  - What about using proportion of successes? (if we have mult. obs at each value of x)

    Don't want predictions outside possible range $[0, 1]$

  - What about using odds of success?

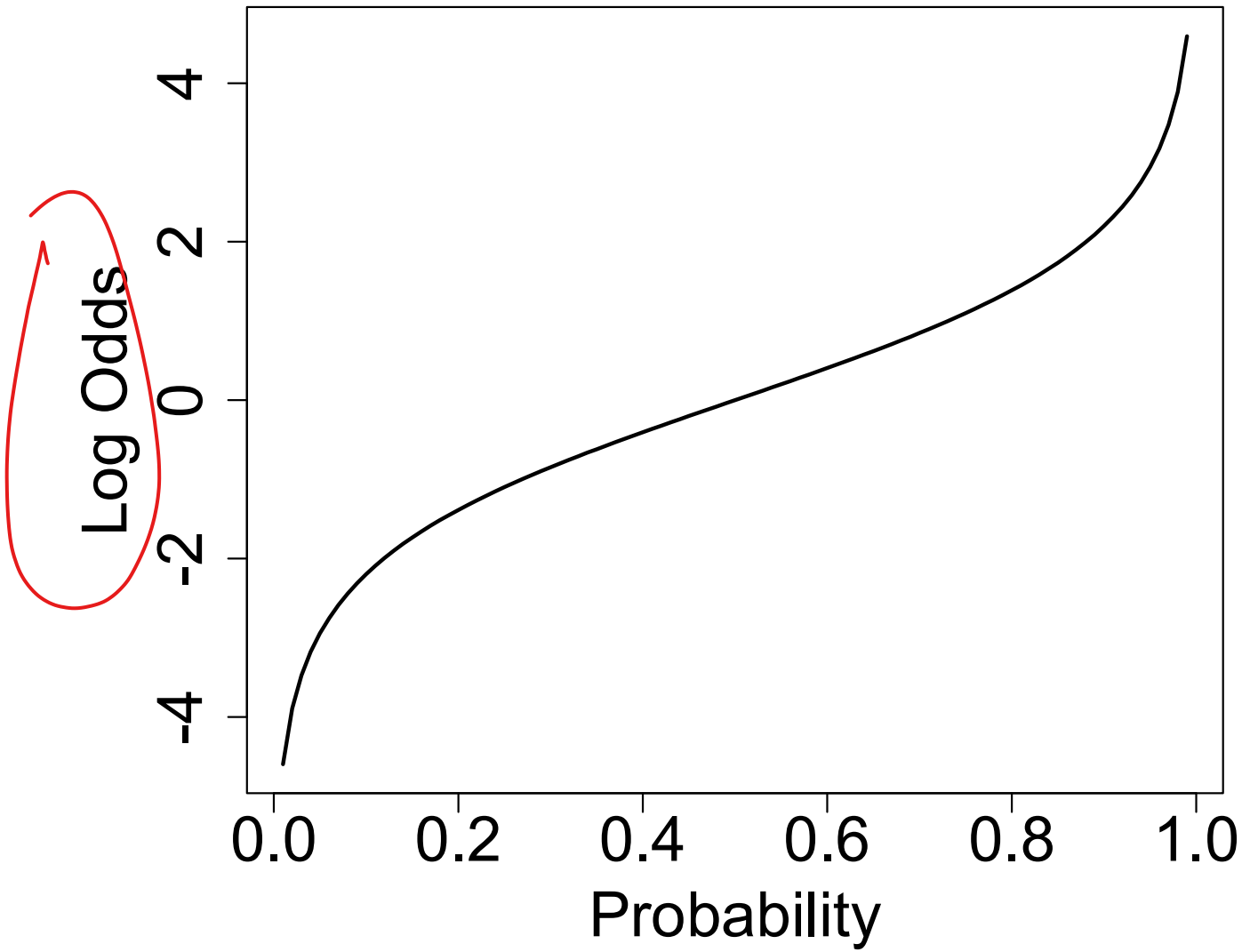
    $$odds = \frac{p}{1-p}$$

    - Worried about negative values

      [regression can give us negative values for $p \Rightarrow$ odds would still be negative]

    - Not typically a linear relationship between odds and the X's.

# From Probability to Log Odds



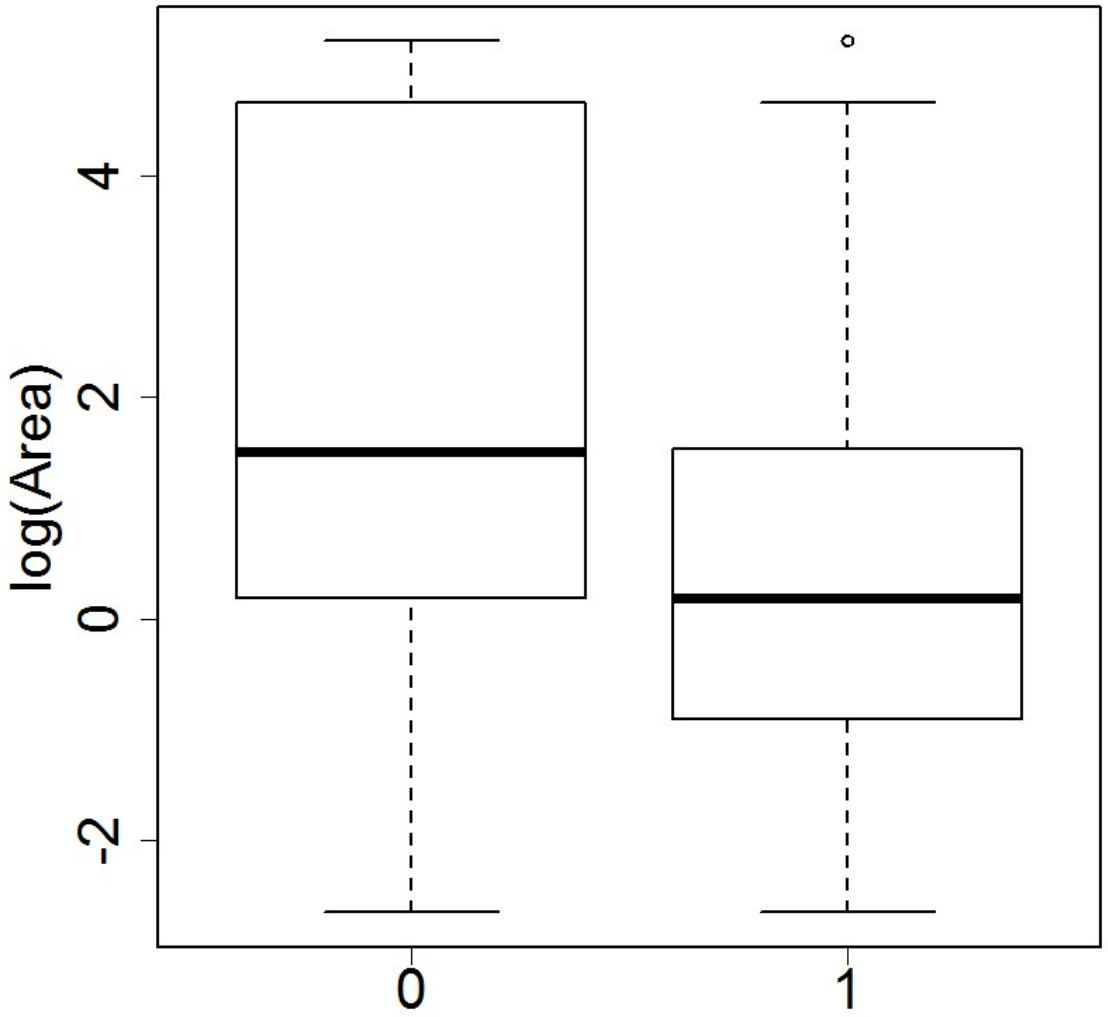
# From Probability to Log Odds

| Probability | 0.1 | 0.25 | 0.5 | 0.75 | 0.9 |
|---|---|---|---|---|---|
| Odds | 0.111 | 0.333 | 1 | 3 | 9 |
| Log(Odds) | -2.2 | -1.1 | 0 | 1.1 | 2.2 |

