STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

February 6-8, 2018

STA 303/1002: Class 10- Logistic Regression

- Case Study III: The Donner Party Example
  - Comparing models
  - Wald vs Likelihood Ratio Tests
  - Effect Plots
  - Related R packages and functions
- Case Study IV: Binomial Logistic Regression

Logistic Regression: Testing whether single β's are zero

WALD CHI-SQUARE PROCEDURES

- Hypotheses: $H_0 : \beta_j = 0$ (X_j has no effect on log-odds)
  $H_a : \beta_j \neq 0$
- Test Statistic: $\boxed{z = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}}$

  where
  - $\widehat{\beta}_j$- maximum likelihood (ML) estimate and
  - $SE(\widehat{\beta}_j)$- estimated standard error from the numerical procedure that generated the MLE.
  - By standard large-sample results, MLE's are normally distributed. Thus, for large n, under H_0, z is an observation from an approx. $\mathcal{N}(0, 1)$ distribution.
- 95% Confidence interval: $\boxed{\widehat{\beta}_j \pm 1.96 SE(\widehat{\beta}_j)}$

Examples: Testing whether β's are zero

Using R output:

	Age	*Sex*
Test statistic	$(-0.078/0.0373)^2$	4.47
P-value	0.036	0.0345
95% CI for β	$-0.078 \pm 1.96(0.0373)$ =(-0.15, -0.0055)	(0.117, 3.078)
CI for Odds ratio	$(e^{-0.15}, e^{-0.0055})$=(0.86,0.995)	(1.124, 21.72)
Conclusion	For the same sex, the odds ratio for a 1-year increase in age is between .86 and 0.995.	

- Note: Both marginal p-values are less that 0.05 and the confidence intervals for the odds ratios do not include 1.
- Hence, we have moderate evidence that both *Age* and *Sex* have an effect on survival over and above each other.
- Recall: If $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi_1$.

Model Assumptions for Binary Logistic Regression

1. Underlying probability model for response is Bernoulli.
2. Observations are independent.
3. The form of the model is correct.
   - Linear relationship between logits and explanatory variables
   - All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs.
   (Recall large-sample properties of MLEs.)

Comparing models: Likelihood Ratio Test

- Idea: Compare likelihood of data under FULL (F) model, $\mathcal{L}_F$ to likelihood under REDUCED (R) model, $\mathcal{L}_R$ of same data.

$$\text{Likelihood ratio}: \frac{\mathcal{L}_R}{\mathcal{L}_F}, \text{ where } \mathcal{L}_R \leq \mathcal{L}_F$$

- Hypotheses: $H_0 : \beta_1 = \cdots = \beta_k = 0$ ≥1 β.

  (Reduced model is appropriate; fits data as well as Full model)

  H_a: at least one $\beta_1, \cdots, \beta_k \neq 0$

  (Full model is better)

- Test Statistic: Deviance (residual),

$$G^2 = -2\log\mathcal{L}_R - (-2\log\mathcal{L}_F) = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right)$$

$G^2 = D_R - D_F$

- For large n, under H_0, G^2 is an observation from a chi-square distribution with k df.

Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether a model with the 3 higher-order polynomial terms and/or interaction terms is an improvement over the additive model.

- Hypotheses: H_0: Additive vs H_a: Higher order model

  $H_0: \beta_3 = \beta_4 = \beta_5 = 0$.
- Test Statistic: $G^2 = 51.256 - 45.361 = 5.895 \sim \chi^2_3$
- Distribution of TS:
- P-value: $P(\chi^2_3 > 5.895) = 0.1168$
- Conclusion: Additive model is satisfactory.