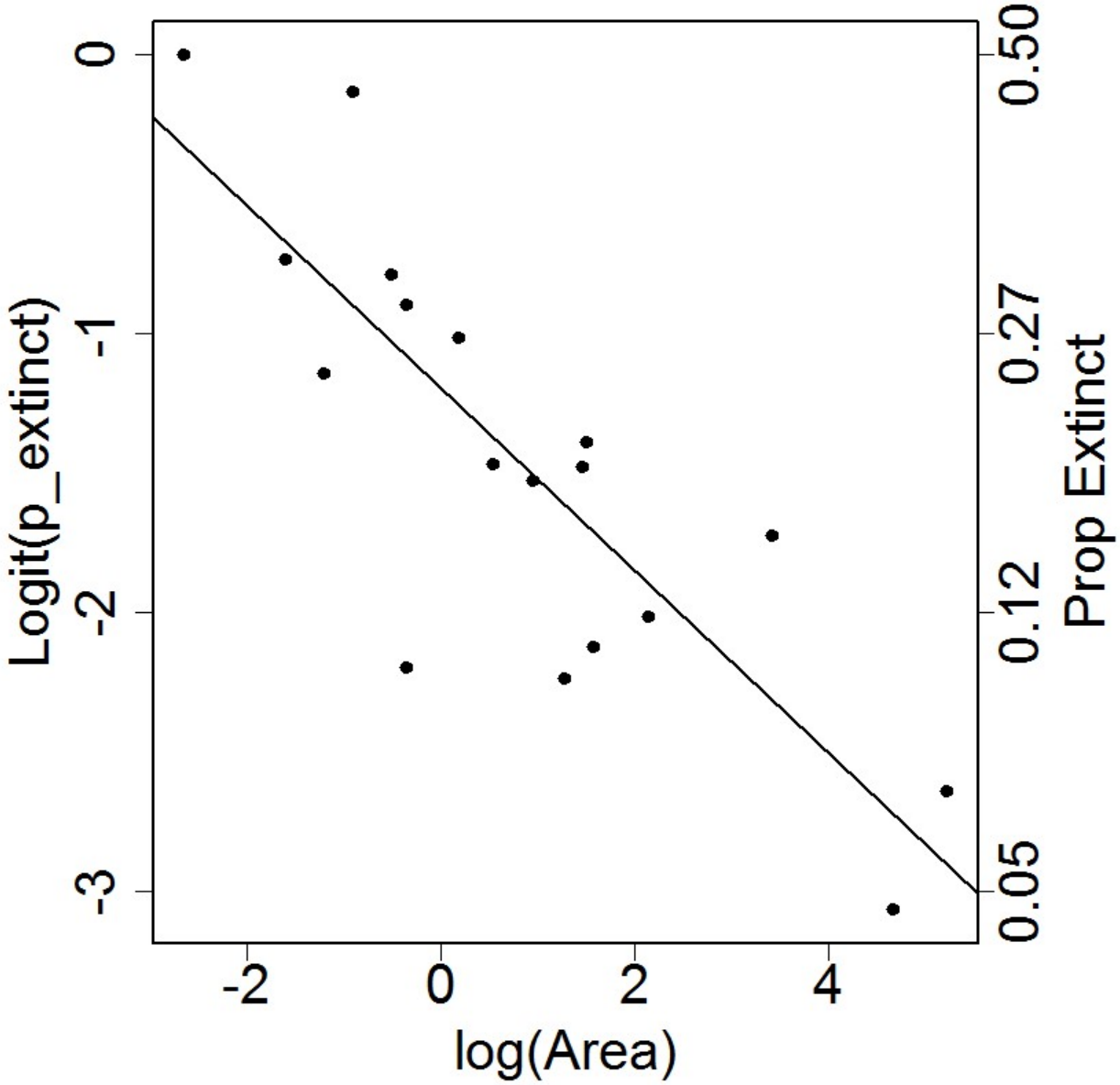
- **Logit(p)** = log odds (p) = $\log\left(\frac{p}{1-p}\right)$

# Example: Bird Extinctions on Islands


# Example: Bird Extinctions on Islands



# Example: Bird Extinctions on Islands



# Example: Bird Extinctions on Islands

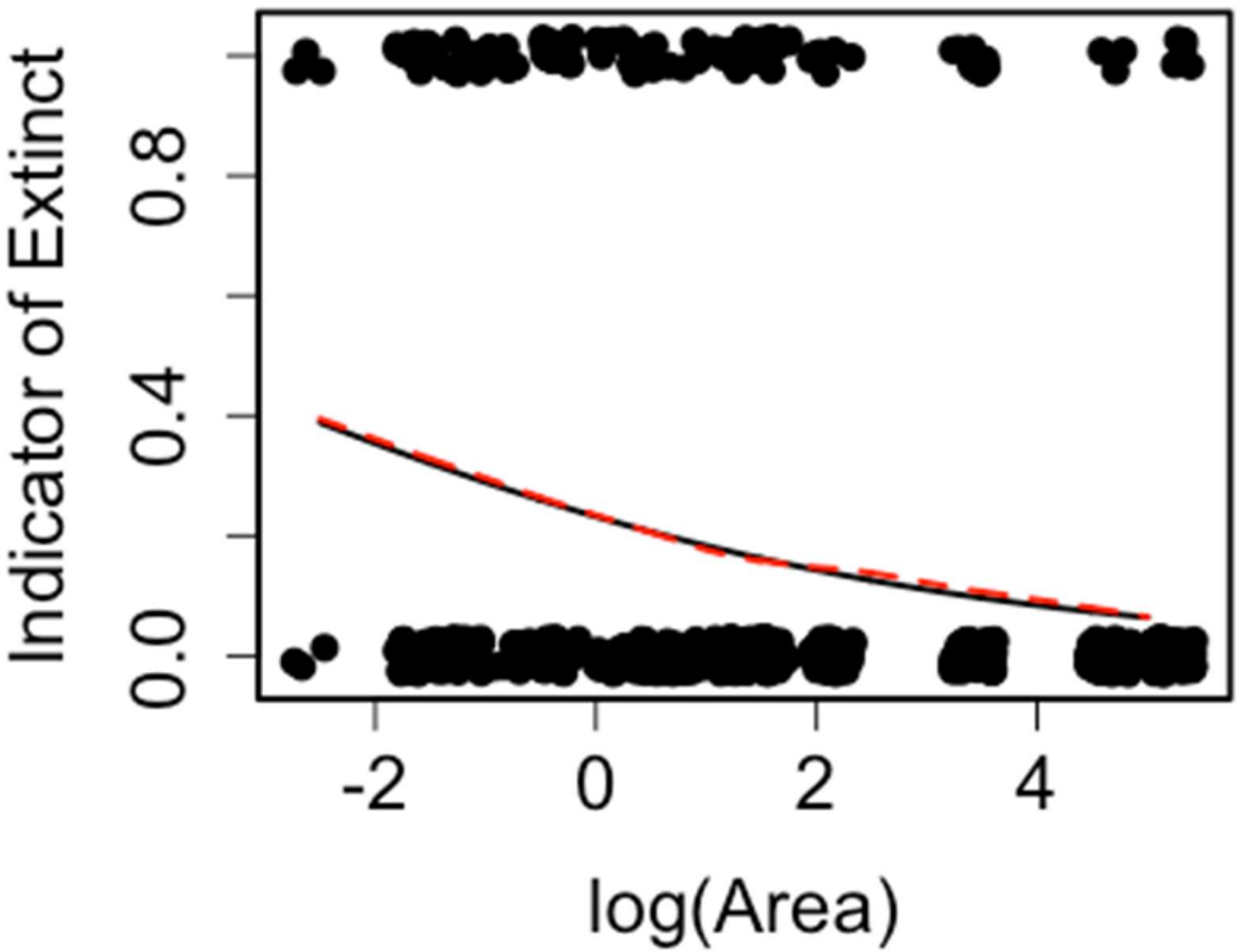
- Model: $\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 x + e$

- Log odds is linear in x; probability has an S-shaped relationship with x.
- X = log(Area), Y = 1 if extinct, 0 if not extinct
- First category in alphabetical order = failure, or 0 = failure.

not t-anywhere.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
xnew        -0.29710    0.05485  -5.416 6.08e-08 ***
```

# Example: Bird Extinctions on Islands



# Example: Bird Extinctions on Islands

If an island has 50 km$^2$, what is the estimated probability that a species will go extinct there?

$$\log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = -1.196 - 0.297\, log(Area)$$

$$\log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = -2.36$$

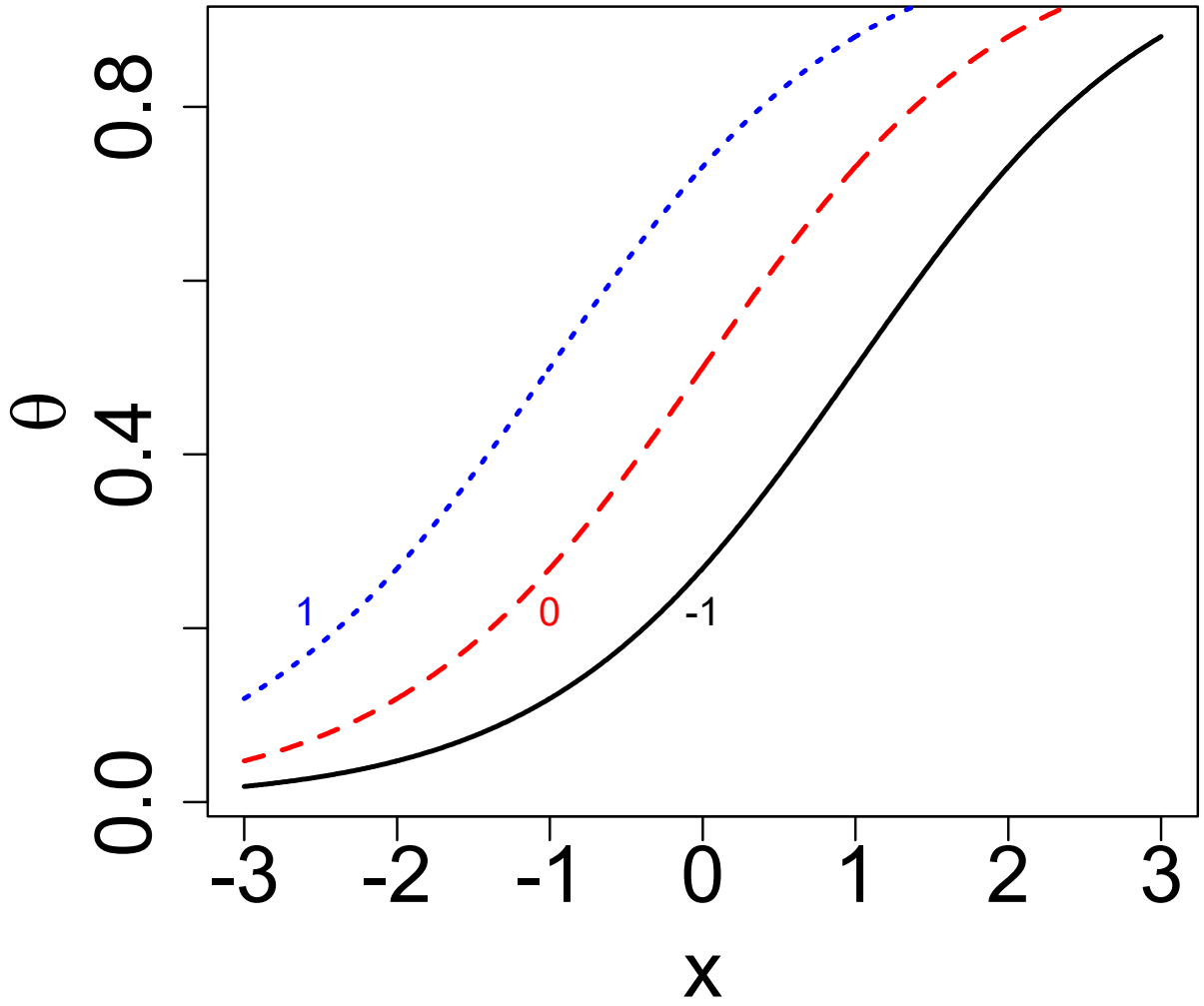
$$\frac{\hat{\theta}}{1-\hat{\theta}} = exp(-2.36)$$

$$\hat{\theta} = \frac{exp(-2.36)}{1+exp(-2.36))} = 0.086$$

The estimated probability that a species will go extinct on an island with 50 km$^2$ is 0.086.

# How $\beta_0$ Affects Model

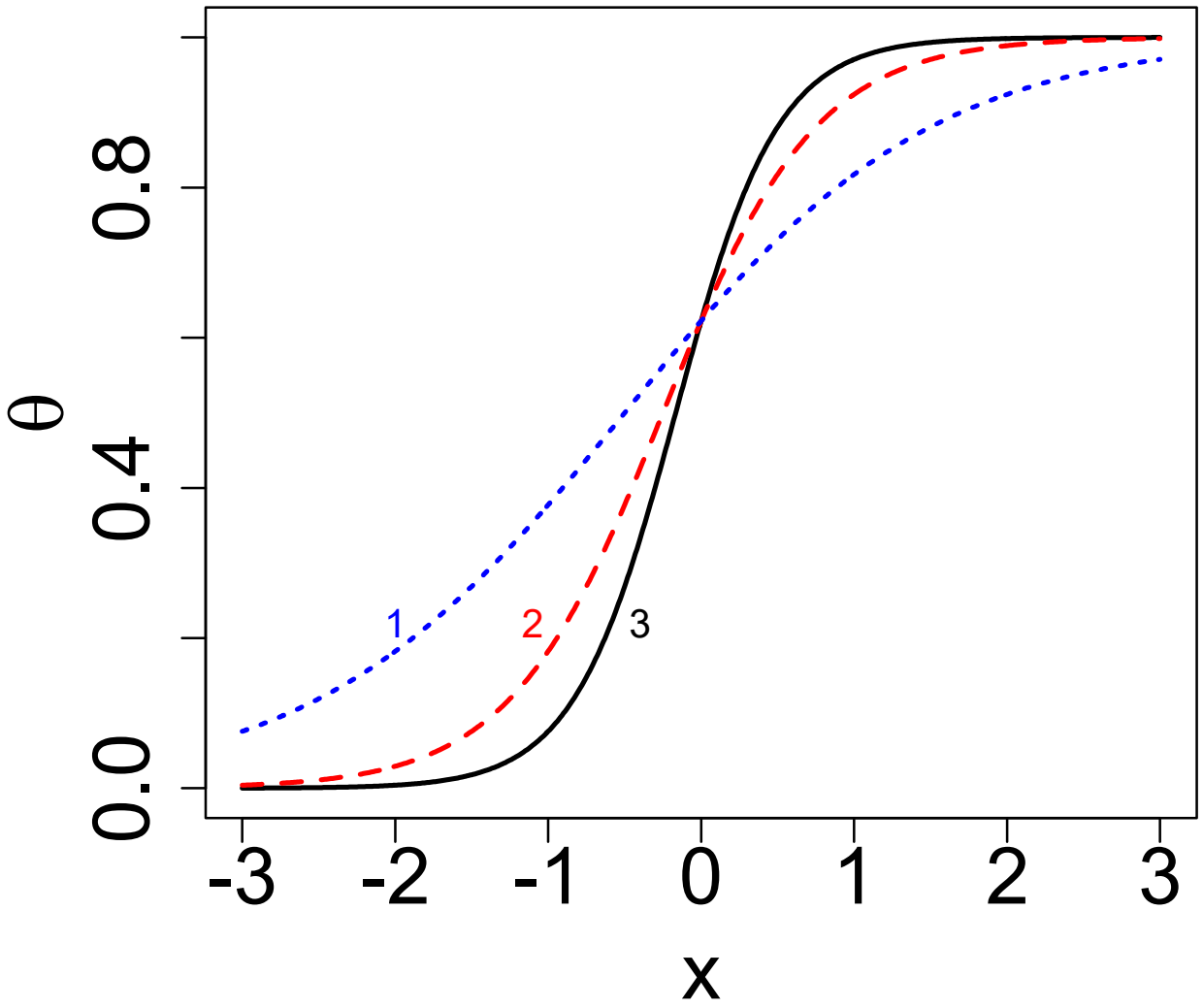
$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 x$




$\beta_0$ as labeled,
$\beta_1 = 1$

- $\beta_0$ determines the vertical location of the curve.
- $\theta = 0.5$ when $x = -\beta_0/\beta_1$

# How $\beta_1$ Affects Model




$\beta_0 = 0.5$,
$\beta_1$ as labeled

- The larger $|\beta_1|$, the steeper the slope .
- If $\beta_1 > 0$, the model has a positive slope.
- At $\theta = 0.5$, the slope of the model predicting probabilities is $\beta_1/4$.

# Interpret Slope

- A study on cereal attempted to predict the probability that a cereal would be classified as a children's cereal rather than adults' cereal, based on its grams of sugar per serving.

$$\log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = -1.647 + 0.158\,sugars$$

$$\frac{\hat{\theta}}{1-\hat{\theta}} = \exp\{-1.647 + 0.158\,sugars\}$$

$$\hat{\theta} = \frac{e^{-1.647 + 0.158(sugars)}}{1 + e^{-1.647 + 0.158(sugars)}}$$

# Interpret Slope

- Usually, we interpret the slope as the predicted amount of increase or decrease in y for a one-unit increase in x.

- What happens to the model when x increases by one unit? Compare the above to:

$$\frac{\hat{\theta}}{1-\hat{\theta}} = \exp\{-1.647 + 0.158\,(sugars + 1)\}$$

$$= \exp\{-1.647 + 0.158\,sugars\}\ \exp\{0.158\}$$

multiplicative factor. ($e^{\beta_1}$)

- So the **odds** of being a children's cereal are predicted to be *multiplied* by exp(0.158) = 1.17 when one gram of sugar is added to the cereal, so the odds are 17% higher. That is, the odds increase by a multiplicative factor of 1.17.

STOP Wed 3/30/22 (Week 10, Lecture 17)

* Friday 4/1/22 went over old Exam 2 & HW 05 solutions. (week 11, lec 28)

+ START Wed 4/6/22 (week 11, lecture 29)

# Interpret Slope

- When $\exp(\widehat{\beta_1})$ is close to 1, subtract 1 or subtract from 1 and interpret as a percent increase:
  - Ex: $\exp(\widehat{\beta_1}) = 1.2$ : The odds are about 20% higher
  - Ex: $\exp(\widehat{\beta_1}) = 0.8$ : The odds are approximately 20% lower

note: not an average or mean anymore b/c were not talking about averages or means.

- When $\exp(\widehat{\beta_1})$ is far from 1, interpret as that integer or fraction:
  - Ex: $\exp(\widehat{\beta_1}) = 3.1$ : The odds are approximately 3 times higher
  - Ex: $\exp(\widehat{\beta_1}) = 0.34$ : The odds are approximately 1/3 as high

# Interpret Slope

- Multiplicative, not additive
- Odds, not chance or probability (or mean, i.e. *on average*)
- Exponentiate slope before interpreting
- Still need context
- Still a statistic, not a parameter (approximately, according to the model)
- Still controlling for (i.e. holding the value constant) other variables *in the model*.

# Ratios

- Ratios larger than 1 have a larger numerator; ratios smaller than 1 have a larger denominator.
- The ratio of male to female students at A&M (Spring semester 2016 data) is 28,717 / 26,375 = 1.09. That is, there are about 9% more men than women at A&M.
- The ratio of female to male students is 0.92; there are about 92% as many women as men.

# What is an Odds Ratio?
# A Ratio of Two Odds!

Ex: Slope is a ratio of odds when $X = V+1$ to $X = x$

(odds ratio) $OR = \dfrac{\hat{\theta}(x+1)/(1-\hat{\theta}(x+1))}{\hat{\theta}(x)/(1-\hat{\theta}(x))} = \dfrac{e^{(\hat{\beta}_0+\hat{\beta}_1(x+1))}}{e^{(\hat{\beta}_0+\hat{\beta}_1(x))}} = e^{\hat{\beta}_1}$

- odds are multiplied by approximately $\exp(\hat{\beta}_1)$ when $x$ increases 1 unit

# What is an Odds Ratio? Physicians Health Study

| | Aspirin | Placebo | Total |
|---|---|---|---|
| **Heart Attack** | 139 | 239 | 378 |
| **No HA** | 54,421 | 54,117 | 108,538 |
| **Total** | 54,560 | 54,356 | 108,916 |

Define success as having a heart attack.

$\theta_{Aspirin} = 0.0025$, $\theta_{Placebo} = 0.0044$

$Odds_{Aspirin} = 0.0026$, $Odds_{Placebo} = 0.0044$

$$Odds\,Ratio = \frac{0.0026}{0.0044} = 0.56$$

# What is an Odds Ratio? Physicians Health Study

$$Odds\,Ratio = \frac{0.0026}{0.0044} = 0.56$$

- Interpret: Our model predicts the odds of having a heart attack to be about half as high if male physicians take an aspirin a day than if they don't.
- Notice the interpretation of a *multiplicative* effect rather than an *additive* effect: The odds of having a heart attack are *multiplied* by about ½ when taking aspirin.

# Odds Ratios

- In a study of a disease in humans, researchers used the independent variable ethnicity to predict probability of the disease. The model was:

$$log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- The variable encoding scheme was as follows:

| | x1 | x2 | x3 |
|---|---|---|---|
| **White Non-Hispanic** | 0 | 0 | 0 |
| **Black** | 1 | 0 | 0 |
| **Hispanic** | 0 | 1 | 0 |
| **Other** | 0 | 0 | 1 |

STOP Wed 4/6/22 (Week 11, Lecture 29)

START Friday 4/8/22 (week 11, lecture 30)

# Odds Ratios

- Below is partial output from the model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3863     0.5000  -2.773 0.00556
x1            2.0794     0.6325   3.288 0.00101
X2            1.7918     0.6455   2.776 0.00551
X3            1.3863     0.6708   2.067 0.03878
```