Case Study 3: Higher Order Model with 3 higher order/interaction terms

```
fitfull<-glm(Status~Age+sex+Age:sex+I(Age^2)+I(Age^2):sex, family=binomial, dat
summary(fitfull)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex + Age:sex + I(Age^2) + I(Age^2):sex,
##     family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3396  -0.9757  -0.3438   0.5269   1.5901
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.318484   3.940184  -0.842    0.400
## Age                  0.183031   0.226632   0.808    0.419
## sexFemale            0.265286  10.455222   0.025    0.980
## I(Age^2)            -0.002803   0.002985  -0.939    0.348
## Age:sexFemale        0.299877   0.696050   0.431    0.667
## sexFemale:I(Age^2) -0.007356   0.010689  -0.688    0.491
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 45.361  on 39  degrees of freedom
## AIC: 57.361
```

$D_f = 45.361$

Testing β's: Wald versus LRT test

	Wald	LRT
Testing whether a single β=0	✓ (easier)	✓
Comparing nested models (>1 β)		✓
Small to moderate sample sizes β near boundary of parameter space		✓ (More reliable in general)

(The two methods are different.)

related to normality and regularity conditions of MLEs.

Case Study III Exercise: Comparing models

Using R output,

Q: Determine whether the effect of *Age* on the odds of survival differ with *Sex*.

- Hypotheses:

H_0: (Additive) $\text{logit}(\hat{\pi}) = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 1_F \iff H_0: \gamma_3 = 0$

H_a: (Interaction) $\text{logit}(\hat{\pi}) = \hat{\gamma}_0 + \hat{\gamma}_1 * Age + \hat{\gamma}_2 1_F + \hat{\gamma}_3 Age * 1_F$

$\iff H_a: \gamma_3 \neq 0$

- Test Statistic:

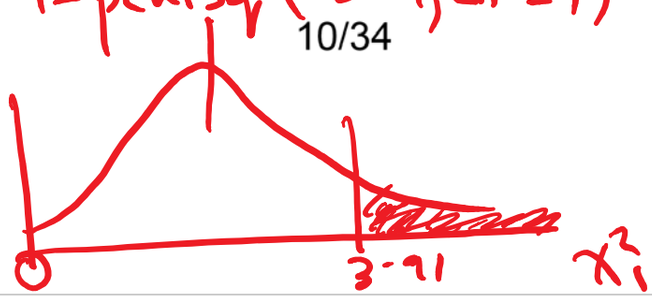
LRT: $G^2 = 51.256 - 47.346 = 3.91 \sim \chi^2_1$

Wald: $z = \hat{\gamma}_3 / \widehat{se}(\hat{\gamma}_3) = -1.714 \sim N(0,1)$.

- Distribution of TS:
- P-value: P-value $= P(\chi^2_1 > 3.91)$ $= 0.048$ (using R)

P-value $= 0.0865$

1-pchisq(3.91, df=1)



- Conclusion:

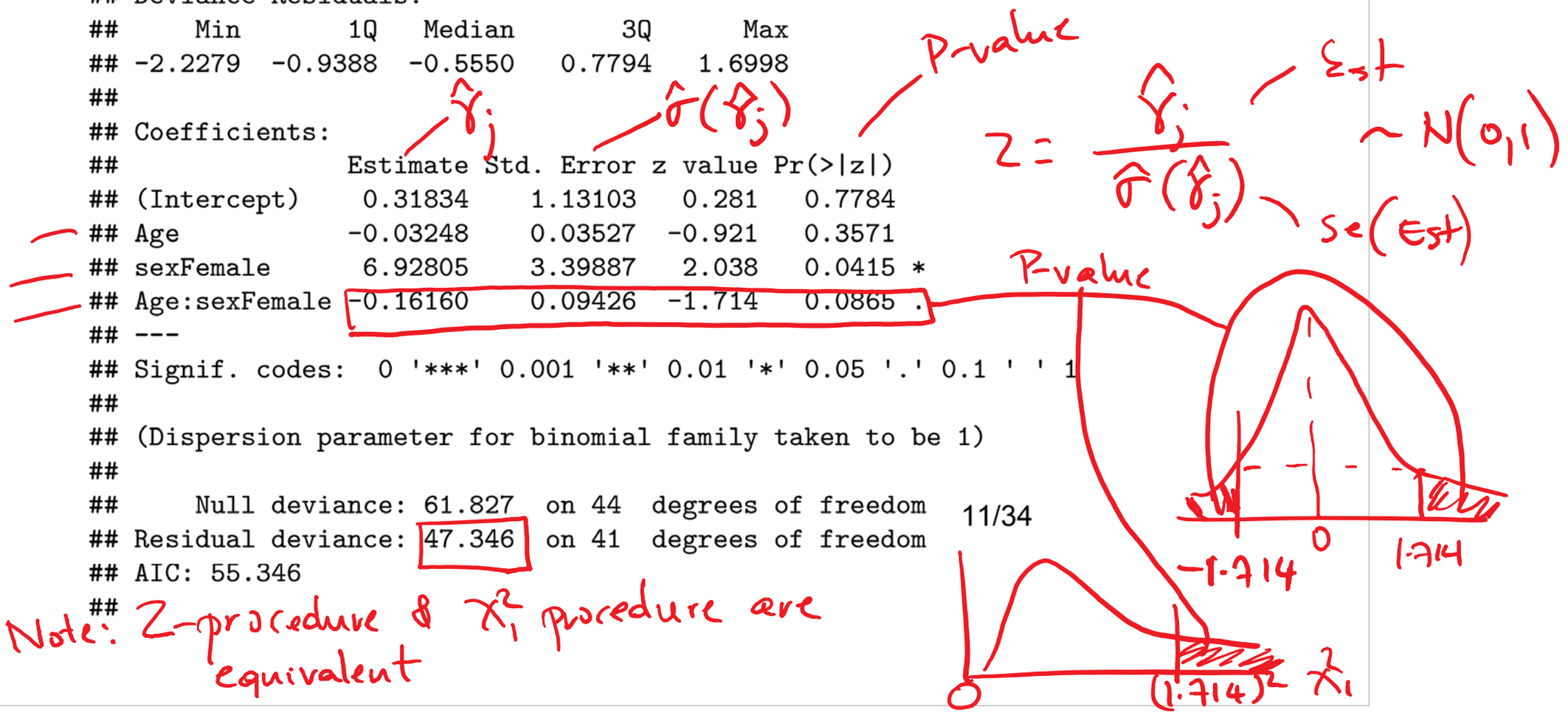
Suggestive but inconclusive evidence of interaction.

Case Study 3: Interaction Model, Age*Sex

```
fitas<-glm(Status~Age*sex, family=binomial, data=donner)
summary(fitas)
```

```
##
## Call:
## glm(formula = Status ~ Age * sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2279  -0.9388  -0.5550   0.7794   1.6998
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.31834    1.13103   0.281   0.7784
## Age            -0.03248    0.03527  -0.921   0.3571
## sexFemale       6.92805    3.39887   2.038   0.0415 *
## Age:sexFemale  -0.16160    0.09426  -1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 47.346  on 41  degrees of freedom
## AIC: 55.346
##
```

Note: Z-procedure & χ^2_1 procedure are equivalent



Comparing models: 'Global' LRT

- Idea: Compares Fitted model to NULL [$\text{logit}(\pi) = \beta_0$] model
- Hypotheses: $H_0 : \beta_1 = \cdots = \beta_p = 0$

  (NULL model is appropriate)

  H_a: at least one $\beta_1, \cdots, \beta_p \neq 0$

  (Fitted model is better)

$G^2 = D_N - D_{Fitted} \sim \chi^2_p$

Case Study III Exercise: 'Global' LRT

Using R output,
Q: Determine whether or not the additive model fits better than the Null model.

- Hypotheses:
- Test Statistic:
- Distribution of TS:
- P-value:
- Conclusion:

Done.
(See previous class)
($p = 0.005$)

Case Study 3: Additive model for Survived

```
fitasf<-glm(Status~Age+sex, family=binomial, data=donner)
summary(fitasf)
```

```
##
## Call:
## glm(formula = Status ~ Age + sex, family = binomial, data = donner)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -1.0441  -0.3029   0.8877   2.0472
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.63312    1.11018   1.471   0.1413
## Age         -0.07820    0.03728  -2.097   0.0359 *
## sexFemale    1.59729    0.75547   2.114   0.0345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

$D_R = 51.256$

STA303/1004 - Class 10 R Markdown

February 6, 2018

Case 3: Deviance test and Estimated Var-Cov of β

```
anova(fitasf, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                    44     61.827
## Age   1   5.5358        43     56.291  0.01863 *
## sex   1   5.0344        42     51.256  0.02485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(vcov(fitasf,digits=3))
```

Var-Cov matrix for $\hat{\beta}_j$'s.