- Report a 95% confidence interval for the odds ratio comparing Hispanic to White.

log odds Hisp: $\log\left(\frac{\hat{\theta}_H}{1-\hat{\theta}_H}\right) = -1.3863 + 1.7918$

W: $\log\left(\frac{\theta_W}{1-\theta_W}\right) = -1.3863$.

$\log(OR) = 1.7918$

$OR = \exp(1.7918)$

95% CI for $\log(OR)$: $1.7918 \pm 1.96(0.6455) = (0.527, 3.057)$

95% CI for OR: (Exponentiate endpoints of the above)

$(e^{0.527}, e^{3.057}) = (1.69, 21.26)$

- What is the odds ratio for comparing Hispanic to Black?

$\Rightarrow$ does not include 1 $\Rightarrow \hat{\beta}_2$ is statistically significantly different from 0.

Log odds Hisp: $-1.3863 + 1.7918$

Log odds Black: $-1.3863 + 2.0794$

$\log(OR) = \log(OR_H) - \log(OR_B)$

$= 1.7918 - 2.0794 = -0.2876$

$OR = \exp(-0.2876) = 0.75$

# Theory: Part A – Multiple values of y at every value of x

# 8.1: A sample of size $m_i$ at every observed value of $x_i$

Binomial distribution:

1. There are m identical trials
2. Each trial has only one of two outcomes (Success or Failure)
3. The probability of success θ is the same for all trials
4. Trials are independent

Let Y = number of successes in m trials of a binomial process. Then Y is binomial with parameters m and θ. We write:

$$Y \sim Bin(m, \theta)$$

The probability that there are j successes in m trials (j = 0, 1, …, m) is given by:

$$P(Y = j) = \binom{m}{j}\theta^j(1-\theta)^{m-j} = \frac{m!}{j!(m-j)!}\theta^j(1-\theta)^{m-j}, j = 1, \ldots, m$$

# 8.1: A sample of size $m_i$ at every observed value of $x_i$

- Mean and variance of Y:

$$E[Y] = m\theta$$
$$Var(Y) = m\theta(1-\theta)$$

- In Section 8.1, we assume we have one predictor variable x with mi measurements at each level of x. In this case:

$$(Y|x_i) \sim Bin(m_i, \theta(x_i)), i = 1, \ldots, n$$

- Notice that we write the probability of success as a function of the value of the predictor variable x.

# 8.1: A sample of size $m_i$ at every observed value of $x_i$

Why not use proportions instead of odds?

Proportion of successes = $y_i / m_i$

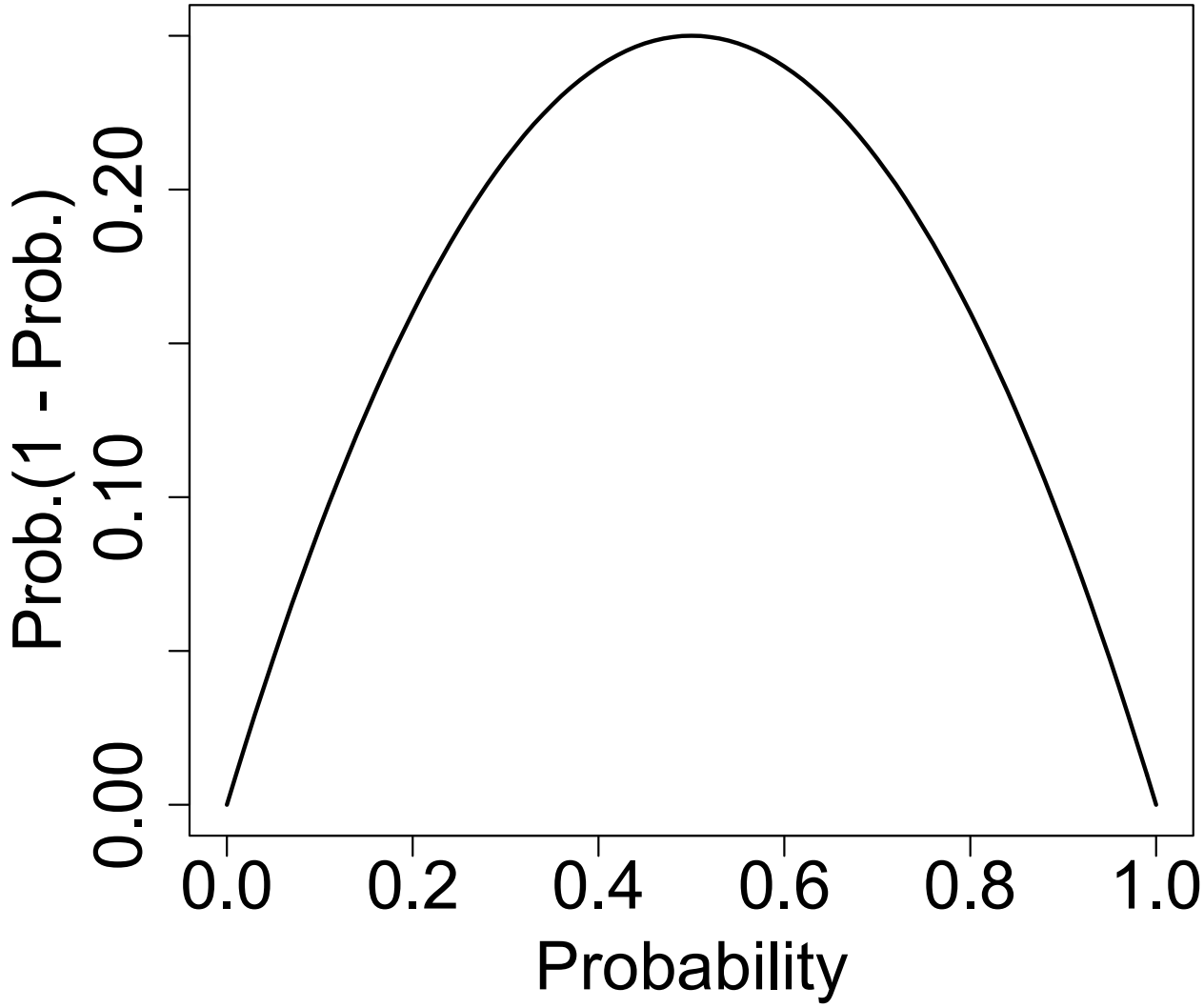
It could be the response since it is an unbiased estimate of $\theta(x_i)$ and it varies between 0 and 1.

BUT: Calculate the mean and variance of the sample proportion:

$$E[y_i/m_i|x_i] = m_i\theta(x_i)$$
$$Var(y_i/m_i|x_i) = \theta(x_i)(1-\theta(x_i))/m_i$$

Variance is not constant! We need our log odds model.

# Relationship between θ and θ(1 – θ)



# Finding Parameter Estimates

- The Likelihood function, denoted L, is the probability of the data, regarded as a function of the unknown parameters with the data values fixed. (Remember when we used least squares, we modeled the sums of squares as a function of the parameters with the data held fixed for regular regression.)[1]

$$f(x_1, x_2, \ldots, x_n|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$$

- We rearrange what is fixed and what is considered random to think of likelihood as the likelihood of that value being the parameter, given the data that we have currently.

$$L(\theta|x_1, x_2, \ldots, x_n) = \Pi_{i=1}^{n} f(x_i|\theta)$$

1 Stat 2: Building Models for a World of Data. Cannon, Ann R. et al. Freeman (2013)

# Finding Parameter Estimates

- The parameters for the logistic regression model are found by maximizing the log-likelihood. This is equivalent to minimizing the deviance, -2 log L.
- For linear regression models, minimizing RSS had closed form solutions; for logistic regression models, we need an iterative method to find estimates such as Newton-Raphson or iteratively reweighted least squares.

# Likelihood

$$
\begin{aligned}
L &= \Pi_{i=1}^{n} P(Y_i = y_i|x_i) = \Pi_{i=1}^{n} \binom{m_i}{y_i} \theta(x_i)^{y_i}(1-\theta(x_i))^{m_i-y_i} \\
log(L) &= \Sigma_{i=1}^{n} \left[ \log \binom{m_i}{y_i} + \log\left(\theta(x_i)^{y_i}\right) + \log\left((1-\theta(x_i))^{m_i-y_i}\right) \right] \\
&= \Sigma_{i=1}^{n} \left[ C + y_i \log\left(\theta(x_i)\right) + (m_i - y_i)\log(1-\theta(x_i)) \right] \\
&= \Sigma_{i=1}^{n} \left[ C + y_i \log\left(\frac{\theta(x_i)}{1-\theta(x_i)}\right) + m_i \log(1-\theta(x_i)) \right] \\
&= \Sigma_{i=1}^{n} \left[ C + y_i(\beta_0 + \beta_1 x_i) - m_i \log(1 + exp(\beta_0 + \beta_1 x_i)) \right]
\end{aligned}
$$

$\theta(x_i) = \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}$ $\Rightarrow m_i \log(1-\theta(x_i)) = m_i \log\left(\frac{1+exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)} - \frac{exp(\beta_0+\beta_1 x_i)}{1+exp(\beta_0+\beta_1 x_i)}\right)$

$= m_i \log\left(\frac{1}{1+exp(\beta_0+\beta_1 x_i)}\right) = -m_i \log(1+exp(\beta_0+\beta_1 x_i))$

# Hypothesis Tests & Confidence Intervals

- The standard error of the parameter estimates is based on the information statistic, the second derivative of the log likelihood.

| | Estimate | Std. Error | z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -1.19620 | 0.11845 | -10.099 | < 2e-16 | *** |
| xnew | -0.29710 | 0.05485 | -5.416 | 6.08e-08 | *** |

- Test whether the area of an island is associated with whether a species goes extinct. (Wald test: note z, not t!)

$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0 \qquad Z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.29710}{0.05485} = 5.416$

Highly confident that log(Area) is associated w/ the probability of gone extinct.

# Hypothesis Tests & Confidence Intervals

- A 95% confidence interval for the parameter $\beta_1$ is:

$$-0.297 \pm 1.96\ (0.055)$$

$$(-0.405, -0.190)$$

$$\exp(-0.405) = 0.67$$

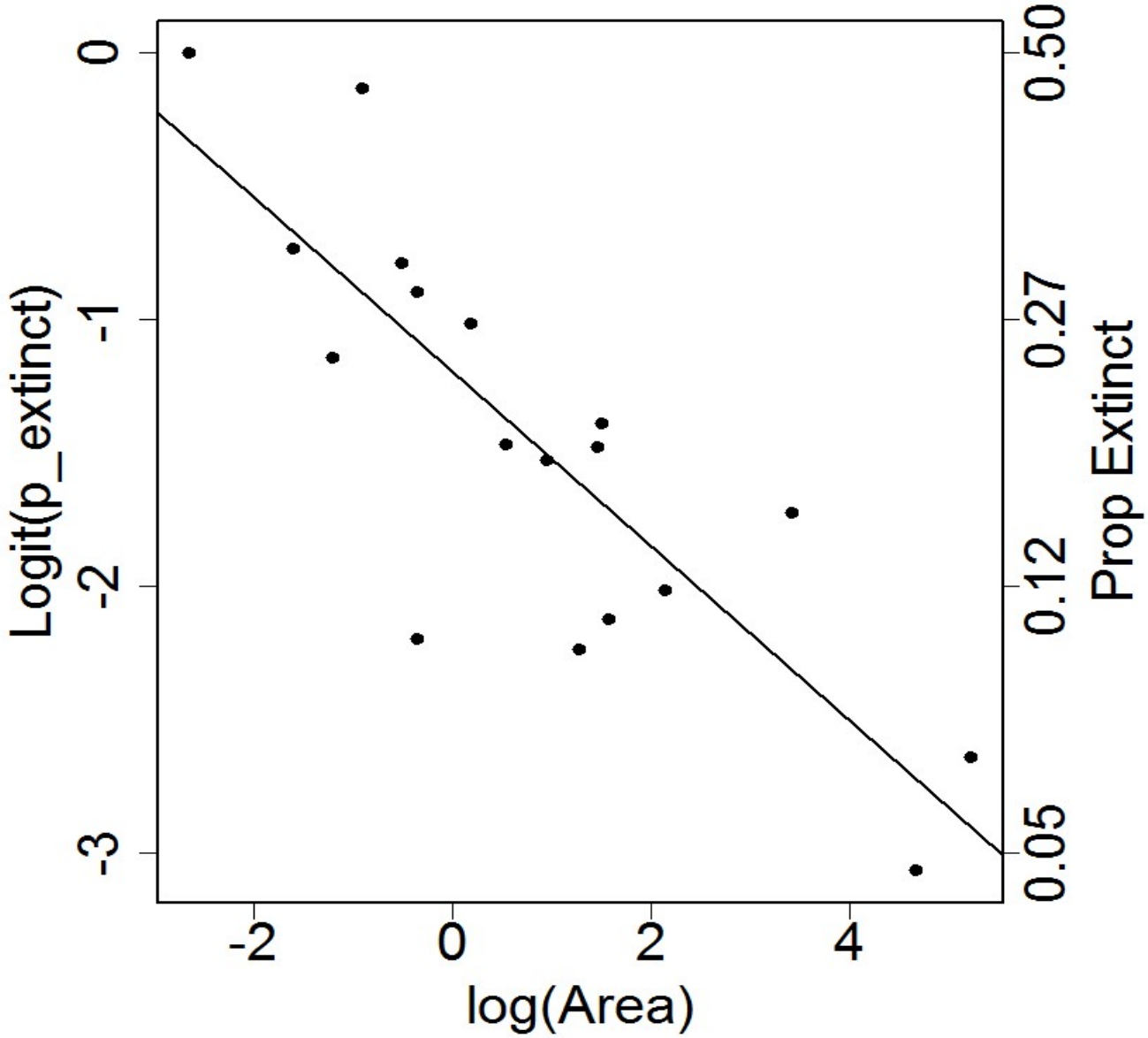
$$\exp(-0.190) = 0.83$$

- Interpretation: I am 95% confident that when log land area increases by 1, the odds that a species goes extinct on that island are between about 2/3 and 4/5 as large. (Or say multiplied by between 0.67 and 0.83 (the odds decrease).)

- Residuals?

Still deals w/ the case where we have multiple obs. for each value of x.

# Deviance


Logit(p_extinct)
0
-1
-2
-3
Prop Extinct
0.50
0.27
0.12
0.05
-2
0
2
4
log(Area)

# Deviance

- In linear models, we measure how well our model works in part by how close the data are to the model. The concept of residual sums of squares is replaced by deviance for logistic regression.
- The saturated model is one with a separate proportion of successes for every value of $x_i$; that is, $\hat{\theta}(x_i) = \frac{y_i}{m_i}$ – has a separate estimate for each value of X.
- Ex: To estimate the number of species that would go extinct on an island with 185.8 square miles, I could use what happened on Ulkokrunni (5 went extinct), which is the saturated estimate. Or I could use the model to predict: 4.5.
  - $\log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) = -1.196 - 0.297\log(185.8)$ = -2.748
  - $\hat{\theta} = \frac{\exp(-2.748)}{1+e^{(-2.748)}} = 0.06$
  - # extinct = (75 species) (0.06) = 4.5 species.