```
##             (Intercept)          Age   sexFemale
## (Intercept)  1.23250837 -0.038472741  0.06007099
## Age         -0.03847274  0.001390134 -0.00823197
## sexFemale    0.06007099 -0.008231970  0.57073339
```


$\hat{\sigma}^2(\hat{\beta}_j) = \left(\widehat{se}(\hat{\beta}_j)\right)^2$

Eg, $\sqrt{0.57} = 0.7557$.

Case 3: Confidence Intervals for β's

```
cbind(bhat=coef(fitasf), confint.default(fitasf))  # 95% CI for betas
```

```
##                    bhat      2.5 %       97.5 %
## (Intercept)  1.63312031 -0.5428002  3.809040837
## Age         -0.07820407 -0.1512803 -0.005127799
## sexFemale    1.59729350  0.1166015  3.077985503
```

$\hat{\beta}_j \pm 1.96\, se(\hat{\beta}_j)$

Compare to results on p.4

```
exp(coef(fitasf)) # exponentiate estimated betas, get odds ratios
```

```
## (Intercept)         Age   sexFemale
##   5.1198252   0.9247757   4.9396452
```

```
exp(cbind(OR=coef(fitasf), confint.default(fitasf))) #CI for odds ratio
```

```
##                   OR     2.5 %     97.5 %
## (Intercept) 5.1198252 0.5811187 45.1071530
## Age         0.9247757 0.8596067  0.9948853
## sexFemale   4.9396452 1.1236716 21.7146143
```

Case 3: Wald tests in R

Computes Wald chi-squared test for 1 or more β coefficients

- R package: aod (Analysis of Overdispersed Data)
- Syntax wald.test(Sigma, b, Terms)
- Sigma: var-cov matrix, extracted from the glm function
- b: coefficients (coef(glm()))
- Terms: specifies which terms in the models are to be tested

β_1 Age

β_2 I_F.

```
library(aod) # Analysis of Overdispersed Data
wald.test(Sigma=vcov(fitasf), b=coef(fitasf), Terms=2:3)
```

β_j.

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 6.9, df = 2, P(> X2) = 0.032
```

Compare to $\chi^2_{2,0.95} = 5.99$ with LRT ($p = 0.005$)

Case 3: Wald tests in R

```
# Testing interaction, Refer to interaction model
# summary(fitas)
# Testing a single beta
wald.test(Sigma=vcov(fitas), b=coef(fitas), Terms=4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 2.9, df = 1, P(> X2) = 0.086
```

$H_0: \gamma_3 = 0$

$H_a: \gamma_3 \neq 0$

Compare with p. 10 & 11

Case 3: Estimated probability of survival

$y_i = 0, 1$ (Observed)

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{Age} + \beta_2 1_F$$

$$\downarrow$$

$$\ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 1_F \quad \Rightarrow \hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 1_F}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \text{Age} + \hat{\beta}_2 1_F}}$$

(Estimate).