# Deviance

- Deviance measures the difference between the log likelihood from the saturated model (S) and the log likelihood from our model (M).

$$G^2 = 2\left[\log(L_S) - \log(L_M)\right]$$

$$= 2\Sigma_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right]$$

(# of islands)

- yi: Raw number that actually went extinct

- $\hat{y}_i$= Predicted number that went extinct from logistic model

$$= m_i\hat{\theta}_i = m_i \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

# Deviance

| Island | Area | SpeciesRisk | Ext | Saturated: $y_i/m_i$ | Logistic Model: θ-hat |
|---|---|---|---|---|---|
| Ulkokrunni | 185.8 | 75 | 5 | 0.067 | 0.060 |
| Maakrunni | 105.8 | 67 | 3 | 0.045 | 0.070 |
| Ristikari | 30.7 | 66 | 10 | 0.152 | 0.099 |
| Isonkivenletto | 8.5 | 51 | 6 | 0.118 | 0.138 |

# Deviance

- When each $m_i$ is large enough, the deviance statistic can be used as a chi-squared goodness-of-fit test for the logistic regression model.

- We wanted residual sums of squares to be small because we wanted the model to fit the data well. We also want deviance to be small because we want the model to fit the data well. If deviance is large, it means the saturated model has a very different fit to the data from our model of interest.

- Ho: The logistic regression model is appropriate.
  Ha: The logistic regression model is inappropriate.

- $G^2$ has the $\chi^2$ distribution with n – p – 1 degrees of freedom, where n is the number of *binomial* samples, not the total sample size!

- Beware sample sizes that are too *large*.

START Monday 4/11/22 (week 12, lecture 31)

# Deviance

- The model summary output gives:

→ assumes $\beta_1 = \beta_2 = \dots = \beta_p$ (similar to overall F test, testing if any of our X's have an association w/ the response.

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

→ for testing if model is appropriate.

- The p-value is found by $P(G^2 > 12.062)$ from a chi-squared distribution with 18 – 1 – 1 = 16 degrees of freedom: 0.74.

# Deviance

- We can also use deviance to test whether two nested models are significantly different. For example, we could test whether our model is equivalent to the null model:
- $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$
- The difference between the deviances for the two models is compared to a $\chi^2$ distribution with the difference between the degrees of freedoms for the two models.
- This **doesn't** give the same result in general as the Wald z-test for whether $\beta_1 = 0$.

# Deviance

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

- 45.338 – 12.062 = 33.276,  df = 17 – 16 = 1

p-value < 0.0001

- We have strong evidence that log area of an island is somewhat associated with whether a species goes extinct.

- Notice that the p-value for the Wald test is not the same as that of the deviance test.

See slide 58

# Deviance: $R^2$ for logistic regression

- Recall that for linear regression:

$$R^2 = 1 - \frac{RSS}{SST}$$

- Since the deviance is a generalization of the residual sum of squares in linear regression, one version of $R^2$ for logistic regression is:

$$R^2_{dev} = 1 - \frac{G^2_{H_A}}{G^2_{H_0}}$$

- So for the cereal data, $R^2 = 1 - 12.062 / 45.338 = 0.73$.

- This still has the issue of increasing when we add useless predictors to the model.

# Pearson goodness-of-fit statistic

- Alternative measure of deviance: Pearson $\chi^2$ statistic

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i/m_i - \hat{\theta}(x_i)\right)^2}{\widehat{\text{Var}}(y_i/m_i)} = \sum_{i=1}^{n} \frac{\left(y_i/m_i - \hat{\theta}(x_i)\right)^2}{\hat{\theta}(x_i)\left(1-\hat{\theta}(x_i)\right)/m_i}$$

- Same degrees of freedom as deviance: df = n – p – 1.
- Same requirement: the statistic has the $\chi^2$ distribution as long as each of the $m_i$ are large enough. If this is true, $G^2$ and $\chi^2$ are similar; we prefer $G^2$ if they yield different conclusions.[1]

McCullagh & Nelder (1989) p. 398: Distributional properties for deviance residuals are closer to the residuals from a Gaussian linear regression model.

# Residuals for Logistic Regression

- Three Types of Residuals:
  1. **Response residuals**
  2. **Pearson Residuals**
  3. **Deviance Residuals**

# Response Residuals

- Response residuals are the difference between the observed and the fitted proportions:

$$r_i = y_i / m_i - \hat{\theta}(x_i)$$

where $\hat{\theta}(x_i)$ is the $i^{th}$ fitted value from the logistic regression model.

- The variance of $y_i/m_i$ is not constant, so response residuals are difficult to interpret in practice.

# Pearson Residuals

- The problem of nonconstant variance is overcome by Pearson residuals, the square root of the individual contributions to the Pearson $\chi^2$ statistic.

$$r_{Pearson\,i} = \frac{y_i/m_i - \hat{\theta}(x_i)}{\sqrt{\hat{\theta}(x_i)\left(1-\hat{\theta}(x_i)\right)/m_i}}$$

# Standardized Pearson Residuals

- Pearson Residuals still don't account for the variance of the model estimate $\hat{\theta}(x_i)$, so we correct for that:

$$sr_{Pearson\,i} = \frac{y_i/m_i - \hat{\theta}(x_i)}{\sqrt{(1-h_{ii})\hat{\theta}(x_i)\left(1-\hat{\theta}(x_i)\right)/m_i}} = \frac{r_{Pearson\,i}}{\sqrt{(1-h_{ii})}}$$

# Deviance Residuals

- Deviance residuals are to the deviance statistic $G^2$ as Pearson residuals are to the $\chi^2$ Pearson statistic.

$$G^2 = \sum_{i=1}^{n} r^2_{Deviance\,i}$$

$$r_{Deviance\,i} = \text{sign}\left(y_i/m_i - \hat{\theta}(x_i)\right) g_i$$

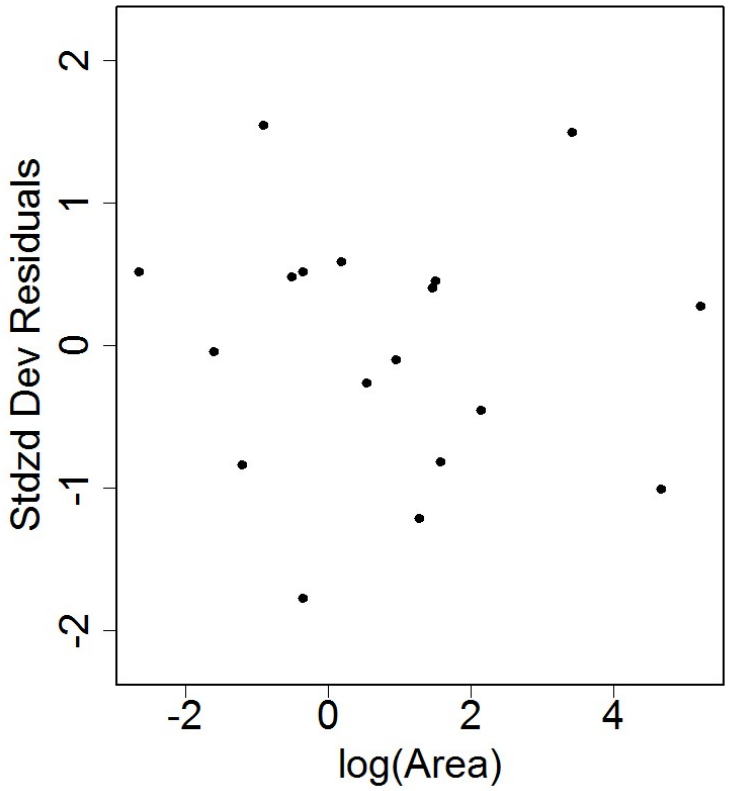
$$G^2 = \sum_{i=1}^{n} g_i^2$$

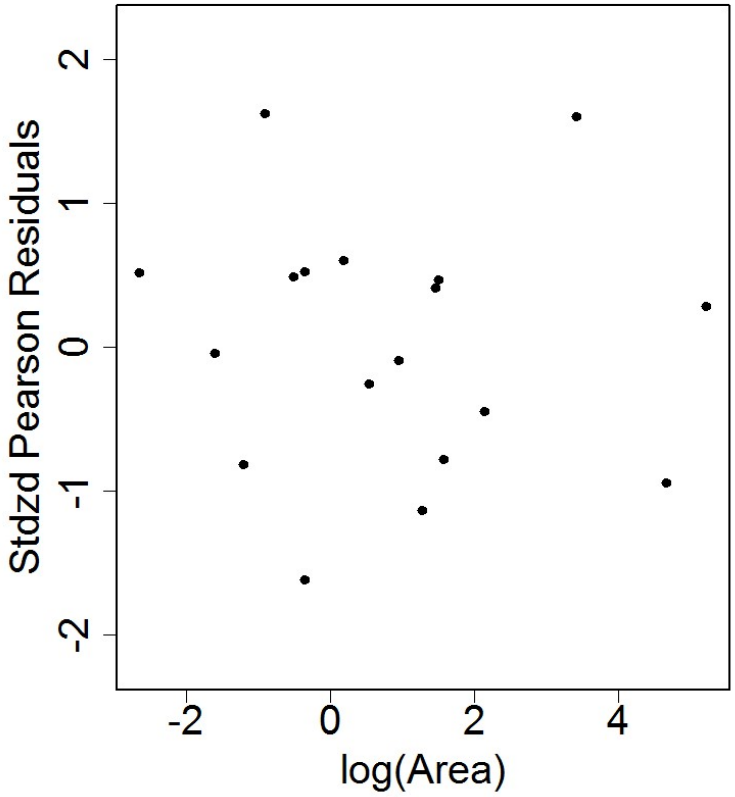
- Standardized deviance residuals are defined to be:

$$sr_{Deviance\,i} = r_{Deviance\,i} \Big/ \sqrt{1 - h_{ii}}$$

# Which residuals are best?

- **Pearson residuals** are most popular, but **deviance residuals** are actually **preferred**; their distribution is closer to that of least squares residuals.