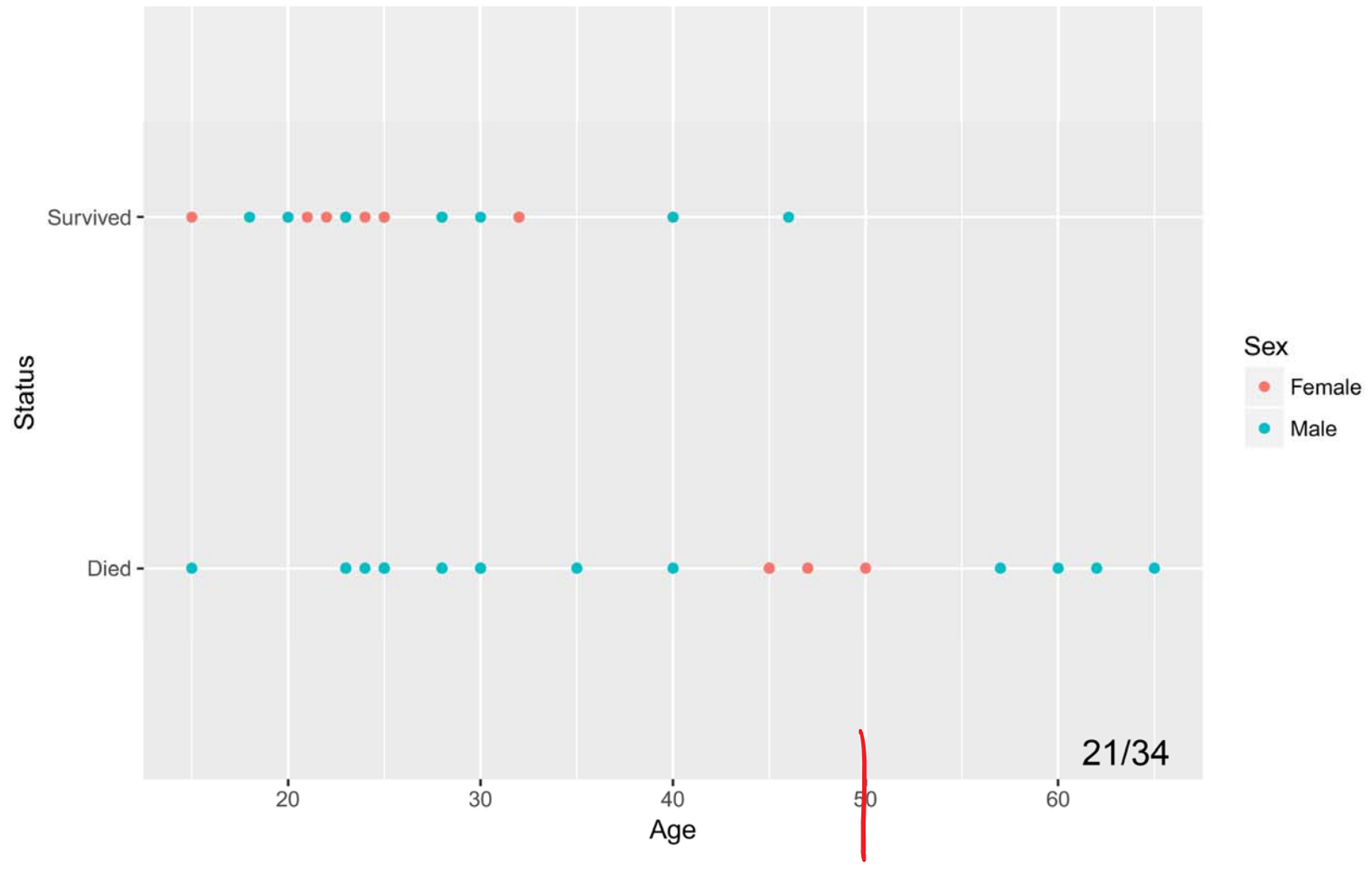
```
phats<-predict.glm(fitasf, type="response") # predicted probability of survival
phats[1:5]
```

```
##         1         2         3         4         5
## 0.4587010 0.5255405 0.1831661 0.3289359 0.3643360
```

$0 < \hat{\pi} < 1$

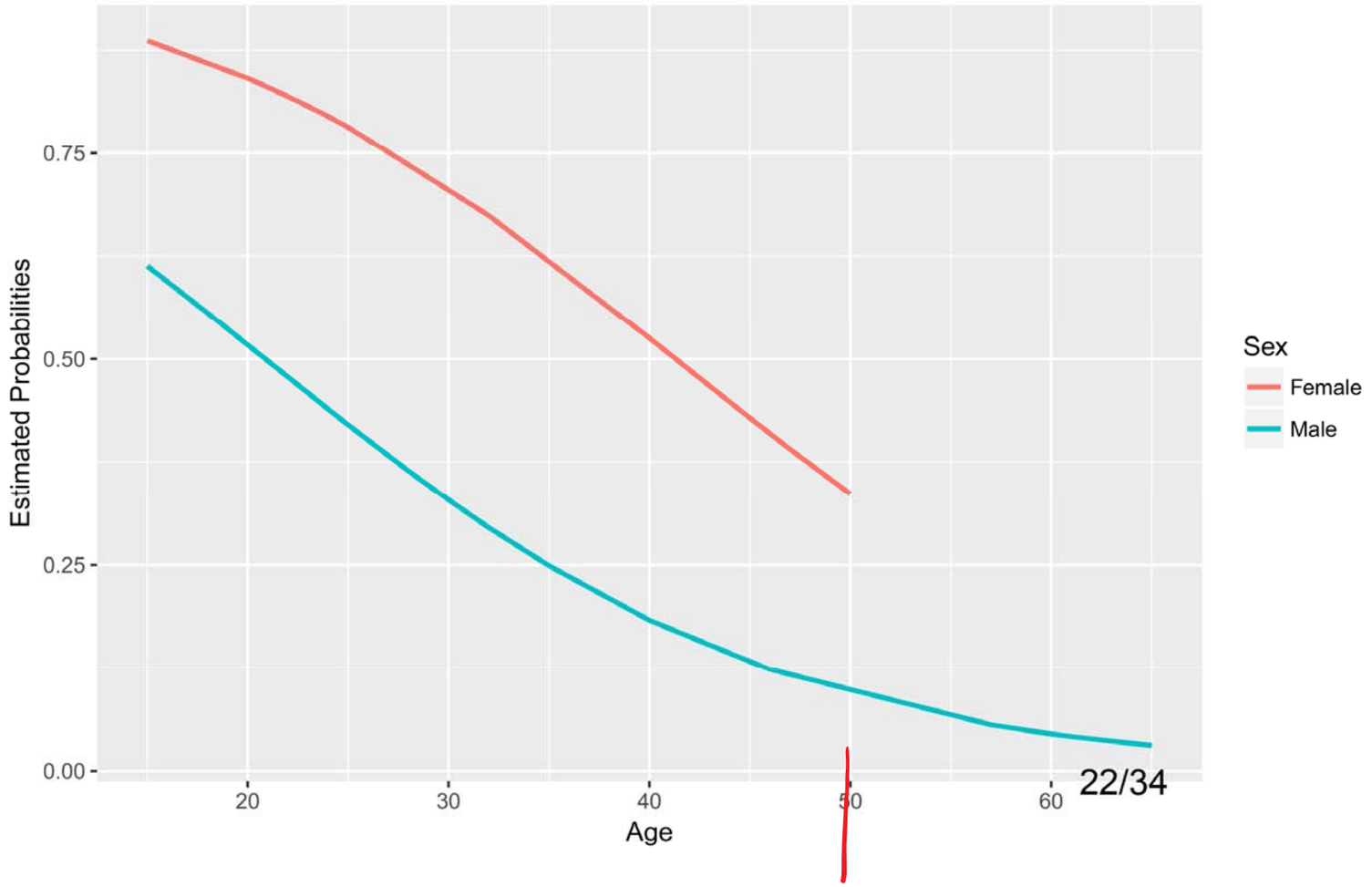
Case 3 Plots: Of Data

```
#contrasts(Status)
library(ggplot2)
ggplot(donner, aes(x=Age, y=Status, color=Sex))+geom_point()
```



Case 3 Plots: Additive Logistic Regression Model

```
ggplot(donner,aes(x=Age, y=phats))+ylab("Estimated Probabilities")+
  geom_line(aes(color=Sex), size=1)
```



Plot

$\hat{p}(\text{"SURVIVED"}) = 1 - \hat{p}(\text{"DIED"})$

$\text{ODDS}(\text{"SURVIVED"}) = \frac{1}{\text{ODDS}(\text{"DIED"})}$

Q: How would the plot of estimated probabilities change if we modelled probability of death rather than survival?

Over 50yrs

Q: Should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50?

Yes, since there were no women older than 50. the model may not extend.