# Theory: Part B – One value of y at every value of x

# 8.2 Binary Logistic Regression

- It is more common that we have only one observation at many values of the predictor variable. (E.g. many predictors!) Such data are called **binary**.
- Goodness of fit measures are problematic and plots of residuals are difficult to interpret. (Two U-shaped curves!)

# Compare Fits

- If we fit data with the assumption that all the $m_i$ equal 1, the parameter estimates, standard deviations, Wald z-scores, and p-values are all equivalent.

- The difference is in the deviance:

  - Binomial Fit:

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

  - Binary Fit:

```
    Null deviance: 578.01  on 631  degrees of freedom
Residual deviance: 544.74  on 630  degrees of freedom
```

- AIC values are also different.

# Binary Deviance

- In the case that $m_i = 1$, the log likelihood function is:

$$\log(L) = \sum_{i=1}^{n} \left[ y_i \log\left(\theta(x_i)\right) + (1 - y_i) \log\left(1 - \theta(x_i)\right) + \log \binom{1}{y_i} \right]$$

- So for the saturated model, the log-likelihood function is:

$m_i = 1$ (see slide 42):

$\log(l_{sat}) = \sum_{i=1}^{n} [c + y_i \log(\frac{\theta_{x_i}}{1-\theta_{x_i}}) + \log(1-\theta(x_i))$

saturated model: $\hat{\theta}_{sat} = y_i$

$\sum_{i=1}^{n} [c + y_i \log(y_i) + (1-y_i) \log(1-y_i)]$

- If $y_i = 0$:

$c + 0 + 0 = c$ } deviance is constant $\Rightarrow$ not useful.

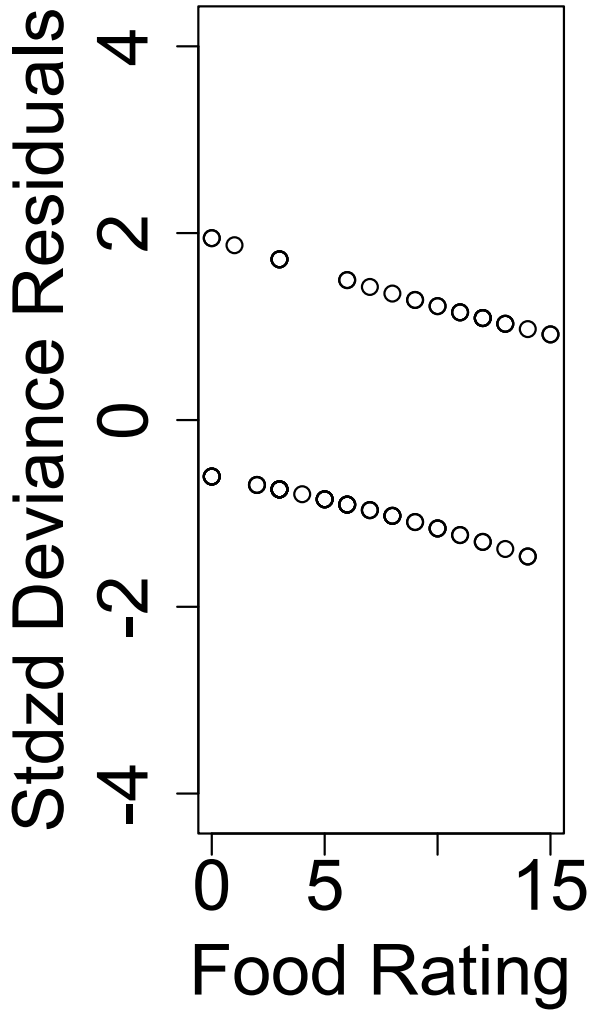
- If $y_i = 1$:

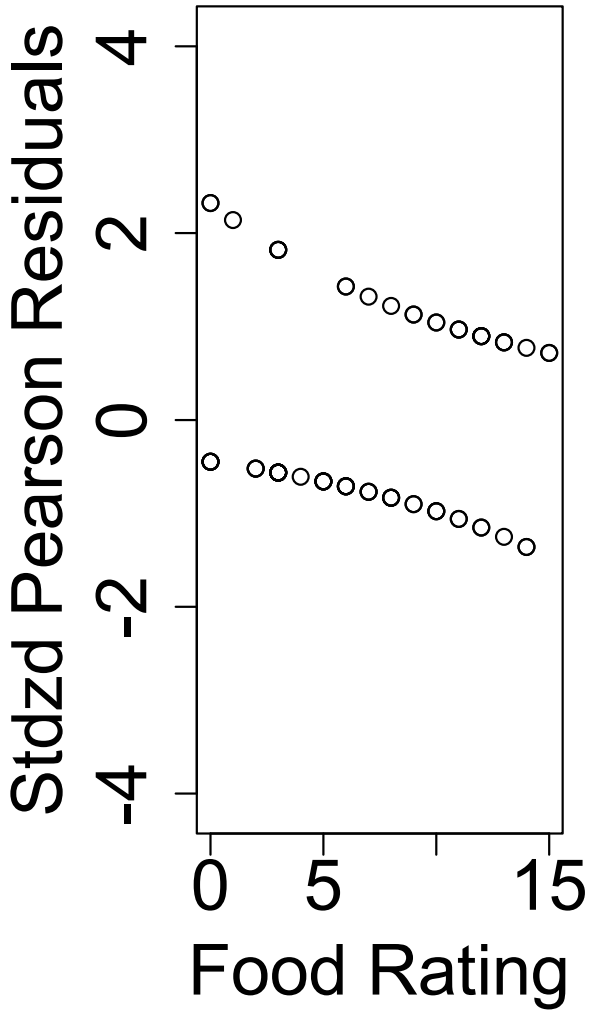
$c + 0 + 0 = c$

# Binary Deviance

- So in the case where mi = 1, the deviance between the saturated model and the current model only depends on $\log(L_M)$.
- Deviance doesn't provide an assessment of the goodness-of-fit of the model! It also doesn't have a $\chi^2$ distribution.
- However, we *can* use deviance to compare two models; the difference between two deviances still has an approximate $\chi^2$ distribution.