Other Model Fit Statistics

- Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty.
- Useful for comparing models with same response and same data
- Extends from normal regression to GLMs

1. Akaike's Information Criterion (AIC)

$$AIC = -2 \log \mathcal{L} + 2(p + 1)$$

2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2 \log \mathcal{L} + (p + 1) \log N$$

where

- p-number of explanatory variables, and
- N=sample size

Model Fit Statistics: AIC and BIC

- Smaller is better!
- BIC applies stronger penalty for model complexity than AIC

- AIC Rule of Thumb:
  - One model fits **better** than another if difference in AIC's > 10
  - One model model is essentially **equivalent** to another if the difference in AIC's < 2

Using AIC: Case Study III Example

- Fitted models are based on same response and data.
- Based on AIC, choose a 'best' model.

BIC (fittedmodel)

Model	Variables	AIC	BIC
1	{age,sex}	57.256	62.676
2	{age,sex,age*sex,age^2,age^2*sex}	57.361	68.201
3	{age,sex,age*sex,age^2 }	55.830	64.863
4	{age,sex,age*sex}	55.346	62.573

Results:

- Difference in AIC between 1 and 3 is within 2
- There is some indication that 2 is worse than 3 and 4.
- Choose Model 1 (the simplest)

Related R packages and functions

- Packages:
  - `aod`: analysis of over-dispersed data
  - `ggplot2`: graphics
  - `Sleuth3`: data sets for Ramsey and Schafer's text
  - `effects`: effects displays for GLM and other models
- Functions:
  - `confint()`
  - `coef()`
  - `vcov()`
  - `wald.test()`
  - `AIC()`
  - `BIC()`

Binomial Logistic Regression

Suppose $Y \sim \text{Binomial}(m, \pi)$

- Y-binomial count of the number of "successes"

$$P(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m$$

  - Link to Bernoulli:
    $Y = \sum_{i=1}^{m} X_i$ if X_i's are independent Bernoulli(π) r.v.s.
    Assume that π is the same for each Bernoulli trial.

- Mean: $\text{E}(Y) = m\pi$
- Variance: $\text{Var}(Y) = m\pi(1-\pi)$

Suppose $Y \sim \text{Binomial}(m, \pi)$

- Consider modelling

$$P = \frac{Y}{m}$$

  - the proportion of "successes" out of m independent Bernoulli trials.
- where,
  - $\text{E}\left(\frac{Y}{m}\right) = \pi$
  - $\text{Var}\left(\frac{Y}{m}\right) = \frac{\pi(1-\pi)}{m}$

Case Study IV Data Example

- Data: counts of bird species for 18 Krunnit Islands off Finland.

i	x_i area	m_i nspecies	y_i nextinct
ISLAND	AREA	ATRISK	EXTINCT
Ulkokrunni	185.8	75	5
Maakrunni	105.8	67	3
Ristikari	30.7	66	10
Isonkivenletto	8.5	51	6
...			
Tiirakari	0.2	40	13
Ristikarenletto	0.07	6	3

$p_i = \frac{y_i}{m_i}$

- AREA- area of island in km^2, x_i
- ATRISK- number of species on each island in 1949, m_i
- EXTINCT- number of species no longer found on each island in 1959, y_i

Case Study IV: Model

- π_i- probability of 'extinction' of each island.
  Assume that this is the same for each species of bird on a particular island.
- *Assume species survival is independent.* Then

$$Y_i \sim Binomial(m_i, \pi_i)$$

- Unlike Case III- Donner party binary logistic example, we can estimate π_i from the data.

Case Study IV: Model

- Observed response proportion:

$$\bar{\pi}_i = \frac{y_i}{m_i}$$

- Observed or Empirical logits: (S-"saturated")

$$\log\left(\frac{\bar{\pi}_{S,i}}{1-\bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

- Proposed Model: $\boxed{\log\left(\frac{\pi_{S,i}}{1-\pi_{S,i}}\right) = \beta_0 + \beta_1 Area_i,}\ i = 1, \ldots, 18$
- AIM:
  - Learn how to create nature preserves that help endangered species.
  - Are large or small preserves better?