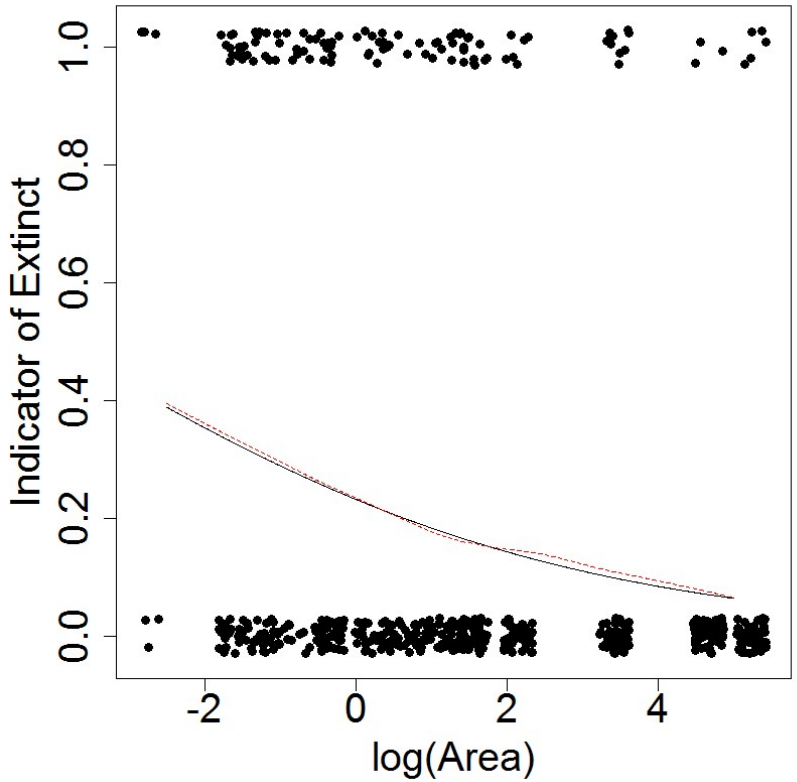
# Binary Residuals



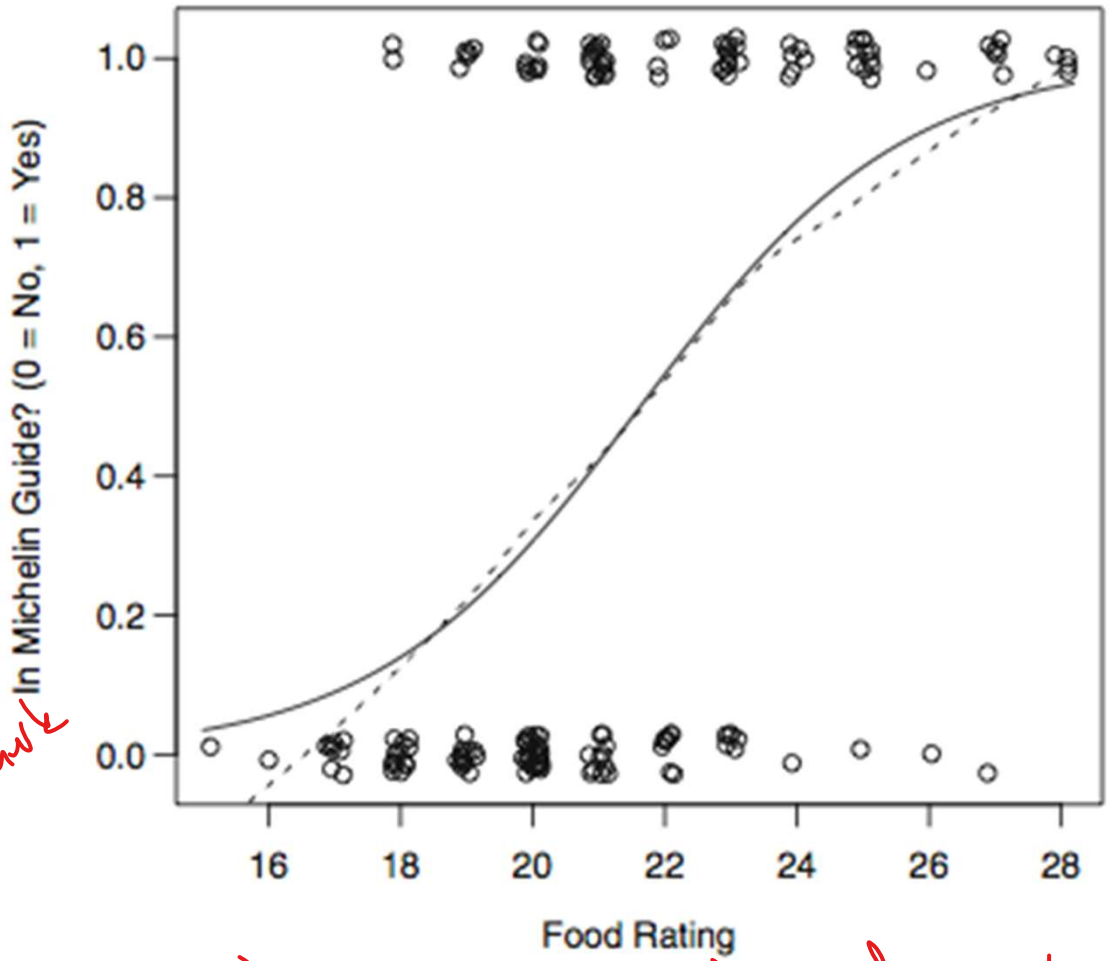



# Binary Residuals

- Residual plots are problematic when the data are binary.
- Instead of examining residual plots, compare the fitted model to a nonparametric fit.



# Binary Residuals




@ 37 min mark → went over related code island. ←

STOP Monday 4/11/22 (Week 12, Lecture 31)

START Wednesday 4/13/22 (week 12, lecture 32)

START w/ islands code

START here around 13 min mark.

of the x-vars in particular.

# Transformations, Marginal Model Plots, Outliers

# Transformations

- Do we need to transform y? Why or why not?
  - when we have outliers.
  - when relationship btwn log(odds) & x-vars is non-linear.

- Reasons to transform x:
  - Linear or quadratic relationships between x's and log odds. or log(odds) is a func. of log(x) for gamma (or other right-skewed dists) and $x^2$ for normal.
  - Constant variance (if we have fanning pattern in plot of log(odds) vs x.)

# Transforming Predictors for Binary Data

$E[y] = \sum y_i P(y_i)$

- Why transform predictor variables?
- Quick review: Suppose 30% of Dalmatians are deaf. If I randomly select 1 Dalmatian, how many are expected to be deaf? $x_i \sim Bin(n, p)$

$E[x] = np$

$n=1 \Rightarrow E[x] = \mu = p$

$$\begin{aligned}\theta(x) &= E[Y|X=x] \\ &= 1 \times P(Y=1|X=x) + 0 \times P(Y=0|X=x) \\ &= P(Y=1|X=x)\end{aligned}$$

# Transforming Predictors for Binary Data: Binary Predictor

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

- First suppose the predictor is a dummy variable:

$$\begin{aligned}
\frac{\theta(x)}{1-\theta(x)} &= \frac{P(Y=1|X=x)}{P(Y=0|X=x)} \\
&= \frac{P(Y=1 \cap X=x)}{P(Y=0 \cap X=x)} \\
&= \frac{P(X=x|Y=1)\,P(Y=1)}{P(X=x|Y=0)\,P(Y=0)}
\end{aligned}$$

- Take logs of both sides:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right)$$

constant in $X$.

# Transforming Predictors for Binary Data: Continuous Predictor

- When X is binary:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{P(X=x|Y=1)}{P(X=x|Y=0)}\right)$$

- Similarly, when X is continuous:

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) + \log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right)$$

- We ignore the first term either way when discussing transformations of X.

# Transforming Predictors for Binary Data: Normal Predictor

- When f(x|Y = j), j = 0, 1, is a normal density (possibly with two different means and variances):

$$f(x|Y=j) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right], j = 0, 1$$

- Then that piece of the log odds we were worried about:

$$\begin{aligned}
\log\left(\frac{f(x|Y=1)}{f(x|Y=0)}\right) &= \log\left(\frac{\sigma_0}{\sigma_1}\right) + \left[-\frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_0)^2}{2\sigma_0^2}\right] \\
&= \log\left(\frac{\sigma_0}{\sigma_1}\right) + \left[-\frac{x^2}{2\sigma_1^2} + \frac{x^2}{2\sigma_0^2} + \frac{2\mu_1 x}{2\sigma_1^2} - \frac{2\mu_0 x}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2} + \frac{\mu_0^2}{2\sigma_0^2}\right] \\
&= \log\left(\frac{\sigma_0}{\sigma_1}\right) + \left(\frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_1^2}{2\sigma_1^2}\right) + \left(\frac{2\mu_1}{2\sigma_1^2} - \frac{2\mu_0}{2\sigma_0^2}\right)x + \left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right)x^2 \\
&= C + \beta_1 x + \beta_2 x^2
\end{aligned}$$

If equal variances. i.e. $var(x|y=0) = var(x|y=1)$ then this coef $= 0$.

shows that the log odds is a function of $x$ & $x^2$ $\Rightarrow$ when fitting a linear regression model, If my data were normal w/ different variances for $y=1$ & $y=0$, it makes sense to include both an $x$ term & $x^2$ term.

# Transforming Predictors for Binary Data: Normal Predictor

- Conclusions:
  1. When x is normal, log odds are a quadratic function of x.
  2. When the variances are equal, the log odds is a linear function of x with:

$$\beta_1 = \left( \frac{\mu_1 - \mu_0}{\sigma^2} \right)$$

* If x is right skewed (like gamma) a natural thing to do is add x & log(x)

If x is binary, the log odds is just a function of x

# Transforming Predictors for Binary Data: Multivariate Normal

- After some more math… when we have p predictors that are multivariate normal, with different covariance matrices for Y= 0 and Y = 1, then the log odds are a function of $x_i$, $x_i^2$, and $x_i x_j$ ($i, j = 1, \ldots, p; i \neq j$).

1. If the variance of $x_i$ is different for Y = 0 and Y = 1, add a quadratic term in $x_i$.
2. If the regression of $x_i$ on $x_j$ has a different slope for Y = 0 and Y = 1, add the interaction $x_i x_j$.

# Interactions

- Recall: an interaction between $x_i$ and $x_j$ means the relationship between $x_i$ and y is different depending on the value of $x_j$.
- That means the relationship between $x_i$ and $x_j$ will be different depending on the value of y.
- Plot $x_i$ and $x_j$, fitting separate slopes for the values of y (0 and 1). The farther apart the slopes, the more important it is to fit an interaction.

# Transforming Predictors for Binary Data: Poisson Predictor

- Distribution of X (possibly with different means again):

$$P(X = x|Y = j) = \frac{e^{-\lambda_j}\lambda_j^x}{x!}, j = 0, 1$$

- Starting over again with the piece of the log odds that varies with X:

$$\log\left(\frac{P(X = x|Y = 1)}{P(X = x|Y = 0)}\right) = (\lambda_0 - \lambda_1) + x\log\left(\frac{\lambda_1}{\lambda_0}\right)$$

- Again, we end up with log odds being a linear function of x.

# Marginal Model Plots

- Residual plots are difficult to interpret; instead we use marginal model plots.
- Same concept as for multiple linear regression: compare nonparametric estimates of (for every variable $x_i$):

$$E[Y|x_1] \text{ and } E\left[\hat{Y}|x_1\right]$$

- If they agree, we conclude that $x_i$ is modeled correctly by our model. If not, then $x_i$ is not modeled correctly by our model.

# Leverage

- Obtained from weighted least squares approximation to the MLEs.
- Average leverage = (p+1)/n; cutoff = 2(p+1)/n.